1-2011

Summative Confidence

Paul Cristian Gugiu

Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Social Statistics Commons, and the Statistics and Probability Commons

Recommended Citation

Gugiu, Paul Cristian, "Summative Confidence" (2011). Dissertations. 413.
https://scholarworks.wmich.edu/dissertations/413

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
SUMMATIVE CONFIDENCE

by

Paul Cristian Gugiu

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Interdisciplinary Ph.D. in Evaluation

Western Michigan University
Kalamazoo, Michigan
December 2011
Often the singular goal of an evaluation is to render a summative conclusion of merit, worth or feasibility that is based on multiple streams of multidimensional data. Exacerbating this difficulty, conducting evaluations in real-world settings often necessitates implementation of less than ideal study designs. This reality gets further complicated by the standard method for estimating the precision of results via the confidence interval (CI). Traditional CIs offer a limited approach for understanding the precision of a summative conclusion. This dissertation develops and presents a unified approach for the construction of a CI for a summative conclusion (SC).

This study derived a formula for estimating the SC and CI that unpacks the multiple pieces of the summative conclusion and accommodates the following study elements: the Type I Error rate; the number, variance, and correlation among the values used to formulate the conclusion; the performance benchmarks for critically important values; the sample size and the amount of measurement error for each value; and the amount of weight accorded to each value, all of which are under varying levels of control by the evaluator. Statistical and psychometric proofs for each of the underlying theories were presented along with Monte Carlo simulations demonstrating how each affect SC.
Methods were derived to fill gaps in the literature for removing sampling error and measurement error from a composite variable, constructing CIs for ordinal variables, determining the distribution of a composite variable generated from variables measured with different scales or that conform to dissimilar distributions, expanding the law of total covariance to accommodate two predictors, and computing a nonparametric reliability estimate and CI. SAS code is presented for generating non-normal correlated data and constructing CIs for ordinal variables. As a result, evaluators can now construct CIs for their summative conclusions, which will help the field of evaluation gain wider acceptance in the scientific community.
ACKNOWLEDGEMENTS

The subject of this dissertation has evolved slowly over the course of three years. However, its completion would not have been possible without the support I received from faculty, colleagues, friends, and family. To all I say a wholehearted “Thank you.”

I am deeply appreciative of the time and knowledge each of the members of my dissertation committee have invested in me during the various stages of my education. I am grateful to my mentor and the chairperson of my dissertation committee, Dr. Brooks Applegate, who readily shared the great depth and scope of his knowledge. He inspired me to grow as a methodologist and psychometrician and continues to do so to this day. I am grateful to Dr. Warren Lacefield who introduced me to measurement theory. The idea for my dissertation would not have occurred to me had I not participated in his courses. I am grateful to Dr. David Hartmann for patiently waiting for several years as I refined my ideas as well as for stressing the importance of incorporating qualitative data into my dissertation.

To my colleagues, I thank you for challenging me to refine my ideas and supporting me when I needed it. I am grateful to Dr. Chris Coryn and Dr. Daniela Schröeter for their constant friendship throughout our time in the Interdisciplinary Ph.D. Program in Evaluation (IDPE) at Western Michigan University. I recall fondly the many conversations we had during our formative years in the IDPE and how we hoped to contribute to the field of evaluation. I am particularly grateful to Dr. Coryn who provided me with financial assistance over the years in the form of evaluation contract work. I am also grateful to Jason Bodnar who pointed out that mathematical-statistics held the key to solving many of the issues with which I was grappling early on in my dissertation.
Numerous friends and mentors have provided me with encouragement over these many years. I am grateful to Dr. Michael Scriven who taught me what it means to be an evaluator. Dr. Scriven encouraged his students to embark upon a dissertation that had the potential to change the discipline. I hope that Summative Confidence meets this high bar. I am thankful to Dr. Daniel Stufflebeam for sharing his experiences as an evaluator, some of which I have included in my dissertation. I am appreciative of the counseling Dr. Marianne Di Pierro provided me during times of stress. I am grateful to Mary Ramlow for all the support she has given me over the course of my doctoral studies. Truly, Mary “makes the trains run on time” in the IDPE. I am also truly appreciative of the numerous classmates and friends who over the years have supported and encouraged me throughout this process.

My deepest respect goes to my family. To my parents, Marin and Aurelia Gugiu, thank you for instilling in me the value of education. Thank you for encouraging me to work hard and for pushing me to be the best at everything I do. And, thank you for financially assisting me as I completed my dissertation. To my aunts, uncles, nieces, and nephews, thank you for believing in me. Your unwavering love and support has strengthened me over the years. Above all, I am grateful to my loving wife, Dr. Mihaela Ristei Gugiu. Your love and support made all this possible. In particular, thank you for the sacrifices you have made in the past year that enabled me to focus my attention on completing my dissertation.

I dedicate this dissertation to my parents and family. I hope it makes you proud.

Paul Cristian Gugiu
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>2</td>
</tr>
<tr>
<td>Background</td>
<td>10</td>
</tr>
<tr>
<td>Logic Underlying Summative Confidence</td>
<td>10</td>
</tr>
<tr>
<td>Factors that Impact Summative Confidence</td>
<td>12</td>
</tr>
<tr>
<td>Illustrative Example: Recommending a Faculty Member for Tenure</td>
<td>14</td>
</tr>
<tr>
<td>Dissertation Chapters and Objectives</td>
<td>25</td>
</tr>
<tr>
<td>Relevance</td>
<td>28</td>
</tr>
<tr>
<td>Delimiters</td>
<td>30</td>
</tr>
<tr>
<td>II. HISTORICAL ANTECEDENTS</td>
<td>32</td>
</tr>
<tr>
<td>Statistical Theory</td>
<td>34</td>
</tr>
<tr>
<td>Measurement Theory</td>
<td>46</td>
</tr>
<tr>
<td>Quantifying Qualitative Data</td>
<td>50</td>
</tr>
<tr>
<td>III. MONTE CARLO SIMULATIONS</td>
<td>60</td>
</tr>
<tr>
<td>Simulating Data</td>
<td>60</td>
</tr>
<tr>
<td>Study Design</td>
<td>62</td>
</tr>
</tbody>
</table>
Table of Contents—Continued

CHAPTER

Testing Normality of Composite Variables ................................................. 64
Generating Nonnormal Data ................................................................. 65
Generating Correlated Data .................................................................... 74
Generating Covariance Matrices ............................................................. 80
Monte Carlo Simulations of the Central Limit Theorem ......................... 85
Model 1 (Independence) ........................................................................... 86
Model 2 (Dependence) ............................................................................. 91
Future Research on Testing for Normality .............................................. 95
Summary ................................................................................................. 97

IV. IMPLICATIONS OF MATHEMATICAL STATISTICS ............................... 99

Expectation Theory .................................................................................. 100
Definition of Expected Value ................................................................. 100
Properties of Expected Value ................................................................. 101
Definition of Variance and Covariance .................................................. 106
Properties of Variance and Covariance .................................................. 107

Limitation of Employing SAS (or Any Other Software) to Compute Property 24 ................................................................. 114

Distribution Theory .................................................................................. 116
Discrete Random Variables ...................................................................... 118
Continuous Random Variables ............................................................... 127
Transformation Theory ............................................................................ 130
Table of Contents—Continued

CHAPTER

Polychotomous ................................................................. 133
Measurement Scales ............................................................ 141
Order Statistics and Quantiles ............................................... 145
Other Transformations ........................................................ 164
Sampling Theory .................................................................... 165
Implications for Summative Confidence ................................. 168
Confidence Intervals .............................................................. 173
Constructing a Basic Confidence Interval .............................. 174
Constructing a Distribution-Free CI on the Population Median .. 184
Algorithm for Constructing a Basic Summative Confidence Interval .. 199

V. IMPLICATIONS OF MEASUREMENT THEORY ...................... 204
Classical Test Theory ............................................................. 206
Fundamental Theory ............................................................ 207
Potential Limitations of CTT ................................................ 211
Implications for Summative Confidence ................................. 215
Classical Reliability ............................................................. 221
Stability and Equivalence ..................................................... 222
Internal Consistency ............................................................ 224

VI. IMPLICATIONS AND CONCLUDING REMARKS ....................... 233
The Summative Confidence Algorithm .................................. 234
Hypothetical Example of Summative Confidence .................... 241
Table of Contents—Continued

CHAPTER

The Implications of Summative Confidence ............................................. 245
Contributions to Evaluation, Statistics, and Psychometrics ..................... 247
Future Research and Final Remarks ...................................................... 249

REFERENCES .......................................................................................... 251

APPENDICES

A. Glossary of Key Terms ........................................................................ 257
B. SAS Code for Generating GLD Lambda Values ................................. 265
C. SAS Code for Generating Positive Definite Correlation Matrices .............. 268
D. SAS Code for Illustrating the Central Limit Theorem (Model 1) ............... 270
E. SAS Code for Illustrating the Central Limit Theorem (Model 2) .............. 281
F. SAS Code for Illustrating Transformation Theory .............................. 288
G. SAS Code for Illustrating Simple Random Sampling .......................... 297
H. SAS Code for Constructing Parametric and Nonparametric Confidence
   Intervals .............................................................................................. 306
I. SAS Code for Testing Classical Test Theory ....................................... 326
J. Extension of the Law of Total Covariance ......................................... 336
### LIST OF TABLES

1. Simulated Skewness (and Standard Error) by Sample Size and Distribution ........................................... 71
2. Simulated Kurtosis (and Standard Error) by Sample Size and Distribution ............................................. 72
3. Stepwise Multiple Regressions of Deviations in Simulated Skewness and Kurtosis .................................................. 73
4. Example of a Theoretical and its Nearest Positive Definite Correlation Matrices ............................................. 76
5. Average Deviation in Correlation Coefficients by Sample Size and Distribution ........................................... 77
6. Multiple Regressions of Deviations in Simulated Correlations ................................................................. 79
7. Empirical Tests of Simulated Covariance Structure ................................................................. 85
8. Number of $k$-Tuples for $N$, $k \leq 10$ ........................................................................................................ 125
9. 2008 Beijing Olympics: Men’s Marathon Results (Top 21 Finishers) ....................................................... 157
10. Ten Simulated Variables Measured on a Likert Scale Based on $\text{Mult}(25; 0.20, 0.25, 0.30, 0.25)$ ........................................................................................................................................... 162
11. Unadjusted Coverage Probabilities and Mean CIs for Three Probability Distributions ................................................................. 190
12. FPC Adjusted Coverage Probabilities and Mean CIs for Three Probability Distributions ................................................................. 192
13. Coverage Probabilities and Mean CIs for $X \sim \text{Mult}(1; 0.15, 0.2, 0.3, 0.22, 0.13)$ ......................................................... 196
14. Coverage Probabilities for Categorical Distributions with Varying Number of Classes ................................................................. 198
15. Comparison of Coverage Probabilities Between the Discrete and Continuous CI Methods ................................................................. 198
16. Hypothetical Case Based on the Tenure Review Example ........................................................................ 243
LIST OF FIGURES

1. Example of the Values Used to Determine Whether a Faculty Should be Recommended for Tenure ............................................................................................................. 17
2. Observed Probability Density Functions Versus the Normal Distribution .... 68
3. Normality and Homoscedasticity Tests of Deviations in Correlations .......... 78
4. Likelihood of Attaining an Approximately Normal Composite Variable Based on Skewness and Kurtosis, Where the Constituent Variables Were Independent ................................................................................................................ 87
5. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Sample Size and Number of Independent Constituent Variables Aggregated, Where Average Skewness Equals 1 and Average Kurtosis Equals 5. 88
6. Probability Density Functions for Four Composite Variables Generated by Averaging 30 Constituent Variables With a Skewness of 2 and Kurtosis of 10 ... 89
7. Probability Density Functions for Four Composite Variables Generated by Averaging 150 Constituent Variables With a Skewness of 2 and Kurtosis of 10 . 90
8. Likelihood of Attaining an Approximately Normal Composite Variable Based on Skewness and Kurtosis, Where the Constituent Variables Were Dependent 92
9. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Sample Size and Number of Dependent Constituent Variables Aggregated, Where Average Skewness Equals 1, Average Kurtosis Equals 5, and Average Correlation Equals 0.5 ................................................................................................................ 93
10. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Sample Size and Correlation Simulated for the Constituent Variables, Where Average Skewness Equals 1 and Average Kurtosis Equals 5 ................. 94
11. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Number of Variables Aggregated and the Correlation Simulated for the Constituent Variables, Where Average Skewness Equals 1, Average Kurtosis Equals 5, and Average Sample Size Equals 480 ......................................................... 95
12. Probability Mass Function and Cdf of \( X \sim \text{Mult}(10; 0.35, 0.40, 0.25) \) ................. 124
13. Pascal’s Triangle for Computing the Number of \( k \)-Tuples ................................. 125
14. Pdf of $X \sim N(0,1)$ and Pmf of $Y \sim Bin[1, 1-\Phi(x)]$ .......................................................... 136
15. Pmf of $X \sim Pois(9)$ and Pmf of $Y \sim Mult[1, 1-F_X(\tau_1), F_X(\tau_2)-F_X(\tau_1), 1-F_X(\tau_2)]$ .... 140
16. Relation Between the Trinomial Distribution and Continuous Order Statistics . 148
17. Relation Between the Trinomial Distribution and Discrete Order Statistics .... 151
18. Distribution of Confidence Intervals About the Unknown Population Parameter .......................................................................................................................... 182
19. Illustration of the Notions of Biasedness and Precision .............................................. 183
20. Illustration of $P(X_{(\phi)} \leq x \leq X_{(\phi)})$ for Discrete Order Statistics ...................................... 194
CHAPTER I
INTRODUCTION\textsuperscript{1}

One of the principal lessons impressed upon students in an introductory methodology course is that “a weak design will yield unreliable conclusions.” While this is certainly true, the constraints of conducting research and evaluation studies in real-world settings often necessitate the implementation of weak study designs (Burstein, Freeman, Sirotnik, Delandshere, & Hollis, 1985).\textsuperscript{2} For example, evaluators may have no choice but to collect small sample sizes, include heterogeneous subjects, employ unreliable and invalid instruments, implement procedures that produce high measurement error, or utilize minimal triangulation. In such instances, investigators are left to debate the tradeoff between the costs and impediments associated with modifying a weak design (i.e., one that is likely to produce an incorrect evaluative conclusion) and sacrificing the precision of their conclusions.

The purposes of this dissertation are two-fold: (a) to develop a methodology that can be used to characterize the precision of an evaluative conclusion and (b) to heighten awareness about the factors that contribute to the overall quality of an evaluative conclusion. It is hoped that careful consideration of the limitations of certain evaluation practices will sensitize evaluators to the need to include necessary safeguards in planning studies. Moreover, consideration by decision-makers of the degree of confidence one can place on an evaluative conclusion, herein referred to as Summative Confidence, may alert them to whether

\textsuperscript{1} An earlier version of this chapter was published by the author in Gugiu (2007).
\textsuperscript{2} All references in this dissertation follow the American Psychological Association (APA) style as expressed in the APA Publication Manual (2009).
\textsuperscript{3} Italicization will be used throughout this document the first or second occurrence of a word or phrase whose definition may be found in Appendix A.
action needs to be taken to correct a serious problem, reward a successful program, or seek further evidence of the merit and worth of the *evaluand* (i.e., entity under investigation).

**Statement of the Problem**

One of the most frustrating aspects of evaluation is that well-intentioned evaluators with widely-recognized expertise will often reach different, and at times even contradictory, conclusions when evaluating the same *evaluand*. The evidence in support of this statement is all around us. For example, frequent disagreements occur between journal reviewers (presumably picked due to their expertise) regarding whether to publish or reject an article whose merits they just evaluated. Similarly, the scientific literature is replete with programs that were deemed a success by one evaluation team and a failure by another evaluation team. Of course, often these contradictions occur in different settings and populations. Hence, one can argue that perhaps the programs were never identical. Documenting contradictions for the same *evaluand* is more difficult since rarely are two or more evaluators asked to conduct parallel evaluations on the same *evaluand* because the costs associated with such an undertaking are prohibitive. Fortunately, some evidence of this does exist.

Perhaps the most vivid example comes courtesy of Dr. Daniel Stufflebeam, former president of the American Evaluation Association. In 1971, he was required by the U.S. Office of Education to retain the services of external evaluators to evaluate a federally supported program at The Ohio State University titled the Model Training Program in Evaluation (the predecessor of the Interdisciplinary Ph.D. in Evaluation at Western Michigan University). In personal communications between Dr. Stufflebeam and myself on August 16 and 17, 2007, he described the evaluation of the program, as follows.

“[a]gainst the advice of his federal sponsor, I decided not to select ‘softees’ to do the evaluation, because I wanted the external assessment to be credible to leaders in the
evaluation field. I selected three very tough evaluators⁴, including two who publicly had criticized the program.”

As a result of this selection, he knew that these two evaluators were

“entering with predilections/biases about the program’s work that could color their evaluations and potentially lead to the termination of the million dollar-plus project. One of the evaluators worked in [another] center that was, in effect, a competitor of ours. Moreover she was wedded to a Tylerian view of evaluation, which was somewhat contradictory to our approach. [Another of the evaluators] was on a crusade against formative evaluation and, at the time, was a devotee of experimental design. Our evaluation work was very much formative then, because we were evaluating projects that were in developmental stages and not conducive to strictly controlled, randomized field experiments… Thus, I decided to put the three evaluators to a test of their accuracy, objectivity, and credibility by giving them access to everything they had on the program …[but] insisted that they not talk to each other during their evaluation processes.”

Because he knew there would be an “error term” in their conclusions, he wanted

“them, our government sponsors, and us to hear the evaluators’ independent assessments and consider the variation among these assessments as well as their central tendencies. In this way I thought we would get a fair hearing and receive a richer set of inputs than would have been the case if they went into a backroom and homogenized their reports, thus potentially hiding initial contradictions in their findings. Their reports turned out to be very different from each other and in the predictable ways. [emphasis added]”

Also, he believed the three evaluators were more careful and even circumspect about their evaluations than would have been the case if they had amalgamated their reports.

Contemplating upon the differences in the conclusions of the three evaluators, Dr. Stufflebeam described it as

“three blind persons who felt around, then described the elephant they had been exploring. Their accounts were incomplete and very different. The three evaluators looked at the program from their favored perspective and arrived at very different assessments. [emphasis added] While all three assessments were, overall, quite positive, their identification of strengths and weaknesses were in different sectors. Each saw things that they counted as important but the things they looked for were different, one from the others. One looked for evidence that the program had specific behavioral objectives, another that it was stressing values and summative evaluation, and another that it was teaching evaluators to involve and address the needs of users.”

Reflecting further upon the case, Dr. Stufflebeam noted that he continues to “believe it was important that I insisted that they develop and present independent reports.”

⁴ It is the understanding of the present author that at least one of the three evaluators went on to become president of the American Evaluation Association.
It can be argued, however, that this case study might not generalize to other evaluations since two of the three evaluators were already critical of the program prior to the initiation of the evaluation. Moreover, it is only a single study in which the evaluators were given free-reign to examine any variables they thought pertinent to the evaluation. One may think, if the evaluators were given identical data and a standard grading guideline or rubric, they would reach similar results. This hypothesis was tested in a study in which 30 raters (10 evaluation doctoral students, 10 evaluation practitioners, and 10 evaluation scholars) were asked to metaevaluate 10 program evaluation reports using the *Program Evaluation Standards*\(^5\) (Joint Committee on Standards for Educational Evaluation, 1994), which are endorsed by the American National Standards Institute and numerous evaluation organizations and luminaries. According to Wingate (2009), the results showed uniformly low interrater reliability. In fact, employing the Spearman-Brown prophecy formula, Dr. Wingate projected that one would need 28 evaluators, on average, to rate each evaluation report to attain a reliability coefficient of 0.8 for the four evaluation domains (utility, feasibility, propriety, and accuracy) considered by the Program Evaluation Standards. Lastly, examination of the agreement rates by type of rater revealed no differences. If anything, the doctoral students appeared to have slightly higher agreement rates than their more seasoned counterparts.

The implication of these studies are clear and along with 16 years of professional evaluation experience—as of the time this study was written—form the basis of what the present author has come to call “the dirty, little secret of evaluation.” Namely, two or more evaluators evaluating the same evaluand often reach different and, at times, contradictory evaluative conclusions.\(^6\) Moreover, as the Dr. Wingate’s study illustrates, such discrepancies

---

\(^5\) The *Program Evaluation Standards* (PES) are a set of 30 evaluation principles that are organized in four major standards: utility, feasibility, propriety, and accuracy. A grading rubric was used to score all 30 PES items.

\(^6\) This is not really a secret to seasoned evaluators but it is one that is rarely, if ever, discussed in public forums.
appear to hold even when grading rubrics are employed, thereby leading one to question whether evaluative conclusions are reliable or not (i.e., lack replicability—a basic premise of science). However, if this premise is true, then it calls into question the very foundation of evaluation by suggesting that an evaluation conclusion is not a scientific claim on the part of the evaluator, despite potentially being the product of a scientifically accepted method, but rather reflects their professional opinion. This is a rather ominous conclusion for it communicates to evaluation clients and decision-makers alike that if they are unhappy with an evaluative conclusion, they can always find an evaluator that will provide them with a conclusion that is more pleasing to their predilection. Fortunately, this problem is not new to social sciences and a means for combating it has been developed.

In scientific circles, professional evaluation, henceforth simply referred to as evaluation, conclusions are persuasive to the extent to which they are accurate and lack error (i.e., are precise). One method of expressing the precision of a conclusion is through the use of a confidence interval—a practice recommended by leading research organizations, e.g., the American Psychological Association (APA) (Wilkinson & APA Task Force on Statistical Inference, 1999). Typically, the method used to determine the precision of a result is the size of the interval. Large intervals suggest that a result is imprecise (i.e., has a large amount of error) whereas small intervals indicate the opposite. Similarly, the confidence level associated with an interval communicates the probability of reaching a correct conclusion at any given time if the study were replicated ad infinitum under parallel conditions. Therefore, small intervals that have a low confidence level are not very impressive. Important as confidence intervals (CI) can be for reporting precision and confidence, the analytical method that is used to calculate a CI suffers from one important limitation: it cannot be employed to
calculate a CI for a complex variable\textsuperscript{7} that is derived from multiple data streams. Unfortunately, a large portion of evaluation practice entails the formulation of evaluative conclusions based on numerous criteria or dimensions of merit or worth. Therefore, a significant gap exists with respect to how one can estimate the degree of confidence that should be placed on an evaluative conclusion when such conclusions are the product of a complex synthesis of multiple factors and their indicators.

Further compounding this problem is the data synthesis dilemma. Evaluation practice often requires the synthesis of qualitative and quantitative data into an overall conclusion, i.e. a summative conclusion. However, because separate methods exist for analyzing qualitative and quantitative data, no unifying method has been proposed for calculating the precision of a summative conclusion derived from data, unless of course the qualitative data is quantified somehow. However, the process of transforming one data type into another\textsuperscript{8} complicates the ability of calculating the precision of the summative conclusion. For example, suppose a professor needed to assign a final grade to a student who received a ‘C’ on a term paper and an ‘A’ on a multiple-choice exam. To what degree is the precision of the student’s final grade a function of the weight assigned to each individual grade? In addressing this question, two factors should be considered: the weighting scheme and the

\textsuperscript{7} A complex variable is a composite variable that synthesizes multidimensional data collected from multiple data streams into a single vector whose elements denote the evaluand’s performance on (or their quality with respect to) the criteria of merit or worth (as judged by some method of measurement) that are included in the evaluation. Moreover, since evaluation data often are collected from multiple data streams, the units of analysis (e.g., case identification numbers, people, respondents) for these datasets are different thereby preventing one from meaningfully merging the data and analyzing it in a single step.

\textsuperscript{8} Because it is often simpler to reduce greater detail to less detail, rather than the reverse, quantitatively-oriented analysts transform qualitative data into frequency or indicator data. However, the reverse process is also possible. Qualitatively-oriented analysts may convert quantitative data into qualitative data through a process of interpretation and labeling. For example, a quantitative IQ score of 160 may be interpreted and labeled as superior whereas an IQ score of 70 may be classified as below average.
reliability of the two grades\(^9\). In general, the grades assigned to written assignments are less reliable than those assigned to quantitatively-scored exams (e.g., multiple-choice, true-false) because the proportion of error variance is greater in the scores of the former than in the scores of the latter (Hopkins, 1998, pp. 184-209). Therefore, the final grade will be more precise if the student is assigned a ‘B+’ (multiple-choice exam, which has less error variance, is given more weight) rather than a ‘B-’ (term paper is given more weight).\(^{10}\) Unfortunately, the current methods by which one can construct a composite variable (via means or totals) do not account for the reliability of each constituent variable. Thus, in the example above, although one can surmise that weighing the multiple choice exam more heavily than the term paper will result in a more precise final grade, without the ability to remove the measurement error associated with the unreliability of each testing method, it is difficult to predict how much more narrow the former CI on the final grade will be as compared to the latter CI.

Similar issues arise with regard to sampling. While the family of randomized sampling designs is widely regarded as the ‘gold standard’ for the purpose of generalizing results from a sample to the population (Cook & Campbell, 1979; Kish, 1995), the impact of sampling error on evaluative conclusions appears to have been overlooked in the literature. That is, most researchers and evaluators acknowledge that selecting a small sample out of a much larger population limits the degree to which one may generalize a result to the entire population. However, many fail to recognize that the larger the sampling error (i.e., the likely deviation over many replications of the sample estimate from the true score), the lower the precision of their estimates and thus, their final conclusions. In another words, sample

\(^9\) A third factor is whether the variance of the grades of the multiple-choice exam differs from that of the term paper. This example, assumes that the variances are equal. However, as will be discussed later on, such differences also impact the variance of the composite constructed from averaging the two grades.

\(^{10}\) The increased reliability of objective exams, like multiple choice tests, however, sometimes can come at the expense of validity. That is, if a multiple choice test is more narrowly defined in scope than an essay, it will contain less information than the essay. Hence, one sacrificed validity to gain reliability.
statistics (e.g., means, variances) are only approximations of the population parameter. Thus, unless an evaluator wishes to confine their conclusions to the sample they collected (and many do), they need to account for sampling error to reach conclusions about the true score of the population. A popular method, advocated by statisticians (e.g., Lohr, 1999), is to use the standard error of the mean to construct a CI. However, this method requires knowledge of the population standard deviation, which is typically unknown. Hence, the sample standard deviation is used as a substitute but this statistic is only a good estimator of its population counterpart when the sample size is relatively large. This connotes that CIs constructed based on a sample standard deviation from a small sample size may not generalize well to large populations due to the inaccuracy of the sample statistic. For example, the CI may be either too conservative (i.e., too wide) or worse still, too liberal (i.e., too narrow) and thereby either overestimate or underestimate the probably coverage.

The process of formulating a conclusion may also require comparison against a known or constructed standard. For example, while the ability of two graduate students, one with a 2.95 GPA and one with a 3.00 GPA, may be nearly identical, the conclusions one would reach about each student would differ when compared against a university’s minimum standard of acceptable academic performance (generally set at a 3.00 GPA). In the case of the former student, one would conclude that the student failed to meet the minimum expectation while in the latter case one would conclude the reverse. However, how precise is the conclusion that the latter student’s ability meets or exceeds the minimum expectation? Given their proximity to the standard, it is safe to conclude that one would be less confident in the conclusion that the second student met or exceeded the standard than had they earned

---

11 Note, this is a fixed standard. In contrast, if the members of a graduate admission committee were asked to identify their personal standard of the minimum undergraduate GPA they would consider acceptable for admitting a student to graduate study, it is likely that their standards will differ from one another. Consequently, determining the merit of a candidate for graduate study requires consideration of the variability in standards.
a 3.60 GPA. Therefore, the degree of precision of a conclusion is inversely related to the difference between performance and the standard. While methods exist for calculating a confidence interval for such cases (Crocker & Algina, 1986, pp. 192-212), no method exists for estimating the impact of such cases on complex variables.\(^\text{12}\)

Finally, commonsense dictates that the more information one knows about an evaluand, the more confident one may be in the conclusion formulated about it. Similarly, the wider the array of methods used to collect information about the evaluand and the data sources from which information is collected (i.e., triangulation), the greater the accuracy of one’s conclusions. For example, if one wishes to know the weight of an object, one could simply weigh the object on a scale. If the scale was error-free, only one weigh-in is necessary. However, because scales are not perfectly accurate (i.e., measurement error exists), one can average the results of several weigh-ins to improve the accuracy and precision of the composite variable.\(^\text{13}\) In many evaluation scenarios, one may need utilize several instruments to measure the construct, particularly if it is multidimensional. This of course raises the question, should one have more confidence in a conclusion formulated from instruments that measured unique dimensions of the latent construct or from instruments that measured highly correlated dimensions? It is important to note that one can perform a confirmatory factor analysis or structural equation model to address this question when the units of analysis are the same for all the variables in the data set. In fact, such analyses can even be performed when the latent variables are measured with different scales and conform to

\(^{12}\) It is important to note that the phrase ‘complex variable’ is used in the broadest sense possible. There most certainly are methods for constructing a CI for composite variables that are measured on the same scale (e.g., Likert) and come from the same data source (e.g., the average 20-items collected from a classroom of students). However, constructing a CI for a construct measured from two or more data sources using different scales (e.g., a survey administered to student and faculty measuring faculty performance that uses different scales, such as 5-point Likert scale and a True/False format) is more challenging.

\(^{13}\) Note, the composite variable carries the assumption that the underlying construct measured (in this example, weight) by the estimates (in this case, weigh-ins) is unidimensional and that the factor loadings for each estimate on to the construct are equal (McDonald, 1999). Otherwise, the accuracy of the composite will suffer.
different probability distributions. However, these analyses rely on the ability to estimate a variance-covariance matrix, which is only feasible when the units of analysis for all the data streams are the same. In the case of complex variables, the units of analysis are different for at least some of the variables. Hence, a variance-covariance matrix cannot be estimated for the entire dataset. As a result, a method is needed for expressing the exact relationship between the precision of a summative conclusion and the amount and quality of information used to formulate it.

**Background**

**Logic Underlying Summative Confidence**

One may think of Summative Confidence as the statistical degree of confidence that one can place on an evaluative conclusion, which was derived from a synthesis of the performance of the evaluand on multiple criteria of merit and worth. More specifically, it refers to the band of error surrounding a conclusion given a specified level of confidence. Therefore, if one were to replicate the evaluation ad infinitum, a distribution of conclusions would form around the true or “correct” conclusion. Summative Confidence refers to the band of uncertainty, placed around a conclusion, which is believed to include the true conclusion with a certain probability. Clearly, smaller confidence bands indicate that the evaluative conclusion was estimated with greater precision whereas larger bands indicate the reverse, at a fixed level of confidence. Alternatively, one can think of the interval as fixed and the confidence level as variable in which case, Summative Confidence refers to the proportion of intervals that would contain the true conclusion if the evaluation were repeated infinitely. For example, a teacher who calculates that there is a 99 percent chance the true performance (i.e. ability) of one of her students falls between an “A-” and an “A+” can feel very confident about her grading scheme and according the student an “A” for the
course. However, had there been a 99 percent chance that the student’s true performance fell between a “C” and an “A” then the teacher should feel less confident about her grading scheme and according the student a “B” given the relative size of the two CIs.

One may wonder given the vast methodological variability that exists across studies, can a methodology be developed that can be applied in every study? The foundation of Summative Confidence rests upon two principles. First, everything that can be operationalized, can be measured\(^\text{14}\) with some degree of accuracy and precision. For example, a doctor checking a patient for high blood pressure (hypertension) could look for typical symptoms such as severe headaches, fatigue or confusion, vision problems, chest pain, difficulty breathing, irregular heartbeat, and blood in the urine (Chang, 2005). However, since a large proportion of the people afflicted by this disease have no symptoms, a diagnosis of hypertension based on the presence or absence of symptoms alone is likely to be error-prone. A more accurate and precise diagnosis of hypertension can be obtained by using a sphygmomanometer. Therefore, evaluands measured with reliable instruments\(^\text{15}\) and methods yield conclusions that contain less measurement error.

Second, the degree of measurement error in a summative conclusion is a function of the measurement error of all of the elements used to formulate the conclusion. This leads to one of the basic principles of Summative Confidence, which should be familiar to all computer programmers: “garbage in, garbage out.” Stated more formally, if the criteria of merit and standards from which the summative conclusion is formulated are measured with a high degree of error, then little faith (i.e., confidence) should be placed on the summative

\(^{14}\) According to the famous philosopher and mathematician René Descartes (1644), “if something exists, it exists in some amount. If it exists in some amount, then it is capable of being measured.”

\(^{15}\) It is important to note that although an instrument “operationalizes” a construct, it and of itself may not be a perfect measure of that construct. For example, does a specific IQ test truly measure the construct intelligence? Probably not. However, for the sake of simplicity this nuance will be ignored in the remainder of this study.
conclusion. However, some reprieve may be gained from triangulation. While little confidence should be placed in a conclusion derived from data containing a high degree of measurement error, hope does exist for conclusions derived from several indicators that were measured with a small or moderate amount of error because reliable information is cumulative. Hence, a composite measure generally provides a more precise explanation of the construct than its constituent parts.\textsuperscript{16} A second basic principle of Summative Confidence is that the more information one has upon which to base a conclusion, the more confident one can be in the conclusion as long as the information is precise.

Factors that Impact Summative Confidence

The Summative Confidence of an evaluative conclusion is contingent upon the measurement error introduced into the evaluation by the choices an evaluator makes regarding sampling scheme, instrument selection, and methodological design. For example, to the extent to which sampling error is largely due to a small sample size or heterogeneity, the wider the CI will be (Hays, 1994). To the extent to which instruments are unreliable or poor agreement is attained between raters of qualitative data, the standard error of the conclusion will be large (Crocker & Algina, 1986). To the extent to which few values\textsuperscript{17} are measured or greater weight is assigned to poorly measured values, Summative Confidence will be negatively affected.

\textsuperscript{16} This is not always the case. For example, if the information provided by a set of indicators is redundant, then a composite measure of these indicators will not be more accurate than any individual indicator. However, this assumes not only that each indicator is perfectly correlated with each other, but that they have identical correlations with the construct. If one of the indicators had a stronger correlation with the construct and was perfectly correlated with the other indicators (think of a Venn diagram in which the other indicators are a subset of this indicator which, in turn, is a subset of the construct), then this indicator would be the most accurate measure of the construct.

\textsuperscript{17} The term “values” will frequently be used throughout this dissertation to denote criteria of merit or worth that are pertinent to formulating an evaluative conclusion for a dimension on which the evaluand is measured.
More specifically, the degree of confidence one can place on an evaluative conclusion depends upon the following 11 factors classified in 6 categories:

a. **Alpha.** The probability the true value of the parameter being estimated falls outside of the estimated CI. That is, alpha refers to the Type I Error rate of the summative conclusion.

b. **Values.** The number, variance, organizational structure, and correlation among the criteria of merit and worth used to formulate an evaluative conclusion about the performance of an evaluand.

c. **Standards.** A standard denotes a performance benchmark for a critically important value, which demarks acceptable from unacceptable or excellent from less than excellent performance. The standard can be either a single fixed point (e.g., a cut-score) and/or an interval\(^{18}\) due to two cut-scores.

d. **Sample size.** The size of the sample taken from the population of impactees or decision-makers for a specific measurement.

e. **Measurement error.** The magnitude of discrepancy between an observed score and its true value or score.

f. **Weighting scheme.** The amount and variability of the importance accorded to each micro- and macrovalue.

It is important to note, many of these factors are interrelated. For instance, the consequences of failing a standard are similar to weighting the performance of the failed value more heavily than the performance of other values. Likewise, these factors may be

---

\(^{18}\) An interval may be required to represent a set of standards for a critically-important value if no consensus is reached as to which of the standards is most appropriate to employ. For example, in an evaluation of the performance of a college, a group of stakeholders may provide an evaluator with multiple standards for what they believe constitutes minimum academic performance for college students, assuming academic performance is a critically-important value. At this point, an evaluator can choose to select to employ a single standard (e.g., average standard) in the evaluation or alternatively employ an interval (e.g., range or CI on the standards).
reorganized according to the level of control an evaluator exerts over them. For example, evaluators can exert a great deal of control over design factors, such as alpha, values, standards, sample size, sampling error, and weighting scheme, and relatively less influence over sampling characteristics, such as the measurement error (when instruments with adequate reliability do not exist) and heterogeneity (i.e., variance among stakeholders). However, it is certainly the case that evaluators do exert some influence over the latter factors based on their selection of measurement instruments, procedures, and study participants (e.g., selecting participants likely to be very similar).

**Illustrative Example: Recommending a Faculty Member for Tenure**

To more fully appreciate the complexity of the factors that contribute to Summative Confidence, a realistic illustration may be helpful. Suppose a university provost was interested in evaluating the university’s tenure review process by calculating the Summative Confidence of a randomly selected case. Examination of the case revealed that the decision was reached after an exhaustive deliberation about the applicant’s performance on numerous values, including research, teaching, service, professional accolades, academic interests, and collegiality. The provost also learned that prior to the start of the process, a panel of faculty members deliberated on which factors would be critically important to their decision, the weight that would be assigned to critically and non-critically important factors, the standards that would be used to judge acceptable performance on the factors identified as critically important, and the standard that would be used to arrive at a decision based on a synthesis of all the data. Finally, to ensure the ratings of faculty members were not unduly influenced by “stronger” members within the group, all ratings were anonymous.

The tenure review process began with a meeting between five tenured faculty members from within the department and five randomly-selected, tenured faculty members
from outside-the-department. During the first meeting, the faculty generated and agreed upon a list of criteria upon which to judge the merits of the candidate. This decision was the first of several decisions that impacted the precision of the final decision of whether or not to recommend the candidate for tenure. Although the process of deliberating over criteria and their importance is common, most evaluators treat the agreed upon decisions derived from such processes as unequivocal when, in fact, unanimous agreement does not always exist. Clearly, the greater the disagreement over the values that should be considered in the evaluation, the lower the likelihood that the same conclusion could be replicated by a different panel of faculty or even by the same faculty at a different point in time. Likewise, the lower the agreement over which values should be considered critically important, the weight that should be applied to each value, and the level at which a standard for a critically important value should be set, the lower the probability that the final decision could be replicated. From a Summative Confidence perspective, the most accurate procedure would be for the department or university to devise a uniform policy or for the faculty to take steps to increase consensus (i.e., inter-rater reliability) amongst themselves on these matters. Of the two alternatives, policy decisions are likely to improve Summative Confidence to a greater extent because they place greater limits on rater disagreement.

In addition to the aforementioned factors, the faculty’s decisions regarding the number of values selected, the structure within which these values were organized, and the degree of redundant information shared by the values influenced the precision of the final decision. The common attribute underlying each of these factors is information. As stated previously, the more information (e.g., number of values examined) one has upon which to base a conclusion, the more confident one can be about the conclusion reached, if the information is unique (i.e., contributes unique variance). However, evaluations are rarely
based on an unstructured list of values. Instead, values are organized into micro- and macrovalues. Figure 1 presents the list of values that were used by the faculty to render a decision about the candidate’s tenure. Macrovalues are represented by a rectangle while the microvalues are organized underneath the macrovalue with which they are associated. Furthermore, these values were organized into critically and non-critically important values, with greater weight assigned to the former group of values.

As may be apparent from the figure, some macrovalues were measured with a greater number of microvalues than other macrovalues (e.g., Research versus Teaching). Thus, the precision of the conclusions reached about these values should exceed the precision of conclusions derived from imprecisely measured values, all other factors being equal. The organization of the values should also influence precision because more microvalues were used to measure performance of non-critically important values than were used to measure performance of critically important values. That is, the impact of organizational structure on Summative Confidence is mediated via other factors (e.g., number of values, weighting scheme). Furthermore, the candidate’s performance on the latter group of values weighed more heavily on the final decision than the former group of values. Hence, one can improve confidence by using microvalues with lower levels of measurement error, increasing the number of indicators and methods used to measure a macrovalue (i.e., triangulation)\(^\text{19}\), and assigning more weight to precisely measured macrovalues.

Although somewhat counterintuitive, another method of improving precision is by reducing the amount of redundant information within and between microvalues and

\(^{19}\) Macro-values are generally composed of multiple dimensions. Hence, increasing the number of micro-values with which a macro-value is measured will improve Summative Confidence, provided these micro-values measure distinct dimensions of the macro-value.
Figure 1. Example of the Values Used to Determine Whether a Faculty Should be Recommended for Tenure

Note: The numbers in the top right-hand corner of each box represent the weight assigned to the respective macro-value.
macrovalues. For example, if the three collegiality indicators are highly associated with each other they would contribute less unique information to the precision of the conclusion about collegiality than the three teaching indicators—which are probably only modestly related with each other—contribute to the precision of the conclusion about teaching, all other factors (e.g., weights, measurement error, scaling) being equal. Similarly, a high association between macrovalues, e.g., the two critically important macrovalues, will lower the degree of confidence of the synthesized group-level conclusion (e.g., all critically important values) compared to unassociated macrovalues.

Despite the implicit suggestion embedded in the previous two paragraphs, the relationship between values and Summative Confidence is not linear. If it was then one could simply improve the Summative Confidence of a conclusion by adding unrelated microvalues. For example, the faculty could have increased the Summative Confidence of their conclusion regarding the candidate’s teaching ability by adding microvalues such as shows up to class on time, turns grades in on time, liked by students, fails few students, and so forth. While each of these criteria is related with teaching, none of them are strong indicators of teaching proficiency. An even more extreme example would occur if the faculty included completely unrelated criteria (e.g., well groomed, attractive, tall) which would have altered both the conclusion and its precision. Therefore, although microvalues should be unrelated, they also need to be valid indicators of the macrovalue which they purport to measure. However, this dual standard is difficult to attain. In many instances, the best one

\[20\) In the case of survey development, one is taught to deliberately include related items in order to improve the accuracy and precision of the construct measured. However, the gain in precision associated with the addition of related items is entirely due to the unique information contributed by these items. That is, a measure with 10 items that are perfectly correlated with each other does not improve the precision with which the construct is measured beyond simply using the item with the lowest variance.

\[21\) Roughly speaking, this is analogous to increasing the coefficient of determination \((R^2)\) by including an excessive number of variables in the statistical model.
can hope for is a set of indicators that are marginally associated with each other and moderately associated with the macrovalue.

Another factor that would likely impact the precision of the final conclusion is the panel’s decision on a weighting scheme. Assume a 100 point weighting scheme was used in which 60 points were allocated to the critically important values and 40 points were allocated to the non-critically important values. Furthermore, assume the variances of microvalues located within a macrovalue are equal. The faculty also decided to redistribute these points evenly to every macro- and microvalue underneath the two groups. Some of the implications of these decisions include: critically important values had 1.5 times the impact on the precision of the summative conclusion than their counterparts, the Teaching macrovalue had a greater impact on the overall precision than the Research macrovalue because of the lower number of indicators with which it was measured, the Accolades microvalue had a greater impact on precision than any individual Research microvalue due to the distribution of weights among microvalues, and so forth. Finally, because the weighting scheme was not prescribed by the department or university, it would have a profound impact on the replicability of the decision, if there was great variability between the weighting schemes each faculty member generated before agreeing on the final scheme. Moreover, if the variances of microvalues are not equal, one will need to standardize the variables before creating a composite variable, assuming one wants each microvalue to contribute equally to the composite variable.

In addition to their agreement on a weighting scheme, the faculty agreed upon a set of standards for some of the critically important values. Specifically, they decided to not recommend the faculty for tenure if she did not have at least one publication per year in a peer reviewed journal and presented at a conference once every two years. The impact of a
standard on Summative Confidence cannot be summarized in a single statement. For one, the magnitude of the impact will depend upon the type of standard set. Scriven (2007) and Davidson (2005) have identified three types of standards: soft-hurdle, hard-hurdle, and bar. Essentially, these standards differ in the penalty exerted on the summative conclusion. However, all of them require ignoring some information about the performance of the evaluand on one or more dimensions of merit. For example, if the panel set a soft-hurdle on the frequency of conference presentations and the candidate failed this standard, the faculty would have to ignore all of the candidate’s presentations, essentially giving the candidate no credit for their performance on this microvalue. The penalty for failing a hard-hurdle is even more stringent. The faculty would have to ignore all of the candidate’s performance on the Research macrovalue. Likewise, in the case of the last standard, failure of a bar would result in the failure of the entire evaluand.

Clearly, these penalties can have a significant impact on a summative conclusion but what affect do they have on Summative Confidence? The impact on Summative Confidence of failure on a standard is similar to the impact failure has on the conclusion in that the impact of failing a soft-hurdle will be smaller than the impact of failing a hard-hurdle, which, in turn, will be smaller than the impact of failing a bar. More specifically, the Summative Confidence of a value on which the evaluand failed the standard is a function of only the evidence that supports failure (i.e., evidence of positive performance on the dimension(s) impacted by the standard is ignored). Hence, in the case of soft-hurdles, the confidence level associated with concluding the evaluand failed a specific criterion is a function of the precision with which the criterion is measured. Similarly, in the cases of hard-hurdles and bars, the confidence level associated with concluding the evaluand failed the macrovalue or evaluand is a function of the precision with which the composite of failed criteria are...
measured. Therefore, one may be more confident in concluding the evaluand failed when a greater number of criteria support this conclusion.

Even when the evaluand does not fail a performance standard, its affect on Summative Confidence may be observed through its impact on the measurement error. According to one of the principles of classical measurement theory, the reliability of a criterion-referenced test is a function of the discrepancy between one’s performance and the cutoff score (Crocker & Algina, 1986). The closer one’s performance is to the cutoff, the lower the reliability of the decision reached based on the test. Hence, it stands to reason, the closer one’s performance is to a standard, the lower the Summative Confidence. This, invites the possibility of setting really low standards in order to increase the Summative Confidence of a conclusion, which, of course, connotes that the standard was not really a standard at all. Moreover, the gain in confidence would occur at the expense of validity and such sacrifices should never be made.

Another factor, and perhaps the most important, that impacts Summative Confidence is measurement error. Although this factor has been mentioned on several occasions, the nature of its relationship with Summative Confidence has yet to be specified other than to state the two concepts are inversely related. Measurement error refers to the discrepancy between an observation and the true value of the entity being measured (Crocker & Algina, 1986). It is expressed either as the standard error of measurement, the standard error of estimation, the standard error of prediction, or as the reliability of a measure or method. The standard error of a measurement is an estimate of the average discrepancy between an observed score and the true score. Similarly, the standard error of estimation refers to the average discrepancy between an observed score and the predicted value. The standard error of prediction refers to the average discrepancy between an observed score and the observed
score on a parallel measure. Reliability, on the other hand, is the degree to which a method consistently reproduces the same result. Therefore, lower standard errors and higher reliabilities are each indicative of greater measurement precision.

Returning to the tenure review example, the measurement error of the collegiality macrovalue is the discrepancy between the candidate’s true collegiality and the degree of collegiality they were rated as possessing in their interactions with staff, students, and other faculty. Considering individuals may interact with people in a variety of ways, it would not be surprising if the estimate of the candidate’s collegiality had a modest amount of measurement error. Furthermore, despite the fact that only one indicator exists to measure the candidate’s professional accolades, this microvalue, most likely, would produce a more accurate estimate of the respective macrovalue than the synthesis of the three collegiality microvalues. The reason for this is because the measure of the candidate’s accolades would only require the counting of their awards—a list of which would not be difficult to obtain and verify. Therefore, little to no measurement error should exist, assuming agreement exists on what an accolade is. However, if the panel wanted to consider the prestige of each award, measurement error would be introduced into the estimate due to potential disagreements over the prestige of each award.

Measurement error exists whenever interpretation is necessary to transform data (in order to derive meaning) from one type into another, providing the transformation represents a loss of information. In such instances, measurement error refers to the degree of agreement over a set of interpretations or ratings. Two types of errors appear in the literature. Inter-rater reliability refers to the degree of consistency in the ratings of the same entity made by several raters, whereas intra-rater reliability, commonly called test-retest reliability, refers to the degree of consistency in the ratings of the same entity made by a
single rater. Both estimates presume the conditions under which ratings are made are as similar as possible; otherwise, the degree of consistency between ratings and raters would be a function of the precision of the instrument or method that produced the rating as well as any contextual factors that might influence the rating. Unfortunately, while it is fairly common practice for evaluators to report the reliabilities of their instruments and methods, no evidence exists that these estimates are utilized to adjust the confidence intervals of the parameters they calculate. This is particularly true whenever qualitative analysis is conducted because no probabilistic method exists for estimating a CI for a finding, unless the data is quantified and statistical procedures are conducted.

One way of combating heterogeneity, sampling error, and virtually any other factor that weakens Summative Confidence, is by increasing the sample size. Given a constant sampling design and a large enough sample, virtually any level of precision or confidence can be attained. However, while theoretically one can improve confidence up to 100 percent by adding to the sample, practical limits (e.g., cost) make this level virtually impossible to attain.

Finally, one of the most important factors to consider in a Summative Confidence analysis is alpha—the confidence level at which the Summative Confidence analysis is conducted. Setting alpha to 10 percent indicates that the analysis will calculate the confidence interval for the summative conclusion such that if the study was conducted ad infinitum, 90 percent of the calculated intervals would contain the true evaluative conclusion. However, alpha and the width of the confidence interval are inversely related. The lower the alpha, the wider the confidence interval will be and vice versa.

Understanding the confidence level of a decision potentially has great implication. To illustrate this point further, consider the following information: the tenure review panel

---

22 Although there are qualitative analogues to reliability, these methods cannot give the probability a qualitative finding is reliable without turning to established statistical procedures, which require quantitative inputs.
set a decision standard of 70 percent on the summative conclusion. In other words, they recommended the candidate for tenure if, and only if, the candidate’s overall performance score was 70 percent or higher. If the candidate received a score of 77 with a 90 percent CI that ranged from 74 and 79 percent then the provost could feel reassured in the reliability of the tenure review process. However, what if the Summative Confidence analysis produced a 90 percent CI that ranged from 60 and 85 percent? In this situation, the provost would have reason to question the reliability and validity of the process and ensuing decision. An alternative, and equally valid, method of interpreting Summative Confidence is to calculate the probability for a given confidence interval. For instance, the provost may not be as interested in knowing the confidence interval around the performance estimate as much as knowing the probability that the decision reached is correct. In this example, a decision to recommend for tenure is correct if the candidate’s true performance is 70 percent or higher. Therefore, the provost may wish to know, “what is the probability that the candidate deserves to be granted tenure (i.e., has a performance score of 70% or higher)?” If the probability turns out to be over 90 percent then the provost may conclude that the decision to grant the candidate tenure would likely be replicated by a different committee reviewing the same information (i.e., the decision is reliable). Yet another way of reporting the confidence one can place in a conclusion is to express it in the form of a reliability coefficient. For example, the reliability of correctly classifying the unknown true score of a candidate with an observed score of 77 given a 70 percent standard is 0.85. Notice then that all three alternative expressions are interrelated and can be used to express Summative Confidence.

---

23 Note then that confidence intervals do not have to be symmetrical around the summative conclusion. In fact, the only time a confidence interval is symmetrical is when the summative conclusion is equal to the median of the underlying distributions (e.g., 50 percent in this example).
So what has Summative Confidence taught the provost about her university’s tenure review process? One would imagine quite a lot. This case demonstrated that at every step during an evaluation, an evaluator is faced with choices that affect the precision of their conclusions. Even without actual data to compute the Summative Confidence of the case she reviewed, the provost would be able to gain insight into how she might improve the process. The biggest obstacle toward attaining a precise summative conclusion is variability. Therefore, one method of reducing variability is to standardize as much of the tenure review process as possible. For example, the university or each department could develop a tenure review policy and enforce the implementation of this policy. The policy should regulate which values will be examined in the review, the structure of these values, the validity of the structure, the organization of values into critically and non-critically important groups, the degree of association between values, the standards that would be used to judge acceptable performance for critically important values, the rubric that would be used to grade performance on values, the number of internal and external faculty that would serve on the panel, the methods and data sources that would be used to measure performance on each value, and the weighting scheme that would be used in data synthesis. Additionally, she should recommend that the policy address measurement error. For instance, the policy could require that the panel undergo training in coding qualitative data derived from documents, interviews, observations, and so forth. Finally, she should recommend that the panel gather as much input (i.e., increase the sample size) as possible, using a systematic and reliable data gathering process, on subjective values (e.g., collegiality).

Dissertation Chapters and Objectives

This dissertation has set forth several ambitious objectives. To the best knowledge of this author, no study has ever been published that reported the CI of a summative evaluative
This claim is supported by (1) several keyword searches in 44 scholarly databases that failed to produce a single article wherein a confidence interval was calculated for the type of complex variables associated with summative evaluations, and (2) discussions with several evaluators of noted importance, including past presidents of the American Evaluation Association. However, the statistical and psychometric foundations underlying Summative Confidence date back more than a century.

Chapter 2 will present a summary of the scientific literature upon which Summative Confidence is founded. The chapter is partitioned into three sections. The first section will summarize the historical antecedents that inspired this author to find a way to integrate them into a unified theory and method. Key statistical concepts, formulas, theoreticians and dates will be presented, when known. The purpose therein is to illustrate how these theories can be integrated into what eventually will become the master formula underlying Summative Confidence. The second section will follow suit from the first. However, it will focus on the historical antecedents provided by psychometric theory. Along the way, both sections will highlight gaps in the scientific literature so as to foreshadow the theories that will need to be derived. Lastly, the third section will focus on issues pertaining to the use of qualitative data in Summative Confidence. Specifically, it will attempt to ameliorate the concerns frequently raised by qualitative researchers at the thought of quantifying qualitative data.

Although complex ideas can be more readily grasped when accompanied by a real-world example, given the sheer number of statistical and psychometric concepts presented herein, the scope of this study precluded the author from amassing a volume of real-world studies. Instead, Monte Carlo simulations will be employed to generate data that can be used to illustrate both principles and the computational mechanics. Chapter 3 will introduce readers to Monte Carlo methodology, including advanced topics such as generating random
variables with a specific skewness, kurtosis, and correlation matrix. The purpose of this chapter is to provide readers with the confidence that the more than 100 simulations employed in subsequent chapters are in fact generating random data with known population parameters. The fact that the population parameters are known—an attribute absent from real-world data—allows one to determine how well statistical and psychometric theory, and sample statistics are able to approximate these values.

Chapter 4 will begin to lay the statistical theory upon which Summative Confidence is founded. Statistical proofs are provided for select concepts, particularly proofs of new concepts that were derived by the present author. Proofs for omitted concepts can be found in standard textbooks on statistical theory. Hence, the principles discussed therein are not a matter of opinion, but rather, are established statistical laws. The chapter is composed of six sections. The first section will formally define the concepts of expected value and variance for single and composite variables. The second section will introduce readers to standard discrete and continuous probability distributions. The third section will present several options evaluators can use to transform dissimilar distributions into a common distribution en route to constructing a composite variable. The fourth section will discuss how to adjust the basic variance formulas to account for sampling error. The fifth section will present three different methods for constructing a basic confidence interval, including two nonparametric methods. Lastly, the sixth section will present a statistical algorithm for constructing a basic Summative Confidence interval.

While the previous chapter provides the statistical theory foundation necessary to derive the basic formula underlying Summative Confidence, it assumes that measurement error is nonexistent. However, this assumption is false since no instrument measures with perfect accuracy or precision. Hence, measurement error must be integrated into the basic
Summative Confidence formula. Chapter 5 lays down the psychometric theory upon which Summative Confidence is founded. As was the case for the previous chapter, proofs for new theorems are presented therein so as to reinforce the fact that Summative Confidence is derived from mathematical law and not the expressed opinion of the present author. The chapter is partitioned into two sections. The first section focuses on exploring the impact of measurement error on variance, from which will emerge the master formula that can be used to construct a Summative Confidence interval or to compute a Summative Confidence reliability coefficient. Since measurement error can be accounted for by the parallel reliability coefficient, the second section will present various indices that have been historically used to measure parallel reliability, including one developed by the author.

Finally, Chapter 6 will summarize the aforementioned chapters culminating in the master formula and algorithm necessary for conducting a Summative Confidence analysis. A comparison will be made between the list for factors presented in this chapter and the terms of the master formula to clearly delineate how each factor is accounted for. Only 8 of the 11 factors are accounted for in the master formula. Although their likely mathematical impact on Summative Confidence can be surmised, construct validity and the variability in the weighting scheme and standards were left out in order to limit the scope of the present study to a reasonable level. A list of the contributions by the present study along with a discussion of the implications of Summative Confidence for the evaluation discipline will be presented. Finally, the chapter will conclude with a list of research areas that have yet to be integrated into Summative Confidence.

Relevance

The purpose of a summative evaluation is to examine the performance of an evaluand on a set of values and to compare this performance to relevant standards in order
to render a summative conclusion. However, without knowing the amount of measurement error that impacted the conclusion, an evaluator cannot gauge the precision of the conclusion nor can a decision-maker determine whether actions are warranted to address the issues that produced the conclusion. Furthermore, in situations in which funding allocation or the viability of the evaluand is to be decided upon by decision-makers, it is reasonable that they would need and want to consider the quality of the evaluative conclusions prior to reaching a decision. Thus, the premise of this dissertation is that evaluators must begin to report the precision of their conclusions and to the extent possible, take steps during the planning phase to ensure that adequate confidence will be attained for each conclusion.

Unfortunately, no studies have ever been published, to the best knowledge of this author, that report the CI of a complex variable despite the fact that much, albeit not all, of the theoretical foundation underlying Summative Confidence has been known for over a century. For example, basic principles of expectation and variance as well as the impact of measurement error, sampling error, and inter-rater reliability, to name a few of the relevant factors, on evaluation and research results are routinely ignored. At most, evaluators may include a list of limitations in their narrative. However, they do not assess the mathematical impact of these errors on either their results or ability to generalize beyond their sample. Finally, researchers and evaluators are often told that they should triangulate their results in order to improve the validity of their conclusions. However, while it stands to reason that more information is better, a question that has yet to be addressed is “How much data is enough?” Clearly, the answer to this question depends upon both the quality and quantity of information employed in the formulation of evaluative conclusions.

This dissertation is clearly relevant to the discipline of evaluation. In a world in which billions of dollars are spent annually on conducting evaluations, the need to maintain
high standards can be overwhelming. To date, poor evaluations can be unearthed through a *metaevaluation*—an evaluation of one or more evaluations for the purpose of determining the merit and worth of the original evaluation(s) (as opposed to the evaluands). However, the cost and time of properly conducting a metaevaluation can be considerable, at times even comparable to the cost and time of the original evaluation. Furthermore, few evaluators have the necessary expertise to conduct such studies. Consequently, according to Dr. Michael Scriven (personal communication, February 28, 2007), the author of metaevaluation, the proportion of metaevaluations conducted is significantly less than 1% of all completed evaluations. Although a Summative Confidence analysis is not intended to replace a metaevaluation, it can act as a barometer of the quality of the evaluation. Even better, it is more cost effective than a metaevaluation.

Furthermore, a Summative Confidence analysis has no data restrictions. It can be used with quantitative, qualitative, and mixed method designs. Its only restriction with regard to qualitative data is that steps must be taken to measure the reliability of interpretations and ratings, and to develop a classification system or grading rubric. However, these constraints are not overly burdensome requirements considering they should—although they are often not—already be conducted as part of an evaluation. The fact that such reliability analyses are often omitted calls into question the validity of the conclusions reached. After all, one cannot establish validity without first establishing reliability. Finally, the scope under which this analysis can be conducted spans every evaluation sub-disciplines (e.g., metaevaluation, personnel evaluation, policy evaluation, product evaluations, program evaluation).

**Delimiters**

It is important to note that Summative Confidence is a method for determining the probability that a result will replicate given parallel conditions. Consequently it is related to
validity because replicability is a necessary but not sufficient condition for establishing validity. That is, if identical design conditions are established, then one should expect to observe similar results. The results produced by a Summative Confidence analysis, however, do not imply that the data and/or methodologies used to collect the data were valid or complete. Nor does it imply that the list of values and standards were complete or valid for addressing the purpose of the evaluation. It also does not suggest that the weighting scheme and scoring rubric were appropriate. These are all factors that must be independently validated by the evaluator. Summative Confidence is simply an expression of the precision of a conclusion within a specific probability given that the same methodologies, data sources, and procedures are used to replicate the evaluation. If the methods used to formulate a conclusion are invalid, then the confidence level surrounding the conclusion is also invalid. Therefore, a necessary precondition to conducting a Summative Confidence analysis is the validation of the methods employed by the evaluation. A secondary precondition is the determination of the suitability of the data for conducting a Summative Confidence analysis. However, these preconditions fall outside the scope of this study.
CHAPTER II
HISTORICAL ANTECEDENTS

Historical antecedents can be found for the statistical algorithm underlying Summative Confidence. Some of the statistical and psychometric theory upon which it is founded can be traced back hundreds of years, if not farther. This chapter will present the key formulas and concepts without proofs, which will be covered in greater depth in subsequent chapters. The primary purpose herein is to provide an overview of the knowledge bases one needs to synthesize into a coherent theory in order to be able to derive the statistical formula underlying Summative Confidence. That is, the purpose is not to assemble the existing knowledge into a master formula but rather to show that much of the theoretical framework for developing Summative Confidence already exists. The secondary purpose is to identify gaps in existing knowledge that will need to be filled prior to the synthesis of a coherent theory.

The goal of Summative Confidence is to account for the various types of variability that can impact a decision. Although the previous chapter outlined 11 potential factors, this list is likely to grow over time. Presently, this study will focus on 8 factors: (1) alpha level, (2) number of values (variables)\textsuperscript{24}, (3) variance of each variable, (4) correlation among the variables, (5) the impact of a fixed standard on the reliability of the corresponding result, (6) sample size of each variable, (7) amount of measurement error for each variable, and (8) weight assigned to each variable. The variance associated with standards and weights, and

\textsuperscript{24} Since the remainder of this study will focus on statistical and psychometric theory, henceforth, the term variable will be employed in place of value, so as to eliminate confusion.
organizational structure will be omitted from this study. However, the statistical theory presented in subsequent chapters will most likely permit readers to derive the statistical formula associated with the first two omitted factors, whereas the latter factor cannot be incorporated into a master formula. Instead, one will need to manually determine how the organizational structure of the values impacts the weights of each individual value.

The chapter is divided into three sections. The first two sections will provide a brief introduction to select topics in statistical and measurement theory, whereas the last section will delve into the quantification of qualitative data. Although the last section deviates from the foundation of the theoretical framework needed to buttress Summative Confidence, in some ways, it is just as important since the vast majority of evaluations entail the collection and analysis of at least some qualitative data. Therefore, the omission of qualitative data from Summative Confidence would render the method of little interest to the majority of evaluators. While the solution for dealing with qualitative data is rather simple—quantify the data, compute a reliability coefficient, and then treat it as usual in data analysis—it has been met with great resistance within the community of qualitatively-oriented researchers and evaluators (Stenbacka, 2001). Therefore, the third section will evaluate the merits of the objections—e.g., Stenbacka (2001) and Miller (2008)—frequently raised to the notion of quantifying qualitative data. It will be argued that the mechanism by which qualitative data is quantified emerges from qualitative theory and not quantitative practice. Moreover, qualitative research cannot exist absent a system of classification. Thus, the only additional steps needed to utilize qualitative data in a Summative Confidence analysis entail classifying units of qualitative data based on specific definitional parameters (i.e., a coding scheme), evaluating their merit based on an agreed scheme for assigning value to a category, and computing the reliability of the procedure employed to classify the qualitative data.
Statistical Theory

At its heart, Summative Confidence is just a confidence interval. Therefore, it is only natural to begin the review of the literature there. In general, researchers rarely have access to all the data in a population. When they do have such access, they can compute parameters (e.g., mean, median, variance) with 100% confidence that they represent the entire population. In the vast majority of instances, however, evaluators only have access to a portion of the population data. Hence, they can never be 100% certain that their sample statistics represent the population parameter. Confidence intervals were developed by Neyman (1937) to provide researchers with a method for reporting the uncertainty associated with using sample data to estimate population parameters. No doubt most readers are acquainted with the basic parametric formula for a $100(1-\alpha)\%$ CI on the sample mean $\bar{X} \pm z_{1-\alpha/2}\sigma_{\bar{X}}$ or more appropriately, $\bar{X} \pm t_{1-\alpha/2,n-1}SE$, where the population standard error is $\sigma_{\bar{X}}=\sigma/\sqrt{n}$, the sample standard error is $SE=S/\sqrt{n}$, and $\sigma$ and $S$ denote the population and sample standard deviation. Notice then that the basic CI formula incorporates three of the aforementioned factors. Namely, the Type I Error rate denoted by alpha ($\alpha$), the sample size ($n$), and the variance ($\sigma^2$ or $S^2$), which is equal to the square of the standard deviation ($\sigma$ or $S$).

But, how does one incorporate the other five factors?

A logical place to start is with the composite score and its corresponding variance. In the case of a single variable, a composite score cannot be computed since there is only one variable. However, one can estimate the variance associated with employing a point estimate of the central tendency—like the mean, sum, or median—to estimate the corresponding population parameter. Generally, researchers prefer to work with either the mean or sum because the Central Limit Theorem describes the sampling distribution of these estimators and statistical inference on them is simpler than on the median. However, as will be
discussed later, the median is needed in some circumstances. In the case of compound variables, especially the complex variables encountered in evaluation studies, one needs to compute an estimate of the central tendency (multivariate centroid) across all the variables of interest. It is only rational then to surmise that the composite score of a set of variables can be expressed by the mean or total of these (constituent) variables (i.e., $\bar{X}=k^{-1}\sum X$, or $T=\sum X$, where the composite score is a function of $k$ variables labeled $X_i, i \in \{1,\ldots,k\}$).

Moreover, one can incorporate a weighting scheme by multiplying each constituent variable by its corresponding weight. Hence, the weighted mean $\bar{X}_w$ and total $T_w$ of $k$ variables are given by $\bar{X}_w=k^{-1}\sum w_i X_i$, or $T_w=\sum w_i X_i$, where $w_i$ denotes the weight assigned to the $i^{th}$ variable.

In a similar fashion, if variance is the key to constructing a CI for an individual variable then perhaps the variance of a composite, such as the mean or total, is a function of the individual variances. In fact, this turns out to be the case. It has long been known that the variance of a composite total is equal to the sum of the variances (denoted as Var or $\sigma^2$) plus twice their covariances (denoted as Cov) (Crocker & Algina, 1986; Meyer, 1970; Larson, 1973; Ross, 2002). That is, for $k$ dependent variables (labeled $X_i$ and $X_j$), the composite variance is given by $\text{Var}(T)=\sum \sigma^2_{X_i}+2\sum \text{Cov}(X_i,X_j)$, where $\text{Cov}(X_i,X_j)=\rho_{ij}\sigma_i\sigma_j$ and $\rho_{ij}$ denotes the population correlation between variables $X_i$ and $X_j$. Then, for $k$ independent variables, the composite variance is given by $\text{Var}(T)=\sum \sigma^2_{X_i}$, since the correlation between two independent variables is zero. Moreover, one may incorporate a weighting scheme by multiplying each individual variance by the square of its corresponding weight and the covariances by the product of the corresponding weights (Ross, 2002; Larson, 1974). That is, for $k$ dependent variables, the variance of the weighted composite total is given by $\text{Var}(T_w)=\sum w_i^2\sigma^2_{X_i}+2\sum w_i w_j \text{Cov}(X_i,X_j)$, where $w_i$ and $w_j$ denote the weight assigned to the $i^{th}$ and $j^{th}$ variables, respectively. Although the origin of this property is difficult to trace back,
according to some scholars (Heyde & Seneta, 1972), it was known as early as 1853 by Irénée-Jules Bienaymé.

The variance of the composite mean, however, differs from that of the composite total and is frequently omitted from applied statistics textbooks. However, it is well-known in mathematical statistics (Larson, 1974; Ross, 2002) that it is equal to the sum of the variances and covariances divided by \( k^2 \). Therefore, the variance of the weighted composite mean for \( k \) dependent variables is given by

\[
\text{Var}(\bar{X}_w) = \frac{\sum w_i^2 \sigma^2 X_i + 2 \sum \sum w_i w_j \text{Cov}(X_i, X_j)}{k^2}.
\]

Note, this formula computes the variability within the composite mean across all the sampling units. That is, given 100 cases and 10 variables, the formula computes the variance among the 100 cases after the 10 variables are averaged—resulting is a single vector with 100 values. Furthermore, since evaluators are primarily interested in rendering a conclusion about the evaluand (e.g., program) and about individual sampling units (e.g., people), they typically report summary sample statistics, like the grand (sample) mean. However, sample statistics may deviate from population parameters. For example, if the 100 cases were drawn from a population of 1,000 cases, the sample grand mean computed across the 100 cases and 10 variables may differ from the population grand mean computed across the 1,000 cases and 10 variables. The weighted composite variance \( \text{Var}(\bar{X}_w) \) represents the variance of the sampling distribution that would be generated if one were to repeatedly randomly draw, ad infinitum, sample data from a population and compute a sample grand mean.

Fortunately, the impact of sample size (i.e., number of sampling units) on variance is quite simple. The variance of the mean of a single variable, known as the sampling error variance \( \sigma^2_{\bar{X}} \), is equal to the variance divided by the sample size, \( \sigma^2_{\bar{X}} = \sigma^2 / n \) (Larson, 1974; Ross, 2002). From here, it is reasonable to conclude that the sampling error variance of a composite mean \( \bar{X}_w \) (i.e., the sampling error variance for the grand mean) generated from \( k \)
independent constituent variables (i.e., covariance is zero) with sample size \( n_i \) is equal to
\[
\text{Var}(\bar{X})_i = \left( \sum w_i^2 \sigma^2_{X_i} / n_i \right) / k^2.
\]
However, the sampling error variance of a composite mean generated from \( k \) dependent constituent variables is not immediately obvious because of a gap in the statistical literature. Namely, what is the sampling error covariance of a composite mean? Intuitively, one may guess that it should be equal to \( \text{Cov}(X_i, X_j) / n \) since \( \text{Cov}(X, X) = \text{Var}(X) \) and the sampling error variance is equal to \( \sigma^2_{\bar{X}} = \sigma^2 / n \). Hence, the sampling error covariance of \( X_i \) and \( X_j \) must reduce to \( \sigma^2_{\bar{X}} \) when \( X_i \) is identical to \( X_j \) (i.e., \( X_i = X_j \)). However, this is simply an educated guess and in absence of a proof holds little value.

Suppose though that the error variance of a composite mean \( \bar{X}_i \) generated from \( k \) dependent constituent variables is in fact
\[
\text{Var}(\bar{X})_i = \left\{ \sum w_i^2 \sigma^2_{X_i} / n_i + 2 \sum \sum w_i w_j \text{Cov}(X_i, X_j) / n_i \right\} / k^2,
\]
where \( i < j \) and the second term assumes that \( n_i \) is the shared sample size between variables \( X_i \) and \( X_j \). Then, the formula incorporates 5 of the 8 factors: the number of variables (denoted by \( k \)), the variance of each variable (denoted by \( \sigma^2_{X_i} \)), the correlation among variables (embedded in the covariance terms), and the weighting scheme (denoted by \( w_i \) and \( w_j \)).

A sixth term can be incorporated by adjusting each variance estimate for sampling error. That is, just as the sample mean and total are the best guess of the population mean and total, so too is the case with the sample variance. It is well-known that one can adjust the sample variance \( S^2 \) by multiplying it by the finite population correction (fpc), \( 1 - n/N \), where \( n \) denotes the sample size and \( N \) denotes the population size (Kish, 1995; Lohr, 1999; Hays, 1994). Hence, the fpc-adjusted weighted error variance for \( k \) independent variables is given by
\[
\text{Var}(\bar{X})_i = \left\{ \sum w_i^2 \left( 1 - n_i/N \right) S^2_{X_i} / n_i + 2 \sum \sum w_i w_j \left( 1 - n_i/N \right) R_{ij} S_{X_i} S_{X_j} / n_i \right\} / k^2,
\]
where \( R_{ij} S_{X_i} S_{X_j} \) denotes the sample covariance, which, in turn, is comprised of the sample correlation between the \( i \)th and \( j \)th variables, and the sample standard deviations \( S_i \) and \( S_j \).

A search of the statistical literature did not yield a formula for the error covariance of a mean.
By omitting the proofs of each of the aforementioned statistical properties and making an intuitive guess at what the error covariance of a mean might be equal to, one can arrive fairly quickly at a formula that incorporates most of the eight factors. Specifically, the error variance can be substituted into the CI formula by replacing $\sigma^2$ or SE with the square root of $\text{Var}(\bar{X}_e)$, which connotes that a $100(1-\alpha)\%$ CI on the weighted composite sample mean is given by $\bar{X}_w \pm z_{1-\alpha/2} \sqrt{\text{Var}(\bar{X}_e)}$ or more appropriately, $\bar{X}_w \pm t_{1-\alpha/2, n-1} \sqrt{\text{Var}(\bar{X}_e)}$, when $n$ is small ($<40$). Notice, this confidence interval definition also incorporates alpha. Hence, the only factors not incorporated at this point are measurement error and the impact of a fixed standard, which will both be covered in the next section. However, the confidence interval definition makes an assumption that requires further examination. Namely, by employing the $z$ and $t$ statistics, the CI assumes that the distribution underlying the weighted composite is normal. Yet, it is widely-known by experienced evaluators that most of the variables they employ are not normally distributed. In fact, a study by the present author of 18 evaluations containing a total of 1,798 variables found that very few variables ($<5\%$) passed a normality test. Therefore, constructing a Summative Confidence interval on the composite requires deeper knowledge of distribution theory than simply assuming it is normally distributed or only that the distribution of the composite be known.

The study of statistical distributions dates back hundreds of years. To date, dozens of distributions have been discovered and employed in practice. However, based on the experience of this author, evaluators are likely to only encounter a handful of distributions in their practice. Specifically, the Bernoulli, binomial, multinomial, uniform (discrete and continuous), Poisson, and normal probability distributions appear with great frequency in evaluation studies. Discovered by the Swiss mathematician James Bernoulli and published posthumously in his *Ars Conjectandi* in 1713 (Ross, 2002), the Bernoulli distribution is the
probability distribution underlying a single trial or process in which there are only two possible outcomes: success or failure. Similarly, the binomial distribution is the probability distribution underlying the sum of multiple Bernoulli trials or processes (i.e., a count of the number of successes). Not surprisingly, the origin of this distribution also traces back to Bernoulli (Grinstead & Snell, 1997). In contrast, the origin of the multinomial distribution has been lost to history although it is clear that its origin arises from an extension of the binomial distribution. Specifically, the multinomial distribution is encountered whenever one sums two or more categorical distributions—a trial or process in which there are two or more possible outcomes (e.g., Likert scales, ordinal data, nominal data).

Similarly, the exact origin of the uniform distribution has also been lost to history. However, the discrete uniform distribution is encountered whenever all the elements of a sample space are equally probable (e.g., rolling a fair die, selecting a card from a well-shuffled deck), whereas the continuous uniform distribution is encountered whenever all the points in a sample space interval are equally probable (e.g., amount of time or distance between two points). Discovered by the French mathematician Siméon D. Poisson in 1837 (Ross, 2002), the Poisson distribution transpires whenever a Poisson process occurs in which one counts the number of Bernoulli trials that occur within a continuous interval of measurement (e.g., time, length, area, volume). Generally, the distribution is employed whenever rare events occur within a continuous process. Lastly, the normal distribution was discovered by the French mathematician Abraham DeMoivre in 1733 but was later popularized by the German mathematician Karl Friedrich Gauss (Walpole, Myers, Myers, & Ye, 2007; Ross, 2002). Without doubt, the normal distribution is the most widely used probability distribution thanks to (1) the frequency with which naturally occurring variables conform to it (e.g., height, weight, intelligence), (2) the fact that it is the limiting distribution for several of the
aforementioned distributions as sample size goes to infinity, and (3) its role in the Central Limit Theorem (CLT) (Ross, 2002).

A more in-depth discussion of each of these distributions will be presented in Chapter 4. Suffice to say, however, that given their long history, the formula, properties, and application of each of these distributions are well-known to statisticians and can be found in virtually any probability textbook, including Bartoszynski and Niewiadomska-Bugaj (2008), Grinstead and Snell (1997), Larson (1974), Meyer (1970), Ross (2002), and Walpole, Myers, Myers, and Ye (2007). Unfortunately, knowledge of distribution theory, in and of itself, does not enable one to construct a confidence interval for a complex variable, with one exception. If all the constituent variables conform to the same distributional family then one generally knows the distribution of the composite. For example, if all the constituent variables conform to a Bernoulli distribution than the sum of these variables must conform to a binomial distribution. Similarly, the sum of two or more binomial random variables also conforms to a binomial random variable, the sum of two or more Poisson random variables conforms to a Poisson distribution, and the sum of two or more normal random variables conforms to a normal distribution. This is known as the convolution of probability distributions (Hogg, McKean, & Craig, 2005). However, in general, these relationships do not hold if one computes the mean of two or more random variables from the same distributional family to form a composite (e.g., the mean of two binomial distributions does not conform to a binomial distribution). Given that evaluators are more likely to be interested in the mean, rather than the total, a more complex process by which one can determine the distribution of the composite mean is needed.

However, the problem is far more complicated than this because rarely do all the constituent variables in an evaluation conform to the same distributional family. Hence, what
one really needs to know is, what is the distribution of the composite mean when the constituent random variables conform to different distributions? Moreover, what if the distribution of one or more of the constituent random variables is unknown? At first glance, the CLT appears to address to these question. Traditionally, the CLT states that the mean of a set of independent and identically distributed (iid) random variables with finite means and finite nonzero variances (i.e., greater than zero but less than infinity) will be asymptotically (approximately) normally distributed as the number of random variables approaches infinity (Ross, 2002). Moreover, the CLT can also be extended to only independent random variables (i.e., the identically distributed condition is relaxed) (Ross, 2002). Furthermore, it has even been extended to correlated random variables (Godwin & Zaremba, 1961; Cocke, 1972). The question, of course, is at what point does the CLT kick-in? More specifically, given a specific set of conditions, how many constituent variables does one need to aggregate in order for the composite variable to be reasonably normal? This question has yet to be addressed by the literature and so, will be taken up in the next chapter.

Despite the ostensible insolubility of this conundrum, humans are generally able to evaluate apples and oranges in order to formulate an evaluative conclusion or decision. They do so, of course, by comparing the evaluands (in this case the two fruit) on one or more common properties. For example, when deciding whether to purchase an apple or an orange, a consumer may consider their price, tastiness, ripeness, the time of season, the store, and so forth. The ability to combine these variables into a composite rests upon whether a common scale can be found for all of the variables. For instance, one can directly compare the apple to the orange to determine which of the two rates higher on desired properties. Such a comparison would result in a prefer or not prefer decision for each property. Statistically, this is akin to a series of Bernoulli trials, which can then be combined
into a composite variable. That is, by transforming the underlying distribution of each random variable into a common distribution (binomial), one can then proceed as usual.

Although an infinite set of transformations are conceivable, the most common transformations can be grouped into one of several classes. The prior transformation can best be classified as dichotomization for it partitions the original probability distributions into two parts. Alternatively, an evaluator may elect to partition the original probability distribution into several parts, known as polychotomization, thereby generating a series of categorical distributions. As mentioned before, a composite variable can be generated from these transformed constituent variables, which would conform to a multinomial distribution. Alternatively, one can transform any continuous distribution into another distribution via the probability integral transformation method (Larson, 1974; Bartoszynski & Niewiadomska-Bugaj, 2008). However, given the complexity this method entails, it is unlikely many evaluators will elect to traverse this road. Regardless of the transformation chosen, the process of transforming the original probability distributions into a common distribution will result in the loss of some information. This is simply the price one must pay for the ability to add apples and oranges in such a way so as to produce an interpretable result.

Besides the fact that the specific nature of this transformative process has not been extensively discussed in the statistical literature26, evaluators are faced with a potentially thornier issue. In the interest of preserving as much of the original information of a random variable as possible, evaluators are likely to utilize polychotomous transformations. Although such transformations are known to unequivocally produce categorical distributions27, the

---

26 Mathematical-statistics textbooks teach how to transform a random variable with a known distribution into another random variable, but tend to focus on one-to-one transformations. None of the aforementioned transformations were one-to-one, except for the integral transform method.

27 A categorical distribution does not imply that the random variable has a nominal level of measurement although it can. Variables with an ordinal level of measurement also conform to a categorical distribution.
level of measurement associated with these distributions has been the subject of great debate (Velleman & Wilkinson, 1993). The theory of measurement scales was introduced by Stanley Stevens (1946), who classified measurement scales into four types: nominal, ordinal, interval, and ratio. Distinctions between these four measurement scales are governed by the presence or absence of 10 mathematical properties, enumerated in McDonald (1999, p. 409). Building progressively upon each other, nominal scales only require the existence of an equivalence rule, ordinal scales require the additional existence of a dominance rule, interval scales require the additional existence of a combination rule, and ratio scales require the addition of a null object (McDonald, 1999).

This means that for nominal scales, one-to-one transformations\(^{28}\) (Bartoszynski & Niewiadomska-Bugaj, 2008) and frequency counts\(^{29}\) are the only permitted mathematical operations. For ordinal scales, order preserving (monotone) transformations\(^{30}\) (Lord & Novick, 1968; McDonald, 1999), frequency counts, and quantiles (e.g., the median) are the only permitted mathematical operations. That is, given a numeric scale, statistics that require the addition, subtraction, multiplication, or division of scale points by anything other than a constant are strictly forbidden; therefore, the dilemma of how to handle polychotomization. For interval scales, only linear transformations are permitted for the interval scale because

---

\(^{28}\) If a transformation \(x \rightarrow g(x)\) is one-to-one then every value of \(x\) maps onto a single value of \(y\) and each value of \(y\) is associated with a single value of \(x\).

\(^{29}\) Frequency counts, however, require a slight change in variables. An entity or element can either be classified in a nominal category or it cannot. Once one counts the number of entities or elements that can be classified in a nominal category, one has changed the variable. In the case of the former, the variable is binary and conforms to a categorical distribution, whereas in the case of the latter the variable is a count and most likely conforms to a Poisson distribution.

\(^{30}\) A monotone transformation is one that satisfy either condition \(u>v \Rightarrow g(u)>g(v)\) or \(u<v \Rightarrow g(u)>g(v)\), where \(u>v\) denotes a strictly increasing function and \(u<v\) denotes a strictly decreasing function (Casella & Berger, 2002). That is, the transformation function \(g\) either preserves the order of the original values or it completely reverses their order. If a monotone transformation \(x \rightarrow g(x)\) is also strictly increasing or decreasing (i.e., no values are ever equal) then there exists a one-to-one relationship between the values \(x\) for random variable \(X\) and \(y\) for random variable \(Y\), where \(Y=g(X)\). Generally, this only occurs for continuous random variables.
“they preserve the equality of differences of scale scores” (Lord & Novick, 1968, p. 21). Therefore, frequency counts, quantiles, addition, and subtraction of scale points are all permissible mathematical operations but multiplication and the ratio between scale points is not meaningful. However, multiplication and division of differences between scale points is meaningful since a true zero for these differences exist. For ratio scales, multiplicative transformations are permitted since they preserve the equality of ratios between scale points (Lord & Novick, 1968). Hence, all mathematical operations are permissible for ratio scales. Finally, it is important to note, that while one can always transform high order scale to lower order scale (e.g., transform an interval scale to ordinal), the reverse is not possible.

Not surprisingly, researchers have sought a way to rescue certain measurement scales, like the Likert, from being relegated to the status of an ordinal scale ever since Stanley Stevens introduced his four levels of measurement. The motivation of these researchers is driven by the fact that few mathematical operations can be performed on ordinal measures, unlike interval and ratio measures. Generally, they argue or otherwise assume that Likert scales fall somewhere between true ordinal and interval scales since subjects self-center their responses and in so doing satisfy the equidistance property that differentiates between ordinal and interval levels of measurement. However, this assumption is never scrutinized by

---

31. This implies that equal distances with respect to the property of interest exist between point intervals of equal length on the metric scale. For example, the difference between 80°F and 100°F Fahrenheit is equal to the difference between 0°F and 20°F Fahrenheit since $100°F−80°F=20°F−0°F$. In contrast, because ordinal scales, like the Likert, do not assign labels, symbols, or numeric values in such a way as to satisfy the observed identity, they fail to satisfy the conditions required of interval scales.

32. Whenever a transformation $g^*(x)$ does not have a true zero then $g^*(x)=a(g(x))+β$, where the scale unit (e.g., Fahrenheit degrees) $a>0$ and $β$ is unknown (Bartoszynski & Niewiadomska-Bugaj, 2008). Hence, multiplication and division between scale points, $g^*(x)\times g^*(y)$ and $g^*(x)/g^*(y)$ respectively, is not inherently meaningful.

33. Notice that $g^*(x)−g^*(y)=[a(g(x))+β]−[a(g(y))+β]=a[g(x)−g(y)]$. Since the term $β$ drops out, the ratio of two differences is interpretable since the value zero is interpretable (it denotes the absence of any difference).
verifying, via polytomous item response theory (IRT) models\textsuperscript{34}, that the distances between the threshold parameters associated with scale points are equal.

Despite the clear violations of the statistical limitations associated with ordinal measures, researchers have continued to employ them in their research and treat them as interval (Lord & Novick, 1968; McDonald, 1999; Bartoszynski & Niewiadomska-Bugaj, 2008). In large part, this inclination has been driven by the dearth of statistical procedures capable of properly handling ordinal scales. Given present knowledge, polychotomization of dissimilar distributions in order to generate a common distribution will leave evaluators with two poor options. First, one can assume that equidistance exists between the scale points of the transformed distribution in which case they can proceed as usual and construct a CI on the composite variable. Given enough scale points, say more than 15 (Jöreskog & Sörbom, 1999), it may be safe to treat an ordinal scale as though it were interval. Unfortunately, the amount of work that would be required to manually transform a distribution into 15 categories is considerable, which is compounded by the fact that most evaluations employ a large number of variables. Moreover, polychotomization to this extent is likely only possible if the original distribution is continuous. Furthermore, employing more than five to seven scale points may have an adverse impact on the reliability of the instrument.

The second option is to construct a nonparametric CI based on continuous order statistics (Hogg, McKean, & Craig, 2005). Software, like SAS PROC UNIVARIATE with the CIQUANTDF option (SAS Institute, 2007), can perform this analysis. This analysis requires replacing the composite mean or total with the median, the appropriate measure of location for ordinal measurement. However, this analysis was developed for continuous data. Hence, any CI constructed on the transformed distribution following polychotomization is

\textsuperscript{34} Examples of polytomous IRT models include the partial credit model, generalized partial credit model, and graded response model (Embretson & Reise, 2000).
likely to be inaccurate, if the distance between the scale points of the transformed variable are not equidistant with respect to the latent (raw) distribution. What is needed is a nonparametric method for constructing a CI on the sample median of discrete or ordinal data. The theory of order statistics has come a long way since it was first developed in the 1960s and the theory of discrete order statistics now exists (Arnold, Balakrishnan, & Nagaraja, 2008) that may allow one to derive a statistical method for constructing nonparametric CIs for discrete and ordinal data.

Even if one could derive a method for constructing a nonparametric CI for discrete and ordinal variables, one is still not able to compute $\text{Var}(\bar{X})$ due to the absence of estimates for the error variance. Unlike parametric CIs that compute the boundaries of the CI based on the standard error, nonparametric CIs compute these boundaries based on order statistics. Fortunately, a solution has been proposed by McKean and Schrader (1984) for estimating the error variance for continuous variables from the boundaries of the nonparametric CI. However, if the error variance can be computed from the boundaries of a continuous nonparametric CI then surely it can also be estimated from the boundaries of a discrete or ordinal nonparametric CI. Given these estimates then one can combine them in the usual way in order to compute the composite error variance.

Measurement Theory

The previous treatise is established on a false premise that measurement error does not exist. Yet, since random, and sometimes even systematic, error exists all around us, no instrument is capable of measuring the properties of an object with perfect precision or accuracy. However, in the natural sciences, the instruments have become so sophisticated that measurement error can be virtually ignored. For example, recently physicists have uncovered evidence that the electron may be nature’s “perfect sphere” by employing an
instrument to measure its diameter within less than $10^{-28}$ of a centimeter (Hudson, Kara, Smallman, Sauer, Tarbutt, & Hinds, 2011). While not all natural science research enjoys such precision, in general, the measurement error in the natural sciences is so low that, for all practical purposes, it can be ignored (i.e., assumed to equal zero). Needless to say, the social sciences do not enjoy such advantages. Therefore, a method must be developed by which measurement error can be removed from variance estimates.

The study of measurement error was put forth by Charles Spearman over 100 years ago (Alexopoulos, 2007) when he introduced a concept, which has come to be known as classical test theory (CTT). Over the past century, this theory has been advanced by other luminaries, including Gulliksen (1950), Lord and Novick (1968), Nunnally (1978), and McDonald (1999). While numerous formulas can be found in CTT, all are derived from the basic model, $X=T+E$. That is, an observed score $X$ is equal to the sum of an unknown true score $T$ and random error $E$, where $T$ and $E$ are independent. Hence, the observed score variance can be decomposed into the sum of the true score variance and the error variance, $\sigma_X^2=\sigma_T^2+\sigma_E^2$. If one defines the true score to equal the mean of the sampling distribution of observed scores that are generated by repeatedly measuring an object with parallel tests, one can then derive the average correlation between parallel tests $p_{XX'}$, known as the parallel test reliability, to equal $p_{XX'}=\sigma_T^2/\sigma_X^2$. This estimator, in turn, can be employed to remove the error variance for the observed score variance since $\sigma_T^2=p_{XX'}\sigma_X^2$.

The exact mechanism for doing so, however, depends on the type of CI one wants to construct. Three distinct CIs can be computed: one on the true score to capture its corresponding observed score, one on the predicted true score to capture its corresponding actual true score, and one on the predicted true score to capture its corresponding observed score. Not all of these CIs, however, have practical value. For example, if one knew the true
score of a variable, why would one be interested in constructing a CI to capture the observed score? The nuanced differences between these three CIs has led one psychometrician to note that many researchers and even psychometricians employ incorrect point estimates (i.e., they use the observed score rather than the predicted true score) and so, misinterpret their CIs (Harvill, 1991). However, other psychometricians have argued that the gain from using the predicted true score instead of the observed true score is small, particularly for highly reliable tests (Hopkins, 1998; Feldt & Brennan, 1988).

Although theoretical development of CTT appears to have slowed down in the past few decades, gaps can still be found in the literature. The predicted true score, for example, is simply a linear regression estimate of the true score given the observed score. Specifically, the predicted true score $\hat{T}$ given an observed score $X$ is equal to $\hat{T} = \rho_{XX'} X + (1 - \rho_{XX'}) E(X)$, where $E(X)$ denotes the expected value (mean) of random variable $X$. Thus, for very reliable measures, the predicted true score reflects the observed score to a greater extent than the mean, and vice versa. Clearly then when a measure is completely unreliable, the best guess of a person’s true score is the mean of the variable. Notice, however, that the predicted true score relies on only a single observed score. However, occasionally two or more observed variables may contain information about the predicted true score of a composite. Although one can incorporate information from multiple variables into a single score (e.g., confirmatory factor analysis [CFA]), a parallel method does not exist for complex variables wherein the units of analysis are different for multiple data streams. Moreover, even if a new estimator can be derived, what estimator does one use to estimate the sampling error

---

35 Since CFA requires the input of a variance-covariance matrix, or raw data that can produce such a matrix, it cannot be used on the data whose units of analysis, across a set of variables, are different because of a variance-covariance matrix cannot be computed.
Lastly, how does one incorporate reliability into \( \text{Var}(\bar{X}_e) \) so as to remove measurement error?

Assuming solutions for these questions can be derived, thereby expanding CTT, one still needs to determine the parallel test reliability. Given the multitude of existing reliability estimators, one must decide which one to employ. Existing methods for estimating parallel reliability can generally be classified into one of three groups (stability, internal consistency, interrater) based on the type of reliability one seeks to establish (Juni, 2007) although a fourth group (criterion) can be added to this list. Some methods (i.e., test-retest, alternate form, test-retest with alternate form) focus on the stability of responses over time (Crocker & Algina, 1986; Hopkins, 1998). Hence, they require two administrations of the same test or a parallel test. In contrast, methods that focus on the internal consistency of the items that comprise a single test only require a single administration. The most popular coefficients used to measure internal consistency, include the split-half method developed by Charles Spearman and William Brown (Crocker & Algina, 1986), coefficient alpha (Cronbach, 1951), KR20 (Kuder & Richardson, 1937), and coefficient omega (McDonald, 1999).

Without doubt, the most popular of these methods is coefficient alpha although coefficient omega has started to gain traction in the past decade. The split-half method has lost favorability due to the work of Brownell (1933), who showed that the number of potential split-half reliability coefficients that can be estimated is astronomically high. However, it is this author’s contention that the split-half method should be revisited. When combined with sampling theory, a bootstrap split-half method can be developed, which can be used to accurately estimate the expected value (mean) of all the possible split-half

---

36 Since the observed scores used to predict the composite true score can be correlated, it is more appropriate to use the term covariance than it is to use the term variance. In fact, the composite predicted term will need to not only account for the unique information contributed by each observed score used as a predictor, it will need to account for the shared contribution as well.
reliability coefficients. Moreover, when combined with order statistics and the Spearman-rank correlation, a nonparametric CI can be computed for the reliability estimate—note, this feature is not available for any of the previous reliability estimates. Hence, the estimate can be employed for ordinal data. Although an ordinal alpha and omega (Zumbo, Gadermann, & Zeisser, 2007) does exist, these coefficients assume that a normal latent distribution exists, whereas the method proposed above makes no such assumption.

Quantifying Qualitative Data

Given that a large portion of the data collected by evaluators is qualitative, the omission of such data from the purview of Summative Confidence would greatly limit its contribution to the field of evaluation. Moreover, the integration of qualitative and quantitative methods provides “a more complete understanding of the research problem than either approach alone” (Creswell, 2008, p. 527) and has been practiced by mixed methodologists for many years. Although the goal of mixed methodologists is rarely to synthesize all the information into a single conclusion, qualitatively-oriented evaluators have conducted summative evaluations since their onset. In fact, summative evaluation was first proposed by Dr. Michael Scriven, who is a noted qualitative methodologist and evaluator. Thus, there is no insurmountable obstacle to prevent an evaluator from utilizing qualitative data to formulate a summative conclusion. With respect to Summative Confidence, one must first quantify the qualitative data before one can perform the requisite statistical operations. Quantification is a relatively simple, albeit cumbersome, task of classifying units of qualitative data based on specific definitional parameters (i.e., a coding scheme, rules of measurement), evaluating the merit of each category based on an agreed grading rubric for assigning value, and computing the reliability of the procedure employed to classify the qualitative data. The most appropriate reliability analysis emerges from Generalizability
Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001) for it allows one to control for numerous sources of measurement error, particularly variance across raters. However, numerous other methods exist, e.g., see Davey, Gugiu, and Coryn (2010).

That said, in the experience of this author, discussions with colleagues and other eminent evaluation scholars that have broached upon the suitability of quantifying qualitative data have often been met with skepticism and occasionally even objection—usually from qualitative researchers. Frequently, proponents of qualitative research argue that qualitative data does not lend itself to statistical analyses and that the quantification of such data invariably results in a loss of essential information. Perhaps even more alarming has been the wholesale rejection of the importance of reliability among many qualitative researchers. Even in instances in which the reliability of research and evaluation conclusions is acknowledged as a desirable attribute, the concept is often replaced with qualitative notions that bear little resemblance to the original idea underlying reliability. Namely, at some level, reliability must express the likelihood that a result or conclusion would replicate in a future study conducted under parallel conditions. Moreover, the methods employed to determine reliability must be grounded in empirical evidence. That is, it is not enough to say that a conclusion is reliable because a certain method was employed that is highly regarded by other scholars if in practice that method fails to consistently reproduce the same result under parallel conditions.

For some qualitative researchers—e.g., Stenbacka (2001) and Miller (2008), the wholesale rejection of all concepts perceived to be quantitative has extended to research concepts like reliability and validity. According to Stenbacka (2001, p. 552), “reliability has no relevance in qualitative research, where it is impossible to differentiate between researcher and method.” However, this notion is inconsistent with traditional qualitative research because measurement is an indispensable aspect of conducting both quantitative and
qualitative research. With respect to qualitative research, measurement occurs during the
coding process, which is integral to qualitative research as noted by Benaquisto (2008):

“The coding process refers to the steps the researcher takes to identify, arrange, and
systematize the ideas, concepts, and categories uncovered in the data. Coding consists of
identifying potentially interesting events, features, phrases, behaviors, or stages of a process
and distinguishing them with labels. These are then further differentiated or integrated so
that they may be reworked into a smaller number of categories, relationships, and patterns so
as to tell a story or communicate conclusions drawn from the data.” (p. 85)

Clearly, in absence of utilizing a coding process, researchers would be forced to
provide readers with all of the data, which, in turn, would place the burden of interpretation
on the reader. However, while the importance of coding to qualitative research is self-
evident to all those who have conducted such research, the role of measurement may not be
as obvious. In part, this may be a result of misunderstanding the true role of measurement.
According to Stevens (1946) and Lord and Novick (1968), measurement is the process of
assigning numbers, symbols, or codes to phenomena (e.g., objects, events, features, phrases,
behaviors, or properties of objects or events) based on a set of prescribed rules (i.e., a coding
scheme). There is nothing inherently quantitative about this process or, at least, there does
not need to be. Moreover, it does not limit qualitative research in any way.

For example, suppose that a researcher conducts an interview with an informant
who states that “the bathrooms in the school are very dirty.” Now further suppose that the
researcher developed a coding scheme, which, for the sake of simplicity, only contained two
categories: cleanliness and academic performance. Clearly, the informant’s statement
addressed the first category (cleanliness) and not the second. It does not make a difference as
to whether the researcher choses to assign this statement a checkmark for the cleanliness
category or a 1, and an ‘X’ or a 0 for the academic performance category. The researcher
clearly used his or her judgment to transform the raw statement made by the informant into
a code. When the researcher decided that the statement best represented cleanliness and not
academic performance, he or she performed a measurement process. Hence, if one accepts this line of reasoning, qualitative research depends upon measurement to render judgments. Furthermore, three questions may be asked. First, does statement X fit the definition of code Y? Second, how many of the statements collected fit the definition of code Y? And third, how reliable (i.e., intra-rater or inter-rater) are coders in their application of the definition of code Y to all the codeable statements?

Fortunately, not every qualitative researcher has accepted Stenbacka’s notion, in part, because qualitative researchers, like quantitative researchers, compete for funding and therefore, must persuade funders of the accuracy of their methods and results (Cheek, 2008). Consequently, the concepts of reliability and validity permeate qualitative research. However, owing to the desire to differentiate itself from quantitative research, qualitative researchers have espoused the use of “interpretivist alternatives” terms (Seale, 1999). Some of the most popular terms substituted for reliability include confirmability, credibility, dependability, and replicability (Golafshani, 2003; Healy & Perry, 2000; Morse, Barrett, Mayan, Olson, & Spiers, 2002; Miller, 2008; Lincoln & Guba, 1985).

In the qualitative tradition, confirmability is concerned with whether the researcher’s interpretations and conclusions are grounded in actual data that can be verified (Jensen, 2008; Given & Saumure, 2008). Researchers may address this reliability indicator through the use of multiple coders, transparency, audit trails, and member checks. Credibility, on the other hand, is concerned with the research methodology and data sources used to establish a high degree of harmony between the raw data and the researcher’s interpretations and conclusions. Various means can be used to enhance credibility, including accurately and

37 Note, a slight change in the variable is required when dealing with nominal data. While nominal data is binary, this defines the variable of interest to be the count of these units (i.e., the variable of interest changes as does the underlying distribution of the variable).
richly describing data, citing negative cases, using multiple researchers to review and critique the analysis and findings, and conducting member checks (Jensen, 2008; Given & Saumure, 2008; Saumure & Given, 2008). Dependability can be addressed by providing a rich description of the research procedures and instruments used so that other researchers may be able to collect data in similar ways. The idea being that if a different set of researchers use similar methods then they should reach similar conclusions (Given & Saumure, 2008).

Finally, replicability is concerned with repeating a study on participants from a similar background as the original study. Researchers may address this reliability indicator by conducting the new study on participants with similar demographic variables, asking similar questions, and coding data in a similar fashion to the original study (Firmin, 2008).

A review of the qualitative terms used as alternations to the quantitative notion of reliability revealed that they were indirectly associated with quantitative notion. However, although replicability is conceptually equivalent to test-retest reliability, the other three terms appear to describe research processes that are only tangentially related to reliability. Moreover, they have two major liabilities. First, they place the burden of assessing reliability squarely on the reader. For example, if a reader wanted to determine the confirmability of a finding they would need to review the audit trail and make an independent assessment. Similar reviews of the data would be necessary, if a reviewer wanted to assess the credibility of a finding or dependability of a study design. Second, they fail to consider inter-rater reliability, which, in the opinion of this author, accounts for a considerable amount, if not a majority, of the variability in findings in qualitative studies. Inter-rater reliability is concerned with the degree to which different raters or coders appraise the same information (e.g., events, features, phrases, behaviors) in the same way (van den Hoonoord, 2008). In other words, do different raters interpret qualitative data in similar ways? The process of
conducting an inter-rater reliability analysis is relatively straightforward. Essentially, the only additional step beyond the development and finalization of a coding scheme is that, at least two or more raters must independently rate all of the qualitative data. Although collaboration, in the form of consensus agreement, may be used to finalize ratings after each rater has had an opportunity to rate all the data, each rater must work independently of the other to reduce bias in the first phase of analysis. Often, this task is greatly facilitated by use of a database system that, for example, (1) displays the smallest codable unit of a transcript (e.g., a single sentence), (2) presents the available coding options, and (3) records the rater’s code before displaying the next codable unit.

Since qualitative and quantitative researchers approach their work from different epistemologies, it is likely that qualitative researchers, particularly those who prescribe to a constructionist paradigm, may object to the constraint of forcing qualitative researchers to use the same coding scheme for a study, rather than developing their own. Researchers who subscribe to a more positivist paradigm, however, would argue that this is an indispensable process for attaining a reasonable level of inter-rater reliability. An example of the perils of not attending to this issue may be found in a study conducted by Armstrong, Gosling, Weinman, and Marteau (1997). Armstrong and his colleagues invited six experienced qualitative researchers from Britain and the United States to analyse a transcript (approximately 13,500 words long) from a focus group comprised of adults living with cystic fibrosis that was convened to discuss the topic of genetic screening. In return for a fee, each researcher was asked to prepare an independent report in which they identified and described the main themes that emerged from the focus group discussion, up to a maximum of five. Beyond these instructions, each researcher was permitted to use any method for extracting the main themes they felt was appropriate. Once the reports were submitted, they
were thematically analyzed by one of the authors, who deliberately abstained from reading the original transcript to reduce external bias.

On the surface, it was clear that a reasonable level of consensus in the identification of themes was achieved. Five of the six researchers identified five themes, while one identified four themes. Consequently, only four themes are discussed herein: visibility, ignorance, health service provision, and genetic screening. With respect to the presence of each theme, there was unanimous agreement for the visibility and genetic screening themes, while the agreement rates were slightly lower for the ignorance and health service provision themes (83% and 67%, respectively). Overall, these are pretty good rates of agreement. However, a deeper examination of the findings revealed two troubling issues. First, a significant amount of disagreement existed with respect to how the themes were organized. Some researchers classified a theme as a basic structure whereas others organized it under a larger basic structure (i.e., gave it less importance than the overarching theme they assigned it to). Second, a significant amount of disagreement existed with respect to the manner in which themes were interpreted. For example, some of the researchers felt that the ignorance theme suggested a need for further education, other researchers raised concern about the eugenic threat, and the remainder thought it provided parents with choice. Similar inconsistencies with regard to interpretability occurred for the genetic screening theme where three researchers indicated that genetic screening provided parents with choice while one linked it with the eugenic threat.

These results serve as an example of how “reality” is relative to the researcher doing the interpretation. However, they also demonstrate how the quality of a research finding requires knowledge of the degree to which consensus is reached by knowledgeable researchers. This statement, of course, assumes that the reliability of findings across different
researchers is a desirable quality. There certainly may be instances in which reliability is not important because one is only interested in the findings of a specific researcher, and the perspectives of others are not desired—a characteristic of constructionism. That being the case, one may consider examining intra-rater reliability. In all other instances, however, it is reasonable to assume that it is desirable to differentiate between the perspectives of the informants and those of the researcher. In other words, are the researcher’s findings truly grounded in the data or do they reflect his or her personal ideological perspectives? For a politician, for example, knowing the answer to this question may mean the difference between passing or rejecting a policy that allows parents to genetically test embryos.

Although qualitative researchers can address inter-rater reliability by following the method used by Armstrong and his colleagues (1997), the likelihood of achieving a reasonable level of reliability may be low simply due to researcher differences (e.g., the labels used to describe themes, structural organization of themes, importance accorded to themes, interpretation of data). In general, given the importance of reducing the variability in research findings attributed solely to researcher variability, it would greatly benefit qualitative researchers to utilize a common coding rubric. Furthermore, use of a common coding scheme does not greatly interfere with normal qualitative procedures, particularly if consensus is reached beforehand by all the researchers on the scheme that will be used to code the data. Of equal importance is the fact that this procedure continues to permit the researcher to remain the instrument by which data are interpreted (Brodsky, 2008).

---

38 It is important to acknowledge that this reasoning is derived from the present author’s inclination towards a positivist paradigm. In fact, one can argue that all of Summative Confidence is rooted in this paradigm since if evaluators do not agree that an evaluand has merit and worth independent of the evaluator and instruments used to measure the evaluand, then clearly one cannot construct a CI on the point estimate for the complex variable representing the evaluand’s overall merit and worth.
Reporting the results of, to this point, this qualitative process should considerably improve the credibility of research findings. However, three issues still remain. First, reporting the findings of multiple researchers places the burden of synthesis on the reader. Second, judging the reliability of a study requires that deidentified data are made available to anyone who requests it. Third, reporting the findings of multiple researchers will only permit readers to get an approximate sense of the level of inter-rater reliability or whether it meets an acceptable standard.

These considerations notwithstanding, qualitative data is a source of rich data and should not be ignored by quantitative researchers and evaluators. While the process of developing and validating a coding rubric is froth with challenges, they are not insurmountable. From the perspective of Summative Confidence, the steps are simple: (1) develop a coding scheme that accurately reflects the qualitative data making sure that there are enough mutually distinct categories to cover the vast majority of the data; (2) partition the qualitative data into the smallest possible coding units; (3) assign two or more raters to code the data using the agreed upon coding scheme; (4) conduct a reliability analysis of the coding scheme; (5) construct a quantitative dataset making sure to resolve all disagreements, either by consensus or by averaging results; and (6) proceed as usual with data analyses.

Naturally, these steps are considerably more complex than the list may lead on. The development of a coding scheme will undoubtedly require numerous passes through the raw data before it can be finalized. In fact, more likely than not, one will need to code the data several times (step 3) in order to eliminate trivial coding categories, break up large categories into smaller mutually distinct categories, and reduce the size of the “other” category (i.e., find a way to code as much of the data as possible). This is both an iterative and collaborative task. It should never be undertaken by a single person in one or two passes.
through the raw data. The second step will differ from study to study. In some instances, for example, it may be appropriate to code an entire paragraph (or more), whereas in other instances one may need to isolate sentences or even individual words. Step three is critical to being able to obtain an inter-rater reliability estimate. It is important to note that while collaboration and discussion among raters is natural and even desirable from the perspective of attaining agreement on the coding scheme (step 1), coding must be done independently. Several methods exist for establishing reliability (step 4). Prior to analyzing the data, agreement will need to be reached on what constitutes the final dataset (step 5). That is, all disagreements will need to be resolved. Generally, this can be achieved through discussions among the raters. However, more elaborate methods exist (e.g., Delphi method). The last step is to analyze the agreed upon dataset in the usual ways (e.g., compute means, variances, and so forth). Notice then that only steps 4 and 6 break with conventional practice among qualitative researchers and one can argue that step 6 is not really a break at all. After all, many qualitative researchers do report on the number of objects that met the definitional parameters of a specific category. It is important to recognize, however, that the methods for treating qualitative data described above clearly emerge from the author’s proclivity toward a positivist paradigm.
CHAPTER III

MONTE CARLO SIMULATIONS

In addition to theoretical proofs, Monte Carlo simulations will be utilized throughout this study to illustrate the veracity of statistical and psychometric principles, both new and old. The primary purpose of this chapter was to detail the methodology used to simulate random variables with a specific skewness, kurtosis, correlation, and variance-covariance structure. The secondary purpose was to illustrate the ability of Monte Carlo simulations to address pertinent questions. Specifically, the analyses herein will investigate the Central Limit Theorem (CLT) to determine whether the composite average of a finite set of random nonnormal variables is approximately normally distributed. Given that several theorems in subsequent chapters assume normality for certain random processes (e.g., measurement error, sampling distribution), this exercise is also intended to provide readers with reassurance in the robustness of those assumptions. Analyses were performed using SAS 9.2 (Windows platform) and were conducted on a Dell Precision T5400 workstation running Dual Quad 2.0 GHz processors with 6 GB of ram on a 64-bit Vista Business operating system. Despite the power of this workstation, the total non-stop runtime of the analyses performed herein exceeded 3 months and resulted in the simulation of over 9 TB of data.

Simulating Data

A Monte Carlo (MC) simulation is a study that utilizes computer generated data based on random sampling techniques to simulate a population for the purpose of understanding the behavior of a mathematical algorithm (Mooney, 1997; Merriam-Webster,
Hence, MC studies avoid the impracticality of sampling a real population multiple times to assess the properties of a mathematical algorithm. Another way in which to think about a MC simulation is as an experiment in which random sample data are generated to conform to a population with known parameters. Unlike real-world data, one can compare the results of the mathematical computations performed on the sample data to those obtained from computations utilizing the known population parameters. If the theory is correct then these results should be equivalent within a very small margin of error. Of course, this is only true to the degree to which the sample data are randomly drawn from the population of interest without bias.

According to the CLT, the sum (or average) of a number of variables that are iid will be approximately normally distributed as the number of constituent variables grows without bounds provided that each of the constituent variables has finite variance (i.e., the first two statistical moments are estimable) (Meyer, 1970). Notice that no mention of the underlying distribution of the constituent variables appears in this theorem because CLT holds even when nonnormal variables are aggregated. However, it stands to reason that the greater the nonnormality, either skewness or kurtosis, of the constituent variables, the more variables that will need to be aggregated before CLT holds. This chapter seeks to determine this number for a variety of factors. One of these factors is sample size because it is widely known that as sample size increases, sample estimators tend to converge to the population parameters—known as the Law of Large Numbers\(^\text{39}\). Therefore, fewer variables may need to be aggregated when the sample size is large. Another factor investigated is the magnitude of the correlation between variables. According to the CLT, factors must be independent (i.e.,

\(^{39}\) The theorem states that for a sequence of independent and identically distributed (iid) random variables the probability tends to zero that the mean of the sequence differs from the population mean \(\mu\) by more than a margin of error \(\varepsilon\), no matter how small \(\varepsilon\) may be. Mathematically, it is expressed as \(\lim_{n \to \infty} P(\bar{X} - \mu > \varepsilon) = 0\), for any \(\varepsilon > 0\).
uncorrelated) in order for CLT to hold although some authors (Godwin & Zaremba, 1961; Cocke, 1972) have shown that the CLT can be extended to correlated random variables as well. Hence, it may be the case that the higher the correlation between random variables, the more variables one will need to aggregate for CLT to hold. Naturally then the covariance structure of these variables influences the normality of the composite and so merits study.

With regard to two key random processes that will be discussed in the next two chapters (i.e., sampling distribution and measurement error), it is important to demonstrate that CLT holds when normal, nonnormal, independent, and dependent variables are aggregated. Though it is well-known that CLT holds for normal and nonnormal independent and identically distributed variables, it is unclear how well it holds, if at all, for dependent variables. Moreover, if CLT does hold, for what levels of skewness, kurtosis, dependence, and number of variables aggregated is it reasonable to assume the distribution of the composite variable will be asymptotically normal? In the case of sampling distributions, there are generally billions of possible random samples that can be drawn from a population before all the possible combinations are exhausted. Similarly, it is commonly accepted that measurement error is the product of an unknown but very large number of random processes. Thus, if one can demonstrate that a composite variable, constructed from a finite number of dependent variables, is approximately normally distributed then the normality assumption for these two processes will be adequately satisfied.

**Study Design**

Various levels for the aforementioned factors will be simulated to produce a graphical representation of the relationship between the variables and the probability that the

---

40 If one were to repeatedly draw random samples from a population and compute a statistic on the data (e.g., the mean), the values for this statistic would form a probability distribution known as the sampling distribution.
composite variable is approximately normally distributed. To compute this probability, \( v \) random variables were simulated with a sample size of \( n \) and a composite variable was constructed by averaging across the \( v \) variables; thereby producing a single variable of sample size \( n \). The distribution of these \( n \) values was tested to determine whether the composite variable was normally distributed. This process was repeated 1,000 times and a dummy variable was created to record whether the composite variable passed (recorded as 1) or failed (recorded as 0) the normality test. The probability of passing the test was computed by dividing the sum of the dummy variable by the total number of simulations. This probability was used as the dependent measure for all subsequent analyses.

Ten levels for the number of constituent variables were investigated (2, 10, 18, 26, 34, 57, 95, 148, 216, and 300); five levels for skewness (0 to 2 by 0.5); six levels for kurtosis (0 to 10 by 2); seven levels for sample size (10, 150, 300, 460, 630, 810, and 1,000); eleven levels for correlations (0 to 1 by 0.1); and three covariance structures (independence, compound symmetry, unstructured). It is important to note, skewness \( \alpha_3 \) and kurtosis \( \alpha_4 \) are dependent upon each other. Theoretically the relationship between the two variables is expressed by the following range space: \( \alpha_3^2 - 2 < \alpha_4 < \infty \) (Karian & Dudewicz, 2000). However, the method used to simulate variables with specific skewness and kurtosis (discussed below) was limited to the following region: \( 1.8(\alpha_3^2 + 1) - 3 \leq \alpha_4 < \infty \). Consequently, crossing the five levels for skewness with the six levels for kurtosis resulted in 20 distributions. These distributions were, in turn, crossed with the number of variables and sample size and resulted in 1,400 unique combinations. These combinations reflect the number of cells that were simulated in the factorial design.

---

\[ \text{The formulas found in Karian and Dudewicz (2000) use the uncorrected value for kurtosis. Therefore, the constant 3 was subtracted from their formulas to make the kurtosis of the normal distribution equal to zero.} \]
The three covariance models selected reflect different conditions an investigator might encounter during an evaluation study. Model 1 examined independence: the condition under which all of the variables in the study are uncorrelated with each other; thus, meeting CLT assumptions. Model 2 examined compound symmetry: the condition under which the same correlation exists between all the variables. To more fully understand the impact of correlations on the probability of normality, the 1,400 cells (from Model 1) were crossed with eleven levels for correlation, resulting in 15,400 unique cells. Model 3 examined unstructured covariance: the condition under which the correlations between the variables can take on any value (i.e., random). Because the correlation coefficient was random, this model did not increase the number of cells simulated (from Model 1) for the factorial design.

Testing Normality of Composite Variables

Normality was assessed using the Shapiro-Wilk W Test for normality, which is the standard test used to assess normality when sample size is less than 2,000 (SAS Institute, 2007). The Shapiro-Wilk W Test is a univariate test that assesses whether the observations of a variable are sampled from a normal distribution. The W statistic represents “the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance” (SAS Institute, 2007). Moreover, the W statistic has a p-value range from 0 to 1, where low values lead to rejection of the null hypothesis (i.e., the variable is not normally distributed). The power of the Shapiro-Wilk Test, however, is a function of sample size. For small sample sizes, the test has been found to lack the ability (power) to detect large departures from normality. Consequently, it is recommended that one declare a higher alpha level (e.g., 0.15 or 0.20)

42 For example, classical test theory (Chapter 5) assumes that the same correlation exists between parallel tests.
43 Measurement error, for example, is assumed to be the composite of a large number of random variables, whose correlation matrix is unstructured.
than the often-used 0.05 level in order to increase the test’s ability to detect such deviations (SAS Institute, 2007). Due to the small sample size (i.e., \( n=10 \)) simulated for part of the analyses, alpha for rejecting the null hypothesis was set at the 0.15 level.

Generating Nonnormal Data

The task of a MC study is to generate data that resembles the data one would observe if one actually sampled the population of interest. Given that the majority of data that researchers typically collect are not normally distributed (Micceri, 1989), this study utilized a fair amount of nonnormal data. Although a variety of mathematical algorithms have been developed to simulate nonnormal data, two methods appear to be the most popular: Fleishman’s Power Transformation (Fleishman, 1978) and the Ramberg and Schmeiser (1972, 1974) Generalized Lambda Distribution (GLD). Of the two methods, Fleishman’s Power Transformation (Fleishman, 1978) appears to be the method most often used because of its ability to easily generate correlated nonnormal data. The choice between the two methods is one of personal preference, according to Dr. Todd Headrick (personal communication, July 21, 2008), although there are some combinations of skewness and kurtosis that only the GLD method can simulate. Therefore, the GLD method was selected to simulate nonnormal data (Headrick & Mugdadi, 2006) once the simultaneous equations presented in Karian and Dudewicz (2000) are solved.

One method that can be used to solve these simultaneous equations is to algorithmically search for a numerical solution that is within an acceptable level of error \( \varepsilon \) (i.e., critical value) set by the researcher. The Newton-Raphson method (Stewart, 2003) was used to derive a suitable solution by solving for the unknowns.

Presently, only a limited number of software platforms exist for solving the simultaneous equations required for the GLD method. These platforms include R (Su,
2007), Maple (Dr. Karian, personal communication, September 16, 2011), Mathematica and Fortran (Headrick & Mugdadi, 2006), and C (King & MacGillivray, 1999). Unfortunately, these programs require knowledge of programming languages that fall outside the knowledge base of many evaluators. Therefore, a SAS solution that solves these equations and produces valid lambdas was programmed by this author and is provided in Appendix B. Despite the complexity of the algorithm, the code converges very quickly.

Given a solution to the simultaneous equations, the values for the four lambdas are inserted into the GLD function, 

\[ P(u) = \lambda_1 + \frac{u^h - (1-u)^h}{\lambda_2}, \]

where \( u \) is a random uniform variable with range \( 0 \leq u \leq 1 \) (Headrick & Mugdadi, 2006; Karian & Dudewicz, 2000; Freimer, Mudholkar, Kollia, & Lin, 1988; Ramberg & Schmeiser, 1972, 1974). For instance, to generate a normal distribution \( N(0,1) \) of sample size 1,000, one needs to generate 1,000 uniform random values and use this vector as the input to the GLD function along with the following lambdas: \( \lambda_1 = 0, \lambda_2 = 0.1975, \lambda_3 = 0.1349, \) and \( \lambda_4 = 0.1349 \). However, the accuracy of the Newton-Raphson solution depends, to some extent, upon the initial guess because if it is not close to the true root (i.e., point at which the function crosses the X-axis), the algorithm may converge upon lambda values that generate distributions other than the one of interest (Karian & Dudewicz, 2000). In another words, matching the functions of the third and fourth statistical moments does not ensure one will generate the distribution of interest. This, of course, raises the question, how good was the fit between the simulated data and the distribution of interest?

Three methods were used to validate the distributions simulated in this study. First, SAS PROC UNIVARIATE was used to plot the probability density curves for each observed distribution against the normal distribution to examine key shape elements. In
particular, the skewness and kurtosis of each empirical density was compared to the normal distribution and each other. For example, a density with a theoretical skewness of 2 was expected to have greater right skew than a density with a theoretical skewness of 1. Likewise, a density with a theoretical kurtosis of 10 was expected to exhibit greater “peakedness” than a density with a theoretical kurtosis of 5. This visual analysis, which is the preferred method of assessing goodness of fit by many researchers (Karian & Dudewicz, 2000), suggested that the GLD method produced valid distributions (see Figure 2.1-2.20). However, because this method may be prone to subjectivity, each of the lambda values were compared to the lambda values published in Karian and Dudewicz (2000). This comparison revealed no departures between the two set of lambda values. In almost all the comparisons, the values were the same up to four decimal points. As a final check of the goodness of fit, a MC simulation was conducted under varying sample size conditions to determine the deviations of the sample moments from the theoretical moments.

Table 1 shows the skewness and standard errors for seven levels of sample size and 20 distributions. This table illustrates that the GLD method can simulate skewness with minimal deviation from the desired statistical moment. However, the deviation appears to be influenced by sample size. Specifically, as sample size increases, the observed skewness appears to converge (in probability) toward the theoretical skewness. The deviation is more pronounced when one simulates kurtosis. Table 2 illustrates that the deviation between the desired and observed kurtosis is a function of sample size and the magnitude of the desired

---

44 Detailed tables for the four lambda values are available in Karian and Dudewicz (2000) for a wide range of values for skewness and kurtosis. However, as is the case with many reference tables, gaps can be found in the reference range. Additionally, certain situations may call for the ability to simulate random values for skewness and kurtosis. Hence, in instances such as these, one will need to employ a programmatic solution, like the one presented in Appendix B.

45 That is, as N→∞, the value of the simulated observed skewness S_O approaches the theoretical skewness S_T. Stated more formally, Lim_{N→∞}P(|S_O−S_T|>ε)=0, where ε is a critical tolerance level greater than zero.
Figure 2. Observed Probability Density Functions Versus the Normal Distribution
Figure 2—Continued
kurtosis. The larger the kurtosis, the greater the deviation, particularly when sample size is small. However, since these interpretations are based on observations, they are subject to subjectivity. Therefore, more formal tests of these relationships are warranted.

Two stepwise regression analyses, parametric and nonparametric, were conducted with SAS PROC REG to investigate the relationship between the deviation for skewness and kurtosis and three factors: the theoretical values for skewness and kurtosis, and sample size. Although stepwise regressions are notorious for capitalizing on chance and overfitting data (Tabachnick & Fidell, 2001)—therefore, failing to replicate in new samples—this was not a concern because the patterns that emerge from MC studies are stable (i.e., they will replicate in parallel simulation studies), given a large number of replications.\footnote{Since there are 1,000 replications, the Law of Large Numbers dictates that the sample statistics should be fairly close to the population parameters.} Moreover,
stepwise regressions allow one to investigate the importance of adding new predictors to the regression model.

Table 1

Simulated Skewness (and Standard Error) by Sample Size and Distribution

<table>
<thead>
<tr>
<th>Theoretical Skew</th>
<th>Kurt</th>
<th>N=10</th>
<th>N=150</th>
<th>N=300</th>
<th>N=460</th>
<th>N=630</th>
<th>N=810</th>
<th>N=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 0</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2. 0</td>
<td>0</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>3. 0</td>
<td>0</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>4. 0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>5. 0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>6. 0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>7. 0.5</td>
<td>0</td>
<td>0.04</td>
<td>0.01</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8. 0.5</td>
<td>0</td>
<td>0.02</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>9. 0.5</td>
<td>0</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>10. 0.5</td>
<td>0</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>11. 0.5</td>
<td>0</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>12. 1</td>
<td>0</td>
<td>0.02</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>13. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>14. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>15. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>16. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>17. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>18. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>19. 1</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>20. 0.5</td>
<td>0</td>
<td>0.03</td>
<td>0.09</td>
<td>0.01</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The parametric analyses were subject to the linear model assumptions. Namely, the relationship between the predicted (i.e., the deviation) and predictor variables was the same for levels of the predictor variables (i.e., homoscedasticity) and the residuals of the model were independent and normally distributed (Hays, 1994; Tabachnick & Fidell, 2001). The first analysis modeled the relationship between the observed deviations and the aforementioned variables on the raw data. However, the residuals of this model were found to violate the assumption of normality and homoscedasticity, primarily due to several outliers. Examination of the residuals by observations plot (not included) revealed that all of the outliers occurred at N=10. Although excluding this subsample from the analysis improved the model, it failed to normalize the...
residuals. The impacts of several data transformations of the dependent variable were examined, which revealed that a square root transformation of the absolute deviations normalized the data and eliminated heteroscedasticity.\(^47\)

Table 2

<table>
<thead>
<tr>
<th>Skew</th>
<th>Kurt</th>
<th>N=10</th>
<th>N=150</th>
<th>N=300</th>
<th>N=460</th>
<th>N=630</th>
<th>N=810</th>
<th>N=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>2.</td>
<td>0.2</td>
<td>0.05</td>
<td>1.6</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>3.</td>
<td>0.4</td>
<td>0.06</td>
<td>2.9</td>
<td>0.11</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>4.</td>
<td>0.6</td>
<td>0.06</td>
<td>3.6</td>
<td>0.12</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>5.</td>
<td>0.8</td>
<td>0.11</td>
<td>4.3</td>
<td>0.14</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>6.</td>
<td>1.0</td>
<td>0.07</td>
<td>5.1</td>
<td>0.24</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>7.</td>
<td>0.5</td>
<td>0.04</td>
<td>1.7</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>8.</td>
<td>0.5</td>
<td>0.06</td>
<td>2.9</td>
<td>0.12</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>9.</td>
<td>0.5</td>
<td>0.06</td>
<td>2.9</td>
<td>0.11</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>10.</td>
<td>0.5</td>
<td>0.06</td>
<td>3.7</td>
<td>0.16</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>11.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>12.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>13.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>14.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>15.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>16.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>17.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>18.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>19.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>20.</td>
<td>0.5</td>
<td>0.06</td>
<td>4.3</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Unfortunately, the elimination of the subsample reduced the sample size. Moreover, a nonlinear model is only as good as the function used to fit the data and may not replicate in future studies. However, this was not a huge concern due to the Law of Large Numbers. Therefore, in order to recover the “lost” cases and to confirm the previous results, a rank regression was performed on the original data by replacing the values for the dependent and independent variables with their corresponding ranks (or average rank in case of ties) (Iman & Conover, 1979; Conover & Iman, 1981). As a result of the rank transformation, outliers

---

\(^{47}\) Although the GLD is a popular method for generating nonnormal distributions (Fan, Felsővályi, Sivo, & Keenan, 2001), it is clear that a bias exists in the simulated skewness and kurtosis values. However, this bias can be safely ignored for moderate to large sample sizes and small to moderate levels for skewness and kurtosis.
do not have an undue influence on the model fit. More importantly, when the assumptions of the linear model are violated, the rank regression is a more robust and powerful approach than the linear regression, which is not able to minimize the impact of outliers (Conover & Iman, 1981).

Table 3 (A)-(C) presents the results of these three stepwise regressions for deviations in skewness and kurtosis. These results support the subjective impressions derived from examination of the previous two tables. Namely, all three factors were found to contribute significantly to deviations. In the case of deviations in skewness, the magnitude of the desired skewness was found to be the strongest predictor whereas in the case of deviations in kurtosis, the magnitude of the desired kurtosis was found to be the strongest predictor. Moreover, sample size was inversely related to the magnitude of deviations in skewness and kurtosis. Although all these analyses showed that the GLD can simulate desired skewness and kurtosis, caution is needed in interpreting results produced by small sample sizes, especially when the magnitude of the skewness and kurtosis is large. Results for the mean and variance were not presented because these deviations were small.

Table 3

Stepwise Multiple Regressions of Deviations in Simulated Skewness and Kurtosis

(A) Non-normal data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model of skewness deviations</th>
<th>Model of kurtosis deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In sr² Beta SE β p-value</td>
<td>In sr² Beta SE β p-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>-- -- 0.106 0.034 -- 0.002</td>
<td>-- -- 1.040 0.231 -- &lt;0.001</td>
</tr>
<tr>
<td>Skewness</td>
<td>2 0.204 0.128 0.021 0.410 &lt;0.001</td>
<td>3 0.020 -0.243 0.145 -0.086 0.096</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3 0.013 0.008 0.005 0.121 0.080</td>
<td>1 0.395 0.421 0.032 0.668 &lt;0.001</td>
</tr>
<tr>
<td>Sample size</td>
<td>1 0.215 0.000 0.000 -0.464 &lt;0.001</td>
<td>2 0.272 -0.003 0.000 -0.522 &lt;0.001</td>
</tr>
<tr>
<td>R²</td>
<td>0.432 N 140 R² 0.687 N 140</td>
<td></td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.420 Normal (p-value) &lt;0.001</td>
<td>Adj. R² 0.680 Normal (p-value) &lt;0.001</td>
</tr>
<tr>
<td>Model (p-value)</td>
<td>&lt;0.001 Homoscedasticity 0.140</td>
<td>Model (p-val.) &lt;0.001 Homoscedasticity 0.015</td>
</tr>
</tbody>
</table>

Note: ‘In’ refers to the step at which the variable entered the model, sr² refers to the semi-partial r-squared, and β refers to the standardized beta coefficient.
Table 3—continued

(B) Square root normal transformation

<table>
<thead>
<tr>
<th>Variable</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>--</td>
<td>--</td>
<td>0.113</td>
<td>0.018</td>
<td>&lt;0.001</td>
<td>--</td>
<td>--</td>
<td>0.372</td>
<td>0.036</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Skewness</td>
<td>1</td>
<td>0.445</td>
<td>0.101</td>
<td>0.010</td>
<td>&lt;0.001</td>
<td>3</td>
<td>0.045</td>
<td>-0.098</td>
<td>0.020</td>
<td>-0.114</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2</td>
<td>0.158</td>
<td>0.018</td>
<td>0.002</td>
<td>&lt;0.001</td>
<td>1</td>
<td>0.807</td>
<td>0.182</td>
<td>0.004</td>
<td>0.956</td>
</tr>
<tr>
<td>Sample size</td>
<td>3</td>
<td>0.103</td>
<td>0.000</td>
<td>0.000</td>
<td>&lt;0.001</td>
<td>2</td>
<td>0.092</td>
<td>-0.001</td>
<td>0.000</td>
<td>-0.303</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.706</td>
<td>N</td>
<td></td>
<td></td>
<td>120</td>
<td>R²</td>
<td>0.944</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Adj. R²</td>
<td></td>
<td>0.699</td>
<td>Normal (p-value)</td>
<td>0.188</td>
<td>Adj. R²</td>
<td>0.943</td>
<td>Normal (p-value)</td>
<td>0.756</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model (p-value)</td>
<td>&lt;0.001</td>
<td>Homoscedasticity</td>
<td>0.443</td>
<td>Model (p-val.)</td>
<td>&lt;0.001</td>
<td>Homoscedasticity</td>
<td>0.300</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(C) Rank transformation

<table>
<thead>
<tr>
<th>Variable</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>--</td>
<td>--</td>
<td>34.865</td>
<td>5.805</td>
<td>&lt;0.001</td>
<td>--</td>
<td>--</td>
<td>46.090</td>
<td>3.562</td>
<td>&lt;0.000</td>
</tr>
<tr>
<td>Skewness</td>
<td>1</td>
<td>0.517</td>
<td>0.682</td>
<td>0.051</td>
<td>0.055</td>
<td>3</td>
<td>0.035</td>
<td>-0.056</td>
<td>0.031</td>
<td>-0.055</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3</td>
<td>0.035</td>
<td>0.199</td>
<td>0.051</td>
<td>&lt;0.000</td>
<td>1</td>
<td>0.647</td>
<td>0.856</td>
<td>0.031</td>
<td>0.840</td>
</tr>
<tr>
<td>Sample size</td>
<td>2</td>
<td>0.138</td>
<td>-0.375</td>
<td>0.048</td>
<td>&lt;0.001</td>
<td>2</td>
<td>0.201</td>
<td>-0.453</td>
<td>0.030</td>
<td>-0.449</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.690</td>
<td>N</td>
<td></td>
<td></td>
<td>140</td>
<td>R²</td>
<td>0.883</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Adj. R²</td>
<td></td>
<td>0.684</td>
<td>Model (p-value)</td>
<td>&lt;0.001</td>
<td>Adj. R²</td>
<td>0.881</td>
<td>Model (p-value)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generating Correlated Data

While the GLD can be used to generate a univariate nonnormal variable, generating multivariate nonnormal variables requires additional steps. Specifically, one needs to generate a \( n \times v \) matrix (where \( n \) is the sample size, \( v \) is the number of variables, and each element ranges from 0 to 1) that has the desired \( v \times v \) correlation matrix and substitute this matrix for \( n \) (the random uniform variable used by the GLD). Six mathematical steps were adopted from Headrick and Mugdadi (2006) with the help of Dr. Headrick and Jason Davey.

1) Input the desired \( v \times v \) correlation matrix \( A \).

2) Obtain the Cholesky decomposition of \( A \), where \( A = U^\prime U \) and \( U \) is a \( v \times v \) upper triangle matrix known as the Cholesky decomposition and \( U^\prime \) is its transpose.
3) Generate a new $n \times n$ matrix $N$ of standard ($\mu=0, \sigma^2=1$) random normal values\(^{48}\).

4) Calculate a new matrix from the product of step 2 and 3, $X=NU$.

5) Generate a new matrix by calculating the normal cumulative distribution function (cdf) $\Phi$ of the matrix created in step 4, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-u^2/2) \, du.$$ 

6) Input the $\Phi$ matrix into the GLD in place of the random uniform variable.

Although this process requires a modest amount of statistical knowledge, the logic behind these steps is simple. The goal is to generate a set of variables that match the desired correlation matrix $A$ and statistical distribution (identified by the choice of lambdas in step 6). Steps 2 through 4 are designed to generate a set of variables that match matrix $A$. The second step decomposes matrix $A$ into an upper and lower triangle matrix ($U$ and $U'$) whose product equals matrix $A$. However, this decomposition is contingent upon whether the original matrix is positive semidefinite\(^{49}\). The Cholesky matrix $U$ has the nice property that when it is post-multiplied to a matrix of independent variables (e.g., matrix $N$), it produces a matrix (e.g., $X$) whose correlation matrix approximates the original matrix (e.g., $A$). However, an additional step must be executed to transform the values of matrix $X$ to fit the range allowed by the GLD. Consequently, the values of matrix $X$ are transformed into probabilities by calculating matrix $\Phi$. The correlations of matrix $\Phi$ continue to approximate matrix $A$ but now meet the $[0,1]$ range requirement of the GLD. The final step replaces the random uniform variable $\pi$ from the GLD with matrix $\Phi$ and calculates a new matrix whose variables approximate correlation matrix $A$ and the desired statistical distribution.

---

\(^{48}\) Some authors also refer to these values as random normal deviates.

\(^{49}\) A positive semidefinite matrix is a matrix whose eigenvalues are nonnegative (Johnson & Wichern, 1998).
Although statistical theory assures this algorithm will produce correlated nonnormal data (Headrick & Mugdadi, 2006), it does not indicate the degree to which deviations are influenced by sample size or the magnitude of the correlation. Hence, MC simulations were conducted to address these questions. To test the full range of correlation coefficients, a theoretical correlation matrix with values from 0.1 to 1.0 was input into the aforementioned algorithm. Unfortunately, this matrix was not positive semidefinite. Therefore, a separate algorithm, which will be described in the next section, was used to search for the closest positive semidefinite matrix. Table 4 presents both correlation matrices, where the lower triangle represents the theoretical matrix and the upper triangle represents the nearest positive definite matrix. Notice that the values of the upper triangle are very close to those of the lower triangle demonstrating that the method used to convert the negative definite matrix into a positive semidefinite matrix did not substantially alter the magnitude of the correlations. The remaining analyses were performed using the positive semidefinite matrix.

Table 4

Example of a Theoretical and its Nearest Positive Definite Correlation Matrices

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1.000</td>
<td>0.101</td>
<td>0.201</td>
<td>0.301</td>
<td>0.391</td>
</tr>
<tr>
<td>V2</td>
<td>0.100</td>
<td>1.000</td>
<td>0.501</td>
<td>0.600</td>
<td>0.686</td>
</tr>
<tr>
<td>V3</td>
<td>0.200</td>
<td>0.500</td>
<td>1.000</td>
<td>0.800</td>
<td>0.880</td>
</tr>
<tr>
<td>V4</td>
<td>0.300</td>
<td>0.600</td>
<td>0.800</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>V5</td>
<td>0.400</td>
<td>0.700</td>
<td>0.900</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5 presents the average deviation for each of the 20 distributions and 7 sample sizes. This table shows two subtle trends: deviations increase as a function of nonnormality but appear to converge (in probability\(^{50}\)) to the theoretical matrix as sample size increases. (Not included in this table, is the impact of the magnitude of the correlation coefficient on

\(^{50}\) That is, as \(N\to\infty\), the observed correlation \(\rho_0\) approaches the theoretical correlation \(\rho_T\). Stated formally, \(\lim_{N\to\infty} P(|\rho_0 - \rho_T| > \varepsilon) = 0\), where \(\varepsilon\) is a critical tolerance level greater than zero.
these deviations.) However, since these interpretations are based on observations, they are subject to subjectivity. Therefore, more formal tests of these relationships are warranted.

Table 5
Average Deviation in Correlation Coefficients by Sample Size and Distribution

<table>
<thead>
<tr>
<th>Skew</th>
<th>Kurt</th>
<th>N=10</th>
<th>N=150</th>
<th>N=300</th>
<th>N=460</th>
<th>N=630</th>
<th>N=810</th>
<th>N=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0</td>
<td>0.009</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>2.</td>
<td>0</td>
<td>0.011</td>
<td>0.004</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>3.</td>
<td>0</td>
<td>0.012</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>4.</td>
<td>0</td>
<td>0.013</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
<td>0.013</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>5.</td>
<td>0</td>
<td>0.014</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
<td>0.015</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>6.</td>
<td>0</td>
<td>0.014</td>
<td>0.015</td>
<td>0.016</td>
<td>0.017</td>
<td>0.017</td>
<td>0.016</td>
<td>0.017</td>
</tr>
<tr>
<td>7.</td>
<td>0.5</td>
<td>0</td>
<td>0.012</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>8.</td>
<td>0.5</td>
<td>0</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
<td>0.006</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>9.</td>
<td>0.5</td>
<td>0.012</td>
<td>0.008</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>10.</td>
<td>0.5</td>
<td>0.013</td>
<td>0.011</td>
<td>0.012</td>
<td>0.012</td>
<td>0.013</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>11.</td>
<td>0.5</td>
<td>0.014</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
<td>0.015</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>12.</td>
<td>1</td>
<td>0.015</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.010</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>13.</td>
<td>1</td>
<td>0.014</td>
<td>0.009</td>
<td>0.010</td>
<td>0.011</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>14.</td>
<td>1</td>
<td>0.014</td>
<td>0.011</td>
<td>0.013</td>
<td>0.013</td>
<td>0.013</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>15.</td>
<td>1</td>
<td>0.014</td>
<td>0.013</td>
<td>0.014</td>
<td>0.015</td>
<td>0.015</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>16.</td>
<td>1.5</td>
<td>0.021</td>
<td>0.017</td>
<td>0.018</td>
<td>0.018</td>
<td>0.018</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>17.</td>
<td>1.5</td>
<td>0.017</td>
<td>0.015</td>
<td>0.016</td>
<td>0.016</td>
<td>0.017</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>18.</td>
<td>1.5</td>
<td>0.016</td>
<td>0.015</td>
<td>0.017</td>
<td>0.017</td>
<td>0.017</td>
<td>0.016</td>
<td>0.017</td>
</tr>
<tr>
<td>19.</td>
<td>2</td>
<td>0.025</td>
<td>0.024</td>
<td>0.026</td>
<td>0.026</td>
<td>0.026</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>20.</td>
<td>2</td>
<td>0.021</td>
<td>0.022</td>
<td>0.023</td>
<td>0.024</td>
<td>0.024</td>
<td>0.023</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Parametric and nonparametric analyses were used to investigate the factors that contributed to deviations between the observed and theoretical correlations. Specifically, three set of analyses were performed to investigate the relationship between these deviations and the theoretical skewness and kurtosis, sample size, and the magnitude of the desired correlation. First, a stepwise regression analysis was conducted with SAS PROC REG to investigate the incremental significance of each independent variable. This analysis revealed a significant relationship between the absolute deviation and three of the factors (skewness, kurtosis, and sample size), $R^2=0.37$, $p<0.001$. Unfortunately, although the QQ plot of the residuals showed that the residuals were normal, a plot of the residuals by predicted scores showed that the residuals were “textbook” heteroscedastic (see Figure 3). Consequently, the
estimated variances and standard errors of the regression coefficients will either overestimate or underestimate the true variance and standard errors of the regression coefficients; and thus, the null hypothesis may be accepted when, in fact, it should be rejected, and vice versa.

![Graph](image)

Figure 3. Normality and Homoscedasticity Tests of Deviations in Correlations

To further examine the factors that contributed to deviations in correlations, a linear analysis was conducted using SAS PROC MIXED with maximum likelihood variance estimation. Unlike the previous analysis, the only statistical assumptions underlying this analysis are model linearity and normally distributed residuals. As a final model test, a stepwise rank regression was conducted, using SAS PROC REG, on the ranks of the dependent and independent variables. The results of all three analyses were found to support the subjective impressions derived from examination of the previous table. Table 6 (A)-(C) presents the results of these three regression models. Examination of these results revealed that both skewness and kurtosis contributed significantly to deviations in correlations.
Sample size was also found to be a significant predictor in the parametric but not the nonparametric analyses. However, its contribution was minor as witnessed by the low semi-partial r-squared. Interestingly, the magnitude of the desired correlations was found to be unrelated to deviations in correlations.

Table 6

Multiple Regressions of Deviations in Simulated Correlations

(A) Stepwise regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>--</td>
<td>--</td>
<td>0.003</td>
<td>0.000</td>
<td>--</td>
<td>0.000</td>
</tr>
<tr>
<td>Skewness</td>
<td>2</td>
<td>0.212</td>
<td>0.004</td>
<td>0.000</td>
<td>0.318</td>
<td>0.000</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1</td>
<td>0.154</td>
<td>0.001</td>
<td>0.000</td>
<td>0.418</td>
<td>0.000</td>
</tr>
<tr>
<td>Sample size</td>
<td>3</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.047</td>
<td>0.029</td>
</tr>
<tr>
<td>Correlation</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.369</td>
<td>N</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td></td>
<td>0.367</td>
<td>Normal test (p-value)</td>
<td></td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>Model (p-value)</td>
<td></td>
<td>&lt;0.001</td>
<td>Homoscedasticity test</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

(B) Mixed linear model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta</th>
<th>SE</th>
<th>β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.003</td>
<td>0.001</td>
<td>--</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.004</td>
<td>0.000</td>
<td>0.318</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.001</td>
<td>0.000</td>
<td>0.418</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sample size</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.047</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.002</td>
<td>0.916</td>
</tr>
<tr>
<td>R²</td>
<td>0.369</td>
<td>N</td>
<td></td>
<td>1400</td>
</tr>
</tbody>
</table>

(C) Stepwise rank regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>In</th>
<th>sr²</th>
<th>Beta</th>
<th>SE</th>
<th>β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>--</td>
<td>--</td>
<td>211.293</td>
<td>26.046</td>
<td>--</td>
<td>0.000</td>
</tr>
<tr>
<td>Skewness</td>
<td>2</td>
<td>0.153</td>
<td>0.264</td>
<td>0.023</td>
<td>0.257</td>
<td>0.000</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1</td>
<td>0.198</td>
<td>0.474</td>
<td>0.023</td>
<td>0.465</td>
<td>0.000</td>
</tr>
<tr>
<td>Sample size</td>
<td>3</td>
<td>0.002</td>
<td>-0.040</td>
<td>0.022</td>
<td>-0.039</td>
<td>0.068</td>
</tr>
<tr>
<td>Correlation</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.353</td>
<td>N</td>
<td></td>
<td></td>
<td>1400</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td></td>
<td>0.351</td>
<td>Model (p-value)</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

79
Although all the aforementioned analyses showed that the method developed by Headrick and Mugdadi (2006) for simulating correlated variables with the GLD is very good, caution is needed in interpreting results produced for really nonnormal distributions. However, given that the magnitude of the bias detected in these analyses is fairly small (<0.026), the impact on the interpretability of the results obtained from simulations in the remainder of this study will be negligible, particularly for moderate to large sample sizes.

Generating Covariance Matrices

Modeling real-world evaluation studies requires that one simulate the measurement of constructs—represented by latent variables in statistical models. Although the measurement of unobservable variables is a problem common to all sciences, it is especially common in the social and educational sciences because many of the constructs of interests cannot be directly measured by a single indicator (e.g., academic ability, psychological well-being) (Thum, 2005). In large part, this is true because the definition of many of these constructs is broader and their measurement is poorer than their natural science cousins. Consequently, the measurement of social and educational constructs often requires the use of structural equation modeling or triangulation (of data, investigator, theory, or method) to overcome the intrinsic weaknesses and biases of using a single source, theory, or method. The problem becomes considerably more challenging as one shifts from measuring a single construct to measuring the relationship among several constructs (Thum, 2005).

Covariance structural models (variance-covariance models) are a statistical means of modeling the relationships of the indicators used to measure constructs. Statistically, these models may be represented by a square $v \times v$ matrix $A$, where $v$ is the number of indicators, diagonal elements represent the variance of an indicator and off-diagonal elements represent
the covariance between two indicators. When all the variables are standardized, the variance-covariance matrix is known as a correlation matrix.

Numerous covariance structures may be utilized to model different situations. This study examined the impact of three covariance structures on the likelihood that a composite variable is approximately normally distributed. The simplest covariance structure is independence. An independent covariance structure is one in which the off-diagonal elements are equal to zero. This structure represents situations in which all of the indicators are statistically independent of each other (i.e., correlations equal zero). The second simplest covariance structure to model is compound symmetry. This structure occurs whenever the covariances of all the elements are equal to the same value. Therefore, an independent covariance matrix is really a special case of a compound symmetry structure in which the off-diagonal covariances are equal to zero. Finally, the most complex covariance of these three structures (from the perspective of the number of parameters estimated) is an unstructured matrix, which permits off-diagonal elements to take on any value.

Simulating independent and compound symmetry covariance structures was fairly straightforward. In both cases, the six step method for generating correlated data was used to simulate variables whose correlations with each other were equal to the desired correlation. Combined, these two structures examined the impact of 11 levels of correlations, ranging from 0 to 1 by 0.1, on the likelihood that a composite variable was normally distributed. Only nonnegative correlations were simulated because statistical theory dictates that the impact of negative correlations should be the same as their corresponding positive correlations. In contrast to the case with which one can simulate independent and compound symmetry covariance structures, simulating unstructured covariance matrices proved more challenging because not only must one simulate a covariance structure with
random off-diagonal elements but this structure must also be positive semidefinite to meet the requirements for performing a Cholesky decomposition.

Two methods were employed for converting negative definite matrices into positive definite or semidefinite matrices. In the first method, negative eigenvalues\(^{51}\) were replaced with positive random values ranging between zero and the maximum positive eigenvalue of the unstructured matrix. A new covariance matrix \(A\) was constructed by pre- and post-multiplying a diagonal eigenvalues matrix \(E\) by the eigenvector matrix \(V\) and its transpose \(V'\) (i.e., \(A = VEV')\) (Johnson & Wichern, 1998). Although replacing negative eigenvalues with random positive eigenvalues results in a new matrix that does not resemble the original matrix, this was not a problem since the goal was to simulate a random positive definite matrix. Of course, frequently one does want the new matrix to resemble the original matrix as closely as possible; therefore, a second method was employed.

In the second method, negative eigenvalues were replaced with zeros and the matrix was standardized. Replacing negative eigenvalues with zeros converts a negative definite matrix into a positive semidefinite matrix but, unfortunately, also shrinks the variance in the matrix. The impact of this procedure on the coefficients of the new matrix is a function of the number of negative eigenvalues in the original matrix relative to the total number of variables and the magnitude of the positive eigenvalues. The fewer the number of negative eigenvalues relative to the number of positive eigenvalues, and the larger the magnitudes of the positive eigenvalues, the smaller the deviations between the original negative definite matrix and the new positive semidefinite matrix will be. This is readily observed from the matrix formula above because the eigenvector matrix \(V\) and its transpose \(V'\) remain unchanged. Thus, deviations between the original and new matrix \(A\) are solely a function of

---

\(^{51}\) A negative definite matrix is a matrix with one or more negative eigenvalues.
the diagonal eigenvalues matrix $E$. Moreover, since each off-diagonal element in covariance matrix $A$ is a linear combination of the eigenvectors and eigenvalues then the previous conclusion must hold true. However, although this method will produce a positive semidefinite matrix, the diagonal elements will be greater than unity. Consequently, to generate a valid correlation matrix, one must standardize the new matrix by dividing each element by the product of the corresponding standard deviations.\textsuperscript{52} Although both methods produce nonnegative definite matrices, the first method is preferable to the second because it is computationally simpler to implement. Appendix C provides SAS code for generating positive definite correlation matrices.

Validating the method for simulating covariance matrices, however, requires more than demonstrating that it can generate desired correlations with minimal error. One must also show that the various covariance structures are in line with theoretical expectations and therefore, differ from each another. Consequently, a series of parametric and nonparametric statistical tests were developed to validate the covariance structures generated by the method described above. Twenty variables—one for each of the distributions included in the study—were simulated 1,000 times under a high sample size condition ($N=20,000$) to minimize the impact of the aforementioned deviations on the covariance structures. Thus, a total of 190 unique pairwise correlations were simulated 1,000 times. Confidence intervals were calculated for each correlation estimate, using Fisher’s $Z$ transformation, and an indicator variable was calculated based upon comparison to the expected correlation. A new variable was created by summing the values of the indicator across the number of simulations, where frequencies ranged from 0 to 1,000 and comparing this number to the number expected to be observed if the correlation estimates were associated with the

\textsuperscript{52} Note, this method may also be employed to correct negative eigen values of real correlation matrices.
theoretical structure. For instance, for 1,000 simulations and alpha 0.01, one would expect that 990 sample CIs would contain the theoretical rho, if the sample covariance structure matched the theoretical covariance structure. Using a Pearson Chi-square goodness-of-fit test, one can calculate the probability that the difference (or more extreme differences) between the observed and expected frequencies is consistent with the theoretical structure. In another words, the null hypothesis for this test is that the observed covariance structure is associated with the theoretical covariance structure. Hence, a low $p$-value indicated that there is significant departure between the observed and theoretical covariance structures.

This test was pretty straightforward for the independent and compound symmetry covariance structures because the theoretical correlation was constant. However, the pairwise correlations for the unstructured covariance structures vary. Hence, testing whether a covariance structure was unstructured was more challenging because no definitive tests of randomness exist. Instead, nonparametric tests must be combined to minimize the chance of missing a pattern in the data (Wang, 2003), where evidence of a pattern indicates the structure is not random. Three tests with different assumptions were selected to test for randomness. The simplest, and perhaps most popular test, was the Wald-Wolfowitz Runs Test (Wang, 2003). Another test for randomness performed was the Runs-Up and Down (RUD) Test (Wang, 2003). Unlike the Runs Test, which compares observations to the median, the RUD Test examines the pattern of sequential differences between observations and assigns a plus when $X_i - X_{i+1} > 0$ and a minus when $X_i - X_{i+1} \leq 0$. Finally, the Rank version of the von Neumann’s Ratio Test (Bartels, 1982) for randomness was performed on the pairwise correlations. This test ranks the correlations and then calculates the sum of the squared sequential differences, $\sum (R_i - R_{i+1})^2$. 
Table 7 presents the statistical conclusions and \( p \)-values for each of these analyses. Examination of the results revealed that the three simulated covariance structures met theoretical expectations and could be mutually distinguished (at \( \alpha=0.01 \)) from each other. One will note that the \( p \)-values were not significant for both the independent and compound symmetry structures when the true structure was independent. This is not surprising considering that independence is a special case of compound symmetry. Therefore, the statistical conclusion of independence was reached by virtue of the fact that the CIs included zero. However, when a compound symmetry covariance structure with a rho of 0.1 was simulated, the null hypothesis for the independent covariance Chi-square test was rejected. Overall, these tests demonstrated the validity of the method for simulating covariance matrices.

Table 7

Empirical Tests of Simulated Covariance Structure

<table>
<thead>
<tr>
<th>True state of the covariance structure</th>
<th>simulated Structure</th>
<th>Statistical Conclusion</th>
<th>Chi-square tests</th>
<th>Nonparametric tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent (IN)</td>
<td>IN</td>
<td>( &gt;0.999 )</td>
<td>( &gt;0.999 )</td>
<td>( &lt;0.001 )</td>
</tr>
<tr>
<td>Compound symmetry (CS) ((\rho=0.1))</td>
<td>CS</td>
<td>( &lt;0.001 )</td>
<td>( &gt;0.999 )</td>
<td>( &lt;0.001 )</td>
</tr>
<tr>
<td>Unstructured (UN)</td>
<td>UN</td>
<td>( &lt;0.001 )</td>
<td>( &lt;0.001 )</td>
<td>0.809 ( &gt;0.999 )</td>
</tr>
</tbody>
</table>

Monte Carlo Simulations of the Central Limit Theorem

As was discussed in the introduction of this chapter, several factors are likely to impact the Central Limit Theorem. It was hypothesized, for example, that a composite variable formed by averaging a set of highly skewed or kurtotic random variables is less likely to be approximately normally distributed than averaging a set of mildly skewed or kurtotic random variables. It was also argued that due to the Law of Large Numbers, composite variables that have a large sample size are more likely to be approximately normally
distributed than their small sample size counterparts. Yet another hypothesis proposed was that the greater the number of constituent variables aggregated, the greater the likelihood the composite variable would be approximately normally distributed. Lastly, since the CLT strictly states that the constituent variables should be independent, it was argued that the greater the correlation between constituent variables, the lower the likelihood the composite variable would be approximately normally distributed.

Model 1 (Independence)

The first model directly tested the Central Limit Theorem. According to the CLT, a composite variable formed by averaging (or summing) a set of independent and identically distributed random variables with finite variance will be asymptotically normally distributed as the number of variables aggregated grows without bounds. Notice then that no mention is made of the underlying distribution of the constituent variables. Hence, a series of nonnormal variables were generated with skewness ranging from 0 to 2 and kurtosis ranging from 0 to 10. As a result of the fact that only five levels of skewness and six levels of kurtosis were generated. PROC G3GRID was employed to interpolate values within this ranges in order to be able to generate smoother surface plots.\(^{53}\)

Figure 4 presents the relationship between skewness, kurtosis, and the probability the composite variable passed the normality test. Notice, the edge facing readers is jagged. Despite the fact that SAS interpolated values for the entire range space, only the values for which \(1.8(x_3^2+1)^{-3} \leq x_4\) were retained in keeping with the GLD constraint for values of skewness \(x_3\) and kurtosis \(x_4\). Examination of the figure confirmed the hypothesized relationship. Namely, as the values for skewness and kurtosis deviated from normality (i.e., \(x_3=x_4=0\)), the likelihood the composite variable passed the normality test decreased.

\(^{53}\) Interpolation was used for all the analyses herein in order to generate smooth surface plots.
Figure 4. Likelihood of Attaining an Approximately Normal Composite Variable Based on Skewness and Kurtosis, Where the Constituent Variables Were Independent

The second set of hypotheses tested were the relationships between sample size, the number of constituent variables aggregated, and the probability the composite variable would pass the normality test. It was hypothesized that as sample size increased the Law of Large numbers would kick-in (i.e., stabilize the parameter estimate of the composite variable), which, in turn, would improve the rate at which the composite variable converged upon the normal distribution. Examination of Figure 5 clearly shows that this relation did not hold. In fact, as sample size increased, the likelihood that the composite variable passed the test decreased. One potential reason for this may be that the Shapiro-Wilk W Test is sensitive to sample size. That is, perhaps as the sample size increases, so does the likelihood that the test will reject the null hypothesis (i.e., the sample is drawn from a normal distribution). Examination of the point biserial correlation $r_{pbis}$ between the dichotomous variable measuring whether the composite variable passed or failed the Shapiro-Wilk test and sample size lends credence to this hypothesis. For nonnormal constituent variables (i.e., $\alpha_3 \neq 0$ and $\alpha_4 \neq 0$), a statistically significant negative correlation was found ($r_{pbis} = -0.230$), whereas for
normal constituent variables, (i.e., $\alpha_i=\beta_i=0$), a nonstatistically significant correlation was found ($r_{pbis}=-0.001$). Of course, these analyses only show that the Shapiro-Wilk test is sensitive to detecting potential nonnormality. They do not provide any information as to the degree to which the composite variables violated normality.

Figure 5. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Sample Size and Number of Independent Constituent Variables Aggregated, Where Average Skewness Equals 1 and Average Kurtosis Equals 5

To shed further light on the nature of the relationship between sample size and the Shapiro-Wilk test, an additional MC study was conducted to examine the $p$-value produced by the normality test under various sample size conditions. A composite variable was generated by averaging 30 nonnormal random variables with a skewness of 2 and kurtosis of 10. Three thousand composite variables were generated whose values were sorted and then averaged, thereby generating an average composite variable for a given sample size. The ensuing $p$-values and probability density functions for four sample sizes are presented in Figure 6. Notice that despite the obvious deviation from normality for the N=50 condition, the Shapiro-Wilk test failed to reject the null hypothesis ($p>0.15$) and hence, concluded that
the composite variable was normally distributed. Similarly, the test failed to reject the null hypothesis for the N=150 and N=250 conditions despite what appears to be a fairly normal observed probability distribution. However, for the N=500 condition, the Shapiro-Wilk test rejected the null hypothesis ($p \leq 0.15$) and hence, concluded that the composite variable was not normally distributed. What is important is not whether the composite variable was in fact normally distributed but rather, the fact that the Shapiro-Wilk test rejected the null hypothesis for the last condition but not the prior two conditions for what appears to be trivial changes in the probability density functions. Hence, the test is sensitive to sample size.

![Composite Normal Distribution](image)

**Figure 6.** Probability Density Functions for Four Composite Variables Generated by Averaging 30 Constituent Variables With a Skewness of 2 and Kurtosis of 10

It is conceivable, however, that the appropriate decision may have been to reject the null hypothesis because there is a noticeable, albeit small, discrepancy between the observed probability distribution and the normal curve. However, this discrepancy can be largely erased by aggregating more variables. Figure 7, for example, represents the probability
distribution of a composite variable generated by aggregating 150 random nonnormal variables with a skewness of 2 and kurtosis of 10. Notice, the observed and expected probability distributions are virtually indistinguishable from each other. Yet, despite this, the Shapiro-Wilk test rejected the null hypothesis for the N=2,000 condition but not for the other conditions. The result of this analysis for small sample sizes is in-line with prior research (SAS Institute, 2007), which found that the test lacks the power to detect large departures from normality. However, this analysis revealed that for large sample sizes, the test may be overly sensitive leading to rejection of the null hypothesis for trivial differences.

While this analysis calls into question the observed probability levels for the tests performed in this section, these results do not necessarily invalidate the remainder of study. The purpose of this section was to investigate the nature of the relationship between various factors and the probability that the composite variable is normally distributed. The flaws
described above notwithstanding, the Shapiro-Wilk test generates a p-value that is the best index currently available (for N≤2,000) for determining whether a variable is normally distributed.\textsuperscript{54} Moreover, since the sample size values were independent of the other factor levels then, while the probabilities reported herein may be underestimated, the nature of the remainder of the relationships of interest were not affected. With respect to the number of constituent variables aggregated, it is clear from Figure 5 that consistent with the CLT, as the number of variables aggregated grows, the composite variable approaches the normal distribution. Hence, a finite number of constituent variables exists for which a composite variable will attain a desired likelihood of being normally distributed. For example, according to Figure 5, given a sample size of about 1,000, one needs to aggregate about 300 variables in order to have a probability of about 75 percent that the composite variable is normally distributed. In all likelihood, given the results for the previous analysis, the number of constituent variables needed may be considerably lower. Moreover, herein the constituent variables had a skewness of 2 and kurtosis of 10. However, in practice, one rarely encounters such nonnormal variables. In fact, a study by the present author of 1,798 variables collected as part of 18 evaluation studies revealed an average skewness of 0.43 and kurtosis of 1.69. Hence, an evaluator would need to aggregate considerably less than 300 variables to attain a probability of 75 percent or higher that the composite variable is normally distributed. SAS code for the analyses contained in this section may be found in Appendix D.

Model 2 (Dependence)

The purpose of the second model was to investigate the degree to which the CLT may be expanded to include the aggregation of dependent and identically distributed random variables. Further, analysis (not included) revealed a similar relationship for N>2,000 between sample size and passing or failing the Kolmogorov-Smirnov test, which is the preferred method (SAS Institute, 2007) for testing normality when N is large.

\textsuperscript{54} Further, analysis (not included) revealed a similar relationship for N>2,000 between sample size and passing or failing the Kolmogorov-Smirnov test, which is the preferred method (SAS Institute, 2007) for testing normality when N is large.
variables with finite variance. It was hypothesized that averaging correlated random variables would reduce, but not preclude, the likelihood that the composite variable would be approximately normally distributed. Hence, in addition to the study design employed for Model 1, the second model examined the impact of 11 levels of rho (population correlation) on the probability of passing or failing the Shapiro-Wilk test. Each of the previous analyses were replicated in order to facilitate comparisons between the independent and dependent models. Once again, PROC G3GRID was employed to interpolate values within the ranges of the factors in order to be able to generate smoother surface plots.

Figure 8 presents the relationship between skewness, kurtosis, and the probability the composite variable passed the normality test. Examination of the figure confirmed that as the values for skewness and kurtosis deviated from normality, the likelihood the composite variable passed the normality test decreased. Comparison to Figure 4 also revealed that while the nature of these relations were similar, the overall level at which the composite variable passed the normality test was considerably lower when dependent variables were aggregated.

Figure 8. Likelihood of Attaining an Approximately Normal Composite Variable Based on Skewness and Kurtosis, Where the Constituent Variables Were Dependent
In large part, these low rates are most likely due to the negative impact sample size has on passing the Shapiro-Wilk test. However, there should be no doubt that aggregating dependent constituent variables further reduces the likelihood that the composite variable will be normally distributed, which is readily observed in Figure 9. Notice that as sample size increased, the likelihood of passing the normality test decreased, whereas as the number of dependent constituent variables increased, so did the probability that the composite variable would pass the normality test. However, a greater number of correlated constituent variables must be aggregated in order to attain the same probability of passing the normality test as would occur had independent constituent variables been aggregated.

The impact of sample size was explored further by assessing the role compound symmetry played in passing the normality test. Figure 10 illustrates that as the correlation increased, the probability of passing the normality test decreased. Furthermore, the combination of large samples and highly correlated constituent variables yielded the lowest
probability levels. However, it is important to recall that due to problems with the Shapiro-Wilk test, the true probability of a composite variable being normally distributed is considerably higher than these observed levels. Figure 11 illustrates the relationship between the probability of passing the normal test and the number of constituent variables aggregated given various correlation levels. Notice that as the number of variables aggregated slightly increased, so too did the probability of passing the normality test. Clearly then, given a fixed correlation and enough constituent variables, one can attain any probability for the composite variable. In practice, the average correlation between variables are likely to be lower than the average value of 0.5 used in these simulations. Hence, only a modest number of variables would need to be aggregated for a composite variable to be normally distributed. SAS code for the analyses contained in this section may be found in Appendix E.

Figure 10. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Sample Size and Correlation Simulated for the Constituent Variables, Where Average Skewness Equals 1 and Average Kurtosis Equals 5
Figure 11. Likelihood of Attaining an Approximately Normal Composite Variable Based on the Number of Variables Aggregated and the Correlation Simulated for the Constituent Variables, Where Average Skewness Equals 1, Average Kurtosis Equals 5, and Average Sample Size Equals 480

Future Research on Testing for Normality

Unfortunately, the results from Models 1 and 2 failed to identify the specific conditions under which it is reasonable to assume the distribution of a composite variable is asymptotically normal, given specific levels for skewness, kurtosis, the average correlation among the constituent variables, and the number of variables aggregated. In large part, this failure occurred because the Shapiro-Wilk test was found to lack the statistical power to detect large departures from normality for small sample sizes yet was overly sensitive when it came to detecting trivial departures from normality for large sample sizes. It may also be the case that these problems were an artifact of the statistical framework underlying the test. Specifically, the Shapiro-Wilk test aims to demonstrate that the estimate (e.g., mean) found for one group is superior (greater) than that found for another group. Yet, what one is really interested in knowing is whether the population distribution from which the sample is drawn is “reasonably close” to the normal distribution so as to permit researchers to perform
A statistical test that seeks to determine whether the null and alternative hypotheses are equal given a margin of error, rather than different, is known as equivalence testing (Merklin & Hirtz, 2007; Skaff & Sloan, 1998). Unlike tests of superiority\textsuperscript{55}, for equivalence tests the null hypothesis ($|\mu_1-\mu_2|\geq\Delta$) is that the two population parameters are different from each other, whereas the alternative hypothesis ($|\mu_1-\mu_2|<\Delta$) is that the two population parameters are equal to each other, given a margin of error $\Delta$ determined by the researcher \textit{a priori}. Developing a statistical test that measures the goodness of fit of a sample distribution to a hypothesized distribution, assuming that such a test does not already exist, is beyond the scope of this study. However, it is not difficult to imagine how such a test should operate. In fact, Figure 6 provides clues as to how such a test can be developed. Suppose that “reasonably close” is operationalized as a 1 percent difference in the nonoverlapping probabilities of the unknown population distribution and the hypothesized distribution. Then, the area under the lower of the two curves (distributions) denotes the probability in common to both distributions. Hence, since the area under a probability curve must sum to unity (Ross, 2002) then 1 minus the shared probability denotes the amount (probability) of nonoverlap between the unknown population distribution and the hypothesized distribution. If this difference, in turn, is less than the margin of error (say, $\Delta=0.01$) then the two distributions will be statistically equivalent and one can perform statistical inference using the hypothesized distribution.

\textsuperscript{55} According to Merklin and Hirtz (2007, p. 315), “many researchers often incorrectly conclude that the failure to reject the null hypothesis in a standard hypothesis test (such as the $t$ test) is ‘proof’ that the null hypothesis is true and hence that the populations are ‘equivalent.’ This erroneous inference neglects the possibility that the failure to reject the null is often merely indicative of a Type II error, particularly when the sample sizes being used are small and the power is low.”
It is difficult to predict the results that would emerge if the simulations performed for Models 1 and 2 were replicated and an equivalence test replaced the Shapiro-Wilk test. Due to the nature of equivalence testing, the sensitivity to large sample sizes would be eliminated since trivial departures from normality would no longer trigger the rejection of the null hypothesis. However, it is conceivable that for small sample sizes, the test might fail to reject the null hypothesis (population distributions are different) even when the differences between the two distributions are trivial. Therefore, it is conceivable that the specific impact of the aforementioned factors on the normality of a composite variable will be more representative of reality than the results reported herein. However, the pattern of the results found in this study will continue to hold true. Namely, the greater the nonnormality or dependence of the constituent variables, the more variables one would need to aggregate before the composite would be approximately normally distributed.

Summary

The primary purpose of this chapter was to validate the Monte Carlo methodology used in the next two chapters to investigate the validity of the algorithm. The first half of this chapter presented the methods that would be employed to simulate variables with known values for skewness, kurtosis, and the correlation coefficient. The GLD method was used to generate nonnormal data with inconsequential deviations between the theoretical and observed skewness and kurtosis. Moreover, it was shown that by adopting the method proposed by Headrick and Mugdadi (2006), the GLD can produce correlated nonnormal variables with only minor deviations between the theoretical and observed correlations. Furthermore, it was shown that these method can be employed to generate specific covariance matrices for nonnormal random variables. Although some caution is warranted
when sample size is small, the analyses herein clearly demonstrated that it is possible to simulate random variables with known parameters with very little error.

The construction of a parametric confidence interval requires a distributional assumption. In fact, in the next two chapters, it will be necessary to assume that the sampling distribution and measurement error are normally distributed. In the case of the former, a MC will be used to illustrate the veracity of the assumption. However, although the latter assumption is frequently stated as a matter of convention and convenience, no further proof is ever offered for its veracity. Hence, the secondary goal of this chapter was to illustrate how a MC study can be employed to investigate this question and others like it.

Since measurement error can be construed as the amalgamation of a vast number of unknown processes (variables), the second half of the chapter focused on the likelihood that a composite variable formed by averaging a set of constituent variables was normally distributed. According to the CLT, as the number of variables that are independent and identically distributed grows without bounds (i.e., approaches infinity), the composite variable formed from their average becomes asymptotically normally distributed. Therefore, Model 1 sought to examine the CLT under its specified assumptions. Since the CLT makes no assumptions about the distribution of the constituent variables, various nonnormal distributions were employed herein. As expected, the more variables aggregated, the higher the likelihood their composite was normally distributed. The limits of the CLT were then examined in Model 2 by relaxing the independence assumption. These results confirmed that the composite of dependent random variables is also asymptotically normally distributed, particularly as the number of variables grows without bounds. Therefore, the assumption that measurement error is normally distributed is reasonable given that it is the product of a vast number of random variable, which in all likelihood are only mildly correlated.
Chapter IV
Implications of Mathematical Statistics

The primary purpose of evaluation is the judgment of the merit and worth of an evaluand. Typically, summative evaluations are designed to address questions such as “How good is the overall performance of the evaluand?”, “Have the needs of impactees been met?”, or “Are the results worth the cost?” For each of these questions, an evaluator needs to identify relevant values upon which to judge the performance of the evaluand, valid standards against which to compare the performance of the evaluand, and a scoring or grading rubric that can be used to synthesize the performance of the evaluand into a summative conclusion. Informative as the answer to these questions may be, however, they do not reveal the range of possible answers that may be supported by the data. For such knowledge, one needs to turn to confidence intervals.

This chapter will examine the mathematical relationship between confidence intervals and specific factors mentioned in the first chapter. The chapter is divided into six sections. The first section, Expectation Theory, will formally define the concepts of expected value and variance for single and composite variables. Namely, how can one construct a CI for a simple composite variable. The second section, Distribution Theory, will introduce readers to standard discrete and continuous probability distributions. The third section, Transformation Theory, will present several options evaluators can use to transform dissimilar distributions into a common distribution en route to constructing a composite variable. The fourth section, Sampling Theory, will discuss how to adjust the basic variance formulas to account for sampling error. The fifth section, Confidence Intervals, will present
three different methods for constructing a confidence interval. Lastly, the sixth section will present a mathematical algorithm for constructing a basic Summative Confidence interval.

**Expectation Theory**

**Definition of Expected Value**

One of the most important concepts in probability theory is that of the expectation of a random variable. The expected value of a random variable $X$, denoted as $E(X)$, is the mean of the variable over a large number of repetitions (Ross, 2002; Meyer, 1970; Grinstead & Snell, 1997). That is, if an infinite number of independent replications of a random variable $X$ occurs, then the probability variable $X$ will take on a specific value $x$ is equal to the proportion of replications in which $x$ occurs, $P_X(x)$. Given enough replications, every possible value, $x_i$ of variable $X$ will occur with probability $P_X(x_i)$. Therefore, the best guess of the true value of $X$ is its expected value. For a discrete random variable, the expected value is defined as $E(X) = \sum_{i=1}^{\infty} x_i P_X(x_i)$, provided the series converges—i.e., the discrete sum ($\sum$) of the crossproduct of the absolute values ($x_i$) of variable $X$ and their corresponding probability, $P_X(x_i)$, is less than infinity ($\infty$). For a continuous random variable, the expected value is defined as $E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx$, provided the improper integral (i.e., an integral with a range space of $[-\infty, \infty]$) is finite—i.e., the continuous sum ($\int$) of the crossproduct of the absolute values of variable $X$ and their corresponding probability density function (pdf) at $x$, $f_X(x)$, is less than infinity. Stated more simply, the expected value for both discrete and continuous random variables is the sum of the weighted average of all the possible values of $X$, where the weight is determined by the probability of the variable at $x$.

---

56 Note, in keeping with standard mathematical notation, capital letters will be used to denote random variables whereas lowercase letters will be used to represent particular values in the range of the random variables. Moreover, while it may be technically more accurate to express the probability that a random variable $X$ will equal a particular value $x$ as $P(X=S=x)$ to acknowledge the functional dependence of $X$ upon the elements of the sample space $S$ associated with the variable, this notation represents a superfluous step that will be omitted.
Properties of Expected Value

The properties of expected value, which will be very useful in subsequent work, are enumerated below. These properties and their proofs can be found in most mathematical-statistics textbooks (e.g., Meyer, 1970; Larson, 1974; Grinstead & Snell, 1997; Ross, 2002; Khuri, 2003; Hogg, McKean, & Craig, 2005). The proofs are also available from the author upon request along with SAS Monte Carlo simulations that illustrate each property.

Definitions. Let \( \mathbb{E} \) represent the expected value of a random variable; let \( X \) and \( Y \) be either two discrete random variables, \( X=\{x_1, x_2, \ldots, x_n\} \) and \( Y=\{y_1, y_2, \ldots, y_n\} \), or two continuous random variables, \( X=\{-\infty < x < \infty\} \) and \( Y=\{-\infty < y < \infty\} \); let \( a \) and \( b \) be constants; let \( g(x) \) denote a real-valued one-dimensional function defined on the domain of \( X \); let \( g(x, y) \) denote a real-valued function defined on the two-dimensional set of points \((x, y)\) that are interior to, or on the boundary of, a square with opposite vertices at \((x_1, y_1)\) and \((x_n, y_n)\) for a discrete random variable; let \( g^{-1}(y) \) denote the inverse function of \( g \) such that \( x = g^{-1}(y) \) for a one-dimensional function and point \((u, v) = g^{-1}(x, y)\) for a two-dimensional function; let \( P_X(x) \) denote the probability mass function (pmf) of random variable \( X \) at value \( x \), i.e., \( P_X(x) = P(X=x) \); let \( f_X(x) \) denote a probability density function (pdf) of random variable \( X \) where every value \( x \) exists in domain \((a, b)\) such that \( f_X(x) = P(a < x < b) \); let \( F_X(x) \) denote the cumulative distribution (density) function (cdf) of a random discrete or continuous variable \( X \) measured at value \( x \), \( F_X(x) = P(X \leq x) \); let \( P_{X,Y}(x_i, y_j) \) denote the joint pmf of two discrete random variables \( X \) and \( Y \) at values \( x_i \) and \( y_j \) respectively, i.e., \( P_{X,Y}(x_i, y_j) = P(X=x_i, Y=y_j) \); let \( f_{X,Y}(x, y) \) denote the joint pdf of two continuous random variables \( X \) and \( Y \) where \( X \) exists in domain \((a, b)\) and \( Y \) exists in domain \((c, d)\) such that \( f_{X,Y}(x, y) = P(a < x < b, c < y < d) \); let \( |\cdot| \) represent the absolute value of the function contained within the vertical bars; let the symbol \( | \) represent a conditional statement such that \( "X|Y" \) denotes \( "X \text{ given } Y" \), let \( P_{X|Y}(x|y) \)
denote the discrete conditional expectation of \( X \) given \( Y = y \); let \( f_{X|Y}(x|y) \) denote the continuous conditional expectation of \( X \) given \( Y = y \); let \( \rho_{XY} \) or simply \( \rho \) denote the correlation coefficient between variables \( X \) and \( Y \); and let the prime ' and double prime '' symbols denote the first and second derivatives of function \( g \).

1. **Expected value of a constant.** If all the values of variable \( X \) are the same, say constant \( a \), then the expected value is the constant, \( E(X=a)=a \).

2. **Adding a constant.** Adding (or subtracting) a constant \( a \) to every value of a variable results in the addition (or subtraction) of the constant from the expected value, \( E(X+a)=E(X)+a \).

3. **Multiplying by a constant.** If the values of a variable are multiplied (or divided) by a constant number, then the expected value is multiplied (divided) by the constant, \( E(aX)=aE(X) \).

4. **Linear transformation.** From properties 2 and 3 it should readily follow that a linear transformation (i.e., \( a+bX \)) of a variable has the same effect on the expected value as adding (subtracting) the constant \( a \) to the product (division) of the expected value of variable \( X \) by constant \( b \). That is, \( E(a+bX)=a+bE(X) \).

5. **One-dimensional function.** If a random variable \( Y \) is equal to the function \( g \) of a random variable \( X \), i.e., \( Y=g(X) \), then the expected value of \( Y \) is simply the product of the values of the function and their respective probabilities. In other words, the expected value of \( Y \) can be determined by knowledge of the pdf of \( X \) without the need to find the pdf of \( Y \). There are times, however, when it may become necessary to compute the expected value of a function from the pdf of \( Y \) rather than \( X \). This is particularly important in instances in which one will need to transform one distribution into another (see Transformation Theory section).
If $Y = g(X)$ is a discrete random variable, based on a single-valued transformation of $X$, with pdf $P_Y(y) = P(Y \leq t)$ defined by $F_Y(t) = \sum_{j=1}^{n} P_Y(y_j)$ for any real number $t$. Then, function $g$ has an inverse $g^{-1}$ such that $x = g^{-1}(y)$, which can be used to determine both the pdf, $P_Y(y) = P_X[g^{-1}(y)]$, and cdf, $F_Y(y) = P(Y \leq y) = P[X \leq g^{-1}(y)] = F_X[g^{-1}(y)]$. Therefore, substituting $g^{-1}(y)$ for $x$ in the definition of expected value yields $E(Y) = \sum_{j=1}^{n} y_j P_X[g^{-1}(y_j)]$.

6. **Multi-dimensional function.** If $Y = g(x_1, x_2)$ is a two-dimensional function of random variables $X_1$ and $X_2$, then the expected value of $Y$ is simply the product of the values of the function and their respective joint probabilities. If $X_1$ and $X_2$ are independent, then their joint pdf is equal to the product of their individual pdfs. Note, the expectation for a two-dimensional function can be readily extended to a $n$-dimensional function.

Alternatively, the expected value of a two-dimensional function can be obtained from its inverse transformation provided a one-to-one relationship exists between the domain and range. If $Y = g(X_1, X_2)$ is a discrete random variable where $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ represent two parametric equations\(^{57}\) that map the point $(x_1, x_2)$ onto point $(y_1, y_2)$, then $g$ has a unique, single-valued inverse transformation $g^{-1}$ such that $x_1 = g_1^{-1}(y_1, y_2)$ and $x_2 = g_2^{-1}(y_1, y_2)$ map the point $(y_1, y_2)$ onto $(x_1, x_2)$. Moreover, the pdfs of $g$ and $g^{-1}$ must be equal because of the one-to-one relationship.

---

\(^{57}\) Parametric equations—where the coordinates of a point, say $(u, v)$—are defined based upon functions that depend upon one or more parameters, e.g., $u = g_1(x_1, x_2)$ and $v = g_2(x_1, x_2)$. Therefore, transforming a set of parametric equations entails eliminating the parameters from the simultaneous equations. Typically, this may be accomplished by solving one of the equations for one of the parameters and then substituting this into the other equation to obtain an equation that involves $x$ and $y$ only. This is known as Gaussian elimination.
one relationship between the domains associated with points \((x_1, x_2)\) and \((y_1, y_2)\). Therefore, the pdf \(P_{Y_1, Y_2}(y_1, y_2) = P_{X_1, X_2}[g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)]\), which connotes that the cdf \(F_{Y_1, Y_2}(y_1, y_2) = P[X_1 \leq g_1^{-1}(y_1, y_2), X_2 \leq g_2^{-1}(y_1, y_2)] = F_{X_1, X_2}[g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)]\). Substituting \(g_1^{-1}(y_1, y_2)\) for \(x_1\) and \(g_2^{-1}(y_1, y_2)\) for \(x_2\) in the definition of expected value, then yields

\[
\mathbb{E}[g(X_1, X_2)] = \sum_{y_1=1}^{\infty} \sum_{y_2=1}^{\infty} g_1^{-1}(y_1, y_2) g_2^{-1}(y_1, y_2) P_{X_1, X_2}[g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)],
\]

where \((y_1, y_2)\) is a single point in the Cartesian plane and \(y_1\) and \(y_2\) denote the axes.

7. Approximate expectation. If an expectation is too difficult to integrate, a suitable alternative may be to use the Taylor series expansion to approximate it. Therefore, if the function is twice differentiable at \(X=\mu\), then the expected value of the function may be approximated by a second-order Taylor polynomial. Although the Taylor series may be extended beyond this order, doing so entails a considerable amount of work for a small gain in precision, especially if the function is well-behaved in the domain of interest.

For a one-dimensional function \(g(X)\) using a second-order Taylor polynomial:

\[
\mathbb{E}[g(X)] \approx g(\mu) + \frac{1}{2} g''(\mu) \text{Var}(X).
\]

For a two-dimensional function \(Z=g(X,Y)\) using a second-order multidimensional Taylor polynomial evaluated at \(Z=g(\mu_X, \mu_Y)\):

\[58\] Whenever a transformation is one-to-one, the probability an element exists in \((\mathcal{E})\) domain \(X (D_X)\) is equal to the probability the reciprocal element exists in domain \(Y (D_Y)\). That is, \(P[(X, Y) \in D_X] = P[(Y, X) \in D_Y]\).

\[59\] Generally, it is easier to determine the derivative of a function than it is to determine its integral. Consequently, if an expectation requires difficult integrations, it may be simpler to calculate an approximation. A Taylor series represents a function \(g\) as an infinite sum of terms centered at point \(a\). Mathematically, it is expressed as

\[
g(X) = \sum_{n=0}^{\infty} \frac{g^{(n)}(a)}{n!}(X-a)^n = g(a) + g'(a)(X-a) + \frac{1}{2} g''(a)(X-a)^2 + \frac{1}{6} g'''(a)(X-a)^3 + \frac{1}{24} g''''(a)(X-a)^4 + \cdots
\]

Moreover, it has been shown that accurate estimates may be obtained with a finite number of terms, providing the function is well-behaved around \(a\) and the range is limited. Therefore, a \(n^{th}\)-order Taylor polynomial denotes a Taylor series in which only the first \(n\) terms are summed. A multidimensional analogue of this theorem also exists. Interested readers are directed to Khuri (2003).
\[ E[Z] \sim Z + \frac{1}{2} \left[ \text{Var}(X) \frac{\partial^2 Z}{\partial x^2} + \text{Var}(Y) \frac{\partial^2 Z}{\partial y^2} + 2 \text{Cov}(X,Y) \frac{\partial^2 Z}{\partial x \partial y} \right]_{Z=g(x,y)}. \]

Note, for independent variables, the covariance term drops out. Furthermore, for linear functions, the Taylor series reduces to just the first term. That is, it reduces to the regular expected value. Therefore, the approximate expected value should only be used for nonlinear functions.

8. **Conditional expectation.** Suppose X and Y are two jointly distributed random variables where X contains limited information about Y. Then the conditional expected value of Y given this information is simply the sum of the product of each value of Y and its corresponding conditional probability given X. This conditional expectation, \( E(Y \mid X) \), in turn is itself a random variable, which has an expected value. Hence, it can be shown that \( E[E(Y \mid X)] = E(Y) \), which is known as the iterated property of expectations. In other words, the mean of the conditional expectation of Y with respect to X is the same as the mean of Y. Moreover, if the relationship between X and Y is linear (i.e., \( \hat{Y} = a + bX \)) then the best linear predictor of Y must minimize the discrepancy between Y and the predicted value \( \hat{Y} \). Hence, it can be proven that the linear relationship between X and Y that minimize \( E[(Y - \hat{Y})^2] \) is given by \( E(Y \mid X) = E(Y) + \rho_{XY}(\sigma_Y / \sigma_X)[X - E(X)] \), where \( \sigma_Y \) and \( \sigma_X \) denote the population standard deviations of Y and X.

9. **Sum of random variables.** An important application of Property 6 occurs when \( g(x_1, x_2) \) is the sum (difference) of two finite random variables. Then the expected value of their sum (difference) is equal to the sum (difference) of their expected values, \( E(X+Y) = E(X) + E(Y) \). Note, this expectation can be readily extended to a n-dimensional function.
10. Mean of random variables. Another important property can be derived, by applying mathematical induction to Property 6. If \( g(\bar{X}) \) represents the mean of a sequence (i.e., a vector) of \( k \) random variables, \( X_1, X_2, \ldots, X_k \), then the expected value of the sample mean is equal to the mean of the sequence of the expected values. That is, \( \mathbb{E}[g(\bar{X})] = \mathbb{E}(X) \), where \( \mathbb{E}(X) \) represents the expected value of the sequence and not the expected value of an individual variable, which is denoted by \( \mathbb{E}(X_i) \).

**Definition of Variance and Covariance**

While the expected value of a variable or function is important for formulating an evaluable conclusion, at least to the degree to which the conclusion is derived from quantitative analysis, the principal concept that underlies Summative Confidence is variance. Variance, denoted as \( \text{Var} \) or \( \sigma^2 \), is a measure of the statistical dispersion of a set of scores around the expected value of a random variable. Mathematically, it is defined as the expected value of the squared difference of each observation and the expected value, i.e., \( \sigma^2_X = \text{Var}(X) = \mathbb{E} \left\{ (X - \mathbb{E}(X))^2 \right\} \), which can also be expressed as the expected value of the square of each observation minus the square of expected value of a variable, i.e., \( \sigma^2_X = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \). For discrete random variables, the computational formula for variance is \( \sigma^2_X = \sum_{x}^\infty (x - \mathbb{E}(X))^2 P(x) \), where \( \mathbb{E}(X) = \sum_{x}^\infty x P(x) \), and for continuous random variables, the computational formula is \( \text{Var}(X) = \int_{-\infty}^{\infty} [x - \mathbb{E}(X)]^2 f(x) \, dx \), where \( \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \, dx \) (Ross, 2002; Larson, 1974).

As these definitions show, the smaller the discrepancy between each observation and the variable’s expected value, the smaller the variance. This implies that the precision of an estimate is a function of the variance with which the variable is measured. However, because variance is reported in squared units, most people prefer to use the square root of the variance (i.e., the standard deviation). A related concept that merits mentioning is covariance.
(Cov), which measures the degree to which two variables vary together. Mathematically, it is expressed as the expected value of the product of the deviation of X from its expected value and Y from its expect value, i.e., $\sigma_{XY} = \text{Cov}(X,Y) = \mathbb{E} \{ [(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \}$, which can also be expressed as the expected value of the product of observations X and Y minus the product of the expected value of each variable, i.e., $\sigma_{XY} = \text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ (Ross, 2002; Larson, 1974; Hogg, McKean, & Craig, 2005).

The following section enumerates the properties of variance and covariance that will be needed later to develop the statistical algorithm necessary to account for several of the factors outlined in the first chapter. These properties and their proofs can be found in most mathematical-statistics textbooks (e.g., Meyer, 1970; Larson, 1974; Grinstead & Snell, 1997; Ross, 2002; Khuri, 2003; Hogg, McKean, & Craig, 2005). Therefore, a proof is only included for the last property since it was derived by the author and it forms the basis upon which Summative Confidence will be built upon. All the other proofs are also available from the author upon request along with accompanying SAS Monte Carlo simulations.

**Properties of Variance and Covariance**

To fully understand how multiple variances can be aggregated or adjusted for methodological characteristics, one must first grasp the properties of variance. In each instance, it is assumed that variance is finite and nonnegative. These assumptions readily follow if the expected value exists (i.e., X is not an infinite series) because the variance must be finite and cannot be negative due to the squaring of deviations.

*Definitions.* In addition to the definitions provided for the properties of expected value, let $\text{Var}$ denote the variance such that $\text{Var}(X)$ and $\sigma^2_X$ represent the variance of X; let
$Cov$ denote the covariance such that $Cov(X,Y)$ and $\sigma_{XY}$ represent the covariance of variables $X$ and $Y$; and let $w_1$ and $w_2$ be constants that represent weights assigned to a variable.

11. **Variance of a constant.** If all the values of a variable are the same, say constant $a$, then no variability exists within the variable, $\text{Var}(X=a)=0$.

12. **Adding a constant.** If a constant $a$ is added (or subtracted) from every value of a variable, its variance will be unchanged, $\text{Var}(X+a)=\text{Var}(X)$.

13. **Multiplying by a constant.** If the values of a variable are multiplied (or divided) by constant $a$, then the variance is multiplied (divided) by the square of the constant, $\text{Var}(aX)=a^2\text{Var}(X)$. Likewise, the covariance of two random variables, which are each multiplied by constants $a$ and $b$, is the product of the constants and the covariance of the two variables, $\text{Cov}(aX,bY)=ab\text{Cov}(X,Y)$. Notice then that the property for variance above is in fact a special case of the covariance property, where the two constants are equal, since $\text{Var}(X)$ is equal to $\text{Cov}(X,X)$.

14. **Linear transformation.** From Properties 12 and 13 it should readily follow that a linear transformation (i.e., $a+bX$) of a variable has the same effect on variance as expressed by Property 13. That is, the variable will equal the product of the squared constant and the variance of $X$. In this case, $\text{Var}(a+bX)=b^2\text{Var}(X)$.

15. **One-dimensional function.** If $Y=g(X)$ then the variance of $Y$ is simply the expected value of the squared discrepancy between $g(X)$ and its expected value, i.e., $\text{Var}(Y)=\text{Var}[g(X)]=\mathbb{E}\{(g(X))-\mathbb{E}[g(X)]\}^2$. If a one-to-one relationship exists between the points that comprise the domains of $X$ (denoted $D_X$) and $Y$ (denoted $D_Y$), then $P(x\in D_X)=P(y\in D_Y)$. In other words, the variance of $Y$ can be determined by knowledge of the pdf of $X$ without the need to find the pdf of $Y$. That is, one
can determine the variance of $Y$ through the use of inverse transformations, where $x = g^{-1}(y)$. For a discrete variable then

$$\text{Var}(Y) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \left[ \left( g_{\delta_1}^{-1} (y_{i,j}) - g_{\delta_2}^{-1} (y_{i,j}) \right) - \mathbb{E}(Y) \right]^2 P_{X_i,X_j} \left[ g_{\delta_1}^{-1} (y_{i,j}) , g_{\delta_2}^{-1} (y_{i,j}) \right].$$

16. **Multi-dimensional function.** If $Y = g(x_1, x_2)$ represents a two-dimensional function, then, like its one-dimensional analog, the variance of $Y$ is simply the squared discrepancy between $g(x_1, x_2)$ and its expected value, i.e.,

$$\text{Var}(Y) = \mathbb{E} \left\{ g(x_1, x_2) - \mathbb{E}[g(x_1, x_2)] \right\}^2.$$ Note, the variance of a two-dimensional function can be extended to a $n$-dimensional function.

Alternatively, the variance of a two-dimensional function, $Y = g(x_1, x_2)$, can be obtained from its inverse transformation provided a one-to-one relationship exists between the two domains. Suppose function $g$ maps point $(x_1, x_2)$ onto point $(y_1, y_2)$ and $g^{-1}$ represents its unique, single-valued inverse transformation that maps point $(y_1, y_2)$ back onto point $(x_1, x_2)$. That is, function $g$ transforms parameters $x_1$ and $x_2$ into $y_1$ and $y_2$, such that $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$, whereas inverse function $g^{-1}$ transforms parameters $y_1$ and $y_2$ back into $x_1$ and $x_2$, such that $x_1 = g_1^{-1}(y_1, y_2)$ and $x_2 = g_2^{-1}(y_1, y_2)$. Thus, $(x_1, x_2)$ and $(y_1, y_2)$ each refers to a single point in a Cartesian plane, where $x_1$ and $x_2$ represent the axes for the former point and $y_1$ and $y_2$ represent the axes for the latter point. Then substituting $g^{-1}(y_1, y_2)$ for $g(x_1, x_2)$ in the definition of variance for a two-dimensional discrete variable yields

$$\text{Var}(Y) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \left[ g_{\delta_1}^{-1} (y_{i,j}) - g_{\delta_2}^{-1} (y_{i,j}) - \mathbb{E}(Y) \right]^2 P_{X_i,X_j} \left[ g_{\delta_1}^{-1} (y_{i,j}) , g_{\delta_2}^{-1} (y_{i,j}) \right].$$

17. **Approximate variance.** Representing complex functions as a power series (i.e., sum of infinitely many terms) is a useful analytical strategy for integrating functions that do not have elementary antiderivatives and for approximating functions by
polynomials (Stewart, 2003). Hence, similar to its role in approximating expected value, a Taylor polynomial evaluated at $X=\mu$ can be used to estimate variance.

For a one-dimensional function $g(X)$ with a first-order Taylor polynomial $T_1$:

$$\text{Var}[g(X)] \approx \left[ g'(\mu) \right]^2 \mathbb{E}(X-\mu)^2 = \left[ g'(\mu) \right]^2 \text{Var}(X).$$

For a one-dimensional function $g(X)$ with a second-order Taylor polynomial $T_2$, where $T_2 = g(\mu) + (X-\mu)g'(\mu) + 0.5(X-\mu)^2g''(\mu)$ and $\mathbb{E}(T_2) = g(\mu) + 0.5g''(\mu)\sigma^2_X$:

$$\text{Var}[g(X)] \approx \sigma_X^2 \left[ g'(\mu) \right]^2 + \mathbb{E}(X-\mu)^3g'(\mu)g''(\mu) + \frac{1}{4} \left[ g'(\mu) \right]^2 \text{Var}[X-\mu]^2.$$ 

For a two-dimensional function $Z = g(X,Y)$ with a $T_1$ polynomial, where $T_1 = g(\mu_x,\mu_y) + (X-\mu_x)\frac{\partial}{\partial X}g(\mu_x,\mu_y) + (Y-\mu_y)\frac{\partial}{\partial Y}g(\mu_x,\mu_y)$ and $\mathbb{E}(T_1) = g(\mu_x,\mu_y)$:

$$\text{Var}[Z] \approx \text{Var}(X) \left[ \frac{\partial Z}{\partial X} \right]^2 + \text{Var}(Y) \left[ \frac{\partial Z}{\partial Y} \right]^2 + 2\text{Cov}(X,Y) \left. \frac{\partial Z}{\partial X} \frac{\partial Z}{\partial Y} \right|_{Z=\hat{g}(\mu_x,\mu_y)}.$$ 

Note, if $X$ and $Y$ are independent, then the covariance term drops out. Furthermore, for linear functions, the Taylor series reduces to Property 16. Thus, the approximate variance should only be used for nonlinear functions.

18. **Conditional variance.** Suppose $X$ and $Y$ are two jointly distributed random variables, where $X$ contains limited information about $Y$. Then the conditional variance of $Y$ given this information follows naturally from the definition of variance:

$$\text{Var}(Y \mid X) = \mathbb{E} \{ (Y - \mathbb{E}(Y \mid X))^2 \mid X \} = \mathbb{E}(Y^2 \mid X) - [\mathbb{E}(Y \mid X)]^2.$$ 

Furthermore, it can be shown that $\text{Var}(Y) = \text{Var}[\mathbb{E}(Y \mid X)] + \mathbb{E}[\text{Var}(Y \mid X)]$\(^{60}\). Hence, while the mean of the conditional expectation of $Y$ with respect to $X$ is the same as the mean of $Y$.

---

\(^{60}\) This formula is known as the law of total variance, variance decomposition formula, or conditional variance formula. It states that the variance of random variable $Y$ is equal to the sum of the variance of the conditional expectation, $\text{Var}[\mathbb{E}(Y \mid X)]$, and the average conditional variance, $\mathbb{E}[\text{Var}(Y \mid X)]$. 

110
(Property 8), knowing X decreases the error variance of Y, if the two variables are related, since $\text{Var}(Y) \geq \mathbb{E}[\text{Var}(Y|X)]$.

If the relationship between X and Y is linear then the variance of the expected value of Y given X is equal to the product of the coefficient of determination $\rho_{XY}^2$ and the variance $\sigma_Y^2$, $\text{Var}[\mathbb{E}(Y|X)] = \rho_{XY}^2 \sigma_Y^2$. The error variance, $\mathbb{E}[\text{Var}(Y|X)] = \text{Var}(Y-\hat{Y})$, is given by the variance of Y, $\sigma_Y^2$, minus the best linear predictor of Y given X, which is equal to the product of $\sigma_Y^2$ and the coefficient of alienation (i.e., one minus the coefficient of determination), $\mathbb{E}[\text{Var}(Y|X)] = \sigma_Y^2(1-\rho_{XY}^2)$. The square root of this expression is known as the standard error of the estimate.

19. Sum of random variables. One of the most important properties of variance, with regards to Summative Confidence, pertains to the sum of a set of random variables. Known as early as 1853 by Irénée-Jules Bienaymé (Heyde & Seneta, 1972), by the time Sir Ronald Fisher (1918) included it in his landmark paper it was already widely known among statisticians. For dependent variables, the variance of a sum is equal to the sum of the variances plus twice the covariance, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$, whereas for independent variables the covariance term drops out, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$. Note, the variance of a two random variables can be extended to a $n$ random variables.

20. Difference between random variables. For independent variables, the variance of a difference is identical to its counterpart from Property 19, whereas for dependent

---

$^{61}$ $\mathbb{E}[\text{Var}(Y|X)] = \mathbb{E}\{\mathbb{E}[(Y-\mathbb{E}(Y|X))^2|X]\}$. If we let $Z = Y - \mathbb{E}(Y|X)$, then, by the property of iterated expectations, $\mathbb{E}[\mathbb{E}(Z|X)] = \mathbb{E}(Z)$. Hence, $\mathbb{E}[\text{Var}(Y|X)] = \mathbb{E}\{\mathbb{E}[(Y-\mathbb{E}(Y|X))^2]|X\} = \mathbb{E}[\mathbb{E}[(Y-\hat{Y})^2]|X] = \text{Var}(Y-\hat{Y})$ since $\hat{Y}$ is the predicted value given by the linear representation of $\mathbb{E}(Y|X)$, i.e., Property 8.
variables, the covariance term is subtracted from the sum of the variances of each variable, \( \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y) \).

21. **Weighted composite.** If weights are assigned to either variable, they are treated as constants. Therefore, from Properties 13, 19, and 20, the variance of a weighted sum (or difference) is the sum of the weighted variances and (the difference of the) covariances, \( \text{Var}(w_1X \pm w_2Y) = w_1^2\text{Var}(X) + w_2^2\text{Var}(Y) \pm 2w_1w_2\text{Cov}(X,Y) \). Note, if the variables are independent, then the sum and the difference of weighted variables will be equal.

22. **Mean of a set of random variables.** Another important extension of Property 16 may be obtained, by mathematical induction and the application of Property 13, for the average of a sequence of \( k \) random variables, \( X_1, \ldots, X_k \). The variance of the mean of this distribution of sample variables is equal to the sum of the variances and covariances divided by \( k^2 \),

\[
\text{Var}(X) = \left[ \sum_{i=1}^{k} \text{Var}(X_i) + 2 \sum_{i=1}^{k} \sum_{j=1}^{k} \text{Cov}(X_i, X_j) \right] / k^2.
\]

The fact that the variance of the mean decreases as \( k \) increases is one of the most important properties in statistics (cf. Central Limit Theorem) and is a key contributor to Summative Confidence.

23. **Mean of a single random variable.** A perhaps not so obvious extension of Properties 13 and 16 underlies a very important theorem in statistics. If \( n \) independent measurements (i.e., sample size \( N=n \)) are made of the same variable using the same method of measurement then they will have the same underlying distribution. That is, the values for the variable will be independent and identically distributed (iid). Hence, the variance of the sample mean is equal to the variance of all the measurements divided by the sample size (i.e., the number of independent measures). That is, \( \text{Var}(\bar{X}) = \text{Var}(X) / n \). The square root of this
property is known as the standard error of the mean. It is used to estimate the expected amount of sampling error in the mean composite, perform hypothesis tests, and construct a CI for the sample mean. Similarly, the covariance of two sample means is equal to the average covariance of the two variables divided by their sample size, $\text{Cov}(\bar{X}, \bar{Y}) = \mathbb{E}[\text{Cov}(X, Y)]/n$.

24. **Error variance of a set of random variables.** Properties 22 and 23 can be combined to estimate the amount of variance in the mean of a distribution of sample variables. If $k$ random variables each with $n_i$ independent measurements are averaged then the error variance of the distribution of sample means is equal to the sum of the error variances and covariances divided by $k^2$. Since this property was derived by the author and does not seem to have been presented in the literature, its proof is included below.

For $k$ independent variables with sample size $n_i$:

\[
\text{Var}[\bar{X}] = \text{Var}\left[\frac{1}{n} \sum_{i=1}^{k} \bar{X}_{i}\right] = \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}[\bar{X}_i] = \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}(X_i).
\]

For $k$ dependent variables with sample size $n_i$:

\[
\text{Var}[\bar{X}] = \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}[\bar{X}_i] = \frac{1}{k^2} \sum_{i=1}^{k} \text{Cov}[\bar{X}_i, \bar{X}_j] + \frac{1}{k^2} \sum_{i<j}^{k} \text{Cov}[\bar{X}_i, \bar{X}_j]
\]

\[
= \frac{1}{k^2} \sum_{i=1}^{k} \text{Var}[\bar{X}_i] + 2 \sum_{i<j}^{k} \text{Cov}[\bar{X}_i, \bar{X}_j]
\]

\[
= \frac{1}{k^2} \sum_{i=1}^{k} \frac{\text{Var}(X_i)}{n_i} + 2 \sum_{i<j}^{k} \rho_{i,j} \sigma_{i} \sigma_{j}
\]

\[
= \frac{1}{k^2} \sum_{i=1}^{k} \frac{\text{Var}(X_i)}{n_i} + 2 \sum_{i<j}^{k} \rho_{i,j} \sqrt{\frac{\text{Var}(X_i)}{n_i} \cdot \frac{\text{Var}(X_j)}{n_j}}.
\]
\[
\frac{1}{k^2} \left[ \sum_{i=1}^{k} \frac{\text{Var}(X_i)}{n_i} + 2 \sum_{i<j}^{k} \frac{\rho_{ij} \sigma_{X_i} \sigma_{X_j}}{n_i n_j} \right]
\]

\[
\frac{1}{k^2} \left[ \sum_{i=1}^{k} \frac{\text{Var}(X_i)}{n_i} + 2 \sum_{i<j}^{k} \frac{\text{Cov}(X_i, X_j)}{n_i n_j} \right]
\]

Note, \( \text{Cov}(\bar{X}_e, \bar{X}_e) = \text{Cov}(X_e, X_e) \) because the sum of a mean value is equal to the product of the mean and the number of times it was summed.

\[
\text{Cov}(\bar{X}_e, \bar{X}_e) = \mathbb{E}[^{\bar{X}_e} \bar{X}_e] - \mathbb{E}[\bar{X}_e] \mathbb{E}[\bar{X}_e]
\]

\[
= \mathbb{E} \left[ \sum_{i=1}^{n} \frac{X_{e_i}}{n} \sum_{j=1}^{m} \frac{X_{e_j}}{m} \right] - \mathbb{E} \left[ \sum_{i=1}^{n} \frac{X_{e_i}}{n} \right] \mathbb{E} \left[ \sum_{j=1}^{m} \frac{X_{e_j}}{m} \right]
\]

\[
= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{E}[X_e X_e] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_e] \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}[X_e]
\]

\[
= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{E}[X_e X_e] - \frac{1}{n} \mathbb{E}[X_e] \frac{1}{m} \mathbb{E}[X_e] = \text{Cov}(X_e, X_e).
\]

This property may be further expanded by combining it with Property 21.

\[
\text{Var}(\mu_{\bar{X}_e}) = \frac{1}{k^2} \left[ \sum_{i=1}^{k} \frac{\mu_i^2 \text{Var}(X_i)}{n_i} + 2 \sum_{i<j}^{k} \frac{\mu_i \mu_j \text{Cov}(X_i, X_j)}{n_i n_j} \right]
\]

This property forms the basic framework of Summative Confidence for it reveals how the error variances from different variables may be combined to generate the overall error variance of a summative conclusion. Furthermore, it shows how the relative importance of variables may be taken into account.

**Limitation of Employing SAS (or Any Other Software) to Compute Property 24**

All the properties of expected value and variance were investigated via Monte Carlo simulations (available from the author). SAS PROC UNIVARIATE was used to calculate descriptive statistics, which were then compared against hand-calculated figures predicted by each for the aforementioned properties. These comparisons readily demonstrated that all of
the properties held true, including Property 24. Given that SAS was able to perform all the analyses, some readers may rightfully wonder whether existing software can be used to output the figures needed to conduct a Summative Confidence analysis and if so, whether Summative Confidence is not a fancy term for an existing analysis. Therefore, a Monte Carlo simulation was designed to illustrate the limitation of existing software.

Two datasets were simulated representing variables where the unit of analysis is the same for the variables. In the first dataset, three independent variables (|r| <0.02) with 4,500 values were simulated from a normal distribution with means (variances) approximately equal to 59.981 (4.177), 10.075 (25.362), and 19.873 (100.86). A fourth variable was then set equal to the average of these three variables. Therefore, in accordance with Properties 10, 22, and 24, the expected value, variance, and standard error of the composite mean were nearly equal to 29.976 [= (59.981+10.075+19.873)/3], 14.489 [=(4.177 +25.362+100.86)/3²], and 0.057 {=√[(4.177+25.362+100.86)/(3²*4500)]}, respectively. As expected, these estimates were similar to those produced by SAS. Moreover, examination of the histogram and goodness of fit indices for the composite variable revealed that it was normally distributed.

The second dataset attempted to simulate a summative conclusion—i.e., a complex variable computed from multiple data streams (i.e., unit of analysis are different for each variable). Three independent variables (|r| <0.04) were simulated from a normal distribution with means (variances; sample sizes) approximately equal to 60.025 (σ²=4.062; N=2,000), 10.064 (σ²=26.601; N=2,500), and 19.834 (σ²=100.43; N=4,500). Notice then, the only substantive difference between the first and second datasets was the sample size of the first two variables. In accordance with the previous analyses, the expected value, variance, and standard error of the composite mean were approximately equal to 29.974 [= (60.025+10.064+19.834)/3], 14.566 [= (4.062+26.601+100.43)/3²], and 0.062 {=√[(4.062/2000)+...]}.
26.601/(2500+100.43/4500)/3^2}, respectively. Naturally, the expected values were nearly identical to the previous estimate whereas the variance and standard error were slightly larger in the second dataset due to the smaller sample size for two of the variables. However, these values differ drastically from those reported by SAS (i.e., $23.873$ mean, $87.53$ variance, and $0.139$ standard error). Examination of the histogram and goodness of fit indices for the composite variable revealed that it was not normally distributed. So what went wrong?

The problem lies in the fact the sample sizes of the constituent variables were not equal. The composite variable was generated by averaging variables with missing values, which resulted in biased estimates. Hence, some of the values of the mean were more precise because they were produced by averaging three variables whereas other values only reflected one variable. Unfortunately, the issue cannot be resolved by averaging only the cases where all three variables have a non-missing response. Although averaging the 2,000 cases with non-missing responses produced comparable estimates (i.e., $\mu=29.934$, $\sigma^2=14.628$, and $\sigma_{\bar{X}}=0.086$), this method always produces inflated error variance estimates due to sample size differences. Worse still, outside of a simulation, one can never be sure of the magnitude of sampling bias that may arise between differences in the two samples (complete responses versus incomplete responses). Moreover, evaluation is a field in which evaluators must synthesize numerous variables from multiple streams into a summative conclusion. Thus, in this author’s experience, rarely are the sample sizes of these variables equal, so estimating the statistical parameters needed to construct a CI for a summative conclusion requires manual computation of these parameters or the aid of a yet-to-be developed computer program.

**Distribution Theory**

To this point, little attention has been accorded to the distributional nature of random variables. Although the previous section focused exclusively on the normal
distribution, dozens of distributions exist with the list growing longer over time. A probability distribution of a random variable produced by an experiment is a mathematical relation by which probabilities are assigned to all the events that constitute the sample space of a random variable. An experiment $\mathcal{E}$ is typically regarded as a controlled trial wherein the magnitude of a set of dependent variables is manipulated by the experimenter to ascertain the relationship between these variables and a set of independent variables. In statistical contexts, however, an experiment refers to “any process, possibly under partial control that we may observe and whose behavior in the future is not totally determined because it is influenced, at least in part, by chance” (Bartoszynski & Niewiadomska-Bugaj, 2008, p. 2).  

The sample space $\Omega$ refers to the set of all possible outcomes that occur as a result of an experiment, whereas an event $\omega$ is a subset of the sample space. A random variable is a real-valued function of the events of a sample space. Hence, a statistical distribution represents the mathematical relation believed, either because of a definitional premise, observational or inferential analysis, or historical assumptions, to accurately describe the assignment of probabilities to all the events that make up the sample space of a random variable. Moreover, the probability function of a random variable can be used to make probability statements about the values the variable may attain when the experiment is performed.

This section will define several well-known statistical distributions evaluators are likely to encounter and define their expected value and variance. These distributions, which can be found in virtually all probability and mathematical-statistics textbooks, are organized into discrete and continuous distributions depending upon whether the number of possible values (events) within the sample space is finite or countably infinite (discrete), or not

---

62 Unless otherwise stated, this section will adopt the statistical rather than the methodological definition.

63 The distributions presented herein have been known for hundreds of years and thus, are treated as common knowledge (i.e., presented without citation). Interested readers are directed to Meyer (1970), Larson (1973, 1974), Grinstead and Snell (1997), Ross (2002), and Hogg, McKean, and Craig (2005) for further reading.
Therefore, a random variable is called discrete if its sample space is a set of real numbers and continuous if its sample space is an interval on the real number line, where the probability of observing any single value within the space is zero.

**Discrete Random Variables**

Some discrete random variables occur with such frequency in practice, they have been given special names. This section will define several of the most important discrete distributions that evaluators are likely to encounter in their practice. However, the list herein is not exhaustive. In fact, statistical distributions such as the geometric, hypergeometric, and negative binomial (Pascal) have been purposely omitted since, in the experience of this author, they rarely occur in evaluation settings.

*Uniform distribution.* The most basic of all probability distributions occurs when all the events in the sample space are equally probable. In other words, if an experiment is conducted (e.g., rolling a fair die, selecting a card from a well-shuffled deck of 52 playing cards, selecting an individual at random from a population), the probability of all of the events (e.g., value of the top face of the die or playing card) is equal to the inverse of the number of events in the sample space. Therefore, if an experiment $\mathcal{E}$ is conducted and the outcomes of random variable $X$ generate a sample space $\Omega$ that contains $n$ events ($\omega$) in interval $[a, b]$, such that $\Omega=\{\omega: a \leq \omega \leq b\}$, and all the events are equally probable for every $\omega \in \Omega$ then the uniform pmf is defined by $P(X=\omega)=P_X(\omega)=1/n$ and zero otherwise, and its cdf is defined by $P(X \leq \omega)=F_X(\omega)=(\omega-a+1)/n$. A random variable $X$ that conforms to a discrete uniform distribution is typically denoted as $X \sim U(n)$. The mean and variance of a discrete uniform random variable with parameter $n$ is given by $E(X)=(n+1)/2$ and $\text{Var}(X)=(n^2-1)/12$, respectively. Suppose then a department interested in hiring a new
faculty member invites five candidates for an interview following which all the candidates are ranked (no ties) by the hiring committee. What are the expected value and variance of this ranking practice? Since \( n = 5 \), \( \mathbb{E}(X) = (5+1)/2 = 3 \) and \( \text{Var}(X) = (5^2 - 1)/12 = 2 \).

**Bernoulli distribution.** Named after a Swiss mathematician, another simple probability distribution is frequently encountered whenever an experiment, called a Bernoulli process, consists of repeated, independent and identical Bernoulli trials each of which produce one of two possible outcomes (i.e., a dichotomy). Typically, one of the outcomes is arbitrarily labeled as “success” while the other is labeled as “failure.” Furthermore, across repeated trials, the event defined as success has a constant probability \( p \) and the event defined as failure has a probability \( q = 1 - p \), which connotes the trials must be independent or else probabilities \( p \) and \( q \) would change from trial to trial. Therefore, if the sample space of \( \mathcal{E} \) consists of two events, such that \( \Omega = \{ \omega: 0,1 \} \) where 0 denotes failure and 1 denotes success, then a random variable \( X \) produced by \( \mathcal{E} \) will have a Bernoulli pmf defined by \( P(X=1) = P_X(1) = p \), \( P(X=0) = P_X(0) = 1 - p = q \), and zero otherwise, and a cdf defined by \( P(X<0) = F_X(0) = 0 \), \( P(0 \leq X < 1) = F_X(0) = q \), and \( P(X \leq 1) = F_X(1) = 1 \). A random variable \( X \) that conforms to a Bernoulli distribution is typically denoted as \( X \sim \text{Bin}(1,p) \). The expected value and variance of a Bernoulli random variable \( X \) is \( \mathbb{E}(X) = p \) and \( \text{Var}(X) = p(1-p) \), respectively.

As an example, let us return to the tenure example discussed in Chapter 1. The decision of whether to recommend a faculty for tenure or not is, in fact, a Bernoulli process for it results in a binary decision (recommend, do not recommend). Suppose that historically the panel of faculty members evaluating a candidate’s application for tenure results in a favorable recommendation 60% of the time. What are the expected value and variance of such a Bernoulli process? Since \( p = 0.6 \) then \( \mathbb{E}(X) = 0.6 \) and \( \text{Var}(X) = 0.6 \times 0.4 = 0.24 \).
Binomial distribution. Another popular probability distribution occurs when one sums a set of random variables produced by a Bernoulli process. That is, a binomial random variable X is produced when $E$ is conducted in which X represents the total number of successes that occurred across $n$ repeated, independent Bernoulli trials with probability $p$ of success. Moreover, the sample space of $E$ consists of all nonnegative integers less than or equal to $n$, $\Omega=\{\omega: 0,1,\ldots,n\}$, where X has a binomial pmf defined by $P(X=\omega)=P_X(\omega)=\binom{n}{\omega}p^\omega(1-p)^{n-\omega}$, and zero otherwise, since there are $\binom{n}{\omega}$ possible combinations of $\omega$ successful outcomes out of $n$ Bernoulli trials, and a cdf defined by $P(X\leq\omega)=F_X(\omega)=\sum_{k=0}^{\omega}\binom{n}{k}p^k(1-p)^{n-k}$. A random variable X that conforms to a binomial distribution is typically denoted as $X \sim Bin(n,p)$. The expected value of $X \sim Bin(n,p)$ is $E(X)=np$ and the variance is $Var(X)=np(1-p)=npq$.

Once again, let us return to the tenure example discussed in the first chapter. Suppose that each tenure review committee member must independently determine whether recent changes occurred in the number, quality, and impact of a candidate’s publications or presentations. Moreover, assume this determination entails a Bernoulli process (positive change, no change or negative change), which entails six independent determinations. Furthermore, suppose review panels historically reach a conclusion of favorable change 75% of the time. If a composite variable is created by summing the number of instances of favorable change across the six criteria, what are the expected value and variance of such a binomial process? Since $n=6$ and $p=0.75$, $E(X)=6*0.75=4.5$ and $Var(X)=6*0.75*0.25=1.125$.

Poisson distribution. Named after the French mathematician Siméon D. Poisson, one of the most frequently encountered probability distributions transpires when an experiment, called a Poisson process, consists of counting the number of Bernoulli trials that occur

---

$^{64}$ Note, $\binom{n}{\omega} = \frac{n!}{(n-\omega)!\omega!}$, where $n$ factorial, $n!$, denotes $n*(n-1)*\ldots*2*1$, $\omega$! denotes $\omega*(\omega-1)*\ldots*2*1$, and so on.
within a continuous interval of measurement, such as time, length, area, or volume. A Poisson process is defined by (1) the presence of independent and identical Bernoulli trials with an average success rate $\lambda$ per unit interval; (2) the dependence of the distribution of the number of events that occur during an interval (composed of one or more subintervals) on only the length of the interval $\ell$ and the probability $p$ of a single event occurring in a given subinterval, but not on the nature of the event being counted; (3) the reliance of the probability of a single event on the length of the subinterval; and (4) the ability to select a sufficiently small enough subinterval to ensure the probability is negligible that two or more events occur in any subinterval. Therefore, the success rate per unit interval is equal to the number ($n$) of nonoverlapping subintervals that comprise the interval times the probability the event occurs in a given subinterval divided by the number of subintervals that comprise the unit interval, $\lambda = np/\ell$. Given the reliance of the Poisson process on a Bernoulli process, it should not come as a surprise to most readers that the Poisson distribution can be used to approximate the binomial distribution. In fact, it can be proven than as $n \to \infty$ and $p \to 0$, the limiting distribution of the binomial distribution is the Poisson distribution.

A Poisson random variable $X$ is produced when an $\mathcal{E}$ is conducted in which $X$ represents the number of successes that occurred across $n$ repeated, independent Bernoulli trials observed within a continuous interval of measurement of length $\ell$ with a success rate of $\lambda = np/\ell$. The sample space of $\mathcal{E}$ consists of all nonnegative integers, $\Omega = \{\omega: 0,1,2,3,\ldots\}$, where $X$ has a Poisson pmf defined by $P(X=\omega) = P_X(\omega) = e^{-\lambda} \lambda^\omega / \omega!$, and zero otherwise, and a

---

65 For example, if on average there are 14 events per week, how many events should one expect in a 6 hour period? Since the time interval 'week' is greater than the interval 'hours,' it stands to reason that the subintervals should be measured in hours (although smaller subintervals can be defined). Therefore, the average success rate $\lambda$ per hour is equal to 14 events per week divided 168 hours per week or one-twelfth of an event per hour. Moreover, in a 6 hour period, one would expect to observe half an event, $\lambda \ell = 6/12 = 0.5$, on average.
66 According to Larson (1974), the Poisson approximation is quite accurate when $n \geq 20$ and $p \leq 0.05$ and very good when $n \geq 100$ and $np \leq 10$. 
cdf defined by \( P(X \leq \omega) = F_X(\omega) = \sum_{k=0}^{\omega} \frac{e^{-\lambda \ell} (\lambda \ell)^k}{k!} \). Furthermore, in cases in which the Poisson distribution is used to approximate the binomial distribution, the pmf and cdf are defined by \( P(X=\omega) = p_X(\omega) = \frac{e^{-\lambda \ell} (\lambda \ell)^\omega}{\omega!} \) and \( P(X \leq \omega) = F_X(\omega) = \sum_{k=0}^{\omega} \frac{e^{-\lambda \ell} (\lambda \ell)^k}{k!} \), respectively. A random variable \( X \) that conforms to a Poisson distribution is typically denoted as \( X \sim P \text{oi}(\lambda) \), where the expected value and variance of \( X \) are \( \mathbb{E}(X) = \lambda \) and \( \text{Var}(X) = \lambda \).

Returning to the tenure example, one finds several variables that are likely to have a Poisson distribution. For example, the number of publications and presentations attributed to a candidate are likely the product of a Poisson process since (a) each submission of a manuscript or presentation is associated with a Bernoulli process, (b) the number of publications and presentations are determined by the length of the time interval examined, and (c) a small enough time interval can be defined to ensure the probability of two or more publications or presentations in the same period is negligible. Therefore, if we assume that, on average, candidates publish two articles a year, what is the probability that a candidate would publish 10 or more articles in three year span? Since \( P(X \leq \omega) = \sum_{k=0}^{\omega} \frac{e^{-\lambda \ell} (\lambda \ell)^k}{k!} \) then \( P(X \geq \omega) = 1 - F_X(\omega - 1) = 1 - \sum_{k=0}^{\omega-1} \frac{e^{-\lambda \ell} (\lambda \ell)^k}{k!} \approx 1 - 0.916 = 0.084 \). In other words, there is reason to believe that a candidate who publishes 10 or more articles in a three year period is an exceptional candidate.

**Multinomial distribution.** While the Bernoulli and binomial distributions focus on trials that can only take on one of two discrete values, evaluators regularly employ measurement scales that permit random variables to take on one of several discrete values. Suppose an experiment produces \( N \) independent and identical trials in which the random variable produced by each trial conforms to a categorical distribution\(^{67}\) with \( k \) classes (response

\(^{67}\)The categorical distribution is the generalization of the Bernoulli distribution to more than two classes where the sample space (e.g., \( \Omega = \{\omega: 1,2,\ldots,n\} \), \( \Omega = \{\omega: 'none of the time', 'some of the time', 'all of the time'\}), \( \Omega = \{\omega: \)
categories) and probabilities \( p_1, p_2, \ldots, p_k \). A multivariate distribution can then be used to calculate the probability of observing a set of outcomes or response classes (i.e., the probability that class 1 will equal \( n_1 \) trials, class 2 will equal \( n_2 \) trials, and so on to class \( k \) will equal \( n_k \) trials) across all the multinomial trials. That is, suppose an \( \mathcal{E} \) produces \( N \) trials each of which can result in \( k \geq 2 \) possible outcomes \( \Theta = \{ \theta : 1, 2, \ldots, k \} \) with probability \( p_\theta \) such that \( \sum_{\theta=1}^{k} p_\theta = 1 \), then a random variable \( X \) representing the number of trials resulting in outcome \( \theta \), where \( \sum_{\theta=1}^{k} n_\theta = N \) and \( \Omega = \{ \omega : n_1, n_2, \ldots, n_k \} \), will have a multinomial pmf defined by

\[
P_X(\Omega) = P(X_1 = n_1, \ldots, X_k = n_k) = \frac{N!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},
\]

and zero otherwise, since there are \( \binom{N}{n_1, \ldots, n_k} \) possible combinations of \( \omega \) successful outcomes out of \( N \) multinomial trials, and a cdf defined by

\[
F_X(\Omega) = P(X_1 \leq n_1, \ldots, X_k \leq n_k) = \sum_{n_1, \ldots, n_k} \frac{N!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k},
\]

where the sum is taken over all \( k \)-tuples for which \( \sum_{\theta=1}^{k} n_\theta = N \) and

\[
\binom{N}{n_1, \ldots, n_k} = \frac{N!}{n_1! \cdots n_k!}.
\]

A random variable \( X \) that conforms to a multinomial distribution is typically denoted as \( X \sim \text{Mult}(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k) \) or as \( X \sim \text{Mult}(N; p_1, p_2, \ldots, p_k) \). The expected value of \( X \sim \text{Mult}(n; p_1, \ldots, p_k) \) is \( \mathbb{E}(X) = Np_\theta \) and its corresponding variance is \( \text{Var}(X) = Np_\theta(1-p_\theta) \).

Figure 12 illustrates the pmf of a multinomial distribution and its associated cdf, where \( N=10 \), \( p_1=0.35 \), \( p_2=0.40 \), and \( p_3=0.25 \). Note, four-dimensions were needed to display the trinomial distribution \( (k=3) \). In fact, \( (k+1) \)-dimensions are needed to display a multinomial distribution with \( k \) classes.

'1: strongly disagree', '2: disagree', '3: agree', '4: strongly agree') is composed of a finite number of mutually exclusive events, typically referred to as classes, that may be assigned a numerical label.
Since the multinomial distribution will be a critical component employed to derive a CI for discrete and ordinal variables, greater focus will be paid to explicating the intricacies of this distribution, as compared to the previous distributions. Specifically, examining a worked-out example may shed further light on how to compute its cdf. However, before endeavoring to calculate a multinomial cdf by hand, it is important to know the number of \( k \)-tuples that exist since it can grow at an incredible rate. A review of several textbooks on probability theory (Ross, 2002; Bartoszynski & Niewiadomska-Bugaj, 2008; Larson, 1974; Grinstead & Snell, 1997) and scholarly databases revealed that a solution to this question does not appear in the literature. However, programmatically, one can determine the number of \( k \)-tuples given \( N \) trials and \( k \) classes. Table 8 provides the number of \( k \)-tuples for values of \( N \) (number of trials) and \( k \) less than or equal to 10.
Table 8

Number of \( k \)-Tuples for \( N, k \leq 10 \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>36</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>10</td>
<td>20</td>
<td>35</td>
<td>56</td>
<td>84</td>
<td>120</td>
<td>165</td>
<td>220</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>15</td>
<td>35</td>
<td>70</td>
<td>126</td>
<td>210</td>
<td>330</td>
<td>495</td>
<td>715</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>21</td>
<td>56</td>
<td>126</td>
<td>252</td>
<td>462</td>
<td>792</td>
<td>1,287</td>
<td>2,002</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>28</td>
<td>84</td>
<td>210</td>
<td>462</td>
<td>924</td>
<td>1,716</td>
<td>3,003</td>
<td>5,005</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>36</td>
<td>120</td>
<td>330</td>
<td>792</td>
<td>1,716</td>
<td>3,432</td>
<td>6,435</td>
<td>11,440</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>45</td>
<td>165</td>
<td>495</td>
<td>1,287</td>
<td>3,003</td>
<td>6,435</td>
<td>12,870</td>
<td>24,310</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>55</td>
<td>220</td>
<td>715</td>
<td>2,002</td>
<td>5,005</td>
<td>11,440</td>
<td>24,310</td>
<td>48,620</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>66</td>
<td>286</td>
<td>1,001</td>
<td>3,003</td>
<td>8,008</td>
<td>19,448</td>
<td>43,758</td>
<td>92,378</td>
</tr>
</tbody>
</table>

Note, if the table is rotated 45° clockwise so the upper half (highlighted in red) begins to resemble a triangle and then ones are added along the sides, a familiar pattern begins to emerge. Figure 13 illustrates the first 11 rows of Pascal’s triangle.

![Pascal's Triangle](image)

Figure 13. Pascal’s Triangle for Computing the Number of \( k \)-Tuples

Since the entries in Pascal’s triangle represent the coefficients of the binomial expansion, then the number of \( k \)-tuples is a function of the binomial theorem. More directly, the number of \( k \)-tuples is equal to \((N+k-1)!/N!(k-1)!\). Therefore, when \( N=9 \) and \( k=7 \), the number of \( k \)-tuples will equal \((9+7-1)!/9!(7-1)!)=5,005\) as shown in the previous table. Given that classes will generally be used to represent discrete categories (e.g., Likert scale points) whereas the number of trials will be used to represent sample size, one can see an immediate problem. Namely, as the sample size becomes large (even modestly so), the
number of \(k\)-tuples dramatically rises. Thus, calculating the multinomial cdf, is cumbersome and typically requires a computer. Although SAS currently does not have a multinomial function, one was programmed by the author (available upon request) to generate the multinomial pdf and cdf for all \(k\)-tuples. However, since these functions require the computation of factorials, after a certain point \((N>150)\), even computers may not be able to compute the factorial.

Returning to the tenure example in Chapter 1, one finds several variables that may be measured in such a way as to necessitate the use of the multinomial distribution. For example, the impact of each publication on the research community could be measured using an ordinal scale wherein ‘A’ denotes that the publication had a good deal of impact (>10 citations), ‘B’ denotes a fair amount of impact (6-10 citations, inclusive), and ‘C’ denotes little or no impact (\(\leq 5\) citations). Furthermore, suppose historically 35% of a candidate’s publications garner more than 10 citations, 40% garner more than 5 citations, and the remaining 25% garnering less than or equal to 5 citations. What are the expected value and variance a typical candidate with 10 publications will have for each grading category? What is the likelihood a candidate will have more than five articles with 10 citations or more? First, since \(E(X_\theta) = Np_\theta\) then \(E(X_1) = 10 \times 0.35 = 3.5\), \(E(X_2) = 10 \times 0.4 = 4\), and \(E(X_3) = 10 \times 0.25 = 2.5\). Moreover, since \(\text{Var}(X_\theta) = Np_\theta(1-p_\theta)\) then \(\text{Var}(X_1) = 10 \times 0.35 \times 0.65 = 2.275\), \(\text{Var}(X_2) = 10 \times 0.4 \times 0.6 = 2.4\), and \(\text{Var}(X_3) = 10 \times 0.25 \times 0.75 = 1.875\). For the second question, one needs to sum the probabilities for all 3-tuples where \(X_1 > 5\). Therefore,

\[
P(X_1 > 5) = \sum_{n_1, n_2, n_3} \frac{N!}{n_1! n_2! n_3!} \prod_{\theta=1}^{k} p_{\theta}^{n_\theta} = \sum_{n_1, n_2, n_3} \frac{10!}{n_1! n_2! n_3!} \prod_{\theta=1}^{3} p_{\theta}^{n_\theta} = \sum_{n_1, n_2, n_3} \frac{10!}{n_1! n_2! n_3!} \left[ (10 - n_1 - n_2) \cdot p_1^{n_1} p_2^{n_2} p_3^{10-n_1-n_2} \right] 
\]

\[
= \frac{10!}{6!3!4!} (0.35)^5 (0.25)^5 + \frac{10!}{6!3!2!2!} (0.35)^6 (0.4)^1 (0.25)^3 + \frac{10!}{6!3!2!2!} (0.35)^6 (0.4)^1 (0.25)^3 + \frac{10!}{6!3!2!2!} (0.35)^6 (0.4)^1 (0.25)^3 
\]

\[
+ \frac{10!}{6!4!4!} (0.35)^6 (0.4)^1 (0.25)^3 + \frac{10!}{7!0!3!} (0.35)^7 (0.25)^3 + \frac{10!}{7!1!2!} (0.35)^7 (0.25)^3 + \frac{10!}{7!1!2!} (0.35)^7 (0.25)^3 
\]

126
In other words, there is reason to believe a candidate who has 6 or more articles published that have received more than 10 citations each is an exceptional candidate.

Continuous Random Variables

As is the case with discrete random variables, some continuous random variables occur with such frequency in practice, they have been given special names. This section will define the two most important continuous distributions evaluators are likely to encounter in their practice. Continuous distributions such as the exponential, gamma, and beta have been omitted since their contribution to evaluation practice is insignificant in this author’s experience. Interested readers may consult introductory probability and mathematical-statistics textbooks for further details on these distributions or those defined herein.

Uniform distribution. The analog to the discrete uniform distribution is the continuous uniform distribution, also known as the rectangular distribution. This distribution occurs whenever all the points in the sample space interval \([a,b]\) are equally probable. For example, the amount of time or distance between two points, geometric probabilities, and \(p\)-values, represent random variables that conform to a uniform distribution. A random variable satisfies the equiprobable property when the probability over its sample space is equal to the reciprocal of the length of the interval and zero otherwise. That is, if an \(\mathcal{E}\) is conducted in which \(\Omega = \{\omega: -\infty < a < b < \infty\}\) and every \(\omega \in \Omega\) is equiprobable, then the \(\mathcal{E}\) will produce a continuous uniform random variable \(X\) with a probability density function, \(f_X(\omega)\), defined by \(f_X(\omega) = (b-a)^{-1}\) and a cdf, \(F_X(\omega)\), defined by \(F_X(a \leq \omega \leq b) = (\omega-a)/(b-a)\). A random variable \(X\) that conforms to a continuous uniform distribution is typically denoted as \(X \sim U(a,b)\). The
expected value and variance of a discrete uniform random variable with parameters \((a, b)\) is 
\[ E(X) = \frac{(a+b)}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}, \]
respectively.

Although examples of the continuous uniform distribution are rarely encountered in social science practice, they perform an important function in certain analyses. Suppose the university provost in the tenure example wanted to compare candidates across departments, so she required review committees to provide her with a composite score for each candidate based on the variables reported in Figure 1. This information could be used to generate a percentile score for each candidate. Assuming there were no ties, the new variable \(X\) would have a continuous uniform distribution with range \((0, 100)\). What are the expected value and variance of \(X\)? Since \(a = 0\) and \(b = 100\) then \(E(X) = 100/2 = 50\) and \(Var(X) = 100^2/12 = 833.33\).

Normal distribution. Originally discovered by the French mathematician Abraham DeMoivre, the normal distribution, as it has come to be known, was later popularized by the famous German mathematician Karl Friedrich Gauss. Today, the normal distribution, also known as the “bell curve” or the Gaussian distribution, is without doubt the most widely used probability distribution. Its popularity among scientists is sustained by its fundamental contribution to the Central Limit Theorem, the frequency with which it occurs in practice, and its ability to estimate the Binomial distribution as \(n \to \infty\). In fact, the normal distribution is often assumed to underlie any process that produces a random variable in which the observations of the variable are symmetric and cluster around the expected value. For example, weight, height, intelligence, and measurement error of physical objects represent just a few of the many random variables that are normally distributed. A random variable \(X\) is said to conform to a normal distribution, with parameters \(\mu\) and \(\sigma^2\), if its pdf is given by

\[
P(X = x) = f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},
\]
for all real numbers, and its cdf evaluated at $w$ is given by

$$P(X \leq w) = \int_{-\infty}^{w} rac{1}{\sqrt{2\pi \sigma}} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} dx,$$

where $\mu$ and $\sigma^2$ represent the expected value and variance of $X$, respectively. Often, this distribution is represented in standardized form, i.e., $\mu=0$ and $\sigma^2=1$, by subtracting the expected value from each observation and dividing by the standard deviation $\sigma$. The pdf and cdf of the standard normal distribution, denoted by $\varphi$ and $\Phi$ respectively, are defined by

$$\varphi_X(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} \text{ and } \Phi_X(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{z^2}{2} \right\} dz,$$

where $z=(x-\mu)/\sigma$ represents a linear transformation of $X$, known as the $z$-score of $X$. A random variable $X$ that conforms to a normal distribution is typically denoted as $X \sim N(\mu, \sigma^2)$. The expected value and variance of a normal random variable with parameters $(\mu, \sigma^2)$ is $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$, respectively.

Several variables in the tenure example may have a normal distribution. For example, the quality of publications or student evaluations may be normally distributed if a continuous scale is used to measure quality and the values observed are symmetric and cluster around the expected value. Suppose publication quality is measured by the journal’s impact score as reported in the *Journal of Citation Reports* and that these scores are normally distributed. What can one say about the candidate if the average impact score (variance) of the journals where they published is 2.38 (0.5) while the expected value and variance of relative journals in the candidate’s field of research are 1.05 and 0.5? Converting the average impact score into a $z$-score, $z=(2.38-1.05)/\sqrt{0.5}=1.88$ allows one to obtain the standard normal probability for $P_X(x \leq z) = \Phi_X(z)$ from a statistical reference table or software. This analysis yielded a probability of 0.97. Hence, a candidate with such high caliber publications is rare indeed.
Transformation Theory

Despite the exposition to expectation and distribution theory, we are still not able to construct even the most basic Summative Confidence CI as the following example will illustrate. Returning to the tenure example from the first chapter, suppose the tenure review panel is asked by the provost to calculate and interpret the expected value and variance for the Research macrovalue (see Figure 1), which is composed of seven criteria (microvalues): the number, quality, and impact of the candidate’s publications and presentations, and recent changes to the number or quality of publications and presentations. How can the panel go about this task?

One way of accomplishing the task, albeit not the best way, is to follow standard analytical practice of being mindful of the underlying distribution of each variable before applying the properties of expected value and variance necessary for combining the variables into a composite variable. As discussed in the previous section, several assumptions can be made concerning the seven microvalues. First, the number of publications and presentations occurring within a fixed period, in all likelihood, is the product of two independent Poisson processes. Second, if the impact of a journal is used to gage the quality of each publication, then the random variable associated with this process is likely to be normally distributed. Third, since no agreed upon method exists for measuring the quality of a presentation, a multitude of methods can be devised. Therefore, suppose each conference organization at which the candidate presented collected information on the quality of the presentation using a survey whose scores were converted into percentiles. In other words, the random variable conforms to a continuous uniform distribution. Fourth, suppose the impact of the number of publications was assessed using the ordinal scale described in the previous section. Moreover, assume an ordinal scale (i.e., categorical distribution) was used to measure the
impact of the presentations based upon a self-reported measure of the likelihood attendees will use the information presented in their own evaluation projects. Lastly, the random variable representing the number of recent changes in the number, quality, and impact of the candidate’s publications and presentations is likely to conform to a binomial distribution if each of the six criteria represents an independent Bernoulli process. In summary, the composite distribution will be composed by either summing or averaging seven criteria consisting of two Poisson distributions, one normal distribution, one continuous uniform distribution, two multinomial distributions, and one binomial distribution.

From a mathematical perspective, constructing a composite variable is simple. One only needs to sum or average the constituent variables (i.e., the seven microvalues). The expected value and variance then can be computed using Properties 10 and 22, respectively. Research and evaluation savvy readers undoubtedly see the flaw in this approach. Namely, since the composite variable is derived from different distributions, which, in turn, are measured on different scales and with different sample sizes, the composite will be uninterpretable. At best, one could interpret the direction of composite values, providing the directionality of all the constituent variables are in agreement with each other (i.e., high values on each variable signify a high attribute on the composite variable, and vice-versa), but the magnitude of a value or comparison between values would be virtually impossible to interpret since the scale and any associated anchors will have been lost. In other words, summing or averaging “apples” and “oranges” does not yield a meaningful composite.

By its very nature, however, summative evaluation requires the synthesis of multiple variables, which are typically composed of different distributions, into a coherent evaluative conclusion. This seemingly impossible dilemma can be resolved by transforming each constituent variable into a common distribution. Although not widely acknowledged, this is
the very heart of evaluation. That is, evaluation is a process by which values with dissimilar underlying distributions are transformed into a common distribution for the purpose of constructing a complex variable that enables an evaluator to describe the evaluand in terms of its merit, worth, or meaningful significance. This definition of evaluation expands upon the one commonly accepted by the majority of evaluators, proposed by Scriven (1991), by clarifying that the “process” of determining the merit, worth, and significance of an evaluand is contingent upon the transformation of dissimilar values into a common metric.

Researchers often point out that one cannot compare apples and oranges. Evaluators, however, retort by noting that consumers are able to compare apples and oranges every time they make a purchasing decision between the two fruit. How are consumers able to make this evaluative decision? Simply put, they intuitively “measure” or rate key properties (values) (e.g., cost, sweetness, ripeness) of each object and then compare these measures. If an object rates higher on all the values, as compared to the alternative decision, then the decision is straightforward. More complicated analyses, however, arise when an object rates highly on some values but not on others. Invariably, such decisions entail the weighting of one value versus another. For example, suppose a consumer prefers the sweetness of apples to that of oranges but an apple costs a dollar while an orange costs half a dollar. Using qualitative reasoning, the consumer may ask themselves the question of whether their taste preference is worth the extra cost. If it is and providing no other values enter into the decision then they will purchase the apple.

Alternatively, the consumer’s question may be addressed quantitatively by transforming dissimilar ratings into a common metric. Suppose cost is measured using a continuous scale that conforms to a normal distribution while tastiness is measured using a five-point Likert scale, where one represents ‘not very sweet’ and five represents ‘very sweet,’
that conforms to an ordered categorical distribution (i.e., an ordinal measure). Clearly, two transformations are needed since cost and sweetness are unrelated dimensions. Numerous transformations may be executed as this section will illustrate. For example, grades measuring merit and worth could be assigned to each variable. Cost can be transformed into grades of worth using the common method of grading on a normal curve (Hopkins, 1998), where an A is assigned to the fruit if its cost is more than $1.5\sigma$ below the average cost of the fruit, a B if its cost is between $-1.5\sigma$ and $0.5\sigma$, a C if its cost is between $-0.5\sigma$ and $0.5\sigma$, a D if cost is between $0.5\sigma$ and $1.5\sigma$, and an F if its cost is more than $1.5\sigma$ above the average cost of the fruit. Conversely, sweetness can be transformed into grades of merit by assigning the grade A to the Likert value of five, a B to the Likert value of four, and so on until F is assigned to the Likert value of one. At this point, cost and sweetness have been transformed from their original distributions to a common distribution that can be interpreted.

This section will examine several methods evaluators can use to transform one distribution into another depending upon the metric desired to represent the composite variable upon which the summative conclusion will be based. A variety of polychotomous and quantile transformations will be discussed. Since the requirements of each evaluation are unique, the choice of which transformation is most appropriate, assuming any are needed, is left up to evaluators. This section will illustrate the choices available and their implications upon expected value and variance. SAS code is provided in Appendix F.

**Polychotomous**

A polychotomous transformation entails the division of an initial probability distribution into two or more mutually exclusive parts or classes. In other words, it transforms the input or raw variable, into an ordered categorical distribution. As the example above illustrates, such a transformation may be used for both continuous and discrete
random variables. For continuous random variables, polychotomization requires the designation of \( k-1 \) cutoff points or thresholds necessary for transforming the continuous distribution into a discrete categorical distribution with \( k \) classes. In other words, the original distribution is partitioned into \( k \) continuous intervals, which are then each assigned to a single class in the categorical distribution that is designated to be the common distribution. Likewise, for discrete random variables, polychotomization entails the association between the elements from the original sample space and each of the \( k \) classes of the categorical distribution. Consequently, it is advisable the number of sample space elements \( \omega \) in the original variable be greater than or equal to \( k \).

One of the obvious implications of such transformations is that of preserving the monotonic or ordinal nature of the original random variable. Thus, only adjacent data points (points within a mutually distinct interval) and classes, in the case of continuous and discrete random variables respectively, may be combined to form the \( k \) classes (groups) of the new variable; otherwise, the transformation is not monotonic and the cdf of the transformed variable does not exist. A second implication of such transformations is that they may lead to the loss of information, unless the number of \( \omega \) is equal to \( k \) in the case of a discrete-to-discrete transformation. In effect, this is the price one must pay for transforming dissimilar distributions into a common distribution. A multitude of methods exist for establishing threshold values, including theoretical, empirical, consensus, and so on. Although to some extent the values at which thresholds are set may be arbitrary, they also depend on the nature of the raw variable and the purpose the transformed variable will serve. However, for the purpose of understanding the implication of using polychotomous transformations, this section will assume that a defendable method was used to establish these thresholds.
Dichotomization. The simplest transformation that may be performed on a random variable is dichotomization (i.e., dividing its distribution into two mutually exclusive classes or groups) based upon a single threshold \(\tau\). Mathematically, dichotomization is simply the application of function \(g\) to random variable \(X\) in order to yield a new random variable \(Y\) that has only two classes, i.e., \(Y=g(X)\). Although any value can be assigned to each class of \(Y\), in practice, one class is assigned the value 1 while the other class is assigned the value 0. Therefore, regardless of the original sample space and distribution of \(X\), the sample space of \(Y\) will be \(\Omega_Y=\{\omega: 0,1\}\) and its pmf will be \(P(Y=1)=P_Y(1)=p\), and zero otherwise. To be precise, the new random variable will conform to a Bernoulli distribution with \(\mathbb{E}(Y)=p\) and \(\text{Var}(Y)=p(1-p)\). The challenge, of course, is to derive \(p\) based exclusively on \(X\) and \(\tau\).

Suppose, for example, all the values of \(X\) greater than \(\tau\) are coded to 1, and zero otherwise. That is, \(Y=1\) if \(X>\tau\) and \(Y=0\) if \(X\leq \tau\). Then based on the identity \(F_Y(y)=F_X[g^{-1}(y)]\) (see Property 5) and the Bernoulli cdf, \(p\) is equal to the difference between the cdf of \(Y\) evaluated at 1 and 0, \(p=F_Y(1)-F_Y(0)=1-P(Y\leq 0)=1-P\{g(X)\leq 0\}=1-P\{\sup[g^{-1}(g(X))\leq g^{-1}(0)]\}\)

\(=1-P(X\leq \tau)=1-F_X(\tau)\). Therefore, \(p\) is equal to one minus the sum of the probabilities in \(X\) less than \(\tau\). Moreover, the expected value and variance of the transformed random variable is \(\mathbb{E}(Y)=1-F_X(\tau)\) and \(\text{Var}(Y)=F_X(\tau)[1-F_X(\tau)]\). Notice that both properties were derived exclusively on the basis of knowledge of \(X\) and \(\tau\). For example, suppose one wishes to transform a standard normally distributed random variable \(X \sim N(0,1)\) into a Bernoulli distributed random variable \(Y\) by dichotomizing the raw variable at \(\tau=1\sigma\) such that \(Y=1\) if \(X>1\sigma\) and \(Y=0\) if \(X\leq 1\sigma\). Then, \(p=1-\Phi_X(1)\approx 1-0.8413=0.1587\). Hence, the expected value

\[68\text{ Since transformation }g\text{ is not one-to-one, numerous values satisfy the inequality }X\leq g^{-1}(0). \text{ However, the goal is to find the maximum value of } p. \text{ Hence, one is only interested in the smallest upper bound (i.e., the supremum, denoted as sup) of the set that satisfies } X\leq g^{-1}(0), \text{ which is } \tau. \text{ Had the problem been reversed and one needed to determine } P[X>g^{-1}(0)], \text{ one would have needed to find the largest lower bound (i.e., the infimum, denoted inf) of the set that satisfies } X>g^{-1}(0), \text{ which is also } \tau.\]
and variance of the transformed variable is $\mathbb{E}(Y)\approx 0.1587$ and $\text{Var}(Y)\approx 0.1587(1-0.1587) = 0.1335$, respectively. Notice, these properties are significantly different than those of the raw variable, where $\mathbb{E}(X)=0$ and $\text{Var}(X)=1$. Figure 14 illustrates the probability density function of $X$ and the probability mass function of $Y$.

![Figure 14. Pdf of $X \sim N(0,1)$ and Pmf of $Y \sim Bin[1, 1-\Phi_X(1)]$](image)

Returning to the tenure example, suppose the review committee decided to dichotomize each of the seven criteria listed under the Research macrovalue (see Figure 1). Moreover, assume the number of publications ($X_1$) and presentations ($X_2$) conform to Poisson distributions with parameters 6 and 10, respectively, i.e., $X_1 \sim \text{Poi}(6)$ and $X_2 \sim \text{Poi}(10)$; the quality of publications ($X_3$) conforms to a normal distribution where $X_3 \sim N(2,0.75)$; the quality of conference presentations ($X_4$) conforms to a continuous uniform distribution ranging between 0 and 100, i.e., $X_4 \sim U(0,100)$; the impact of the candidate’s publications ($X_5$) conforms to a categorical distribution with three classes (H=High, M=Moderate, L=Low) where $X_5 \sim \text{Mult}(1; .35, .40, .25)$; the impact of the conference presentations ($X_6$) conforms to a categorical distribution with five classes where $X_6 \sim \text{Mult}(1; .05, .15, .25, .35, .20)$ and each class represents the self-reported likelihood an attendee would use the information in their
research or evaluation projects (5=Very likely, 4=Likely, 3=Possibly, 2=Probably not, 1=Definitely not); and finally, the number of recent changes in the number, quality, and impact of the candidate’s publications and presentations (X) conforms to a binomial distribution where \( X \sim Bin(6,0.25) \). Furthermore, suppose the threshold levels for \( X \) are set at \( \tau_1=4, \tau_2=8, \tau_3=1.5, \tau_4=75\%, \tau_5=L, \tau_6=3, \) and \( \tau_7=2 \), respectively. What is the expected value and variance of the composite? To calculate the mean and variance of the composite, one must first transform each of the seven variables.

The first two transformations require the use of the Poisson distribution. The pmf of the first transformed variable, \( Y_1 \), is given by \( Y_1=1 \) for \( X_1>4 \) and zero otherwise. Hence, 
\[
\mathbb{E}(Y_1)=p_1=P(X_1>\tau_1)=1-F_{X_1}(4)=1-\sum_{k=0}^{4} \frac{\lambda^k e^{-\lambda}}{k!} \approx 1-0.2851=0.7149
\]
and 
\[
\text{Var}(Y_1)=p_1(1-p_1)=0.7149(1-0.7149)=0.2038.
\]
In a similar fashion, the pmf of the second transformed variable is given by \( Y_2=1 \) for \( X_2>8 \) and zero otherwise, which connotes that 
\[
\mathbb{E}(Y_2)=1-F_{X_2}(8)=1-\sum_{k=0}^{8} \frac{\lambda^k e^{-\lambda}}{k!} \approx 1-0.3328=0.6672
\]
and 
\[
\text{Var}(Y_2)=0.6672(1-0.6672)=0.222.
\]
The third transformation makes use of the normal distribution, where the pmf of \( Y_3 \) is given by \( Y_3=1 \) for \( X_3>1.5 \) and zero otherwise. Thus, 
\[
\mathbb{E}(Y_3)=1-F_{X_3}(1.5)=1-\Phi(-0.5774) \approx 1-0.2819=0.7181
\]
and 
\[
\text{Var}(Y_3)=0.7181(1-0.7181)=0.2024.
\]
The fourth transformation utilizes the continuous uniform distribution, where \( Y_4=1 \) for \( X_4>75\% \) and zero otherwise. The expected value and variance of \( Y_4 \), thus, is given by 
\[
\mathbb{E}(Y_4)=1-F_{X_4}(75)=1-(\tau_4-a)/(b-a)=1-75/100=0.25,
\]
where \( a=0 \) and \( b=100 \), and 
\[
\text{Var}(Y_4)=0.25*0.75=0.1875.
\]
The next two transformations employ the categorical distribution. In the case of the fifth transformation, the pmf is given by \( Y_5=1 \) for \( X_5>L \) (i.e., \( X_5=H \) or \( M \)) and zero otherwise. Thus, the expected value and variance of \( Y_5 \) is 
\[
\mathbb{E}(Y_5)=1-F_{X_5}(L)=1-0.25=0.75 \quad \text{and} \quad \text{Var}(Y_5)=0.75(1-0.25)=0.1875.
\]
In a similar fashion, the pmf of the sixth transformed variable is given by \( Y_6=1 \) for \( X_6>3 \) and zero otherwise, which
means $\mathbb{E}(Y_6)=1-F_{X_6}(3)=1-\sum_{i=1}^3 p_i=1-(0.25+0.35+0.20)=0.2$ and $\text{Var}(Y_6)=0.2(1-0.2)=0.16$.

Finally, the pmf of $Y_7$ is given by $Y_7=1$ for $X_7>2$ and zero otherwise, which means $\mathbb{E}(Y_7) = 1-F_{X_7}(2)=1-\sum_{k=0}^2(\begin{pmatrix} 6 \\ k \end{pmatrix})(0.25)^k(0.75)^{6-k}=1-[(0.75)^6+6(0.25)(0.75)^5+15(0.25)^2(0.75)^4]=1-0.8306 =0.1694$ and $\text{Var}(Y_7)=0.1694(1-0.1694)=0.1407$.

In light of these values, the expected value and variance of the composite may now be computed. Suppose the seven transformed variables are independent. If the composite variable is constructed by averaging the seven new variables then its expect value and variance is given by Properties 10 and 22, respectively. That is to say, $\mathbb{E}(Y)=k^{-1}\sum_{i=1}^k \mathbb{E}(Y_i) = (0.7149+0.6672+0.7181+0.25+0.75+0.2+0.1694)/7=0.4957$ and $\text{Var}(Y)=k^{-2}\sum_{i=1}^k \text{Var}(Y_i) = (0.2038+0.2224+0.2024+0.1875+0.1875+0.16+0.1407)/7^2=0.0266$. One can also construct a composite variable by summing the seven transformed variables. The expected value and variance are then given by Properties 9 and 19; namely, $\mathbb{E}(Y)=\sum_{i=1}^k \mathbb{E}(Y_i)=3.4696$ and $\text{Var}(Y)=\sum_{i=1}^k \text{Var}(Y_i)=1.3039$. Notice, that the two composite variables do not conform to a Bernoulli or binomial distribution despite the fact that the expected value of a set of Bernoulli variables has a range $(0,1)$. Still, the figures are substantially more interpretable than the figures one would obtain by constructing a composite variable from the raw variables. However, dichotomization is not always the best method of establishing a common distribution due to the amount of information lost, particularly for continuous variables.

**Polychotomization.** An obvious way in which more information may be preserved following a transformation is to partition the original distribution into more than two mutually exclusive classes. Hence, polychotomization is a generalization of dichotomization in which the probability distribution of the raw random variable is partitioned into $k$ classes.

---

69 If the first composite variable conformed to a Bernoulli variable then its variance would be approximately equal to $0.25 \approx 0.4957(1-0.4957)$, whereas if the second composite variable conformed to a binomial distribution then its variance would be approximately equal to $1.75 \approx 7\times0.4957(1-0.4957)$.  

138
based on the designation of \( k-1 \) thresholds. As such, polychotomization may be viewed as a function \( g \) applied to random variable \( X \), which yields random variable \( Y \), i.e., \( Y=g(X) \). As was the case before, only adjacent data points or classes of the raw distribution can be combined to form the classes of the transformed random variable in order to preserve monotonicity. Furthermore, as a rule, the number of classes for the transformed variable should not exceed the number of classes of the lowest raw discrete variable. This issue is not a mathematical limitation as much as a substantive one. For example, it is certainly the case that the lowest and highest values of a dichotomous variable can be recoded to the lowest and highest values of a polychotomous variable with \( k \geq 2 \) classes leaving the remainder of the classes with missing data, but no additional information is gained as a result of such a transformation. Moreover, problems are likely to arise whenever one must construct and interpret a composite generated from variables with missing data.

Although any value or label can be assigned to a class, in practice, the first class is assigned the value 1 with sequential classes assigned incremental integers. Similar to dichotomization, classes are defined on the basis of thresholds \( \tau_\theta \), where the values of \( X \) less than or equal to \( \tau_1 \) are coded 1, the values of \( X \) less than or equal to \( \tau_2 \) but greater than \( \tau_1 \) are coded to 2, and so on until the values of \( X \) greater than \( \tau_{k-1} \) are coded to \( k \). That is, \( Y=1 \) for \( X \leq \tau_1 \), \( Y=2 \) for \( \tau_1 < X \leq \tau_2 \),…, \( Y=k \) for \( X > \tau_{k-1} \). Hence, regardless of the distribution of \( X \), \( Y \) conforms to an ordered categorical distribution with a sample space of \( \Theta_Y=\{0; 1,2,\ldots,k\} \) and a pmf of \( P(Y=1)=p_1, \ P(Y=2)=p_2, \) and so on till \( P(Y=k)=p_k=1-(p_1+\ldots+p_{k-1}) \). The challenge is to derive the pmf of \( Y \) based solely on \( X \) and \( \tau_\theta \). To this end, it can be shown \( p_\theta \) is equal to the difference between the cdf of \( Y \) evaluated at \( \theta \) and \( \theta-1 \), \( p_\theta=F_Y(\theta)-F_Y(\theta-1) \). Furthermore, \( F_Y(\theta-1)=0 \) when \( \theta \) is less than the first element of \( \Theta \) and \( F_Y(\theta)=1 \) when \( \theta=k \). Each class is then defined by an inverse transformation of the thresholds boundaries for \( \theta \),
where $p_\theta = P(Y=\theta) = P\{\sup[X \leq g^{-1}(\theta)]\} - P\{\inf[X \leq g^{-1}(\theta)]\} = P[X \leq \tau_1] - P[X \leq \tau_{1-1}] = F_X(\tau_1) - F_X(\tau_{1-1})$

in view of the fact that the supremum and infimum of the inverse transformation of $Y$ evaluated at $\theta$ define the boundaries of class $\theta$.

In the previous section, a method was proposed by which the impact of a publication could be measured. Therein, it was suggested an ‘A’ would be assigned to a publication with more than 10 citations, a ‘B’ would be assigned to a publication with 6-10 citations, and a ‘C’ would be assigned to a publication with 5 or less citations. If we assume the number of citations ($X$) conforms to a Poisson distribution then this coding process, $Y=g(X)$, is an example of polychotomization where $k=3$, $\tau_1=5$ and $\tau_2=10$. Suppose $X \sim \text{Poi}(9)$, what is the proportion of values of $X$ that will be classified in each of the three classes of $Y$?

Given that $F_X(\omega) = \sum_{i=0}^{\omega} e^{-\lambda} \frac{\lambda^i}{i!}$, then $P(Y=C) = P\{X \leq \sup[g^{-1}(C)]\} = P[X \leq \tau_1] = F_X(\tau_1) = P[X \leq 5] \approx 0.116$, $P(Y=B) = P\{\sup[X \leq g^{-1}(B)]\} - P\{\inf[X \leq g^{-1}(B)]\} = F_X(\tau_2) - F_X(\tau_1) = F_X(10) - F_X(5) \approx 0.706 - 0.116 = 0.59$, and $P(Y=A) = 1 - P\{\inf[X \leq g^{-1}(A)]\} = 1 - P[X \leq \tau_2] = 1 - F_X(\tau_2) = 1 - F_X(10) \approx 1 - 0.706 = 0.294$. A quick check of the sum of the probabilities confirms they sum to unity. Moreover, a SAS simulation and ensuing Figure 15 lend further support to the veracity of these results.

<table>
<thead>
<tr>
<th>Raw variable</th>
<th>Transformed variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_1=5$</td>
<td>$59.03%$</td>
</tr>
<tr>
<td>$\tau_2=10$</td>
<td>$29.99%$</td>
</tr>
</tbody>
</table>

Figure 15. Pmf of $X \sim \text{Poi}(9)$ and Pmf of $Y \sim \text{Mult}[1, 1-F_X(\tau_1), F_X(\tau_2)-F_X(\tau_1), 1-F_X(\tau_2)]$
Measurement Scales

The expected value and variance of Y depends upon the nature of the measurement scale (i.e., ordinal, interval, ratio) assumed to underlie the transformed variable. An ordinal transformation, in slight contrast, produces a random variable that conforms to an ordered categorical distribution. However, because the numbers or labels assigned to each class only represent order, rather than quantity, the methods previously presented for calculating the expected value and variance are not permitted and the results of such computations are uninterpretable. Fortunately, one may calculate percentiles because they only require sorting data and reporting the number of values, as a percent of the total, that are less than a specific value. These statistics will be discussed in the ensuing subsection while this subsection focuses on interval and ratio transformations designed to produce a discrete random variable. In the latter case, the expected value of Y is given by \( \mathbb{E}(Y) = \sum_{\theta=1}^{k} y_\theta p_\theta \) which, following the same reasoning used for dichotomization, can be shown to be computed by

\[ \mathbb{E}(Y) = \sum_{\theta=1}^{k} y_\theta[F_X(\tau_\theta) - F_X(\tau_{\theta-1})] \]

Likewise, the variance of Y, \( \text{Var}(Y) = \sum_{\theta=1}^{k} (y_\theta - \mathbb{E}(Y))^2 p_\theta \) can be shown to be

\[ \text{Var}(Y) = \sum_{\theta=1}^{k} (y_\theta - \mathbb{E}(Y))^2[F_X(\tau_\theta) - F_X(\tau_{\theta-1})] \]

Notice the striking similarity between these formulas and the ones in Properties 5 and 15.

Now, suppose the review committee decided to polychotomize each of the three criteria listed under the Teaching macrovalue. Examination of two of the three criteria, number of teaching awards received \((X_1)\) and amount of teaching workload \((X_2)\), suggests that, in all likelihood, they conform to Poisson distributions. Thus, suppose \(X_1 \sim \text{Poi}(1)\) and \(X_2 \sim \text{Poi}(4)\) with thresholds \(\tau_1(1) = 0, \tau_1(2) = 1, \text{ and } \tau_1(3) = 2\) for \(X_1\) and \(\tau_2(1) = 2, \tau_2(2) = 4, \text{ and } \tau_2(3) = 6\) for

---

70 Consider an ordinal scale measuring educational attainment in which 1 represents high school degree, 2 represents associate’s degree, 3 represents college degree, 4 represents master’s degree, and 5 represents doctoral degree. It is certainly not the case that 3 minus 2 equals 1 nor that 4 divided by 2 is 2. However, the expected value and variance necessitate the use of these mathematical operations. Therefore, since the product of such operations is uninterpretable, so are the mean and variance produced from ordinal data.
Furthermore, although we can be fairly certain the third criterion, quality of student evaluations ($X_3$), is measured by numerous questions, assume, for the sake of simplicity, it is measured by a single variable employing a four-point Likert scale, where $\Theta=\{0: 1=\text{strongly disagree}, 2=\text{disagree}, 3=\text{agree}, \text{and} 4=\text{strongly agree}\}$. To keep the number of $k$-tuples to a manageable level, suppose the median student response per course is treated as a trial rather than individual student responses,\(^7\) where $X_3 \sim \text{Mult}(3; 0.2, 0.25, 0.35, 0.2)$. Furthermore, suppose the thresholds, going in reverse order, are set so that $\tau_{3(3)}$ demarks the $k$-tuples for which half or more of the responses are rated ‘strongly agree’, $\tau_{3(2)}$ demarks the $k$-tuples for which half or more of the responses are rated ‘agree’ or higher, and $\tau_{3(1)}$ demarks the $k$-tuples for which half or more of the responses are rated ‘disagree’ or higher. Then, what is the expected value and variance of the mean of the three transformed variables? Since three thresholds were identified for the raw variable $s$, the transformed variables will each have four groups. The pmf of the first transformed variable, $Y_1$, is given by $Y_1=1$ for $X_1=0$, $Y_1=2$ for $X_1=1$, $Y_1=3$ for $X_1=2$, and $Y_1=4$ for $X_1>2$. Hence, since $F_{X_1}(\omega) = \sum_{i=0}^{\omega} e^{-1(i)/\lambda}$, then $p_{1(1)}=F_{X_1}[\tau_{1(1)}]\approx0.3679$, $p_{1(2)}=F_{X_1}[\tau_{1(2)}]-F_{X_1}[\tau_{1(1)}]\approx0.7358-0.3679=0.3679$, $p_{1(3)}=F_{X_1}[\tau_{1(3)}]-F_{X_1}[\tau_{1(2)}]\approx0.9197-0.7358=0.1839$, and $p_{1(4)}=1-F_{X_1}[\tau_{1(3)}]\approx1-0.9197=0.0803$.

Likewise, the pmf of the second transformed variable, $Y_2$, is given by $Y_2=1$ for $X_2\leq 2$, $Y_2=2$ for $2<X_2\leq 4$, $Y_2=3$ for $4<X_2\leq 6$, and $Y_2=4$ for $X_2>6$. Hence, since $F_{X_2}(\omega) = \sum_{i=0}^{\omega} e^{-4(4)/\lambda}$, then $p_{2(0)}=F_{X_2}[\tau_{2(0)}]\approx0.2381$, $p_{2(1)}=F_{X_2}[\tau_{2(1)}]-F_{X_2}[\tau_{2(0)}]\approx0.6288-0.2381=0.3907$, $p_{2(2)}=F_{X_2}[\tau_{2(2)}]-F_{X_2}[\tau_{2(1)}]\approx0.8893-0.6288=0.2605$, and $p_{2(3)}=1-F_{X_2}[\tau_{2(3)}]\approx1-0.8893=0.1107$.

The third transformed variable, $Y_3$, requires slightly more work due to the number of tuples that must be computed. Because there are four trials (i.e., the number of courses with

\(^7\) The number of $k$-tuples dramatically increase as the number of trials increase. For example, if the candidate received 100 student evaluations, then the number of $k$-tuples for a single Likert four-point item would be $(100+4−1)!/100!(4−1)!=176,851$. Hence, the median course rating was used instead of individual responses.
student evaluations) and four classes (i.e., four Likert points), a total of the sum of 35 tuples (= \(4+4-1)/4!(4-1)\)) must be computed. Furthermore, although the description provided for how the thresholds are set is typical of those found in evaluation, it is not precise enough for mathematical work. Therefore, the statement must be decoded. Since there are four possible trials (courses ratings) that may result in outcome \(θ\), \(N=\sum_{i=1}^{4} n_i=4\), the statement ‘half or more of the responses’ translates into two or more trials. This, in turn, implies that the thresholds of \(Y_3\) are \(τ_3(1)\) when \(\sum_{\theta=2}^{4} n_\theta\geq2\), \(τ_3(2)\) when \(\sum_{\theta=3}^{4} n_\theta\geq2\), and \(τ_3(3)\) when \(n_4\geq2\). Consequently, the pmf of \(Y_3\) must be \(Y_3=1\) for \(n_1>2\), \(Y_3=2\) for \(\sum_{\theta=2}^{4} n_\theta\geq2\) and \(\sum_{\theta=3}^{4} n_\theta<2\), \(Y_3=3\) for \(\sum_{\theta=3}^{4} n_\theta\geq2\) and \(n_4<2\), and \(Y_3=4\) for \(n_4\geq2\). Notice, that unlike univariate distributions, the lower and upper bounds are not a single value since multiple combinations of \(n_\theta\) may satisfy these conditions. Hence, utilizing the multinomial cdf, \(F_{X_3}(ω)\), the four probabilities are

\[
p_1 = \frac{4!}{4!} \left(0.2^4\right) + \frac{4!}{3!1!} \left(0.2^30.2\right) + \frac{4!}{2!2!} \left(0.2^20.25\right) + \frac{4!}{1!3!} \left(0.20.35\right) = 0.0272
\]

\[
p_2 = \frac{4!}{4!} \left(0.25^4\right) + \frac{4!}{3!1!} \left(0.25^30.2\right) + \frac{4!}{2!2!} \left(0.25^20.25\right) + \frac{4!}{1!3!} \left(0.250.35\right)
\]

\[
+ \frac{4!}{2!2!} \left(0.25^20.25\right) + \frac{4!}{1!3!} \left(0.250.35\right) \approx 0.2143
\]

\[
p_3 = \frac{2}{2} \frac{4!}{4!} \left(0.35^4\right) + \frac{4!}{3!1!} \left(0.35^30.2\right) + \frac{4!}{2!2!} \left(0.35^20.25\right) + \frac{4!}{1!3!} \left(0.350.35\right)
\]

\[
+ \frac{4!}{2!2!} \left(0.35^20.35\right) + \frac{4!}{1!3!} \left(0.350.35\right) \approx 0.5777
\]
\[ p_i = \sum_{n_1=0}^{2} \sum_{n_2=0}^{2-n_1-n_2} \sum_{n_3=0}^{4} n_1 n_2 n_3 ! (4-n_1-n_2-n_3) \frac{4!}{4!-\text{i}} p_1^2 p_2^2 p_3^2 p_4 \]

\[ = \frac{4!}{4!} (0.2^2) + \frac{4!}{4!} (0.35^1 \cdot 0.2^1) + \frac{4!}{4!} (0.35^1 \cdot 0.2^4) + \frac{4!}{4!} (0.25^1 \cdot 0.2^3) + \frac{4!}{4!} (0.25^1 \cdot 0.35^1 \cdot 0.2^2)
\]

\[ + \frac{4!}{2!} (0.25^2 \cdot 0.2^2) + \frac{4!}{2!} (0.2^2 \cdot 0.2^2) + \frac{4!}{2!} (0.2^1 \cdot 0.35^1 \cdot 0.2^2) + \frac{4!}{2!} (0.2^1 \cdot 0.25^1 \cdot 0.2^2)
\]

\[ + \frac{4!}{2!} (0.2^2 \cdot 0.2^2) = 0.1808. \]

Note, the most challenging aspect in determining these probabilities was not the math but rather identifying the correct bounds for each summation. Fortunately, the SAS program in Appendix F (lines 137-220) only requires the threshold conditions and not both boundaries.

A check of the transformed variables confirms their probabilities sum to unity. The expected value and variance can be computed using the computational formulas for discrete variables:

\[ E(Y_1) = \sum_{i=1}^{4} y_i [F_X(\tau_0) - F_X(\tau_{i-1})] = 1(0.3679) + 2(0.3679) + 3(0.1839) + 4(0.0803) = 1.9766 \]

and \[ Var(Y_1) = \sum_{i=1}^{4} [y_i - E(Y)]^2 [F_X(\tau_0) - F_X(\tau_{i-1})] = (1-1.9766)^2(0.3679) + (2-1.9766)^2(0.3679) + (3-1.9766)^2(0.1839) + (4-1.9766)^2(0.0803) \approx 0.8725; \]

\[ E(Y_1) = 1(0.2381) + 2(0.3907) + 3(0.2605) + 4(0.1107) = 2.2438 \]

and \[ Var(Y_2) = (1-2.2438)^2(0.2381) + (2-2.2438)^2(0.3907) + (3-2.2438)^2(0.2605) + (4-2.2438)^2(0.1107) \approx 0.882; \]

\[ E(Y_3) = 1(0.0272) + 2(0.2143) + 3(0.5777) + 4(0.1808) = 2.9121 \]

and \[ Var(Y_3) = (1-2.9121)^2(0.0272) + (2-2.9121)^2(0.2143) + (3-2.9121)^2(0.5777) + (4-2.9121)^2(0.1808) \approx 0.4962. \]

Second, the expected value and variance of the composite mean can be computed by Properties 10 and 22, respectively. That is, \[ E(Y) = k^{-1} \sum_{i=1}^{k} E(Y_i) = (1.9766 + 2.2438 + 2.9121)/3 = 2.3775 \]

and \[ Var(Y) = k^{-2} \sum_{i=1}^{k} Var(Y_i) = (0.8725 + 0.882 + 0.4962)/3 \approx 0.2501. \]

Note, however, the composite variable was constructed under the assumption that the three variables are independent. Had the transformed variables not been independent, one could still determine the variance of the composite using Property 22 for dependent measures. Of course, this would require adjusting the correlation coefficient. Another
unstated assumption is the existence of equidistance between each class. If this assumption is not met, then the expected value and variance estimates are tenuous at best.

Order Statistics and Quantiles

In most situations, the equidistance assumption is difficult to establish. In the social sciences, data elicited from subject-centered scaling methods (e.g., Likert, frequency, or quality scales) tends to be ordinal. This has led to considerable debate and consternation among researchers and evaluators who wish to properly analyze the data without sacrificing their ability to do so. In one camp are “purists” who believe that mathematical operations such as addition, subtraction, multiplication, and division should not be performed on ordinal data. In the other camp are researchers who wish to rescue certain measurement scales, particularly Likert scales, from being relegated to ordinal status. These researchers argue or otherwise assume Likert scales fall somewhere between true ordinal and interval scales since subjects self-center their responses and in so doing, establish equidistance between response choices. Unfortunately, this is only an assumption that is neither easily tested nor met. Item response theory (IRT) does offer one the possibility of examining the assumption by determining whether the distances between the threshold parameters (boundaries) of a polytomously scored item are equal.

Perhaps the strongest and most frequently stated objection by proponents who favor treating Likert data as interval has been the remark that suitable analytical methods do not exist, particularly when interpretability is an issue. Therefore, evaluators and researchers alike argue that science should not be handcuffed to rigid mathematical principals, frequently

72 An example of frequency scaling is “never”, “seldom”, “sometimes”, “often”, and “always” while an example of a quality scale would be “poor”, “fair”, “good”, and “outstanding”.
stated as “Do not let the perfect get in the way of the good enough.” Not surprisingly, this attitude is also shared by some mathematical statisticians.

“There is nothing unique or objective in assigning the values 4, 3, and 2 to grades A, B, and C, considering that the process of averaging grades usually includes different subjects, grades by different teachers, and with criteria often formulated rather vaguely. Nevertheless, tradition and the practical need to assess students’ performance force one to fix the scoring system for grades, and regard them as a measurement on an interval scale [italic added].” (Bartoszynski & Niewiadomska-Bugaj, 2008, p. 370)

The purpose of the present subsection is to introduce readers to order statistics (Arnold, Balakrishnan, & Nagaraja, 2008; Hogg, McKean, & Craig, 2005) and illustrate how they may be used to measure central tendency and dispersion. Order statistics are a nonparametric method (i.e., no distributional assumptions are made) developed in the past 50 years and can be found in most mathematical-statistics textbooks. Originally, the method was developed for continuous data and was intended to be used whenever distributional assumptions were not met or the distribution could not be identified. However, discrete order statistics (Arnold, Balakrishnan, & Nagaraja, 2008) have also been developed that can be used to analyze discrete data without the need to employ the mathematical operations such as addition or subtraction. Hence, it is the contention of this author that this method offers researchers and evaluators the long searched for alternative to simply assuming ordinal data can be treated as interval.

Before presenting these methods, one must determine whether expected value and variance are appropriate estimates to use when data violate distributional assumptions or are ordinal. With regard to the latter, since it is clear that if one cannot perform arithmetic operations, then estimates such as the mean and variance are not meaningful. Likewise, given a severe enough violation of distributional assumptions or the presence of outliers, analyses are likely to produce misleading results unless one employs a robust estimator (resistant to
outliers), such as the median\textsuperscript{73}. For odd data, the median ($m$) is defined as the point at which half of the observations fall below the point, $P(X\leq m)\geq 0.5$, and half the observations fall above the point, $P(X\geq m)\geq 0.5$. For even data, the median is defined as the average of the two middle (ordered) values. Whenever the median is employed, the measures of dispersion most frequently employed are the range and the interquartile range (IQR), which is defined as the middle 50 percent of the range. Since these estimators are not frequently employed in social science research, this subsection will continue to rely on variance or its approximation.

*Continuous order statistics.* Order statistics are a relatively new subject in mathematical-statistics with the first treatise only dating back to 1962 (Arnold, Balakrishnan, & Nagaraja, 2008). However, the popularity of these statistics has risen, and can now be found in most mathematical-statistics textbook, because some of their properties do not depend upon the distribution from which the random sample is drawn (Hogg, McKean, & Craig, 2005). Essentially, order statistics entail arranging observations randomly sampled from a distribution in ascending order of magnitude. Suppose $X_1, X_2, \ldots, X_n$ represent a random sample from the distribution of variable $X$, then $X_{(1)}$ represents the smallest sample value, $X_{(2)}$ represents the second smallest sample value, and so on until $X_{(n)}$ represents the largest sample value. The ordered values $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are known as the order statistics corresponding to the random values of $X$. For example, suppose a sample of size five is drawn from random variable $X$ and the following observations are recorded: 2, 56, 45, 6, and 0 (i.e., $x_1=2$, $x_2=56$, $x_3=45$, $x_4=6$, and $x_5=0$, where the subscript denotes the order in which the observations were recorded). Then, the order statistics for these observations are $x_{(1)}=0$, $x_{(2)}=2$, $x_{(3)}=6$, $x_{(4)}=45$, and $x_{(5)}=56$. Note, the definition of order statistics did not require the

\textsuperscript{73} Breakdown analyses have found that the sample median achieves the highest possible breakdown point (50\%) whereas for the mean the breakdown point is zero (Hogg, McKean, & Craig, 2005). In other words, it only takes one extreme outlier to corrupt the mean while at least 50\% of the data would need to contain extreme outliers to corrupt the sample median.
observations \((x_i)\) to be independent, identically distributed, or even continuous. However, assumptions are necessary to derive the pdf and cdf of these statistics.

It can be shown\(^{74}\) that the pdf of the \(k^{th}\) order statistic \(X_{(k)}\), denoted as \(P(X_{(k)}=x) = f_{X_{(k)}}(x)\), is given by

\[
f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} \left[F_X(x)ight]^{k-1} \left[1-F_X(x)ight]^{n-k} f_X(x),
\]

where there are \(k-1\) observations of \(X\) with probability \(P(X \leq x) = F_X(x)\), \(n-k\) observations of \(X\) with probability \(P(X > x) = 1-P(X \leq x) = 1-F_X(x)\), one observation of \(X\) with probability \(P(X=x) = f_X(x)\), and \(\binom{n}{k-1,1,n-k}\) ways in which the values of the \(n\) observations can be arranged into the three groups, as illustrated in Figure 16. In other words, the pdf of the \(k^{th}\) order statistics is given by the trinomial distribution—a multinomial distribution with three classes.

![Figure 16. Relation Between the Trinomial Distribution and Continuous Order Statistics](image)

Although the cdf of \(X_{(k)}\) evaluated at \(X=x\), \(F_{X_{(k)}}(x)\), can be obtained by integrating \(f_{X_{(k)}}(x)\), a simpler derivation is attained by realizing the \(k^{th}\) order statistic is less than or equal to \(x\), \(P(X_{(k)} \leq x) = F_{X_{(k)}}(x)\), if, and only if, \(k\) or more of the values of \(X\) are less than or equal to \(x\) and \(n-k\) of the values of \(X\) are greater than \(x\). For example, suppose \(P(X \leq x) = 0.25\), what is the probability \(X_{(3)}\) is less than or equal to \(x\) if a random sample of size six is drawn from \(X\)? The third order statistic can only be less than or equal to \(x\) if three or more of the \(X_i\) observations are less than or equal to \(x\). That is, the condition \(\{X_i \leq x\}\) holds true when \(X_{(3)}\), \(X_{(6)}\), \(X_{(5)}\), or \(X_{(6)}\) are less than or equal to \(x\). Moreover, since \(P(X \leq x) = 0.25\), \(P(X > x) = 0.75\), and 3 out of the 6 observations can be arranged in \(\binom{6}{3}\) ways resulting in the same two groups, the

\(^{74}\) Proofs of the formulas herein can be found in many statistics textbooks, including Arnold, Balakrishnan, and Nagaraja (2008), Bartoszynski and Niewiadomska-Bugaj (2008), Hogg, McKean, and Craig (2005), Ross (2002), and Larson (1973). The proofs can also be made available by the present author upon request.

\(^{75}\) The band that encloses \(X_{(k)}\) ranges from \(x\) to \(x+dx\), where \(dx\) is so small the probability that more than one random variable falls in the interval \([x, x+dx]\) is negligible and \(P(X>x)\) is approximately equal to \(P(X>x+dx)\).
probability \( P(X_{(n)} \leq x < X_{(k)}) \) is equal to the binomial probability of \((\frac{3}{5})(0.25)^3(0.75)^3 \approx 0.1318\).

Summing the binomial probabilities for the remaining three order statistics to the previous probability then yields \( P(X_{(n)} \leq x) \approx 0.1694\). More generally, the probability \( P(X_{(k)} \leq x < X_{(k+1)}) \) then is equal to \((\frac{3}{5})[F_X(x)]^4[1-F_X(x)]^{n-k}\) since there are \(k\) values of \(X\) for which \(P(X \leq x)\), \(n-k\) values of \(X\) for which \(P(x < X) = 1 - P(X \leq x) = 1 - F_X(x)\), and \(\binom{n}{k}\) ways in which \(n\) observations can be arranged into two groups. Since the \(k^{th}, (k+1)^{th}, \ldots, n^{th}\) order statistics all satisfy the condition \(\{X_{(k)} \leq x\}\) then \(P(X_{(k)} \leq x) = P(X_{(k)} \leq x < X_{(k+1)}) + P(X_{(k+1)} \leq x < X_{(k+2)}) + \ldots + P(X_{(n)} \leq x);\)

\[
F_{X_{(k)}}(x) = P(X_{(k)} \leq x) = \sum_{i=0}^{n-k} \binom{n}{i} [F_X(x)]^i[1-F_X(x)]^{n-i} = 1 - \sum_{i=0}^{n-k} \binom{n}{i} [F_X(x)]^i[1-F_X(x)]^{n-i}.
\]

That is, the cdf of the \(k^{th}\) order statistics is the tail probability starting from \(k\) of a binomial distribution with \(F_X(x)\) probability of success and \(n\) number of trials.\(^76\)

Suppose five observations, \(X_1\) to \(X_5\), are randomly sampled from a standard normal distribution with pdf \(f_X(x) = \varphi_X(x)\), cdf \(F_X(x) = \Phi_X(x)\), and order statistics \(X_{(1)}\) to \(X_{(5)}\). What is the probability the third order statistic lies in the interval \([-0.5,0.5]\)? As discussed above, two methods can be used to compute this probability. The first method entails integrating \(f_{X_{(3)}}(x)\), whereas the second method entails subtracting one cdf from another. If \(-\infty < a < X_{(k)} < b < \infty\), then \(P(a < X_{(k)} < b) = F_{X_{(b)}}(b) - F_{X_{(a)}}(a) = \sum_{j=0}^{k-1} \binom{n}{j} [F_X(a)]^j[1-F_X(a)]^{n-j} - \sum_{j=0}^{k-1} \binom{n}{j} [F_X(b)]^j[1-F_X(b)]^{n-j} \).

Method 1: Integrating \(f_{X_{(3)}}(x)\) and evaluating it for \(X_{(3)} \in \{x: \ [-0.5,0.5]\}\)

\[
f_{X_{(3)}}(x) = \frac{n!}{(k-1)!(n-k)!} [F_X(x)]^{k-1}[1-F_X(x)]^{n-k} f_X(x)
\]

\[
= \frac{5!}{(3-1)!(5-3)!} [\Phi_X(x)]^{3-1}[1-\Phi_X(x)]^{5-3} \varphi_X(x) = \frac{5!}{2^2} [\Phi_X(x)]^{3} [1-\Phi_X(x)]^{2} \varphi_X(x).
\]

\(^76\) It is assumed that each random observation has the same probability of not exceeding \(x\) due to being drawn from an identical distribution, and the sample observations are independent. Consequently, under the iid assumption, the cdf of \(X_{(k)}\) evaluated at \(x\) is a binomial random variable with parameters \(n\) and \(p=F_X(x)\).
Therefore,  \( P(-0.5 < X_{(3)} < 0.5) = 30 \int_{-0.5}^{0.5} \left[ \Phi_X(x) \right]^3 [1 - \Phi_X(x)]^2 \varphi_X(x) \, dx \). Substituting 

\( u = \Phi_X(x) \) and \( du = \varphi(x) \, dx \) into the integrand then yields

\[
= 30 \int u^2(1-u^2) \, du = 30 \int u^2(1-2u+u^4) \, du = 30 \left[ \frac{u^3}{3} - \frac{2u^4}{4} + \frac{u^5}{5} \right]_{u=-0.5}^{u=0.5} \]

Substituting \( \Phi_X(x) = u \) back into the integrand and evaluating it at \([-0.5,0.5]\) yields

\[
= 30 \left[ \frac{\left[ \Phi_X(x) \right]^3}{3} - \frac{\Phi_X(x)^4}{2} + \frac{\Phi_X(x)^5}{5} \right]_{u=-0.5}^{u=0.5} \approx 0.651.
\]

Method 2: Utilizing \( F_{X(k)}(x) \) and evaluating it for \( X(k) \in \{ x : [F_X(-0.5), F_X(0.5)] \} \)

\[
P(-0.5 < X_{(k)} < 0.5) = F_{X(k)}[F_X(0.5)] - F_{X(k)}[-F_X(-0.5)] = F_{X(k)}[F_X(0.5)] - F_{X(k)}[F_X(-0.5)]
\]

\[
= \sum_{i=k}^{n} \binom{n}{i} F_X(0.5)^i [1-F_X(0.5)]^{n-i} - \sum_{j=k}^{n} \binom{n}{j} F_X(-0.5)^j [1-F_X(-0.5)]^{n-j}
\]

\[
= \sum_{i=3}^{5} \binom{5}{i} F_X(0.5)^i [1-F_X(0.5)]^{5-i} - \sum_{j=3}^{5} \binom{5}{j} F_X(-0.5)^j [1-F_X(-0.5)]^{5-j} \approx 0.651.
\]

This example illustrates how one could compute the probability of an order statistic falling between a fixed boundary. Moreover, it illustrates how much simpler the second method is to employ than the first. As a means of further explicating the application of order statistics and verifying the previous results, a Monte Carlo simulation was conducted. Five random values were generated from the normal distribution, sorted, and the third observation (i.e., \( X_{(3)} \)) was recorded. A variable was then created to record whether the value of the third observation fell within the interval \([-0.5,0.5]\). This process was repeated one million times and the proportion of times the third variable fell within the interval of interest was found to be approximately 0.6508. In other words, when samples of size five are drawn from a normal distribution, the third order statistic (median) will fall within the interval \([-0.5,0.5]\) approximately 65% of the time.
Discrete order statistics. The previous formulas are only suitable for strictly continuous random variables and should not be employed for discrete random variables. However, since the definition of order statistics does not require continuity, order statistics may be employed to extend the range of applications to discrete data. Suppose a random sample is drawn from an identically distributed and independent categorical distribution with unique values in ascending order \( x_1 < x_2 < \ldots \), pmf \( P(x) \), and cdf \( F(x) \). Then, the ordered statistics of \( X \) are \( X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \) since there exists a chance two or more observations can attain the same value of \( x \). Using logic similar to that used to derive the distribution of the \( k^{th} \) order statistic, one may reason that if \( i \) observations of \( X \) are less than \( x \) and \( j \) observations of \( X \) are greater than \( x \), then \( n-i-j \) observations of \( X \) must equal \( x \) with corresponding probabilities of \( \pi_1 = P(X < x) \), \( \pi_3 = P(X > x) \), and \( \pi_2 = P(X = x) \). Figure 17 illustrates these relations.

![Figure 17](image-url)  
**Figure 17. Relation Between the Trinomial Distribution and Discrete Order Statistics**

Utilizing the multinomial distribution, one can determine the probability that \( x \) falls within the range space of \( X \) by summing all of the individual probabilities as follows,

\[
P[X_{(1)} \leq x \leq X_{(n)}] = \sum_{i=0}^{n} \sum_{j=0}^{n-i} \frac{n!}{i!(n-i-j)!} \pi_1^i \pi_2^{n-i-j} \pi_3^j = \sum_{j=0}^{n} \sum_{i=0}^{n-j} \frac{n!}{i!(n-i-j)!} \pi_1^i \pi_2^{n-i-j} \pi_3^j = 1,
\]

where \( 0 \leq i, j \leq n \) and \( 1 \leq x \leq n \). From this equation, several other probabilities may be derived by suitably adjusting the range of \( i, j \), or both. In the case of the pdf of the \( k^{th} \) order statistic, no more than \( k-1 \) observations may be less than \( x \) and no more than \( n-k \) observations may be greater than \( x \) to fulfill the condition \( X_{(k)} = x \); for the cdf of the \( k^{th} \) order statistic, no more than \( n-k \) observations may be greater than \( x \) to fulfill the condition \( X_{(k)} \leq x \); and so on.
\[ f_{X_0}(x) = P(X_0 = x) = P \left( \text{no more than } k-1 \text{ observations are less than } x \text{ and no more than } n-k \text{ observations are greater than } x \right) \]

\[ = \sum_{i=0}^{k-1} \sum_{j=0}^{n-k} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(0 \leq i \leq k-1\) and \(0 \leq j \leq n-k\)

\[ F_{X_0}(x) = P(X_0 \leq x) = P(\text{no more than } n-k \text{ observations are greater than } x) \]

\[ = \sum_{i=0}^{n-k} \sum_{j=0}^{n-i-j} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(0 \leq i \leq n\) and \(0 \leq j \leq n-k, \text{ and } i+j \leq n\)

\[ P(X_0 \geq x) = P(\text{no more than } k-1 \text{ observations are less than } x) \]

\[ = \sum_{i=0}^{k-1} \sum_{j=0}^{n-i} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(0 \leq i \leq k-1\) and \(0 \leq j \leq n, \text{ and } i+j \leq n\)

\[ P(X_0 < x) = P(\text{no more than } n-k \text{ observations are greater than or equal to } x) \]

\[ = \sum_{i=k}^{n} \sum_{j=0}^{n-i-j} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(k \leq i \leq n\) and \(0 \leq j \leq n-k\)

\[ P(X_0 > x) = P(\text{no more than } k-1 \text{ observations are less than or equal to } x) \]

\[ = \sum_{i=0}^{k-1} \sum_{j=k}^{n-i} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(0 \leq i \leq n-k\) and \(k \leq j \leq n\)

\[ P[X_0 \leq x \leq X_0] = P \left( \text{no more than } b-1 \text{ observations are less than } x \text{ and no more than } n-a \text{ observations are greater than } x \right) \]

\[ = \sum_{i=0}^{b-1} \sum_{j=0}^{n-i-j} \frac{n!}{i!(n-i-j)!j!} \binom{n}{i} \binom{n-i-j}{j} \frac{1}{\binom{n}{i} \binom{n-i-j}{j} \binom{n}{k}}, \]

where \(0 \leq i \leq n-a\), \(0 \leq j \leq b-1\), and \(a \leq b\)

The constraint \(i+j \leq n\) was included for \(P(X_0 \leq x)\) and \(P(X_0 \geq x)\) primarily for programming purposes because the second and first equations for these probabilities, respectively, will not resolve without the constraint since \(n-i-j\) becomes negative. Note, the last formula will prove to be very useful for constructing nonparametric confidence intervals.
for it computes the probability that $x$ falls within the interval formed by the $d^{th}$ and $b^{th}$ order statistics, also known as *coverage probability*. Hence, one may ask, what order statistics will result in a desired (nominal) coverage probability for $x$? This question will be taken up in the section on confidence intervals. In the meantime, since working with the multinomial can be cumbersome, binomial equivalents can be derived for each formula.

\[
    f_{X_{(d)}}(x) = F_{X_{(d)}}(x) - F_{X_{(d-1)}}(x) = \sum_{j=d}^{n} \binom{n}{j} \left[ (1 - \pi_3)^{n-j} - \pi_3 (1 - \pi_1)^{n-j} \right]
\]

\[
    F_{X_{(d)}}(x) = P(X_{(d)} \leq x) = \sum_{j=d}^{n} \binom{n}{j} (1 - \pi_3)^{n-j} = \sum_{j=d}^{n} \binom{n}{j} (1 - \pi_1)^{n-j}
\]

\[
    P(X_{(d)} \geq x) = \sum_{i=d}^{n} \binom{n}{i} (1 - \pi_1)^{n-i} = \sum_{i=d}^{n} \binom{n}{i} (1 - \pi_1)^{n-i}
\]

\[
    P(X_{(d)} < x) = F_{X_{(d-1)}}(x) = \sum_{j=d}^{n} \binom{n}{j} (1 - \pi_3)^{n-j} = \sum_{j=d}^{n} \binom{n}{j} (1 - \pi_3)^{n-j}
\]

\[
    P(X_{(d)} \geq x) = \sum_{i=d}^{n} \binom{n}{i} (1 - \pi_3)^{n-i} = \sum_{i=d}^{n} \binom{n}{i} (1 - \pi_3)^{n-i}
\]

\[
    P[X_{(d)} \leq x \leq X_{(b)}] = \sum_{i=d}^{b} \binom{n}{i} \pi_3^{n-i} - \sum_{i=d}^{b} \binom{n}{i} (1 - \pi_1)^{n-i}
\]

Suppose 15 observations are randomly sampled from a categorical distribution with five classes, where the values of $X$ range from 1 to 5 and $X \sim \text{Mult}(1; 0.1, 0.2, 0.3, 0.25, 0.15)$. What is the probability the population median ($m$) is contained within the interval formed by the seventh and ninth order statistics? That is, find $P(X_{(7)} < m < X_{(9)})$. Since the population median is defined by $P(X \leq x) \geq 0.5$, or alternatively $P(X \geq x) \geq 0.5$, in this example, it is clearly three. Hence, $\pi_1 = P(X < 3) = 0.1 + 0.2 = 0.3$, $\pi_2 = P(X = 3) = 0.3$, and $\pi_3 = P(X > 3) = 0.25 + 0.15 = 0.4$. Employing the multinomial and binomial approaches then yields the same result; namely,

\[
    P[X_{(7)} \leq m \leq X_{(9)}] = \sum_{i=0}^{15} \binom{15}{i} \pi_3^{i} \pi_1^{15-i} = 0.3^{15-j} \cdot 0.4^{j} \approx 0.8897
\]
\[ P\left[ X_{(7)} \leq m \leq X_{(9)} \right] = \sum_{i=0}^{9} \binom{15}{i} 0.4^i (1-0.4)^{15-i} - \sum_{i=0}^{7} \binom{15}{i} (1-0.3)^i 0.3^{15-i} \approx 0.8897. \]

Notice, although both methods yield the same result, the binomial method is considerably simpler to compute. Lastly, as a means of further explicating and verifying these results, a Monte Carlo simulation was conducted. Fifteen random values were generated from the categorical distribution defined above, sorted, and the interval defined by the seventh and ninth order statistics was tested to determine whether it contained the population median (i.e., \( m=3 \)). This process was repeated a million times and the proportion of times \( m \) fell within the interval was found to be approximately 0.8894. In other words, the population median was found in approximately 89% of the sample intervals defined by the seventh and ninth order statistics. Alternatively, the coverage probability for \( m \in (X_{(7)}, X_{(9)}) \) is 89%.

**Quantiles.** Quantiles represent cumulative probabilities taken at constant intervals from the cdf of a random variable. That is, if one divided \( F_X(x) \) into \( q \) mutually exclusive intervals of identical size, the quantiles would mark the boundaries between each interval. Therefore, the \( k^{th} q \)-quantile is the value of \( x \) where \( F_X(x) = P(X \leq x) = k/q \). For example, if \( q=100 \) (percentiles) and \( k=25 \), then the 25\( ^\text{th} \) percentile of random variable \( X \) is the supremum value of \( x \) for which \( F_X(x) \geq 0.25 \) (i.e., the minimum value of \( x \) for which this inequality hold true). Naturally, a large number of quantiles may be defined by varying the magnitude of \( q \). In general, however, it is customary to report five summary statistics, all of which may be expressed by order statistics: the minimum value, first quartile, median (second quartile), third quartile, and the maximum value. Specifically, the minimum value, denoted by the 1\( ^{st} \) order statistic \( X_{(1)} \), is defined as \( \min(X) \); the first quartile represents the smallest value of \( x \) for which \( P(X \leq x) \geq 0.25 \) and is denoted by order statistic \( X_{(0.25(n+1))} \); the median is defined by order statistic \( X_{((n+1)/2)} \) for odd sample sizes and the arithmetic mean of
the middle order statistics, \((X_{(n/2)} + X_{(n/2+1)})/2\), for even sample sizes; the third quartile represents the smallest value of \(x\) for which \(P(X \leq x) \geq 0.75\) and is denoted by order statistic \(X_{q(0.75(n+1))}\); and lastly, the maximum value, denoted by the \(n^{th}\) order statistic \(X_{(n)}\), is defined as \(\text{max}(X)\). This subsection will focus on the median and estimates of its variance since it is a robust estimate of location that may be substituted for the mean.

Since confidence intervals are necessary for computing the variance of the median\(^{77}\), a brief introduction is warranted. A parametric two-sided CI is constructed by using \(\theta \pm T^*\sigma_{\bar{X}}\), where \(T\) represents the approximating distribution and \(\sigma_{\bar{X}}\) represents the standard error, \(\sigma/\sqrt{n}\). Although CIs for the median are constructed using distribution-free methods, variance is still a useful summary statistic for medians and can be particularly important for conducting statistical inference on it. The CI of a median is constructed using quantiles (order statistics) rather than the standard error. Without getting into the specifics of how such an interval is constructed, which will be addressed in a later section, one can use this information to derive the variance and standard error associated with the interval. Suppose the \((1-\alpha)\%\) CI placed around sample estimate \(\theta\) is defined by the interval \((\text{LB}, \text{UB})\), where \(\text{LB}\) and \(\text{UB}\) denote the lower and upper boundaries, respectively, and alpha \(\alpha\) denotes the Type I error rate one is willing to accept. Furthermore, if one assumes the approximating distribution is close to normal,\(^{78}\) then \(T=Z_{1-\alpha/2}\) where \(Z\) represents the inverse standard normal cdf (a.k.a., the critical value or \(Z\)-score) for the probability \(1-\alpha/2\). For example, the critical value for 0.975, which represents \(\alpha=0.05\), is approximately 1.96. Therefore, one can

---

\(^{77}\) Actually, several methods exist including ones based on order statistics (Arnold, Balakrishnan, & Nagaraja, 2008; Evans, Leemis, & Drew, 2006) and bootstrapping (Huang, 1991), to name a few. However, the method presented herein is among the simplest to compute and has been shown to perform very well in Monte Carlo simulations (Price & Bonett, 2001; McKean & Schrader, 1984).

\(^{78}\) Asymptotic theory dictates for large \(n\), the approximating distribution will be close to normal. Furthermore, McKean and Schrader (1984) found no advantage in using the Student's \(t\)-distribution in a Monte Carlo study. Additionally, they noted “the normal has appropriately correct tail weight even for small \(n\), yielding tests which vary between slightly liberal to slightly conservative depending upon the distribution sampled” (p. 757).
say with 95% confidence that the population parameter $\Theta$ can be found within the interval $[\theta-1.96*\sigma_X, \theta+1.96*\sigma_X]$, if an experiment is repeated ad infinitum. Based on the definition of a CI, it is easy to prove that $LB=\theta-\frac{z_{1-\alpha/2}}{2}\sigma_X$ and $UB=\theta+\frac{z_{1-\alpha/2}}{2}\sigma_X$, which connotes that the error variance is $\sigma^2_X=\frac{[(UB-LB)/2z_{1-\alpha/2}]^2}{2z_{1-\alpha/2}^2\sigma_X}$ and the variance is $\sigma^2=\frac{\left[(UB-LB)/2z_{1-\alpha/2}\right]^2}{2z_{1-\alpha/2}^2}$.

$$UB-LB=\theta+\frac{z_{1-\alpha/2}}{2}\sigma_X-\left[\theta-\frac{z_{1-\alpha/2}}{2}\sigma_X\right]=2\frac{z_{1-\alpha/2}}{2z_{1-\alpha/2}^2}\sigma_X$$

$$\sigma_X=\frac{UB-LB}{2z_{1-\alpha/2}} \Rightarrow \sigma_X^2=\frac{(UB-LB)^2}{2z_{1-\alpha/2}^2}$$

Since $\sigma^2_X=\sigma^2/n$ this yields

$$\sigma^2_X=\frac{\sigma^2}{n}=\frac{(UB-LB)^2}{2z_{1-\alpha/2}^2} \Rightarrow \sigma^2=n\left[\frac{UB-LB}{2z_{1-\alpha/2}}\right]^2$$

as was to be proven.

Consider the results for the top 21 finishers, reported in Table 9, of the 2008 Beijing Olympic Men’s Marathon (Runner’s World, 2008). The actual time (in seconds) each runner needed to complete the marathon is recorded in the second column $X$ while the first column $f(X)$ denotes their rank order based on their time. Furthermore, assume that $X$ is measured by a stopwatch (ratio scale) and that $f(X)$ is measured by an ordinal scale, since the process of computing ranks transforms the data from a higher level of measurement to a lower level of measurement. This can be easily verified by observing that the distance in $X$ between the first two ranks is not equal to the distance between the second and third place finishers. Assuming the data in Table 9 represented sample data, what is the sample median of random variable $X$? Since the sample size is odd, the median is the eleventh order statistics, which is equal to 7997. Now, suppose the 95% CI of the median is formed by the 6th and 16th order statistics (7835,8040). What is the approximate sample variance and error variance? Substituting these figures into the aforementioned formulas yields $\sigma^2=n\left[\frac{(UB-LB)/2z_{1-\alpha/2}}{2z_{1-\alpha/2}}\right]^2=21\left[\frac{(8040-7835)/2*1.96}{2z_{1-\alpha/2}}\right]^2\approx57.432$ and $\sigma^2_X=\left[\frac{(UB-LB)/2z_{1-\alpha/2}}{2z_{1-\alpha/2}}\right]^2=\left[(8040-7835)/2*1.96\right]^2$.
≈2735. As a quick check, if one constructs a 95% CI on the median using the standard error, 
\[ \sigma_{\bar{X}} = \sqrt{2735} \approx 52.3, \]
the CI should be approximately equal to the nonparametric 95% CI:
\[ \theta \pm T^* \sigma_{\bar{X}} = (\theta - 1.96 \cdot \sigma_{\bar{X}}, \theta + 1.96 \cdot \sigma_{\bar{X}}) = (7997 - 1.96 \cdot 52.3, 7997 + 1.96 \cdot 52.3) = (7894.5, 8099.5). \]
In this example, however, a noticeable discrepancy between the two CIs can be detected. This discrepancy is due to the critical assumption employed in the construction of the latter CI that was not present for the former; namely, the latter CI assumed the boundary estimates were symmetric about \( \theta \). This assumption is not always reasonable, particularly when the distribution underlying the random variable is not symmetric, as was the case here.

Table 9

2008 Beijing Olympics: Men’s Marathon Results (Top 21 Finishers)

<table>
<thead>
<tr>
<th>Ranks</th>
<th>X (Time (sec))</th>
<th>Reality</th>
<th>Transformed Observations</th>
<th>Likert</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7592</td>
<td>7592</td>
<td>Very fast</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>7636</td>
<td>7636</td>
<td>Very fast</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>7800</td>
<td>7800</td>
<td>Very fast</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>7821</td>
<td>7821</td>
<td>Fast</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>7824</td>
<td>7824</td>
<td>Fast</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>7835</td>
<td>7835</td>
<td>Fast</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>7852</td>
<td>7852</td>
<td>Fast</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>7871</td>
<td>7871</td>
<td>Fast</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>7919</td>
<td>7919</td>
<td>Average</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>7953</td>
<td>7953</td>
<td>Average</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>7997</td>
<td>7997</td>
<td>Average</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>8005</td>
<td>8005</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>8006</td>
<td>8006</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>8013</td>
<td>8013</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>8019</td>
<td>8019</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>8040</td>
<td>8040</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>8062</td>
<td>8062</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>8077</td>
<td>8077</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>8084</td>
<td>8084</td>
<td>Slow</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>8157</td>
<td>8157</td>
<td>Very slow</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>8167</td>
<td>8167</td>
<td>Very slow</td>
<td>5</td>
</tr>
</tbody>
</table>
Now, suppose the values of X were unknown and instead a judge recorded their impression regarding the finish time of the 21 runners he or she observed. The fact that the judge observed the race connotes that their impressions are at least partially grounded in reality. However, the rules by which the judge formed their impressions may not be known to them or be a linear transformation of X. Hence, the labels recorded in column \( g(X) \) will, in all likelihood, be ordinal. In turn, these labels can be transformed into numeric values via a simple order-preserving transformation. What is the sample median of random variable X? Once again, the median is the eleventh order statistic, which now is equal to 3. Suppose the 95% CI of the median is formed by the 9th and 15th order statistics (3,4). What is the approximate sample variance and error variance? This is a more problematical question to address than its previous analog. If one is willing to assume the equidistance assumption is met, then one can proceed as usual. That is, 

\[
\sigma^2 = n\frac{(UB-LB)^2}{2(21-1/2)^2} = 21\cdot\frac{(4-3)/2\cdot1.96^2}{21}\approx1.37 \quad \text{and} \quad \sigma^2_X = \frac{(UB-LB)^2}{2\cdot(21-1/2)^2} = \frac{(4-3)/2\cdot1.96^2}{21}\approx0.065.
\]

This approach has the drawback that for moderate sample sizes or larger, there exists the possibility that the lower and upper bounds will be identical resulting in a variance of zero. This result is rather unappealing considering that, unless all the responses are identical, variability in fact does exist.

Alternatively, this author proposes that the variance estimate should be computed from the numerical values associated with the rank orders (e.g., 1,2,…,n) of the CI boundaries rather than on the associated order statistic. Therefore, if the method for identifying the CI boundaries (to be discussed in a later section) does not violate the mathematical operations permissible for ordinal variables then the variance computed from these boundaries should be approximately equal to the actual variance of the ordinal variable, i.e., \( \text{Var}[g(X)] \) and \( \text{Var}\{h[g(X)]\} \). By this statement, however, it is important to note that no reference was made to the variance of the untransformed variable X, \( \text{Var}(X) \). In all
likelihood, the variance of the ordinal variable is different than the variance of $X$, $\text{Var}(X) \neq \text{Var}[g(X)]$. Moreover, unless a monotone transformation was applied to $X$, one cannot estimate the variance of $X$ even if the method of transformation $g$ were known.

Suppose $R$ denotes the rank order function. Then, the approximate variance and error variance of the rank orders is given by $\sigma_R^2 = n^* \{[R(\text{UB})−R(\text{LB})]/2 \zeta_{1−α/2}\}^2$ and $\sigma_{R(X)}^2 = [R(\text{UB})−R(\text{LB})]/2\zeta_{1−α/2}^2$, respectively. This, in turn, may be combined with order statistics to express the CI of the median. For odd sample sizes, the CI of the median of a random variable $X$ is approximately equal to $\left(X_{(n+1)/2−\lceil \zeta_{1−α/2}n^*\sigma_{R(X)}^2 \rceil}, X_{(n+1)/2+\lceil \zeta_{1−α/2}n^*\sigma_{R(X)}^2 \rceil}\right)$ whereas for even sample sizes it is approximately equal to $\left(X_{n/2−\lceil \zeta_{1−α/2}n^*\sigma_{R(X)}^2 \rceil}, X_{n/2+\lceil \zeta_{1−α/2}n^*\sigma_{R(X)}^2 \rceil+1}\right)$.

Note, the approach described above ensures the variance estimate will never equal zero, unless the lower and upper bounds have identical rank orders. Finally, this method for constructing CIs works equally well for continuous and ordinal data, providing the CI is symmetric.

Returning to the continuous data example, the variance and error variance of the ranks underlying the CI can be computed as follows: $21*[\{(16−6)/(2*1.96)\}^2 \approx 136.7$ and $[(16−6)/(2*1.96)]^2 \approx 6.51$, respectively. A quick check of these figures reveals the rank orders of the original CI, $(6,16)=\{11\pm\lfloor1.96*\sqrt{6.51}\rfloor\}$, which when converted into order statistics returns the CI $(X_{(6)}, X_{(16)})=\{(7835,8040)\}$. For the ordinal data example, the variance and error variance of the rank orders of the CI then are $21*[\{(15−9)/(2*1.96)\}^2 \approx 49.2$ and $[(15−9)/(2*1.96)]^2 \approx 2.34$, respectively. Therefore, the CI of the median is given by the rank orders $11−\lfloor1.96*\sqrt{2.34}\rfloor$ and $11+\lfloor1.96*\sqrt{2.34}\rfloor+1$, which when converted into order statistics returns the CI $(X_{(9)}, X_{(15)})=\{(3,4)\}$. Notice then that since the variance estimate is computed from the CI boundaries when employed to construct a CI it will always reproduce the

---

79 The braces $\lceil \rceil$ denote the lower bound integer of the value contained within (e.g., $\lceil 3.3 \rceil = 3$) whereas the braces $\lfloor \rfloor$ denote the upper bound integer of the value contained within (e.g., $\lfloor 3.3 \rfloor = 4$), which will be employed later on.
original CI, when the $\zeta$-scores are identical. As a result, if the method used to identify the CI boundaries for ordinal variables does not violate the permissible mathematical operations associated with such data, then the variance estimate derived from these boundaries must be representative of the actual variance estimate for the given CI interval.

To clarify, the purpose of the previous examples was not to promote the use of this method over existing methods for constructing a CI but rather to illustrate how a variance estimate can be obtained while conforming to the limitations of ordinal data. The contribution of this method to the construction of nonparametric CIs will be revealed in the penultimate section of this chapter. In the meantime, the question one must address is, what are the most appropriate statistics to replace the expected value and variance of a composite variable since the latter statistics cannot be computed without distributional assumptions and an interval/ratio level of measurement? Clearly, the mean can be replaced by the median because the latter statistic is derived solely based on order and therefore, is a permissible mathematical operation, whereas the former statistic is not, unless one can demonstrate that equidistance exists between scale points (interval or ratio). Furthermore, for composite variables, the grand mean can be replaced by the median of medians or simply, the grand median, which is also a permissible operation. Unfortunately, the median does not share some of the nice properties of the mean. For example, suppose one is asked to compute the grand median of the following $3 \times 4$ data matrix comprised of 3 respondents by 4 variables: 1 1 3 4, 1 2 3 4, 1 2 2 4 (commas are used to separate respondents). Then, the column medians across respondents are 1, 2, 3, and 4, which connotes that the grand median is 2.5. Alternatively, the raw medians across variables are 2, 2.5, and 2, which connotes that the grand median is 2. Yet another method of attaining the grand median is to compute it on all the data, which yields a grand median of 2. In general, the three approaches will yield nearly
the same result. Of the three methods, this author favors the last approach, particularly since the need to average two ordinal values (in the case of an even number of values) can never occur more than once thereby, limiting the number of mathematical operations not permitted. However, situations may arise (e.g., missing values for some variables) in which the use of one of the other methods may be preferable. Fortunately, the degree of disagreement is small, especially when the data come from a moderate to large number of variables and sample size.

The real challenge, of course, lies in deriving an alternative measure for the variance of a composite. Unfortunately, the properties of variance discussed in the first section do not hold for medians. That is, if one calculated the variance for each constituent variable and then applied Property 24, it would not equal the error variance of the grand median. However, since a nonparametric method exists for constructing CIs on the median, this same method can be used to construct a CI on the grand median. Thus, utilizing grand medians and nonparametric CIs, one can analyze ordinal data without violating the mathematical operations to which such data are limited.

Returning to an example from the previous subsection, suppose the tenure review committee chose to create a composite from 10 Likert items that appeared on the student evaluation form, with responses labeled ‘1’ denote ‘strongly disagree’, ‘2’ denote disagree, ‘3’ denote ‘agree’, and ‘4’ denote ‘strongly agree’. (Quotations surrounding the numbers are added to emphasize the ordinal nature of the measurement scale.) Based on the simulated data provided in Table 10, what is the sample grand median and the order statistics of the variable-level medians (i.e., last row) that will yield a coverage probability of at least 95%?
Table 10

Ten Simulated Variables Measured on a Likert Scale Based on Mult$(25; 0.20, 0.25, 0.30, 0.25)$

<table>
<thead>
<tr>
<th>Order</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
<th>$X_8$</th>
<th>$X_9$</th>
<th>$X_{10}$</th>
<th>Composite Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: Random variables $X_1$ to $X_{10}$ denote 10 orthogonal ordinal items. The responses are sorted for the reader’s convenience. Sorting is independent for each column. The shaded region represents a nonparametric 95% CI.

The median response for each item, $X_1$-$X_{10}$, is straightforward and can be produced by all statistical software packages. Manually, one only needs to sort each variable in ascending order and then locate the appropriate order statistic, in this example $X_{(13)}$. The grand median here is 3 (‘agree’) regardless of which method is used to compute it. Symmetrical 95% CIs are provided for each variable. Without going into further detail, they are constructed by finding the boundaries with a 95% coverage probability. For now, it is
only important to note, variability exists in the width of the 10 CIs and only eight of the CIs include the population median \((3)\), i.e., the supremum of the inequality \(F_X(x)\geq 0.5\). The second part of the question asks one to find the order statistics that yield a coverage probability of 95%. As noted earlier, the coverage probability of a median \(m\) is defined by \(P(X_{(a)} < m < X_{(b)})\). Hence, the question can be restated as, what points \((a,b)\) produce a coverage probability of 95%? Since a direct method for calculating \((a,b)\) does not exist, one has to employ a search algorithm to examine the inequality \(P(X_{(a)} < m < X_{(b)})\geq 0.95\). The natural starting point is to set \((a,b)\) equal to the order statistic of the sample median, which if one examines the 10 column medians in this example is given by the fifth and sixth order statistics. Since \(\pi_1 = P(X<3) = 0.2 + 0.25 = 0.45\), \(\pi_2 = P(X=3) = 0.3\), and \(\pi_3 = P(X>3) = 0.25\), the coverage probability for the median is

\[
P[X_{(5)} \leq m \leq X_{(6)}] = \sum_{i=0}^{5} \binom{10}{i} 0.25^i (1-0.25)^{10-i} - \sum_{i=0}^{5} \binom{10}{i} (1-0.25)^i 0.25^{10-i} \approx 0.7187.
\]

Seeing as \(P(X_{(5)} < m < X_{(6)})\leq 0.95\), one must increase \(a\) and \(b\) by a unit and repeat the analysis,

\[
P[X_{(4)} \leq m \leq X_{(7)}] = \sum_{i=0}^{4} \binom{10}{i} 0.25^i (1-0.25)^{10-i} - \sum_{i=0}^{4} \binom{10}{i} (1-0.25)^i 0.25^{10-i} \approx 0.8945.
\]

Once again, the coverage probability is less than 95% and so the process must be repeated,

\[
P[X_{(3)} \leq m \leq X_{(8)}] = \sum_{i=0}^{3} \binom{10}{i} 0.25^i (1-0.25)^{10-i} - \sum_{i=0}^{3} \binom{10}{i} (1-0.25)^i 0.25^{10-i} \approx 0.9722.
\]

Thus, the inequality \(P(X_{(3)} < m < X_{(8)})\geq 0.95\) was satisfied and the population median is found in the interval formed by the third and eighth order statistics of the column medians.

The accuracy of the coverage probabilities reported above can easily be tested via a MC simulation. Twenty-five random observations were generated from the multinomial distribution \(\text{Mult}(25; 0.20, 0.25, 0.30, 0.25)\), sorted within each of the 25 rows, and tested to determine whether the population median \((m=3)\) was contained in the intervals \((X_{(5)}, X_{(6)})\),
The proportion of rows in which the median was found in the interval was then computed. Lastly, this process was repeated a million times and the average proportion of times \( m \) fell within these intervals was computed. The results of these analyses yielded coverage probabilities of approximately 0.7186, 0.8944, and 0.9722 for the previous intervals. Hence, one can be particularly confident that the population median falls within the range defined by the third and eighth order statistics, i.e., \((2, 3)\) or \((\text{disagree, agree})\). That is, the 25 students who completed the student evaluation form for the three courses taught by the candidate had mixed reviews regarding the candidate’s teaching ability.

As a concluding remark to this subsection, the reader’s attention is drawn to the fact that at no point in this analysis was there a need to add or subtract any of the data. Consequently, all the operations performed conformed to the limitations of ordinal data and as such are interpretable. Furthermore, should the need arise to perform statistical inference on the composite, one only need to calculate the standard error of the composite from the order statistics that define its CI. Although three different approaches may be taken to construct the CI, utilizing all the data in one step will result in the tightest CI. For the data above, the width of the CI on the rank orders of the 10 column medians is 5 (\(=8-3\)), which represents 50% of the maximum possible width (\(=5/10\)). Had the CI been constructed on the rank orders of 25 row medians, its width would be 6, representing 24% of the maximum possible width, whereas had it been constructed on the rank orders of the 250 observations, its width would be 3, representing 1.2% of the maximum possible width. In other words, the more data that is used to construct the CI, the tighter its width.

**Other Transformations**

It is important to note that numerous other transformations exist and can conceivably be employed in an evaluation. Of these, one deserves brief mention. One could
employ rank transformations on the original random variables, rescale the new variables by dividing by $n$ (creating a uniform distribution), and then employ the probability integral transformation (Larson, 1974; Bartoszynski & Niewiadomska-Bugaj, 2008) to transform the new variables into another set of random variables that would conform to any desired distribution. This is, in fact, the method used by computer programs to simulate other distributions based upon the continuous uniform distribution. However, since this method is restricted to variables with a continuous strictly increasing cdf, it is unlikely to find much use in the field of evaluation where many of the variables collected are discrete.

**Sampling Theory**

To this point, an unstated assumption has been presupposed in the previous sections. Namely, the data utilized in analyses is sampled from an infinite or extremely large population. A population consists of all the individuals, clusters (e.g., households, schools), areas, and so on, that comprise a group whose members share a defining trait or characteristic. Censuses of large populations, however, are rarely implemented because they are time consuming, expensive, unfeasible when destruction of population elements is required (e.g., eating all the cookies produced by a cookie manufacturer to evaluate their quality), and may even increase error when poorly trained personnel are used in data collection. Consequently, researchers and evaluators often conduct analyses on a sample (subset of units) taken from a target population with the goal of drawing inferences and conclusions about the population based on the results of the sample. Two different sampling approaches may be employed. The first method relies on nonprobabilistic methods, such as haphazard or convenience, quota, and purposeful sampling (Sapsford, 2004; Larsen, 2007).

A haphazard or convenience sample is one in which data are gathered on sampling units (i.e.,

---

80 Herein, the members of a population shall be generically referred to as units or elements.
the members that comprise the target population) based upon no predefined set of rules and drawn, in part or in whole, at the convenience of the researcher or evaluator (e.g., Psychology 100 students, volunteers). In quota sampling, the population is classified into separate subpopulations with the number of units to be sampled determined beforehand based on prespecified targets. However, the decision over the specific sampling units selected is left up to the data collector, who may or may not be given instructions on how to select the units. Purposeful sampling entails the selection of sampling units based upon preestablished selection criteria. For example, key stakeholders or experts may be asked to participate in an evaluation due to their knowledge of the evaluand or expertise in a particular area of interest. Another type of purposeful sampling, known as quota sampling, occurs when sampling units are selected by researchers until a predesignated quota (level) on a desired characteristic is attained (e.g., continuing to recruit participants until there are exactly 50 females and 75 males in the study). In contrast, if members of the target population are asked to recruit other members, this is known as snowball sampling.

Regardless of the nonprobabilistic method employed, these designs share one trait in common. Namely, one cannot determine the selection probability for any of the sampling units. Therefore, the potential for selection bias (omission of certain population elements from the list of potential sampling units) is a serious concern and so, information produced by such designs must be treated with caution. In fact, history is replete with examples in which nonprobability samples, including large samples, failed to generalize to the population. Two famous historical examples include the 1936 and 1948 Presidential elections (Freedman, 2004; Lohr, 1999).

---

81 Selection bias can occur, for example, when an unknown selection method depends upon characteristics associated with the primary outcome, targets only certain groups of sampling units, misspecifies the target population, undercovers the target population, substitutes convenient sampling units for unavailable designated units, results in a high nonresponse rate, and consists of volunteers (Lohr, 1999).
The second method, probability sampling, relies on randomly selecting sampling units from a population based on a known, nonzero probability (Lohr, 1999). Specifically, a list of distinct sampling units (the sampling frame) is compiled from all the elements in the target population and units are then randomly selected from the list. Since the number of units comprising the sampling frame can be counted and each sampling unit is selected at random then the probability of being selected is the same for all the units in the sampling frame. This method is commonly referred to as a simple random sample (SRS).

The impact of the degree to which one may generalize from a nonprobability sample to the target population cannot be mathematically determined without invoking further assumptions (e.g., the sample is representative of the target population). As a result, these sampling methods will be omitted from further discussion. Moreover, although several probabilistic sampling methods exist, this section will be limited to investigating the impact of SRS designs on Summative Confidence since it is the most fundamental and frequently employed probabilistic sampling method. The restriction to a probabilistic sampling design is not intended to suggest problems may not arise in such designs. Issues such as selection bias, response bias, nonresponse, and measurement error may affect probability as well as nonprobability designs (Freedman, 2004). However, probabilistic sampling allows one to use probability laws to estimate how well a sample result may generalize to the target population. Thus, this section will focus on the degree to which the sample mean and total of a composite variable can generalize to the population mean and total of a composite variable. As was the case in previous sections, much of the material presented herein is common knowledge within the discipline of mathematical statistics and thus, will be presented without citation. Proof of the concepts and formulas discussed herein can be found in Kish
Implications for Summative Confidence

Finite population correction. It is well known that the finite population correction\(^{82}\) (fpc), \((1-n/N)\), can be used to adjust the variance of a random variable when sample data is used to estimate population parameters (Lohr, 1999; Kish, 1995). Therefore, as the sampling fraction \(n/N\) tends towards zero (i.e., \(n \ll N\)), the sample estimator converges onto the population estimator. Conversely, as \(n \to N\), the sample estimator converges onto zero. This result should make intuitive sense. If the sample is comprised of exactly the same elements as the population (i.e., \(n=N\)), then no variability exists between the sample and population estimators (e.g., the sample mean equals the population mean). Although the fpc can be ignored when the sampling fraction is small\(^{83}\), the peril of ignoring the fpc in other instances can be readily demonstrated via a Monte Carlo simulation. It stands to reason then that the fpc can be used to adjust Property 23 when one employs sample data to estimate the variance of the population mean.

Accounting for sampling error. Before discussing the results of these analyses, it is important to directly consider the implications of the fpc on the properties of variance presented in the first section. The fpc should be integrated into Property 23 whenever variance estimates are computed from sample data rather than population data. Hence, the variance of the mean of a single random variable computed from a sample is equal to the product of the fpc and the sample variance divided by the sample size. That is,

\[^{82}\text{Some authors (Anderson, Sweeney, & Williams, 1994) use } (N-n)/(N-1) \text{ as the fpc, which is obtained by multiplying } (1-n/N) \text{ by } N/(N-1). \text{ However, this constant is only appropriate when } S_N(Y) = N \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \text{ is used to estimate the population variance } \sigma^2.\]

\[^{83}\text{The fpc is typically ignored when the sampling fraction is less than 5% or when the population size is assumed to be infinite.}\]
Var(\bar{X})=(1-n/N)S_x^2/n. However, since Property 23 was used to derive Property 24, the fpc should also be used to adjust the error variance of a set of random variables whose variance estimates are based on sample data. Therefore, whenever sample data is used, the appropriate variance property for \( k \) independent variables with sample size \( n_i \) is

\[
\text{Var}(\bar{X}) = \frac{1}{k^2} \sum_{i=1}^{k} \left[ \frac{n_i}{N_i} - 1 \right] S_x^2 \left( 1 - \frac{n_i}{N_i} \right) / n_i
\]

while the appropriate variance property for \( k \) dependent variables with sample size \( n_i \) is

\[
\text{Var}(\bar{X}) = \frac{1}{k^2} \sum_{i=1}^{k} \left[ \frac{n_i}{N_i} - 1 \right] S_x^2 \left( 1 - \frac{n_i}{N_i} \right) / n_i + 2 \sum_{i<j} w_i w_j \sqrt{\frac{1}{1 - \frac{n_i}{N_i}} \frac{1}{1 - \frac{n_j}{N_j}} S_{x_i x_j}}
\]

where \( n_i = n_j \) and both samples are drawn from the same population. Notice, the fpc varies as a function of the sample since evaluators utilize information from multiple data streams to formulate evaluative conclusions. In instances in which samples are drawn from the same population then the fpc can be factored out in front of the equation.

**Monte Carlo simulations.** The purpose of the remainder of this section is to illustrate the impact of the fpc on variance. To this end, a random variable \( X \) with population size \( N=10,000 \) was generated, where \( X \sim N(100, 25) \). One thousand samples of size \( n=5,000 \) were repeatedly drawn from the population using SRS. Hence, the sampling fraction was 50%. Comparison of the sample mean (99.89) to the population value (\( \mu=99.89 \)) indicated that the sample estimator was virtually unbiased, as expected. However, differences arose in the variance and standard error based on whether the fpc was used to adjust the estimators. When no adjustment was made, the sample variance and standard error were 24.99 and 0.07 (\( =\sqrt{24.99/5000} \)), respectively. While the former estimate compares favorably with the
population variance (24.99), the latter estimate is noticeably larger than the population standard error (0.05). In contrast, the fpc adjusted variance and standard error estimates were 12.5 ($\approx [1-5000/10000]*24.99$) and 0.05 ($\approx \sqrt{[1-5000/10000]*24.99/5000}$), respectively. However, these results merely illustrate the difference between the variances produced by either ignoring or employing the fpc. Since both estimators are unbiased, what is needed is a method for judging the impact of the fpc.

One way of evaluating the impact of the fpc is to examine the coverage probability of a CI constructed from each of the standard errors and compare it to the nominal coverage probability. The coverage probability is the proportion of sample CIs that contain the population value, if the study was replicated a large number of times under parallel conditions. The nominal coverage probability refers to the desired probability for a CI. That is to say, the nominal coverage probability is set by the researchers, whereas the coverage probability is determined by the data. Hence, the objective is to construct a CI whose coverage probability is close to the nominal coverage probability. Utilizing this evaluative criterion then one would expect to observe a CI constructed based on the unadjusted sample standard error to result in a conservative coverage probability (i.e., the width of the CI would be too large resulting on a larger than desired coverage probability). In contrast, a CI based on the adjusted standard error should be relatively close to the nominal coverage probability.

To test these expectations, a CI was constructed for each of the 1,000 sample means and a new variable was created to record whether each of the CIs contained the population mean. The coverage probability was then computed based on the proportion of CIs that contained the population mean to the total number of simulations conducted. Hence, for a 95% CI, the coverage probability was expected to be approximately equal to the nominal

84 Note, the population figures are close to the desired parameters of $\mu=100$ and $\sigma_{X}=\sqrt{25/10,000}=0.05$. 
probability (i.e., 95%). This analysis revealed that the unadjusted variance and standard error resulted in a coverage probability of 99.6%, whereas the adjusted variance and standard error resulted in a coverage probability of 95.2% thereby, confirming both expectations.

Another criterion for determining the better estimator is efficiency, which answers the question: how spread out about $\Theta$ is the sampling distribution of the estimator $\theta$? The smaller the variance of the sampling distribution, the more “efficient” is the estimator. Since the only difference between the numerator and denominator of this ratio is the fpc, then the ratio of the unadjusted to adjusted error variance is $(1-n/N)^{-1}$. Similarly, one can also compare the ratio of the square root of the variance of a sampling distribution (i.e., the standard error) to determine the relative size of one CI to the other. Hence, this ratio is equal to the $100(fpc^{-1/2}-1)\%$, when converted to a percentage. This analysis indicated that the unadjusted CI was more than 40% larger than the adjusted CI for a sampling fraction of 50%. In other words, the failure to adjust the variance by the fpc produced a conservative CI. In statistical literature, however, the fpc adjustment is only discussed in the context of single variables, so how does the fpc affect the variance of composite variables? According to the fpc adjusted variance property, the only change is the addition of the multiplicative fpc constant to the error variance of each variable. Moreover, when the sample is drawn from the same population, then one only needs to adjust Property 24 by the fpc.

To test this expectation, two simulations were conducted. In the first simulation, four independent random variables were generated, which conformed to a normal distribution $N(100,25)$, from the same population of size $N=100,000$. One thousand samples of size $n=20,000$ were repeatedly drawn from the population using SRS. Hence, the sampling fraction was 20%. As expected, the observed population means, variances, and

---

85 The ratio of variance estimates of the sampling distributions is known as asymptotic relative efficiency.
correlations were nearly identical to the target population values. Therefore, a composite variable was created by averaging the four variables. As predicted by Property 22, the variance of the composite variable was equal to the sum of the variances divided by the square of the number of variables summed (i.e., \([25+25+25+25]/4^2=6.25\)), while, in accordance with Property 24, the standard error was equal to the square root of the variance of the composite divided by the population size (i.e., \(\sqrt{6.25/100000}\approx0.0079\)). As before, the unadjusted variance was nearly identical to the population variance, producing a standard error of 0.0177 (\(\approx\sqrt{6.25/20000}\)). In contrast, the fpc adjusted variance was 5.02, which is nearly equal to the predicted value (6.25*(1−20000/100000=5.0). As a result, the unadjusted standard error was 12% larger than its fpc adjusted counterpart (0.0158), which also conformed with expectation. Lastly, the coverage probability for the unadjusted CI was 97.4%, whereas for the adjusted CI it was 95.3% thereby, confirming the stated expectation.

A second simulation was conducted in which the independence criterion was relaxed and correlation matrix \(\Sigma\) was allowed to exist for the four random variables:

\[
\Sigma = \begin{bmatrix}
1.0 & 0.1 & 0.2 & 0.3 \\
0.1 & 1.0 & 0.4 & 0.5 \\
0.2 & 0.4 & 1.0 & 0.6 \\
0.3 & 0.5 & 0.6 & 1.0
\end{bmatrix}
\]

Due to the large population size, discrepancies between the observed population correlations and the corresponding target correlations were extremely small, \(|R−\rho|<0.005\). Therefore, in accordance with Property 24, the error variance of the composite variable computed from the average of the four dependent variables was

\[
\text{Var}(\bar{X}) = \frac{1}{k} \left[ \sum_{i=1}^{k} S_i^2 + 2 \sum_{i<j} R_{ij} SS_{ij} \right] = \frac{1}{4^2 \times 100,000} \left[ \sum_{i=1}^{4} 25 + 2(25) \right] = 0.0001281,
\]

since \(N=N=100,000\) and \(SS_j=S^2=25\) for all the population elements and so, they can be
factored out. The standard error of the composite was approximately 0.01132 \(=\sqrt{\text{Var}(\bar{X})}\).

Similar to the previous analysis, one thousand samples of size \(n=20,000\) were repeatedly drawn from the population to generate a sampling distribution. The unadjusted composite variance was approximately 12.87 \(\{\text{Var}(\bar{X})=4^{-2}[25+25+25+25+2(25)(.1+.2+.3+.4+.5+.6)]\}\) in accordance Property 22, whereas the fpc adjusted composite variance was approximately 10.296 \(\{(1-\frac{20000}{10000})*\text{Var}(\bar{X})\}\). As a result, the unadjusted standard error (0.02537) was once again nearly 12% larger than its fpc adjusted counterpart (0.02269). Not surprisingly, the coverage probability for the unadjusted CI was conservative (97.6%) whereas the coverage probability for the fpc adjusted CI (95.2%) was nearly identical to the nominal coverage probability of 95%. These examples clearly illustrate the danger of ignoring the fpc when one needs to construct a CI and the sampling fraction is large (traditionally defined by \(n/N>0.05\)). However, this does not tell us how to construct a CI. For this, we now turn to the penultimate section of this chapter.

Confidence Intervals

Considerable effort has been exhausted in the previous sections to illustrate how point estimators may be derived for certain statistical properties. If the estimator is unbiased and has a small variance, then, on average, it will produce an estimate whose value should be close to the unknown population value. However, since a sample is a subset of units drawn from a population, sample statistics are only estimates, educated guesses if you will, of their population analogs. Therefore, differences between sample statistics and their corresponding population values may occur due to sampling error. Without an additional step, one cannot express how confident one should be in an estimate given a particular sample. Confidence intervals were introduced by Jerzy Neyman (1937) to provide researchers with a method for estimating and representing the statistical uncertainty associated with using sample data to
estimate population parameters. The following section will summarize the theory underlying confidence intervals and thus, will form the basis upon which Summative Confidence rests. As was the case with many of the previous sections, the theory underlying CIs is well known in the mathematical-statistics literature. Hence, with one notable exception—the derivation of an ordinal CI, the subject will be treated as common knowledge and presented without citations. Interested readers are directed to Meyer (1970), Larson (1973; 1974), Casella and Berger (2002), Smithson (2003), and Hogg, McKean, and Craig (2005) for further reading. Note, SAS code is provided for select Monte Carlo analyses in Appendix H.

Constructing a Basic Confidence Interval

*Basic theory.* Anytime a sample estimator \( \theta \) is used, one can ask, how well did it estimate the population estimator \( \Theta \)? Since evaluators are primarily interested in knowing the population mean, this question can be further refined to ask, how much sampling error is involved when the value of \( \bar{x} \) is used to estimate the population mean \( \mu \)? The most direct measure of sampling error is \( |\bar{x} - \mu| \). However, since \( \mu \) is generally unknown, an alternative method is needed to infer the amount of sampling error, which, in turn, can be used to estimate \( \mu \). Fortunately, the CLT provides the means for solving this impasse. Recall that the number of distinct possible samples increases geometrically for samples drawn, with or without replacement, from even modest-sized populations. By extension, the number of possible sample means that can be observed also increases at a geometric rate allowing one to generate a sampling distribution of these means. As a result, the normal distribution is the limiting probability distribution of the sampling distribution. In other words, as the number of iid random variables (e.g., the sample means) that are aggregated into a composite, either by summation or averaging, increases without bound, the pdf of the sampling distribution of
the composites will approach the normal distribution as illustrated previously in Figures 6 and 19. Based on Property 23, the variance of the sampling distribution is \( \sigma^2_{\bar{X}} = \sigma^2/n \). Therefore, the sampling distribution conforms to \( N(\mu, \sigma^2/n) \) and its linear transformation, \( Z = [(\bar{X} - \mu)/(\sigma/\sqrt{n})] = n(\bar{X} - \mu)/\sigma \), conforms to a standard normal distribution, \( N(0,1) \). This fact will play an important role shortly, but first, a CI must be formally defined.

A confidence interval is composed of three elements: a margin of error, the confidence level associated with this margin, and in the case of parametric CIs, a distributional assumption for the sampling distribution of the point estimate. The margin of error is an interval defined by a lower (L) and upper (U) confidence limit, whereas the confidence level represents the proportion of such intervals that would contain the unknown population parameter (\( \Theta \)) if the population was repeatedly sampled. Mathematically, given a Type I error rate (\( \alpha \)), a CI is defined by the probability, \( P(\theta_L \leq \Theta \leq \theta_U) \geq 1 - \alpha \), that the unknown population parameter is contained within the confidence interval (\( \theta_L, \theta_U \)) with confidence level \( 100(1 - \alpha)\% \) for all possible values of \( \Theta \). Notice, however, the lower (\( \theta_L \)) and upper (\( \theta_U \)) bounds are defined by estimators and not fixed values, a fact that plays an important role in the interval’s interpretation. The lower confidence limit is the value expected to fall below the \( \theta \)'s true value \( 100(\alpha/2)\% \) of the time, whereas the upper confidence limit is the value expected to exceed the \( \theta \)'s true value \( 100(1-\alpha/2)\% \) of the time.

Since the sampling distribution is approximately normal, then the \( 100(1-\alpha)\% \) CI of the standardized value for the sample mean is bound by the interval \( (-z_{1-\alpha/2}, z_{1-\alpha/2}) \), which, in turn, connotes that the probability

---

86 It is worth noting that the convergence rate is governed by how closely the original distribution resembles the normal distribution. If the sample size and number of iid variables aggregated is small, the sampling distribution may deviate from normality. Thus, one will need to use another method, such as a distribution-free method, to construct a CI whose coverage probability is nearly equal to the nominal coverage probability.

87 Generally, a CI is written as \( P(L \leq \Theta \leq U) \geq 1 - \alpha \). However, this notation tends to obfuscate the fact that the confidence boundaries are defined by estimators and not estimates.
since \(-z_{\alpha/2} = z_{1-\alpha/2}\) due to the fact the normal distribution is symmetric. Recall, the term \(\sigma/\sqrt{n}\), known as the standard error of the mean and denoted by either \(\sigma_{X}\) is the square root of Property 23. Therefore a 100\((1-\alpha)\)% CI for \(\mu\) is frequently written as \(\bar{X} \pm z_{1-\alpha/2} \sigma/\sqrt{n}\) = \(\bar{X} \pm z_{1-\alpha/2} \sigma_{X}\). Notice, since the interval is a function of the sample mean, which is a random variable, the interval is also a random interval. Moreover, the CI for the unknown parameter \(\mu\) has a distribution free of unknown parameters. Since the standard error is a fixed, yet potentially unknown quantity, the critical value \(z_{1-\alpha/2}\), or simply \(z\), is the only factor under the full control of the researcher. Hence, one can set it to a value that produces a probability of \(1-\alpha\). In the case of \(\alpha=0.05\) (95% confidence level), for example, \(z=\pm 1.96\). More generally, \(z\) is equal to the inverse standard normal cdf for the probability \(1-\alpha/2\), \(z_{1-\alpha/2} = \Phi_{X}^{-1}(1-\alpha/2)\).

The challenge in using this method to construct a simple CI is that in most cases the value of the population standard deviation is unknown. Fortunately, the sample variance is an unbiased estimator of the population variance, and thus, one can substitute the sample standard deviation \(S\) for the population standard deviation \(\sigma\). However, what is the sampling distribution of the statistic \((\bar{X}-\mu)/(S/\sqrt{n})\)? The derivation of this distribution may be found in many mathematical-statistics textbooks (Hogg, McKean, & Craig, 2005; Larson, 1974). According to the derivation, the sampling distribution of the aforementioned statistic conforms to a \(t\) distribution. Hence, a 100\((1-\alpha)\)% CI of the standardized value for the sample mean then is bound by the interval \((-t_{1-\alpha/2,\sigma-1}, t_{1-\alpha/2,\sigma-1})\), which connotes that

\[
1-\alpha = P\left[t_{\alpha/2,\sigma-1} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{1-\alpha/2,\sigma-1}\right] = P\left[\bar{X} - t_{1-\alpha/2,\sigma-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\alpha/2,\sigma-1} \frac{S}{\sqrt{n}}\right],
\]

since \(-t_{\alpha/2,\sigma-1} = t_{1-\alpha/2,\sigma-1}\) because the \(t\)-distribution is symmetric. The term \(S/\sqrt{n}\) is the sample
standard error of the mean and is typically denoted by SE. The 100(1−α)% CI can then be expressed as \( \overline{X} \pm t_{1−α/2, n−1}SE \). However, since the \( t \)-distribution rapidly converges upon the standard normal distribution as \( n \to \infty \), statisticians often recommend using \( \overline{X} \pm \frac{z_{1−α/2}}{\sqrt{n}}SE \) to construct a 100(1−α)% CI when \( n > 30 \).

Given the large combination of samples that can be drawn from a population, computing exact confidence intervals based upon the population variance is not always possible or feasible. Consequently, one is forced to rely on asymptotic confidence intervals. That is, researchers typically compute approximate CIs whose accuracy improves as the sample size gets larger. If the coverage probability exceeds 100(1−α)% then the interval is considered conservative, whereas if the coverage probability is less than 100(1−α)% then the interval is considered liberal. In general, conservative intervals are preferred to liberal intervals. As was illustrated in the previous section, sampling error influences the sample standard deviation estimator, so it is important to incorporate the fpc whenever the sampling fraction is large to avoid constructing an overly conservative interval. Hence, when \( \sigma \) is known, the 100(1−α)% CI is equal to \( \overline{X} \pm \frac{z_{1−α/2}}{\sqrt{n}} \sqrt{\frac{n}{N}} \), whereas when \( \sigma \) is unknown, the 100(1−α)% CI is equal to \( \overline{X} \pm t_{1−α/2, n−1}SE \sqrt{\frac{n}{N}} \).

As may be evident from these definitions, the width of the CI is governed by four factors: the desired level of confidence, the amount of variance, the sample size, and the population size. By convention, the confidence level is generally set at either 95% (\( \alpha = 0.05 \)) or 99% (\( \alpha = 0.01 \)). Therefore, researchers generally avoid relaxing this factor (i.e., allowing \( \alpha > 0.05 \)). Likewise, although researchers can exert some influence over the variance (e.g., choice of measurement scales, reduction of measurement error), this control is not absolute. Control is even more limited when it comes to the population size. Once a target population
is identified, its size is fixed, albeit potentially unknown. Therefore, the only factor under direct control of researchers, to the degree to which cost and availability are not constraints, is sample size. Sample size impacts the width of the CI in two ways. First, as the sampling fraction increases \( (n \to N) \), the width of the CI approaches zero due to the shrinking fpc. Naturally, this makes sense because as the sample size approaches the population size, one would expect deviations between the sample and population estimators to decrease. Second, as \( n \) increases, the standard error decreases. Algebraic manipulation of the aforementioned CI definitions, allows one to solve for \( n \). Since the CI is symmetric about the sample mean, then the width \( w \) is equal to

\[
\frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} = \frac{2\zeta_{\alpha/2}}{n} \sqrt{1 - \frac{n}{N}} \Rightarrow \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}} = \frac{2\zeta_{\alpha/2}}{n} \sqrt{\frac{1}{N}} 
\]

\[
\frac{w^2}{4\zeta_{\alpha/2}^2\sigma^2} = \frac{1}{n} \frac{1}{N} \Rightarrow \frac{w^2}{4\zeta_{\alpha/2}^2\sigma^2} + 1 = \frac{1}{nN} \Rightarrow \frac{Nw^2 + 4\zeta_{\alpha/2}^2\sigma^2}{4\zeta_{\alpha/2}^2\sigma^2} = \frac{1}{n}
\]

\[
n = \frac{4N\zeta_{\alpha/2}^2\sigma^2}{Nw^2 + 4\zeta_{\alpha/2}^2\sigma^2} = \frac{4\zeta_{\alpha/2}^2\sigma^2}{w^2 + \frac{4\zeta_{\alpha/2}^2\sigma^2}{N}}.
\]

Similarly, when \( \sigma \) is unknown, then \( n = \frac{4\ell_1^2}{\sigma^2\ell_1} \cdot \frac{S^2}{(w^2 + 4\ell_1^2/(2z_{\alpha/2}^2 \cdot \sigma^2)/N)} \). As \( N \to \infty \), then \( n \approx 4\zeta_{\alpha/2}^2\sigma^2/w^2 \) and \( n \approx 4\zeta_{\alpha/2}^2\sigma^2/S^2/w^2 \). Thus, if the sampling fraction is small, one would need to quadruple the sample size to reduce the CI width by half.

**Distribution-dependent CI.** While the method for constructing a CI is straightforward, it does not specify the source of the variance estimator. For this, one must turn back to the second section in this chapter. Therein, variance was defined as a function of the observations of a random variable, which follows a specific probability distribution. That is, (parametric) variance estimators depend upon the identification of a specific distribution that can be shown to underlie the observed data. Hence, in order to construct a basic CI, one

---

88 The size of many, if not most, research studies is quite large, generally greatly exceeding the ability of the researcher to collect a significant fraction of the data. Hence, the fpc can usually be ignored. However, population sizes in evaluation contexts may be smaller, in which case the fpc should be employed.
must combine distribution theory with CI theory. However, the quality of a CI, with respect
to its coverage probability rests upon the degree to which the sampling distribution is
approximately normal. Certainly, the sampling distribution of a normal random variable is
normal, but how quickly does the sampling distribution of other random variables converge
to the normal? In instances in which the sampling distribution converges slowly, the CI may
be either too liberal or too conservative.

Suppose a random variable $X$ conforms to a discrete uniform distribution, $X \sim U(n)$. Since the variance of a $X$ is $(n^2-1)/12$, the standard error $\sigma_x$ and fpc adjusted $100(1-\alpha)\%$ CI of the sample mean $\bar{X}$ is $\sigma_x = \sqrt{\frac{(n^2-1)}{12n}}$ and $\bar{X} \pm t_{1-\alpha/2,n-1} \sqrt{\frac{(1-n/N)^*(n^2-1)}{12n}}$. Now, further suppose 40 candidates are randomly selected from a pool of 100 applicants, who are ranked 1 through 100 with no ties. How can one construct a CI on the sample mean rank of the 40 randomly selected candidates? A Monte Carlo simulation was conducted to explore this question, in which a million samples of size 40 were randomly selected from a fixed population of size 100. The average sample mean, and fpc adjusted variance and standard error across all the simulations were 48.40, 465.64, and 3.40, respectively.\(^9\) Then, the average fpc adjusted 95% CI is given by $48.4 \pm 2.02 \times 3.4 = (41.53, 55.27)$, where $t_{0.025,40} = 2.02$.

Coverage probability was defined as the proportion of CIs that contained the population mean out of the one million Monte Carlo simulations. Examination of this probability (81.9%) revealed that the CI underestimated the nominal coverage probability, which underscores the importance of taking into consideration the convergence rate of the sampling distribution to normality. It turns out, the sampling distribution of the mean of 40 uniform variables does not sufficiently approximate the normal distribution because uniform

\(^9\) These sample estimates are lower than their theoretical expectations of 50.5, 499.95, and 3.54, respectively, which suggests they are slightly biased, possibly due to the fact the sampling distribution of uniform random variables does not converge to the normal at a rapid rate.
random variables depart substantially from a normal distribution. In fact, one would need to aggregate over 1,000 random uniform variables to generate the classic bell-shaped curve of the normal. Hence, in this example, utilizing $t$-distributions resulted in a very liberal CI.

In contrast, one would expect Bernoulli, binomial, and Poisson random variables to converge more rapidly since most of their observations coalesce around a central region. Monte Carlo simulations were conducted to test this expectation by randomly drawing a million samples of size 40 from a population of size 100. Since the variance of a Bernoulli trial is $p(1-p)$, then $\sigma_{X}^2 = \sqrt{p(1-p)/n}$ and the fpc adjusted 100(1-$\alpha$)% CI on the sample proportion of $X$ is equal to $p \pm t_{1-\alpha/2, n-1} \sqrt{[p(1-p)/n]}$. A random variable $X \sim Bin(1, 0.1)$ was generated, where the population proportion and average sample proportion were 0.1 and 0.097, respectively. Hence, the fpc adjusted variance and standard error of $X$ are equal to $\sigma^2 = (1-40/100)*0.1*(1-0.1) = 0.054$ and $\sigma_{X} = \sqrt{0.054/40} = 0.0367$. Moreover, its fpc adjusted 95% CI is given by $0.097 \pm 2.02*0.0367 = (0.023, 0.171)$. Examination of the coverage probability for the million CIs revealed that 95.5% contained the population proportion. In and of itself, this result suggests that the coverage probability of the binomial should also be approximately equal to the nominal coverage since the binomial is the sum of a set of independent Bernoulli trials. The variance, standard error, and fpc adjusted 100(1-$\alpha$)% CI of a binomial random variable are given by $\sigma^2 = np(1-p)$, $\sigma_{X} = \sqrt{p(1-p)}$, and $np \pm t_{1-\alpha/2, n-1} \sqrt{[p(1-p)]}$. A binomial random variable $X \sim Bin(40, 0.1)$ was generated with a population and average sample number of successful trials both equal to 3.77. Consequently, the fpc adjusted variance, standard error, and 95% CI are equal to $\sigma^2 = (1-40/100)*40*0.1*(1-0.1) = 2.16$, $\sigma_{X} = \sqrt{2.16/40} = 0.232$, and $3.77 \pm 2.02*0.232 = (3.301, 4.239)$. Examination of the coverage probability revealed that 95.1% of the million CIs contained the population number of successful trials.
Lastly, the coverage probability of the Poisson distribution was examined. Since the variance of a Poisson process is \( \lambda \) (which is also equal to the expected success rate), then \( \sigma_x = \sqrt{\lambda / n} \) and the fpc adjusted 100\((1-\alpha)\)% CI is equal to \( \lambda \pm t_{1-\alpha/2, n-1} \sqrt{[(1-n/N)\lambda/n]} \). A Poisson random variable \( X \sim \text{Poi}(10) \) was generated with a population and average sample success rate of 9.6 and 9.585, respectively. Therefore, the fpc adjusted variance, standard error, and 95% CI are equal to \( \sigma^2 = (1-40/100)\times9.585 = 5.751 \), \( \sigma_x = \sqrt{5.751/40} = 0.379 \), and 9.585\(\pm2.02\times0.379 = (8.819,10.35) \). Examination of the coverage probability revealed that 95.2% of the million CIs contained the population success rate. In summary, these results illustrate how well the sampling distribution is approximated by the normal distribution, providing the original distribution does not substantially deviate from the normal.

(Mis)Interpreting Confidence Intervals. Despite the occasional convergence issue, the principal limitation of CIs is interpretability. What one is really interested in knowing is the probability of whether a specific CI contains the population parameter. Unfortunately, despite the fact that CIs are very often misinterpreted in this fashion, they do not provide this information (Smithson, 2003). CIs can only be interpreted in the context of repeated samples, as was illustrated in the previous Monte Carlo simulations. The reason for this is simple. The probability that any given CI contains the population parameter is either zero or one (i.e., a Bernoulli trial)—either the parameter is present in the CI or it is not. Thus, only within the framework of repeated sampling under identical conditions is a CI interpretable and even then the interpretation is less direct than one would like. If one cannot state that the population parameter falls within a specific CI, one would like to say that if repeated samples were taken from the population, the population parameter would fall in the interval \((L, U)\) for 100\((1-\alpha)\)% of the samples. Unfortunately, even this interpretation is incorrect!
Essentially, the confidence level refers to the expected percentage of calculated CIs that would contain the population parameter value if the study was repeatedly replicated under identical conditions (i.e., the coverage probability). In other words, the population parameter is treated as a fixed point while the CI is a random interval because it depends upon the sample mean and variance, both of which are random variables. Hence, a 95 percent confidence level denotes that if a study was replicated \( k \) times under identical conditions (i.e., repeatedly sampling of the same population using the same instruments and procedures), it would produce \( k \) means and CIs of roughly the same size, 95 percent of which would contain the unknown population parameter. Figure 18 illustrates the results of a Monte Carlo in which 100 simple random samples \((n=200)\) were drawn from a larger population \((N=1000)\) and a 95% CI was constructed for each sample mean. Notice, four of the sample CIs failed to capture the population mean. Moreover, although the width of the CIs are roughly the same, some variation does exist.

![Figure 18. Distribution of Confidence Intervals About the Unknown Population Parameter](image_url)
The ramification of these results may be readily understood with respect to the notions of accuracy and precision. An accurate result is one that comes close to matching reality, whereas precision is a characteristic of an instrument and method. Specifically, precision is a measure of the amount of variability with which a parameter is estimated. For example, consider the game of darts in which a person is trying to hit the bullseye. Accuracy is the distance from where the dart lands on the board to the center of the bullseye. Precision, on the other hand, can be thought of as the average distance between each of the darts thrown without consideration of their relation to the bullseye. Darts that land very close to each other exhibit a high degree of precision on the part of the thrower (assuming luck was not a factor) but not necessarily accuracy, unless they landed close to the bullseye. Panel A in Figure 19 illustrates results that are both biased and imprecise (i.e., the worst of both worlds). Panel B illustrates results that are unbiased (the centroid of the darts hits the bullseye) but imprecise. In contrast, Panel C illustrates results that are precise (small degree of variability) but biased (the centroid of the darts is far from the bullseye). Lastly, Panel D illustrates the desired goal of all researchers; namely, results that are unbiased and precise.

\[ \text{Bias}[\theta] = E(\theta) - \Theta \]

Figure 19. Illustration of the Notions of Biasedness and Precision

---

90 The estimation bias of an estimator \( \theta \) is the difference between the expected value of the estimator and the parameter value \( \Theta \), \( \text{Bias}[\theta] = E(\theta) - \Theta \). If, and only if, this difference is zero is the estimator considered unbiased. That is, if one were to repeat the sampling experiment ad infinitum, the estimator would be unbiased if the mean of the sampling distribution equaled the population value. In the panels in Figure 19, bias is equal to the difference between the average estimate, denoted by the dart, and the parameter, denoted by the bullseye. In real-world examples, one should expect estimators to contain some degree of bias because calculating exact estimators based on the complete sampling distribution is unrealistic.
Recall from the previous section that the sample mean and variance $S^2_Y$ are unbiased estimators of their corresponding population parameters. Hence, the natural tendency is to assume the results of a study that employs these estimators must conform to either Panels B or D. While in the long run, this interpretation is correct, it does not necessarily hold true for any one particular sample. As Figure 18 clearly illustrates, not all CIs capture the population parameter. Some CIs produce really high estimates (e.g., samples 26, 35, and 42), whereas other CIs produce really low estimates (e.g., sample 83). Therefore, the conclusion derived from a CI may not necessarily be accurate. However, because the width of sample CIs does not generally vary widely, a CI provides a good estimate of the precision of the conclusion, which, in turn, reflects the quality of the methodology and data used to formulate it. In other words, a CI is an index of the quality of the evaluation design and data collected. Hence, a small CI with a high confidence level indicates that the study can produce a precise estimate of the population parameter. However, even though the CI either contains the population parameter or it does not, because of the high probability of success, in the long run, one can be fairly confident that a random interval will be successful. For a more direct interpretation of the CI, one must employ Bayesian theory to construct a credible interval (Hogg, McKean, & Craig, 2005). However, this topic falls outside the scope of this study.

**Constructing a Distribution-Free CI on the Population Median**

Experienced evaluators have undoubtedly encountered variables whose data did not approximate one of the probability distributions covered in the second section or any other known distribution. Yet, evaluators are expected to arrive at valid evaluative conclusions regardless of the degree to which data may violate the distributional assumptions of standard analytical techniques. The previous techniques for constructing a CI assume that the sampling distribution does not deviate significantly from the normal. Fortunately, in many
instances, the sampling distribution can be shown to be approximately normal; hence, CIs are quite robust with respect to the non-normality of the underlying distribution. However, in the case of serious departures from normality by the distribution underlying the random variable, unknown distributions, or ordinal levels of measurement, the sampling distribution may not accurately approximate the normal resulting in a CI of questionable value (i.e., a CI whose coverage probability is substantially different than the nominal coverage probability). Hence, a need exists for nonparametric techniques for constructing CIs. This subsection will focus on how to construct a CI for the median since it is a robust measure of central tendency. The first method, partially based on one presented in Hogg, McKean, and Craig (2005), should be used when continuous data have an unknown distribution. The second method was developed by the present author and should be used for discrete distributions, particularly when the random variable is measured with an ordinal scale.

Continuous distribution. Nonparametric or distribution-free techniques were developed to reduce the potential for decision error associated with distributional violations. Many of these methods rely upon order statistics since these statistics are not required to be iid. In the third section it was shown that if \( X \) is a continuous random variable and \( X^{(k)} \) is the \( k \)th order statistic, then for \( X^{(k)} = x \) to be true, only one observation must have a value equal to \( x \) (since the probability of ties in a continuous distribution is zero), \( k-1 \) observations must have values less than \( x \), and \( n-k \) observations must have values greater than \( x \). Notice, no distributional assumptions were made for \( X \). These observations can be classified into three groups, each of which is associated with a specific probability: \( P(X=x) = f_X(x) \), \( P(X\leq x) = F_X(x) \), and \( P(X>x) = 1-F_X(x) \). Consequently, although the distribution of \( X \) may be unknown, the distribution of its order statistics \( X^{(k)} \) conforms to the trinomial distribution, 

\[
f_{X^{(k)}}(x) = P(X^{(k)} = x) = \binom{n}{k-1, 1, n-k} [F_X(x)]^{k-1} [1-F_X(x)]^{n-k} f_X(x).
\]
Integrating the pdf \( f_{X,p}(\theta) \) to obtain the cdf \( F_{X,p}(\theta) \) can be difficult, so many statisticians prefer to transform the pdf of \( X \) into a binomial distribution by dichotomizing it at \( \theta \). As a result, \( F_{X,p}(\theta) = P(X_{\theta}) \leq \theta \) if, and only if, \( k \) or more of the values of \( X \) are less than or equal to \( \theta \). That is, the probability the \( k^{th} \) order statistic is at most \( \theta \), \( F_{X,p}(\theta) \), can be computed from a binomial distribution, where the probability of success is defined by \( P(X \leq \theta) = F_X(\theta) \) and the corresponding cdf is given by \( F_{X,p}(\theta) = \sum_{i=0}^{n} \binom{n}{i} [F_X(\theta)]^i [1-F_X(\theta)]^{n-i} \).

Furthermore, the third section illustrated how the difference between two cdfs can be used to determine the probability of the interval defined by two fixed values of \( \theta \), i.e., \( P(a < X_{\theta} < b) = \sum_{i=0}^{k-1} \binom{n}{i} [F_X(a)]^i [1-F_X(a)]^{n-i} - \sum_{i=0}^{k-1} \binom{n}{i} [F_X(b)]^i [1-F_X(b)]^{n-i} \), where \( 1 \leq a < b \leq n \). This method, however, will not yield the correct probability for a CI since the boundary points change from sample to sample.

What one really wants to know is the probability \( P(X_{\theta} < \theta \leq X_{\theta}) \), where \( \theta \) represents the \( p^{th} \) distribution percentile (e.g., population median \( \theta_{0.5} \)). That is, what is the probability the population quantile \( \theta \) falls between two order statistics over repeated samples? Though, in general, the distribution of \( X \) is unknown, the sample percentile \( x_p \) can be used to estimate \( \theta_p \). Thus, it naturally follows that the quantiles of the sample order statistics can be estimated by \( F_X(x_p) = p \), where \( 0 < p < 1 \), provided the observations of \( X \) are iid and drawn from a strictly increasing distribution function (i.e., \( X \) is continuous). That is, given a large enough sample size, a value \( x \) exists such that it is nearly equal to the \( p \)-percentile. Furthermore, given the ordered sequence \( X_{(1)}, \ldots, X_{(b)}, \ldots, X_{(n)} \) then \( p \) can be estimated by the \( k^{th} \) order statistic \( X_{(k)} \), where \( k \in \{1, 2, \ldots, n\} \) such that \( p = k/(n+1) \). In the case of the median \( \theta_{0.5} \), \( k = 0.5(n+1) \), which connotes that \( p = 0.5(n+1)/(n+1) = 0.5 \), as one would expect. Similarly, given a large enough sample size, 25% of the observations of \( X \) will be lower than \( x_{0.25} \), \( F_X(x_{0.25}) = P(X \leq x_{0.25}) = 0.25 \), while 75% of the observation will be greater than \( x_{0.25} \), \( 1-F_X(x_{0.25}) = 1-P(X \leq x_{0.25}) = 0.75 \).
For the conditions \( \{X_{(a)} < x_p\} \) and \( \{x_p < X_{(b)}\} \) to hold, then at least \( a \) of the values of \( X \) must be less than \( x_p \) and fewer than \( b \) of the values of \( X \) must be less than \( x_p \), respectively. In other words, condition \( \{X_{(a)} < x_p\} \) only holds if any of the order statistics \( X_{(a)}, X_{(a+1)}, \ldots, X_{(b)} \) are less than \( x_p \) and condition \( \{x_p < X_{(b)}\} \) only holds if any of the order statistics \( X_{(1)}, X_{(2)}, \ldots, X_{(b-1)} \) are greater than \( x_p \). Hence, since \( a < b \), the two conditions can only be simultaneously satisfied for \( X_{(a)}, X_{(a+1)}, \ldots, X_{(b-1)} \). To put this in the context of a binomial distribution, where the probability of success is \( F_X(x_p) \), the condition \( \{X_{(a)} < \theta_p < X_{(b)}\} \) is equivalent to obtaining between \( a \) and \( b-1 \) successes in \( n \) independent binomial trials. Therefore,

\[
P(X_{(a)} < \theta_p < X_{(b)}) = \sum_{i=a}^{b-1} \binom{n}{i} [F_X(x_p)]^i [1-F_X(x_p)]^{n-i}.
\]

A Monte Carlo simulation was conducted to generate 500 observations from a transformed normal distribution, \( X \sim [N(10, 4)]^{2.5} \). Suppose, however, one had no knowledge of the pdf of \( X \). Graphical analysis would clearly show it to be asymmetrical and not conform to any of the aforementioned continuous distributions. Hence, the natural choice for measuring the central tendency of \( X \) would be the median since it is more robust to the impact of outliers. Even if the pdf were known, determining the population median is quite difficult for it requires finding the upper limit of the integral of the cdf that results in a probability of 0.5. Hence, the simplest estimate of \( \theta_{0.5} \) is the sample median, which in this simulation was 319.6, i.e., the average of \( X_{(250)} \) and \( X_{(251)} \). However, since the sample median is unlikely to be exactly equal to the population median, one can construct a CI around the sample median and then calculate the probability that \( \theta_{0.5} \) lies in the interval. For example,

---

91 Order statistics are ranked from low to high, so it stands to reason that if a higher order statistic is less than \( x_p \), then all lower order statistics must also be less than \( x_p \), and vice versa. Note, \( X_{(a)} \) is excluded from the order statistics that satisfy the condition \( \{x_p < X_{(a)}\} \) since fewer than \( b \) of the values of \( X \) can be less than \( x_p \).
what is the probability the population median $\theta_{0.5}$ lies in the interval $(X_{(240)}, X_{(260)})$, whose realized values are approximately (312.1, 323.7)? Since $F_X(x_{0.5})=0.5$ then $P(X_{(240)} < \theta_{0.5} < X_{(260)})$ = $\sum_{i=240}^{250} [F_X(x_{0.5})] [1-F_X(x_{0.5})]^{500-i} \approx 0.628$. That is, the probability the CI formed by the interval (312.1, 323.7) will capture the population median is 62.8%. Clearly expanding the interval will eventually result in a coverage probability approximately equal to 100(1-\alpha)%.

Fortunately, a direct method is available for obtaining these boundaries based on the formal definition of a CI, $P(\theta_L \leq \theta \leq \theta_U) \geq 1-\alpha$, where $\theta_L = X_{(a)}$, $\theta_U = X_{(b)}$, and $\theta \in \Theta$. Since over repeated samples the probability the population parameter falls within a given CI is at least $1-\alpha$ then the probability of exclusion must be $\alpha$. By extension, the probability the population parameter will be lower (higher) than $\theta_L$ ($\theta_U$) is equal to $\alpha/2$. Hence, the boundaries may be obtained by finding the inverse of the binomial cdf (i.e., the binomial quantile function). That is, if one sets $\theta_L$ equal to the smallest upper bound (supremum) of $a$ that satisfies $\sum_{i=0}^{a-1} [F_X(x_{ij})] [1-F_X(x_{ij})]^{n-i} \leq \alpha/2$ and $\theta_U$ equal to the greatest lower bound (infimum) of $b$ that satisfies $\sum_{i=b}^{n} [F_X(x_{ij})] [1-F_X(x_{ij})]^{n-i} \geq 1-\alpha/2$ then the probability $P(\theta_L \leq \theta \leq \theta_U) \geq 1-\alpha$, ensuring the coverage is slightly conservative. Solving these inequalities in the context of the previous example yields $a=228$ and $b=273$, for $\alpha=0.05$. Hence, a 95% CI for the sample median is given by the order statistics $(X_{(228)}, X_{(273)})$, whose realized values are about (304.7, 333.7). The coverage probability, $P(X_{(228)} < \theta_{0.5} < X_{(273)})$, for this interval was 95.6%, which is slightly more conservative than the nominal coverage.

Rather than locating each boundary independently, this method can be altered so that after the first boundary is located, a search algorithm is used to find the second boundary. For example, if the lower bound is located, one can incrementally increase the upper bound until the first upper bound satisfies $P(\theta_L \leq \theta \leq \theta_U) \geq 1-\alpha$. This method can also be implemented to find the lower bound given an upper bound by incrementally decreasing
the lower bound until the inequality is satisfied. Experience has shown that for $\theta$ less than or equal to the median, one should use the former method, whereas for $\theta$ greater than the median, one should use the latter method. This method for locating the boundaries of the CI will result in the tightest CI possible, when the population is assumed to be infinitely large.

For large sampling fractions, however, the CI tends to be conservative. To illustrate that this point is as true of CIs on the median as it is of CIs on the mean, three probability distributions were generated with a population size of 150. Sample sizes of 10, 30, 50, 70, and 90 were randomly drawn 100,000 times from each population. Coverage probabilities and average 95% CIs were then computed for five quantiles (10%, 25%, 50%, 75%, 90%), i.e., $\theta_{0.1}$, $\theta_{0.25}$, and so on. To explore the impact of skewness and kurtosis on coverage probability, two probability distributions were generated using the GLD based on the normal family. The first distribution was a normal $N(50,25)$, whereas the second distribution had a skewness of 2.6 (right skew) and kurtosis of 11.2, and the third distribution had a skewness of -1.7 (left skew) and kurtosis of 6.0. These deviations from the zero skew and kurtosis values for a normal also exceeded the values for typical evaluation variables. Thus, the distributions test the degree to which the method presented in this section can be used in standard evaluation and research studies. Table 11 presents the results of this analysis, which revealed that the larger the sampling fraction, the more conservative the coverage was for the average CI. Moreover, the coverage probability was only influenced by the quantile and sample size. That is, skewness and kurtosis did not have an adverse impact.

Clearly, the width of a CI shrinks as a function of sample size. However, since for sampling fractions greater than 20% the coverage probability exceeded 98%, sampling error must be incorporated into the method. Unfortunately, the solution is not as simple as multiplying the standard error by the square root of the finite population correction factor.
Table 11

Unadjusted Coverage Probabilities and Mean CIs for Three Probability Distributions

<table>
<thead>
<tr>
<th>n</th>
<th>10% Quantile Coverage (L,U)</th>
<th>25% Quantile Coverage (L,U)</th>
<th>50% Quantile Coverage (L,U)</th>
<th>75% Quantile Coverage (L,U)</th>
<th>90% Quantile Coverage (L,U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(50,10,0)</td>
<td>( \theta_{0.1} = 44.0 )</td>
<td>( \theta_{0.25} = 47.7 )</td>
<td>( \theta_{0.5} = 50.8 )</td>
<td>( \theta_{0.75} = 54.0 )</td>
<td>( \theta_{0.9} = 56.8 )</td>
</tr>
<tr>
<td>10</td>
<td>69.1% (42.4, 57.4)</td>
<td>94.8% (42.4, 57.4)</td>
<td>98.3% (45.7, 55.2)</td>
<td>94.4% (48.9, 57.4)</td>
<td>64.9% (51.4, 57.4)</td>
</tr>
<tr>
<td>30</td>
<td>98.0% (38.9, 48.0)</td>
<td>98.5% (43.6, 49.7)</td>
<td>97.7% (48.4, 53.2)</td>
<td>98.7% (51.9, 56.8)</td>
<td>96.6% (54.0, 59.5)</td>
</tr>
<tr>
<td>50</td>
<td>99.2% (37.3, 46.5)</td>
<td>98.7% (44.4, 49.0)</td>
<td>98.5% (48.8, 52.6)</td>
<td>98.8% (52.6, 55.7)</td>
<td>99.8% (54.4, 60.4)</td>
</tr>
<tr>
<td>70</td>
<td>99.6% (41.2, 46.3)</td>
<td>99.6% (45.4, 49.0)</td>
<td>99.5% (49.1, 52.5)</td>
<td>99.5% (52.7, 55.2)</td>
<td>99.4% (54.7, 58.2)</td>
</tr>
<tr>
<td>90</td>
<td>100% (41.5, 46.1)</td>
<td>99.9% (45.8, 48.8)</td>
<td>99.9% (49.3, 52.3)</td>
<td>99.8% (52.9, 55.0)</td>
<td>99.9% (54.8, 58.1)</td>
</tr>
<tr>
<td>N(50,25,1.12)</td>
<td>( \theta_{0.1} = 45.1 )</td>
<td>( \theta_{0.25} = 46.9 )</td>
<td>( \theta_{0.5} = 48.9 )</td>
<td>( \theta_{0.75} = 52.4 )</td>
<td>( \theta_{0.9} = 55.2 )</td>
</tr>
<tr>
<td>10</td>
<td>67.7% (44.2, 62.2)</td>
<td>95.3% (44.2, 62.2)</td>
<td>98.2% (45.9, 55.2)</td>
<td>95.2% (47.9, 62.2)</td>
<td>69.1% (49.8, 62.2)</td>
</tr>
<tr>
<td>30</td>
<td>97.8% (42.5, 47.3)</td>
<td>98.7% (44.7, 48.4)</td>
<td>97.7% (47.6, 51.3)</td>
<td>98.5% (49.9, 58.3)</td>
<td>97.4% (52.3, 70.5)</td>
</tr>
<tr>
<td>50</td>
<td>99.3% (41.7, 46.1)</td>
<td>98.8% (45.4, 48.1)</td>
<td>98.4% (47.9, 50.5)</td>
<td>98.8% (50.5, 54.9)</td>
<td>99.8% (52.9, 74.9)</td>
</tr>
<tr>
<td>70</td>
<td>99.6% (43.5, 46.0)</td>
<td>99.7% (45.6, 48.1)</td>
<td>99.5% (48.2, 50.4)</td>
<td>99.6% (50.6, 54.0)</td>
<td>99.6% (53.4, 64.0)</td>
</tr>
<tr>
<td>90</td>
<td>100% (43.6, 45.8)</td>
<td>99.9% (45.7, 48.0)</td>
<td>99.9% (48.3, 50.2)</td>
<td>99.9% (50.9, 53.7)</td>
<td>99.9% (53.5, 63.5)</td>
</tr>
<tr>
<td>N(50,25,1.76)</td>
<td>( \theta_{0.1} = 43.9 )</td>
<td>( \theta_{0.25} = 48.1 )</td>
<td>( \theta_{0.5} = 50.8 )</td>
<td>( \theta_{0.75} = 53.4 )</td>
<td>( \theta_{0.9} = 54.6 )</td>
</tr>
<tr>
<td>10</td>
<td>67.7% (40.5, 56.4)</td>
<td>95.3% (40.5, 56.4)</td>
<td>98.2% (45.4, 54.3)</td>
<td>95.2% (49.4, 56.4)</td>
<td>69.1% (51.4, 56.4)</td>
</tr>
<tr>
<td>30</td>
<td>97.8% (34.5, 48.6)</td>
<td>98.7% (42.6, 50.3)</td>
<td>97.7% (49.1, 52.7)</td>
<td>98.5% (51.7, 55.5)</td>
<td>97.4% (53.3, 58.6)</td>
</tr>
<tr>
<td>50</td>
<td>99.3% (31.1, 46.4)</td>
<td>98.8% (44.7, 49.9)</td>
<td>98.4% (49.7, 52.2)</td>
<td>98.8% (52.2, 54.4)</td>
<td>99.8% (53.6, 59.6)</td>
</tr>
<tr>
<td>70</td>
<td>99.6% (38.5, 46.1)</td>
<td>99.7% (45.3, 49.9)</td>
<td>99.5% (50.0, 52.1)</td>
<td>99.6% (52.3, 54.1)</td>
<td>99.6% (53.8, 57.2)</td>
</tr>
<tr>
<td>90</td>
<td>100% (38.8, 45.8)</td>
<td>99.9% (45.6, 49.7)</td>
<td>99.9% (50.2, 52.0)</td>
<td>99.9% (52.5, 54.0)</td>
<td>99.9% (53.9, 57.0)</td>
</tr>
</tbody>
</table>

After all, the CI on the median is produced from order statistics and not the standard error. Moreover, the variance, and by extension the standard error, of order statistics is very complicated to compute. Fortunately, an alternative method exists for computing the approximate variance of a nonparametric CI based solely on knowledge of the CI boundaries. As was proven in the third section, the standard error used to construct a CI can be back-calculated to be \( \sigma_X = (UB - LB) / 2z \). However, this estimate is only appropriate for constructing symmetric CIs. For asymmetric CI, one must compute two standard errors: one for each half of the CI split along the quantile. Therefore, let \( \sigma_{X_L} \) represent the standard error associated with the left side of the CI, \( (L - k) / z \), \( \sigma_{X_R} \) represent the standard error associated with the right side of the CI, \( (U - k) / z \), where \( L \) and \( U \) represent the rank associated with the realized values of the CI boundaries (i.e., LB and UB, respectively), and \( k \) represent the quantile of interest. Then, one can adjust the standard error based CI by the fpc as follows.
\[1 - \alpha = P[\theta_L \leq \theta \leq \theta_U] = P[k - z\sigma_{\bar{X}_L} \leq \theta - k \leq k + z\sigma_{\bar{X}_R}] = P[-z\sigma_{\bar{X}_L} \leq \theta - k \leq z\sigma_{\bar{X}_R}]
\]

Dividing by \(z\), substituting \((L - k)/z\) for \(\sigma_{\bar{X}_L}\) and \((U - k)/z\) for \(\sigma_{\bar{X}_R}\), and eliminating \(z\) yields

\[P\left[-\frac{\theta - k}{z} \leq \frac{\bar{X}_L - k}{z} \leq \frac{U - k}{z}ight] = P(k - L \leq \theta - k \leq U - k)\]

Although the terms \(L, U, k\) are discrete uniform random variables, if one were to standardize the probability by multiplying by the reciprocal of the standard deviation, one could then apply a normal approximation. However, the normal approximation of a discrete distribution requires the use of the continuity correction 0.5. The purpose of this explanation is not to standardize the probability, but to explicate the use of the continuity correction. Hence, incorporating the continuity correction to correct \(L\) and \(U\) then yields

\[P(k - L - 0.5 \leq \theta - k \leq U + 0.5 - k) = P(-L - k + 0.5 \leq \theta - k \leq U + 0.5 - k)\]

Note, the outer terms denote the standard deviation used to construct the CI, so one can adjust them by the square root of the \(fpc\), as was done for the CIs on the mean. This yields

\[P\left[-(L - k + 0.5)\sqrt{1 - \frac{n}{N}} \leq \theta - k \leq (U - k + 0.5)\sqrt{1 - \frac{n}{N}}\right]
\]

which concludes the derivation. Therefore, the new ranks of the CI boundaries are given by \(L^* = k - [(L - k + 0.5)\sqrt{fpc}]\) and \(U^* = k + [(U - k + 0.5)\sqrt{fpc}]\), and the realized values of the CI are given by \((X_{(L^*)}, X_{(U^*)})\). Table 12 illustrates the impact of this adjustment on the previous data.

Comparison of this table to the previous one illustrates the dramatic improvement in the coverage probability. Previously, over 73% of the coverage probabilities exceeded 98% as compared to less than 15% for the \(fpc\) adjusted method. Skewness and kurtosis continue to play no role in affecting coverage. Furthermore, one will note that the coverages are
reasonably good for sample sizes 30 and higher. This makes sense considering that for small sample sizes the normal approximation breaks down. Additionally, one may notice the coverage probabilities are better (i.e., closer to the nominal coverage probability of 95%) the closer one is to the median, which also makes sense given the probabilities are based on the binomial distribution, whose normal approximation improves as \( np > 5 \) and \( n(1-p) > 5 \). Sample size also impacts the estimated quantiles of the sample order statistics because for small sample sizes, one cannot locate a \( x_p \) whose cdf will result in exactly \( p \) (i.e., \( F_X(x_p) \neq p \)). Hence, a decrease occurs in the coverage probability, which also manifests if too many values are tied. However, the latter issue can be fixed by adding a small, say less than 0.001, random number to each value; thereby, eliminating ties without sacrificing interpretability.

Table 12
FPC Adjusted Coverage Probabilities and Mean CIs for Three Probability Distributions

<table>
<thead>
<tr>
<th>( n )</th>
<th>10% Quantile Coverage (L,U)</th>
<th>25% Quantile Coverage (L,U)</th>
<th>50% Quantile Coverage (L,U)</th>
<th>75% Quantile Coverage (L,U)</th>
<th>90% Quantile Coverage (L,U)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n=50,25,0,0 )</td>
<td>( n=50,25,2,6,11,2 )</td>
<td>( n=50,25,-1,7,6,0 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>69.1% (42.4, 57.7)</td>
<td>94.8% (42.4, 57.4)</td>
<td>98.3% (45.7, 55.2)</td>
<td>94.4% (48.9, 57.4)</td>
<td>64.9% (31.4, 57.4)</td>
</tr>
<tr>
<td>30</td>
<td>97.7% (38.9, 47.5)</td>
<td>94.7% (44.6, 49.3)</td>
<td>93.6% (48.8, 52.8)</td>
<td>97.2% (52.3, 56.8)</td>
<td>96.1% (54.3, 59.5)</td>
</tr>
<tr>
<td>50</td>
<td>96.1% (40.4, 46.0)</td>
<td>96.0% (45.4, 48.6)</td>
<td>96.2% (49.0, 52.3)</td>
<td>96.4% (52.8, 55.3)</td>
<td>98.1% (54.6, 58.5)</td>
</tr>
<tr>
<td>70</td>
<td>97.9% (42.2, 45.6)</td>
<td>96.7% (46.3, 48.6)</td>
<td>96.9% (49.5, 52.1)</td>
<td>95.5% (53.3, 54.7)</td>
<td>96.3% (55.2, 57.8)</td>
</tr>
<tr>
<td>90</td>
<td>98.3% (42.7, 45.4)</td>
<td>97.1% (46.7, 48.3)</td>
<td>93.8% (49.8, 51.6)</td>
<td>95.8% (53.5, 54.5)</td>
<td>96.6% (55.5, 57.6)</td>
</tr>
</tbody>
</table>

Discrete distribution. Although computing the CI on the median of an unknown continuous distribution provides evaluators with considerable flexibility, many of the
probability distributions of interest to evaluators are discrete due to the need to transform
dissimilar distributions into a common distribution. Because the previous method required
\( F_X(x_i) = p \) to produce an accurate CI, it cannot be used for discrete data since the cdf of
discrete random variables is a step function rather than a smooth increasing function. Hence,
a one-to-one function does not exist between quantiles and the probability function.
Fortunately, a new method can be derived based upon discrete order statistics. The principal
departure of this method from the previous one is the way in which tied values are handled.
The observations of a discrete random variable \( X \) can be classified into one of three mutually
distinct groups, \( X_{(k)} < x \), \( X_{(k)} = x \), and \( X_{(k)} > x \), with associated probabilities of \( \pi_1 = P(X < x) \),
\( \pi_2 = P(X = x) \), and \( \pi_3 = P(X > x) \). Although these population probabilities may not be known,
sample estimates can be substituted à la the Law of Large Numbers. That is, for large sample
sizes, the sample probabilities provide adequate estimates of their population counterparts.

For event \( \{ X_{(k)} < x \} \) to be true, then \( i \) observations of \( X \) must have values less than \( x \); for event \( \{ X_{(k)} > x \} \) to be true, \( j \) observations of \( X \) must have values greater than \( x \); and for event \( \{ X_{(k)} = x \} \) to be true, \( n-i-j \) observations of \( X \) must equal \( x \). Hence, although the
distribution of \( X \) may be unknown, the distribution of its order statistics \( X_{(k)} \) conforms to
the trinomial distribution, \( f_{X_{(k)}}(x) = P(X_{(k)} = x) = \binom{n}{i,j} \pi_1^i \pi_2^j \pi_3^{n-i-j} \). The cdf of the \( k^{th} \) order statistic
follows from a multinomial cdf with three classes and a maximum limit of the inner sum of
\( n-k \), i.e., \( F_{X_{(k)}}(x) = \sum_{j=0}^{n-k} \sum_{i=0}^{n-j-k} \binom{n}{i,j,k} \pi_1^i \pi_2^j \pi_3^{n-i-j-k} \). In the context of a binomial distribution, if \( k = a \),
then \( F_{X_{(a)}}(x) = \sum_{i=0}^{a} \binom{n}{a} \pi_1^a \pi_2^{n-a} \pi_3^0 \pi_1^i \pi_2^{n-i} \pi_3^0 \). Since a successful event
\( \{ X_{(a)} \leq x \} \) occurs only if any of the order statistics \( X_{(b)}, \ldots, X_{(a)} \) are less than \( x \). Therefore, the
probability of success is \( \pi_2 + \pi_3 = 1 - \pi_1 \). Similarly, if \( k = b \), then \( F_{X_{(b)}}(x) = \sum_{i=0}^{b} \binom{n}{b} \pi_1^0 \pi_2^b \pi_3^{n-b} \pi_1^i \pi_2^{n-i} \pi_3^0 \). Since a successful event \( \{ X_{(b)} \leq x \} \) occurs only if any of
the order statistics \( X_{(b)}, \ldots, X_{(a)} \) are less than \( x \), which has a probability of success of \( \pi_1 \).
Although a bit counterintuitive, the probabilities \( P(X(a) \leq x) \) and \( P(X(b) \leq x) \) are computed by summing the individual probabilities of the order statistics whose values are greater than \( k \), as depicted in Figure 20. Hence, the difference between \( P(X(a) \leq x) \) and \( P(X(b) \leq x) \) then yields

\[
P(X(a) \leq x \leq X(b)) = \sum_{i=a}^{n} (1-\pi_i)^i - \sum_{i=b}^{n} (1-\pi_i)^i = \sum_{i=0}^{b-1} (1-\pi_i)^i - \sum_{i=0}^{a-1} (1-\pi_i)^i.
\]

Figure 20. Illustration of \( P(X(a) \leq x \leq X(b)) \) for Discrete Order Statistics

While this method enables one to calculate the coverage probability of an interval, it does not identify the actual boundaries. As was the case before, what one really wants to know is the values of \( a \) and \( b \) that produce the smallest possible CI about the population value \( x \) with a coverage probability \( P(X(a) \leq x \leq X(b)) \geq 1-\alpha \). That is, one is interested in finding two sample order statistics that capture the population value \( x \) with \( 100(1-\alpha)\% \) confidence. If one were interested in the population median, for example, then one would construct the \( 100(1-\alpha)\% \) CI on the sample median \( X(k) \) since it is the best estimate of the population median. Recall, if the sample size is odd, then \( k=(n+1)/2 \); otherwise, if the sample size is even, the sample median is the average of two sample order statistics, \( X(n/2) \) and \( X(n/2+1) \). A number of different search algorithms can be used to identify the values of \( a \) and \( b \). The simplest method is to set \( a \) to the supremum of \( \sum_{i=0}^{b-1} (1-\pi_i)^i \leq \alpha/2 \) and \( b \) to the infimum of \( \sum_{i=0}^{a-1} (1-\pi_i)^i \geq 1-\alpha/2 \). This method ensures the probability \( P(\theta_L \leq x \leq \theta_U) \geq 1-\alpha \) is slightly conservative. However, it does not always yield the tightest possible CI because \( a \) and \( b \) are computed independently. A slightly better method is to locate \( a \) first, when \( \pi_1 \leq \pi_3 \), and then incrementally decrease the value of \( b \) from \( n \) until the probability \( P(\theta_L \leq x \leq \theta_U) \geq 1-\alpha \) is satisfied. Alternatively, when \( \pi_1 > \pi_3 \), one can locate \( b \) first and then incrementally increase the value of \( a \) from 1 until the probability \( P(\theta_L \leq x \leq \theta_U) \geq 1-\alpha \) is satisfied. A better method is to
set $u$ to the supremum of $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} \leq x$ and $v$ to the infimum of $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} \geq 1-\alpha$ and then find the infimum of the probability $F_{X_0}(x) - F_{X_0}(x) \geq 1-\alpha$ produced by all the pairwise possibilities that result from $a \in (1, u)$ and $b \in (v, n)$. However, this method has a slight tendency to result in overly conservative CIs. To counteract this tendency, one can compare the theoretical coverage probability for $(X_{(a)}, X_{(b)})$ to those of $(X_{(a+1)}, X_{(b)})$, $(X_{(a)}, X_{(b-1)})$, and $(X_{(a+1)}, X_{(b-1)})$ and then select the CI whose coverage is closest to the nominal coverage. This method for locating the values of $a$ and $b$ will produce the absolute tightest possible CI.

Although intuitively one expects the lower bound order statistic to be less than or equal to the upper bound order statistic, this is not always the case, particularly when the sample size is large. For example, if $\alpha=0.05$, $\pi_1=0.35$, $\pi_2=0.3$, $\pi_3=0.35$, and $n=100$, then the supremum of $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} = \sum_{i=0}^{100}(1-0.35)^{0.35}0.35^{100-i} \leq 0.025 = \alpha/2$ is 56, whereas the infimum of $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} = \sum_{i=0}^{100}(1-0.35)^{100-i} \geq 1-\alpha$ is 45. That is, if $a=56$ then $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} \approx 0.0246 \leq 0.025$ and if $b=45$ then $\sum_{i=0}^{n-1}(1-\pi_i)^{\pi_i-\epsilon i} \approx 0.9754 \geq 0.975$. This connotes that the probability coverage is $97.54\% - 2.46\% = 95.08\%$, which is only slightly more conservative than the nominal probability coverage of 95%. Had one limited the lower bound order statistic $X_{(a)}$ from exceeding $X_{(50)}$ (lower bound median) and the upper bound order statistic $X_{(b)}$ from falling below $X_{(51)}$ (upper bound median) then the probability coverage would have been overly conservative ($\approx 99.85\%$). From a substantive standpoint, the CIs are interpreted in the same fashion as regular CIs. That is, the smaller number should be used as the lower bound and the higher number should be used as the upper bound.

A Monte Carlo was conducted to simulate 100,000 random variables measured on a 5-point Likert scale, with class probabilities of $p_1=0.15$, $p_2=0.2$, $p_3=0.3$, $p_4=0.22$, and $p_5=0.13$. Hence, the population median was class three and $\pi_1=0.35$, $\pi_2=0.3$, and $\pi_3=0.35$. Table 13 presents the observed and theoretical coverage probabilities along with the mean CIs for
three alpha-levels ($\alpha=0.1$, $\alpha=0.05$, $\alpha=0.01$). Notice, the theoretical coverage probabilities were extremely close to the nominal coverage probabilities, whereas all but one of the observed coverage probabilities were conservative, albeit a bit too conservative. However, overly conservative CI is problematic for it connotes that a tighter CI could have been used.

Table 13

Coverage Probabilities and Mean CIs for $X \sim Mult(1;0.15,0.2,0.3,0.22,0.13)$

<table>
<thead>
<tr>
<th>N</th>
<th>90% Nominal Coverage</th>
<th>95% Nominal Coverage</th>
<th>99% Nominal Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed coverage (L,U)</td>
<td>Theoretical coverage</td>
<td>Observed coverage (L,U)</td>
</tr>
<tr>
<td>10</td>
<td>91.3% (2.50, 3.48)</td>
<td>92.07%</td>
<td>95.0% (2.34, 3.61)</td>
</tr>
<tr>
<td>20</td>
<td>94.4% (2.77, 3.23)</td>
<td>90.05%</td>
<td>95.4% (2.67, 3.33)</td>
</tr>
<tr>
<td>30</td>
<td>97.5% (2.84, 3.14)</td>
<td>90.13%</td>
<td>99.1% (2.72, 3.27)</td>
</tr>
<tr>
<td>40</td>
<td>99.4% (2.86, 3.14)</td>
<td>91.26%</td>
<td>99.4% (2.87, 3.13)</td>
</tr>
<tr>
<td>50</td>
<td>99.6% (2.91, 3.09)</td>
<td>90.17%</td>
<td>99.6% (2.91, 3.09)</td>
</tr>
</tbody>
</table>

A couple of possible reasons exist for why the CIs were overly conservative. It is conceivable the probability distribution underlying the example just happened to yield conservative CIs by chance alone. Alternatively, it may be due to the fact that the probability distribution was unimodal, where the mode was also equal to the median. That is, perhaps when the distribution has a strong center of the mass (i.e., probability is less spread out), the resulting CI will tend to be conservative. This hypothesis was tested by two additional Monte Carlo analyses with 100,000 simulations conducted for $\alpha=0.1$ and $N=30$. In the first analysis, 5% from each outer class was shifted toward the median, $X \sim Mult(1;0.05,0.25,0.4,0.27,0.03)$ thereby, increasing the center mass without affecting the median or balance of the original distribution. This analysis yielded a coverage probability of 99.6%. In the second analysis, 5% from each inner class was shifted away from the median, $X \sim Mult(1;0.2,0.2,0.3,0.22,0.18)$; thereby, decreasing the center mass without affecting the median or original balance. This analysis yielded a coverage probability of 90.2%, which is very close to the predicted
coverage probability of 90.13%. In combination, the results of these analyses lend support to the hypothesis that the degree to which a CI is overly conservative or not depends upon the size of the center of the mass. The closer the median is to the center mass (i.e., less spread in data), the higher the likelihood the CI will capture the population median.

A third possible explanation resides in the fact that discrete variables generally have smaller variances than continuous variables, assuming sample sizes are the same. Therefore, it is reasonable to surmise that if the responses of discrete variables were more spread out (e.g., the variable was measured by a scale with a wider range), then the CI may be less conservative. A Monte Carlo analysis was conducted to test this hypothesis and to eliminate the dependence on a specific probability distribution. Once again, 100,000 random variables were generated and 95% CIs were constructed from which the probability coverage was computed in the usual way. However, unlike previous simulations, a different probability distribution was randomly generated for each random variable. That is, each of the 100,000 random variables had a unique probability distribution with the only condition that a class, which can be thought to represent a point on a Likert scale, could not have a probability less than 1%. Table 14 presents the results of this analysis for five sample sizes and seven multinomial classes. The results are clearly less conservative than the ones reported in the previous table ($\alpha=0.05$). Moreover, although the results show that the number of classes has an impact on coverage, the impact is mild and virtually disappears after 5-classes. Sample size, however, continues to affect the probability coverage, with larger sample sizes associated with more conservative CIs. This is not unexpected considering that the larger the sample size, the greater the likelihood the sample distribution will approach the population distribution; hence, reducing sampling error and improving coverage.
Table 14

Coverage Probabilities for Categorical Distributions with Varying Number of Classes

<table>
<thead>
<tr>
<th>N</th>
<th>3-classes</th>
<th>4-classes</th>
<th>5-classes</th>
<th>6-classes</th>
<th>7-classes</th>
<th>10-classes</th>
<th>15-classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>96.1%</td>
<td>95.5%</td>
<td>95.4%</td>
<td>95.3%</td>
<td>95.3%</td>
<td>95.2%</td>
<td>95.3%</td>
</tr>
<tr>
<td>40</td>
<td>97.5%</td>
<td>97.1%</td>
<td>96.9%</td>
<td>96.9%</td>
<td>96.8%</td>
<td>96.9%</td>
<td>96.8%</td>
</tr>
<tr>
<td>60</td>
<td>98.4%</td>
<td>98.1%</td>
<td>97.9%</td>
<td>97.9%</td>
<td>97.8%</td>
<td>97.9%</td>
<td>97.8%</td>
</tr>
<tr>
<td>80</td>
<td>98.3%</td>
<td>98.0%</td>
<td>97.8%</td>
<td>97.8%</td>
<td>97.8%</td>
<td>97.8%</td>
<td>97.8%</td>
</tr>
<tr>
<td>100</td>
<td>98.7%</td>
<td>98.5%</td>
<td>98.3%</td>
<td>98.2%</td>
<td>98.2%</td>
<td>98.2%</td>
<td></td>
</tr>
</tbody>
</table>

Although the previous example attempted to evoke the idea that this method can be used for categorical distributions (e.g., Likert or ordinal data), its application is far greater. Not only can the method be used for any discrete distribution, it can also be used for continuous distributions. After all, a continuous distribution is merely a discrete distribution with $n$ classes where the probability of ties for any given class is zero. To demonstrate these points, the discrete method for constructing CIs was compared to the continuous method discussed in the previous section for two discrete and one continuous distributions. Ninety-five percent CIs were constructed on the sample median of three random variables, $X \sim \text{Bin}(30,0.3)$, $Y \sim \text{Poi}(15)$, and $Z \sim [\mathcal{N}(10,4)]^{1.5}$. Table 15 presents the theoretical coverage probabilities, 95% CI, and rank orders of the CI boundaries for various sample sizes.

Table 15

Comparison of Coverage Probabilities Between the Discrete and Continuous CI Methods

<table>
<thead>
<tr>
<th>N</th>
<th>Binomial (X)</th>
<th>Poisson (Y)</th>
<th>Continuous (Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coverage</td>
<td>$(a,b)$</td>
<td>Coverage</td>
</tr>
</tbody>
</table>

Continuous Method

<table>
<thead>
<tr>
<th>N</th>
<th>Coverage</th>
<th>$(a,b)$</th>
<th>Coverage</th>
<th>$(a,b)$</th>
<th>Coverage</th>
<th>$(a,b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>95.7%</td>
<td>(8, 10)</td>
<td>95.9%</td>
<td>(28, 32)</td>
<td>95.9%</td>
<td>(23.2, 33.0)</td>
</tr>
<tr>
<td>40</td>
<td>96.2%</td>
<td>(8, 10)</td>
<td>96.2%</td>
<td>(28, 32)</td>
<td>96.2%</td>
<td>(25.9, 33.0)</td>
</tr>
<tr>
<td>60</td>
<td>96.4%</td>
<td>(8, 10)</td>
<td>96.0%</td>
<td>(28, 32)</td>
<td>96.0%</td>
<td>(27.0, 33.0)</td>
</tr>
<tr>
<td>80</td>
<td>95.5%</td>
<td>(8, 10)</td>
<td>95.5%</td>
<td>(28, 32)</td>
<td>95.5%</td>
<td>(27.9, 33.1)</td>
</tr>
</tbody>
</table>

Discrete Method

<table>
<thead>
<tr>
<th>N</th>
<th>Coverage</th>
<th>$(a,b)$</th>
<th>Coverage</th>
<th>$(a,b)$</th>
<th>Coverage</th>
<th>$(a,b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>96.3%</td>
<td>(9, 9)</td>
<td>95.9%</td>
<td>(29, 30)</td>
<td>95.7%</td>
<td>(24.3, 32.1)</td>
</tr>
<tr>
<td>40</td>
<td>95.2%</td>
<td>(9, 9)</td>
<td>95.1%</td>
<td>(29, 30)</td>
<td>95.1%</td>
<td>(25.9, 32.0)</td>
</tr>
<tr>
<td>60</td>
<td>95.1%</td>
<td>(9, 10)</td>
<td>95.1%</td>
<td>(29, 32)</td>
<td>95.0%</td>
<td>(27.8, 33.1)</td>
</tr>
<tr>
<td>80</td>
<td>95.0%</td>
<td>(9, 9)</td>
<td>95.1%</td>
<td>(29, 30)</td>
<td>95.1%</td>
<td>(27.8, 32.0)</td>
</tr>
</tbody>
</table>
Comparison of the two methods and the previous findings clearly support the use of the discrete method for discrete and continuous distributions. Although both methods were able to attain slightly conservative CIs, the theoretical coverage probabilities for the discrete method were closer to the nominal coverage probability. This is also confirmed by the fact that the width of the CIs produced by the continuous method was consistently larger than their discrete counterparts.\(^{92}\) Although absent from this analysis, the observed coverage probability produced by a Monte Carlo simulation would naturally follow the theoretical coverage, as previously illustrated, because it is a function of the width of a CI. Hence, the continuous method should never be used to construct a CI for a discrete random variable because it will yield an overly conservative CI. What may be somewhat surprising is the fact that the discrete method produced tighter CIs with better coverage (i.e., closer to the nominal coverage) for the continuous random variable than the continuous method. Moreover, although the theoretical coverage probability improves as the sample size increases, it is nearly optimal even for small sample sizes. Furthermore, the same method that was developed in the previous section for incorporating sampling error can also be incorporated into the discrete method. Thus, researchers and evaluators can use this method for constructing a CI on the median regardless of the distribution, data type, or whether the variable was measured on an ordinal, interval, or ratio level of measurement. Simply stated, this is one of the best methods for constructing CIs on the median.

Algorithm for Constructing a Basic Summative Confidence Interval

The principal purpose of this chapter was to outline a framework by which one can construct a basic Summative Confidence interval. Despite the ostensible simplicity of the

\(^{92}\) It is important to recall that this difference cannot be attributed to a difference in the way in which variance was estimated for each CI since both CIs were constructed based on order statistics: one continuous and the other discrete.
question with which the chapter began, “how good is the overall performance of the evaluand,” the solution for deriving a master formula required considerable work. However, in many regards, the formula is quite simple to use and requires very little overhead in the way of mathematical knowledge. The remainder of this section will clarify and formalize the basic algorithm that should be followed when conducting a Summative Confidence analysis. It will be assumed one has already collected the data based upon a theoretical framework.

To begin, prior to constructing a composite variable, the first step is to identify the distribution of each of the constituent variables. If all the variables share a common distribution, then one can produce a composite variable either by summing or averaging all of the constituent variables. In the case of summation, the expected value and variance of the composite total are given by Properties 9 and 19, respectively. If weights need to be applied to any of the constituent variables, then Properties 3 and 9 can be combined to measure the expected value of the weighed total and Property 21 can be employed to measure the weighted variance rather than Property 19. However, since parametric CIs are constructed based on the standard error, Properties 19 (or 21) and 23 can be combined to create a general method for estimating the error variance of the total. If the population size is known for any of the constituent variables, then their respective error variance can be multiplied by the fpc to account for sampling error. A basic Summative Confidence interval can then be constructed based upon the method described in the first part of the last section.

In most instances, however, evaluators are likely to create a composite variable by averaging all the constituent variables. In such cases, the expected value and variance of the composite mean are given by Properties 10 and 22, respectively. If weights need to be applied to any of the constituent variables, then Properties 3 and 10 can be combined to measure the expected value of the weighed mean and Properties 21 and 22 can be combined
to measure the weighted variance. However, since parametric CIs are constructed based on the standard error, Property 24 should be used to construct the Summative Confidence interval. If the population size is known for any of the constituent variables, then their respective error variance can be multiplied by the fpc to account for sampling error. A basic CI can then be constructed using the method described in the first part of the last section. Note, if all the data come from the same data stream measured by the same method then standard software can be employed to construct a Summative Confidence interval. That is, the Summative Confidence interval is equivalent to a standard CI when the constituent variables used to construct the complex variable all come from the same stream. In all other instance, one needs to manually compute the Summative Confidence interval.

These two options describe the best case scenario in which no transformation is required because the constituent variables have the same underlying probability distribution. In most instances, this will not be true. Transformation is the process by which evaluators convert dissimilar underlying distributions into a common distribution for the purpose of constructing a composite variable in terms of merit, worth, or meaningful significance. Ignoring the context in which “merit, worth, or meaningful significance” is defined for now, numerous transformations are available to evaluators for converting dissimilar distributions into a common distribution. The simplest transformation is to dichotomize all the constituent variables, which would produce Bernouli trials whose mean or sum would yield a composite variable with an expected value and variance that conforms to the previous properties. Sampling error can then be incorporated into the variance estimates in the usual way and Property 24 can be used to construct the Summative Confidence interval. Despite its simplicity, this method has a rather unsavory drawback; dichotomization results in the lost of a considerable amount of information, particularly for continuous variables.
Although a better alternative is polychotomization because it reduces the amount of information lost, it potentially introduces another complication. Monotone transformations of random variables measured with an interval or ratio level of measurement yield new random variables in which the distances between adjacent points are equal. Hence, one can employ the aforementioned properties to compute the expected value and variance of the composite variable. Moreover, one can adjust the expected value and variance for weights and sampling error in the usual way to construct a Summative Confidence interval. Unfortunately, equidistance is unlikely to exist for transformations of random variables measured with an ordinal level of measurement. In such situations, one will need to employ the discrete order statistics CI method and then compute the approximate variance.

One of the most important theorems in all of statistics is the CLT, which states that the distribution of a composite variable, created by either summing or averaging a set of constituent variables, will approach the normal distribution given a large enough number of constituent random variables. However, the rate at which the distribution of the composite approaches the normal is a function of the number of constituent variables and the degree to which they deviate from the normal distribution. Hence, at times, sample statistics may deviate substantially from their population counterparts, which, in turn, connotes that the composite sample mean (total) also deviates from the composite population mean (total). In such instances, one is better off utilizing the median as a measure of central tendency, rather than the mean, since the median is robust to outliers. Even better, medians can be computed for discrete, continuous, and ordinally measured variables. Hence, no distributional assumptions are needed to compute the grand median of each constituent variable.

Although medians are simple to compute, constructing confidence intervals for them does require a bit more work than their parametric counterparts. However, the SAS code
provided by the author eliminates the need to manually compute these intervals. Moreover, a method was developed for also accounting for sampling error. Although this method is not perfect because the variance adjustment is based solely on two points (the boundaries of the unadjusted CI), it does a good job of compensating for small to moderate sized sampling fractions. Moreover, both the continuous and discrete methods were shown to work well for small samples, with moderate to large sample producing even more accurate results with respect to the theoretical coverage probability. In practice, however, it was shown that the observed coverage probability can be higher than the theoretical coverage depending upon the distribution of the data. Essentially, one may think of the theoretical as a lower bound estimate of the coverage probability. Hence, the tendency of both the continuous and discrete methods to produce slightly conservative CIs suggests that the probability any given sample CI will contain the population median is at least as great as the nominal coverage probability. These results have great implication for evaluators.

With respect to Summative Confidence, the implications are clear. Whenever the distribution of one or more continuously measured constituent variables is unknown or there is reason to believe that the underlying distribution is seriously violated, then one should use the continuous method to construct the Summative Confidence interval on the grand median. In contrast, when the distribution of the constituent variables are naturally discrete or transformed into discrete random variables, then one should use the discrete method to construct the Summative Confidence interval on the grand median.
CHAPTER V
IMPLICATIONS OF MEASUREMENT THEORY

Instruments, whether mechanical or surveys, can never measure an entity with perfect accuracy or reliability. Changes over time, administration, noise during testing, guessing, and natural fluctuations in the respondent’s psychological or physical state, fatigue, memory, and so on are likely to produce random measurement errors. Likewise, systematic biases or errors in measurement, such as wording of questions, reading level of questions, choice of measurement scales employed, recording of data under a limited set of conditions, rounding errors, imperfect calibration of measurement instrument, and so on are also likely to produce measurement error. The extent to which an observation incorporates random or systematic error is a concern for evaluators and researchers because error not only reduces the likelihood of replicating conclusions derived from such data (attributed to random error) but it also suggests the conclusions derived from repeated study may be incorrect (attributed to systematic error). Moreover, since evaluative conclusions are based upon the synthesis of numerous constituent variables, a real danger exists wherein even small errors in the measurement of these variables may propagate to produce an erroneous evaluative conclusion. One may then ask, are the conclusions of a study random fluctuations, biased, or an accurate representation of the truth?

In the case of random errors, the Law of Large Numbers ensures that if a large enough sample of data are recorded, the expected value of the sample will converge upon the expected value of the population. Hence, the effect of random fluctuations can be diminished by gathering large samples to reduce sampling error. Moreover, one can replicate
a study numerous times to reduce effects such as time, administrative error, and so on. The idea being that although random errors occur within every study, the same set of random errors are unlikely to repeat. Over repeated study, the expected impact of random error on a conclusion should be zero since some errors are likely to suppress results whereas others are likely to enhance results. Unfortunately, few options exist for correcting systematic error, particularly when the error is undetected. Two options exist for detecting systematic errors: (1) check whether measures drift in a particular direction over time or (2) compare measures against known quantities. Unfortunately, neither option is particularly effective in social science research, where results are likely to change over time and known quantities are rarely available. Moreover, the effects of systematic errors cannot be removed by averaging a large number of observations or studies (unless the studies employed different instruments in which case the systematic bias would be minimized). If the degree of systematic error is known one could remove its impact on a conclusion through the use of simple arithmetic. However, when the error magnitude is unknown, the only option is to recalibrate the instrument, assuming this is believed to be the cause of the error.

The present chapter is partitioned into two sections. The first section focuses on exploring the impact of measurement error on variance. Specifically, it integrates classical test theory into the body of knowledge presented in the previous chapter from which will emerge the master formula that can be used to construct a Summative Confidence interval or compute a Summative Confidence reliability coefficient. Since, as will be proven in the first section, measurement error can be accounted for by the parallel reliability coefficient, the second section will present various indices that have been historically used to measure parallel reliability, including one developed by the present author.
Classical Test Theory

Measurement is the fundamental process that underlies all scientific research. Without measurement the only available method by which scientific principles may be unearthed is via thought experiments (e.g., Albert Einstein’s (1996) thought experiment of chasing a beam of light). However, in absence of supporting data, principles derived from such experiments remain in the realm of conjecture. Outside of thought experiments, all other methods of scientific inquiry entail measurement. Building upon the definitions of Stevens (1946) and Lord and Novick (1968), since measurement requires the use of a metric (i.e., a scale), it can be further thought of as a transformative process whereby an unknown (latent) distribution of a random variable is assigned numbers (symbols, labels) corresponding to a new distribution on the basis of a set of prespecified rules. That is, measurement entails the mapping out of a characteristic of a latent distribution onto an observed distribution.

The present section will study the impact of such measurement error on conclusions. As will become evident shortly, measurement theory relies heavily upon expectation theory, which was expounded upon in the previous chapter. Moreover, since the theoretical foundation of measurement theory has been around since Charles Spearman put forth the concept of classical test theory over 100 years (Alexopoulos, 2007), the subject will be treated as common knowledge and presented with limited citations. Interested readers are directed to the seminal works of Gulliksen (1950), Lord and Novick (1968), and Nunnally (1978), and more recently, the work of Crocker and Algina (1986) and McDonald (1999) for further reading. SAS code for Monte Carlo analyses is provided in Appendix I.

---

93 Test theory, psychometric theory, and measurement theory are identical, so they are used interchangeably.
Fundamental Theory

The fundamental principle underlying measurement theory is that all observations contain random and, potentially even, systematic error. As noted earlier, systematic error is particularly problematic when it occurs because it is difficult to identify. Even if one suspects that it occurred, one would not be able to correct the error without knowing its magnitude, which, of course, is generally unknown. As a result, when systematic error does occur, it is combined with random error, so since the two cannot be disaggregated, the discussion of measurement error herein and throughout the literature is restricted to random error.

In classical test theory (CTT), an observation is composed of two components: a *true score* and measurement error. Although the true score is not directly measurable, it is defined as the expected value of the sampling distribution of all possible measurements derived from the same measurement process. Measurement error by definition then is considered to be random and is defined as the difference between the true score and the observed score. Hence, the observed score for a random variable X is comprised of the true score T and measurement error E, where the true score is constant for each sampled unit while the error is random. That is, X=\(T+E\).

On the surface, it would appear this formulation is not very useful were it not for the assumption that errors are random and independently distributed. That is, some errors are positive whereas others are negative, so they tend to cancel each other out in the long-run. Moreover, knowing a person scored below their true score on one measure (test) relays no information as to whether they scored below, above, or equal to their true score on a different parallel measure.\(^{94}\) Hence, the long-run average of error for the \(i^{th}\) individual over repeated testing is zero, \(\mathbb{E}(E_i)=0\), which implies the expected value of their observations X,

---

\(^{94}\) Although CTT does not specifically pertain to people or testing situations, such terms will be used throughout this section to generically represent the object being measured and the measurement process.
must equal their individual true score \( T_i \), 
\[
\mathbb{E}(X_i) = \mathbb{E}(T_i + E_i) = \mathbb{E}(T_i) + \mathbb{E}(E_i) = \mathbb{E}(T_i).
\]
Of course, if \( \mathbb{E}(X_i) = \mathbb{E}(T_i) \) for every individual in the population then the expected value across individuals must also be equal, \( \mathbb{E}(X) = \mathbb{E}(T) \), which connotes that \( \mathbb{E}(E) = 0 \). Since errors are random then \( T \) and \( E \) must be independent, which connotes that the observed score variance, \( \sigma_X^2 \), must equal the sum of the variance of \( T \) and \( E \), 
\[
\sigma_X^2 = \sigma_T^2 + \sigma_E^2.
\]

In summary, classical theory was developed from the following seven properties: (1) \( \mathbb{E}(X) = \mathbb{E}(T) \), the expected value of observed scores is the expected value of their true scores; (2) \( \mathbb{E}(E) = 0 \), the expected value of error is zero; (3) \( \sigma_{TE} = 0 \), the covariance between the true and error scores on the same or different tests is zero; (4) \( \sigma_{EE'} = 0 \), the covariance between errors on different tests is zero; (5) \( \rho_{TE} = 0 \), the correlation between the true and error scores on the same or different tests is zero; (6) \( \rho_{EE'} = 0 \), the correlation between errors on different tests is zero; and (7) \( \sigma_X^2 = \sigma_T^2 + \sigma_E^2 \), the total variance of observed scores is equal to the sum of the variance of true scores and error. These properties arise from the assumption of randomness and independence. The independence assumption, which is formally defined by the cdf \( F(E_i, E_j) = F(E_i)F(E_j) \), can be relaxed slightly to imply linear independence, which is defined by the conditional expectation \( \mathbb{E}(E_i \mid E_j) = \mathbb{E}(E_j) \). Furthermore, although not directly stated, these properties assume an interval or ratio level of measurement and finite variances \( \sigma^2 < \infty \), where the former assumption follows from the fact that parametric correlations and variances require the use of arithmetic operations not permissible to nominal and ordinal variables. In all likelihood, this is a relatively safe assumption since the latent distribution is frequently, but not always, continuous.

The correlation between two parallel tests, known as the parallel test reliability, can be used to estimate the coefficient of determination since 
\[
\frac{\sigma_T^2}{\sigma_X^2} = \rho_{XX'}^2 = \rho_{XY}^2.
\]
It then follows
that the correlation between the true scores and observed scores, known as the **reliability index**, is equal to the square root of the parallel test reliability, i.e., \( \rho_{XT} = \sqrt{\rho_{XX}'} \). Hence, the observed score variance is a function of test reliability, \( \sigma_X^2 = \rho_{XX}' \sigma_X^2 + (1-\rho_{XX}') \sigma_{XX}' \), where the first component represents the true score variance and the second component represents the average measurement error variance. The square root of the second component, known as the standard error of measurement \( \sigma_E = \sigma_X \sqrt{1-\rho_{XX}'} \), is used for constructing confidence intervals. However, it is important to note, these confidence intervals are placed around actual true scores, rather than observed scores.\(^95\) Hence, if the true score is known, a \( 100(1-\alpha)\% \) CI on \( T \), \( T \pm z_{1-\alpha/2} \sigma_E \), can be used to predict their observed score. For example, given the cutoff score associated with a performance standard (which can be regarded as a true score), one could compute a 1-sided CI to identify the observed scores whose true scores are likely to be higher (or lower) than the standard. In general, however, one is unlikely to know the value of a true score, particularly for individuals, and if one did, why would one care about the observed scores?

A more likely scenario is that one would like to estimate an individual’s unknown true score based upon knowledge of their observed score. Fortunately, a linear relationship exists between the observed and true scores thanks to Property 8. Accordingly, given an observed score, the true score can be estimated by

\[
E(T|X) = E(T) + \rho_{XT} \frac{\sigma_T}{\sigma_X} [X - E(X)] = E(X) + \rho_{XT}^{2} [X - E(X)] = E(X) + \rho_{XX'} X - \rho_{XX'} E(X) = \rho_{XX'} X + (1-\rho_{XX'}) E(X),
\]

since \( E(T) = E(X) \) and \( \rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \). Hence, this predicted true score is composed from the individual’s observed score, the group mean, and the parallel test reliability. For highly

\(^95\) A common mistake is to construct the CI on \( X \). However, it follows from the error variance definition, \( \sigma_E^2 = E[Var(X|T)] \), that the error variance is equal to the average variance across the given levels of \( T \).
reliable tests, the predicted true score will reflect the observed score to a greater extent than
the group mean, and vice versa. This implies that if the reliability of a test was zero then the
best estimate of each individual's true score would be the mean, which, in turn, connotes
that the true score variance would be zero. The error variance associated with this
conditional expectation is given by Property 18, \( \sigma^2_e = \text{Var}(T|X) = \sigma^2_t(1-\rho^2_{XX}) = \rho^2_{X|T}\sigma^2_X(1-\rho^2_{XX}) \)
\( = \sigma^2_X(1-\rho^2_{XX}) \). This connotes that the standard deviation of this estimate, known as the
standard error of estimation, is equal to \( \sigma_e = \sigma_X \sqrt{(1-\rho^2_{XX})} \). Hence, given an observed
score, a \( 100(1-\alpha)\% \) CI can be constructed to predict the unknown true score based on the
best linear estimate of the true score, \( [\rho_{XX}X+(1-\rho_{XX})E(X)] \pm z_{\alpha/2}\sigma_e \).

Occasionally, it may be necessary to predict performance on a parallel test given an
observed score. For example, if an individual has an observed score of \( x \) on a test, what is
the most likely score \( y \) they would observe on a parallel test? The solution to this question is
provided by Property 8. Their predicted score is simply the conditional expectation given
their observed score. Hence, given an observed score, the predicted score on a parallel test is

\[
E(X'|X) = E(X') + \rho_{XX} \frac{\sigma_X}{\sigma_X'} [X - E(X)] = E(X) + \rho_{XX} [X - E(X)] = \rho_{XX}X+(1-\rho_{XX})E(X),
\]

since \( E(X) = E(X') \) and \( \text{Var}(X) = \text{Var}(X') \). Note, the predicted true and parallel observed scores
are equal, \( E(T|X) = E(X|X) \) and \( \Delta = X - \hat{X} \) then \( \text{Var}(\Delta) = \text{Var}(X - \hat{X}) \) is
minimized in the same way as Property 18. Namely, \( \sigma^2_{\Delta} = E[(\text{Var}(\hat{X}|X)] = \sigma^2_{\hat{X}}(1-\rho^2_{XX}) = \sigma^2_{\hat{X}}(1-\rho^2_{XX}) \)
since the variances and correlation of parallel tests are equal. It then follows that the
standard deviation, known as the \textit{standard error of prediction}, is equal to \( \sigma_{\Delta} = \sigma_X \sqrt{(1-\rho^2_{XX})} \). A
\( 100(1-\alpha)\% \) CI for a parallel observed score then is given by \( [\rho_{XX}X+(1-\rho_{XX})E(X)] \pm z_{\alpha/2}\sigma_{\Delta} \).
Hence, although the predicted true and parallel observed scores given an observed score are
equal, their variance estimates and CIs are not. Comparing these standard errors, it is easy to verify that they decrease as a function of the parallel test reliability and $\sigma_{d} \geq \sigma_{e} \geq \sigma_{e}$.

Potential Limitations of CTT

While the last Monte Carlo analysis clearly illustrates the ability of CTT to construct CIs of importance to researchers, this capacity depends upon three primary assumptions; namely, each person has the same standard error, measurement errors are normally distributed, and an interval or ratio level of measurement is employed.\(^{96}\) The former assumption is a well-known violation of reality. In fact, IRT proponents frequently criticize CTT for applying the average standard error to all the scores of a particular population (Embretson & Reise, 2000) despite the fact that error is likely to differ from individual to individual and along various points along the score scale (Harvill, 1991). Feldt, Steffen, and Gupta (1985), for example, empirically demonstrated this problem leading the authors to state that “the standard error of measurement computed by the traditional formula for the test as a whole does not adequately summarize the error propensity of many—perhaps most—examinees” (p. 358). As a result of this problem, the Standards for Educational and Psychological Testing (1999) recommends that test publishers provide an estimate of the standard error for each of a number of widely spaced score levels.

It is important then to assess the impact that differences between individual levels of consistency has on CTT, in general, and probability coverage, in specific. To this end, the a Monte Carlo analysis was conducted in which the mean and variance were 100 and 225, respectively. However, the consistency (variability) across cases was allowed to vary widely as long as the average of these variance equaled the expected error variance. Hence, the

---

\(^{96}\) Although additional limitations of CTT have been proposed, this subsection will only focus on these three.
expected true score and error score variances remained unchanged, $\sigma^2_T = \rho_{XX'} \sigma^2_X = 0.35 \times 225 = 78.75$ and $\sigma^2_E = (1 - \rho_{XX'}) \sigma^2_X = (1 - 0.35) \times 225 = 146.25$. The folded normal distribution\(^{97}\), with parameters $\mu = \sigma^2_E$ and $\sigma^2_B = 25$, was employed to simulate different variance estimates for each case. Since it is natural to expect that the between ($\sigma^2_B$) and within ($\sigma^2_W$) variance are independent, their sum must equal the total error variance $\sigma^2_E$. Comparison of the highest to lowest individual between variance revealed a ratio of approximately 148,042 to 1. Despite this enormous difference, the results of the analysis were very similar to the previous analysis. If fact, the coverage probabilities for the three CIs ranged between 93.8% and 94.9%. Hence, it would appear that CTT is quite robust despite wide differences in the consistency of respondents, at least with respect to group level decisions. Individual-level decisions, however, are likely to suffer from the use of constant standard error.

Although the investigation of this issue will be addressed in future research, one potential solution is to obtain variance estimates for each individual based upon the width of the nonparametric CI associated with their order statistic. As stated in the previous chapter, order statistics can be used to construct a nonparametric CI for any observation, not just the median. Since the maximum information for a variable is provided by the median, the width of the CI increases as one moves further away from the median, which connotes that the variance also increases in a corresponding fashion. Therefore, it is reasonable to surmise that variance of any given individual is approximately equal to $\sigma^2 = \eta [(UB - LB)/2 \zeta_{1-\alpha/2}]^2$, where LB and UB represent the lower and upper bound estimates of the individual’s CI. Hence, the standard error of measurement must be $\sigma_E = [(UB - LB) / 2 \zeta_{1-\alpha/2}] \sqrt{\eta (1 - \rho_{XX'})}$; the standard

\(^{97}\) The folded (or half) normal distribution (Leone, Nelson, & Nottingham, 1961) is equal to the absolute value of a normal distribution. This distribution was used to simulate variance since variance cannot be negative and the variance estimates across individuals will form a sampling distribution. Although the mean of this distribution must equal the error variance, otherwise one violates the seventh property of CTT, the variance of the distribution cannot be known without empirical testing. Consequently, a variance estimate was chosen for the folded normal distribution that would produce a wide range of variance estimates between individuals.
error of estimation must be $\sigma_e = \frac{(UB-LB)}{2\sqrt{z_{1-\alpha/2}}} \sqrt{n\rho_{XX}(1-\rho_{XX}^2)}$; and the standard error of prediction must be $\sigma_\Delta = \frac{(UB-LB)}{2\sqrt{z_{1-\alpha/2}}} \sqrt{n(1-\rho_{XX}^2)}$. Furthermore, for $\sigma_e$ and $\sigma_\Delta$, one will need to adjust the predicted true and parallel observed scores by replacing the expected value with the median of $X$, $m(X)$, i.e., $\rho_{XX}X + (1-\rho_{XX})m(X)$. Then, similar to the parametric predicted scores, as the parallel test reliability increases, the predicted true score will reflect the observed score to a greater extent than the group median, and vice versa. This implies that if the reliability of a test was zero then the best estimate of each individual’s true score would be the median, which, in turn, connotes that the true score variance would be zero.

Another objection to the use of CTT is the assumption that measurement errors are normally distributed. This assumption is not so much a function of statistical convenience as it is a recognition of the fact that random measurement errors are generally found to be normally distributed (Harvill, 1991). Recall, Chapter 3 provided an illustration of why measurement error tends to be normally distributed. However, there is no law in nature that dictates this to always be the case. Lord and Novick (1968, p. 22), for example, noted that “in treating data by interval methods...we are, in effect, stipulating a specific distance function for our scale where the underlying measurement process and theory supporting it have not done so. This could be considered as an arbitrary strengthening of our model.” However, they go on to explain that they were forced to do so because “no general theory of [nonparametric] true-score models [was] available” (p. 42) at the time. Hence, the impact of violating the normality assumption and the use of interval levels of measurement on CTT needs to be tested.

To this end, a Monte Carlo analysis was performed in which the GLD was used to simulate nonnormal measurement error (skewness of about 2 and kurtosis of about 10). In another analysis, a nonnormal distribution was simulated for both the true and error scores.
Somewhat to the surprise of this author, the CTT held up extremely well despite the violation of normality. In fact, the observed coverage probabilities for all three of the aforementioned CIs were approximately equal to the nominal probability coverage. In retrospect, however, this should not have been a surprise given that the theory underlying CTT is based on sampling distributions; hence, the CLT ensures that the sampling distributions will be approximately normal even if the underlying probability distributions are not normal. Additionally, a Monte Carlo analysis was conducted to determine whether CTT held for ordinal data. Two categorical distributions were simulated, one for the true scores and one for the error scores. However, two conditions were imposed on the error scores. First, the error scores had to center on zero; otherwise, the expected value or even the median of the error scores would not equal zero.98 Second, it was assumed that the true and error scores were measured on the same scale as the observed random variable, although their exact probability distribution were allowed to differ, but their sum could not violate the boundaries of the observed measurement scale. For example, if random variable $X=T+E$ was measured on a 5-point Likert scale (with scores ranging from 1 to 5) then the observed values of $X$ must be bound to the domain of the Likert scale, i.e., $X \in (1,5)$. This analysis revealed that even though $T$ and $E$ were uncorrelated, their variances did not sum to the variance of $X$, which connotes that $T$ and $E$ are not independent.99 The implication of this violation is that the coverage probabilities of any CI constructed from a standard error will not equal the nominal coverage probability. That is, CTT breaks down for ordinal data.

98 Note, even the regular CTT model breaks down when $E(E) \neq 0$. Specifically, the coverage probability for the observed score given the true score as well as for the true score given an observed score suffers, whereas the coverage probability for parallel observed scores is unaffected.

99 Note, the absence of a correlation, which is a measure of linear dependence, does not imply the absence of dependence. For $T$ and $E$ to be independent then each variable must be able to take on any value in their range. However, if the value of $T$ for each person is fixed, then $E$ cannot take on every value in its range since it cannot take on a value whose sum to $T$ would yield a $X$ outside the interval $(1,5)$, i.e., $T$ and $E$ are dependent.
Further research, however, is necessary to determine if the model breaks down only when T and E are ordinal (i.e., latent class variables) as was simulated here, or whether the model also breaks down when the observed scores for X are ordinal but not their latent true and error scores.

Implications for Summative Confidence

In the previous chapter, Summative Confidence was defined as the aggregation of sampling error across multiple variables. Till now, this method assumed that all the variables were measured with perfect accuracy. However, this is a false premise since we know that all variables are measured with some degree of measurement error. The goal then is to use CTT to remove measurement error from variance estimators whenever there is information regarding the reliability of the constituent variables. Recall then, the total score variance can be decomposed into two components: one that measures variability between true scores and one that measures variability across repeated measures. In general, evaluators are not interested in modeling error variance across repeated measures but rather, they are interested in true score variance. Hence, one can replace the observed error variance in \( \text{Var}(\bar{X}) \) with a reliability adjusted variance estimator.

Determining the appropriate reliability adjusted variance estimator for constructing a CI depends upon the type of CI one wants to construct. For example, if the observed score variance and parallel test reliability are known then the true score variance is given by \( \sigma_T^2 = \rho_{XX} \sigma_X^2 \). Substituting this estimator for the sample variance in \( \text{Var}(\bar{X}) \) would enable one to construct a CI on the sample true score for the complex variable (i.e., sum or average of

100 Technically, it would be more appropriate to limit this discussion to composite variables generated from survey responses, for example, rather than individual variables, since reliability information is more readily available for such measures. However, since it is possible to obtain reliability data even for single items (e.g., test-retest reliability) this distinction is ignored. When the reliability of a measure is unknown and cannot be empirically estimated, one should assume it is equal to unity so as to produce a conservative variance estimate.
the true scores of the constituent variables) in order to capture the population observed score for the complex variable (i.e., sum or average of the observed scores of the constituent variables). However, it is difficult to conceive of an evaluation scenario in which one would wish to know the complex variable observed score given knowledge of its true scores. Hence, this estimate has greater theoretical value than practical significance, especially since the true scores are never known.

A more likely scenario occurs whenever an evaluator is interested in knowing the population true score for a complex variable across a set of measures given knowledge of the corresponding sample observed score for a complex variable. Alternatively, an evaluator may be interested in knowing the population complex variable score for a set of measures given knowledge of the sample complex variable observed score for a set of parallel measures. Using Property 18, it is easy to prove \( \text{Var}[\mathbb{E}(T|X)] = \text{Var}[\mathbb{E}(X'|X)] = \rho_{XX'}^2 \sigma^2_X \).

Substituting this estimator for the variance estimator in \( \text{Var}(uX) \) allows one to compute the Summative Confidence true score, sampling error variance. With this estimator one can then construct a CI for the sample mean of a set of predicted true scores, \( \mathbb{E}(T) \), so as to capture the population mean of either the true score, \( \mathbb{E}(T) \), or the observed score, \( \mathbb{E}(X') \), of a set of parallel measures. In either case, one is only interested in the predicted true score error variance, which for \( k \) independent variables is given by

\[
\text{Var}(uX) = \frac{1}{k} \sum_{i=1}^{k} w_i^2 \rho_{XX'}^2 \frac{\sigma^2_X}{n_i} \cdot \frac{1}{k} \sum_{i=1}^{k} w_i^2 R_{XX'}^2 \left[ 1 - \frac{n_i}{N_i} \right] \frac{S^2}{n_i},
\]

where \( \rho_{XX'} \) and \( R_{XX'} \) denote the population and sample parallel test reliability for

\[101\] According to Property 18, \( \text{Var}[\mathbb{E}(T|X)] = \rho_X^2 \sigma_T^2 = \rho_{XX'} \rho_{XX'} \sigma_X^2 = \rho_{XX'}^2 \sigma_X^2 \) since \( \sigma_T^2 = \rho_{XX'} \sigma_X^2 \). In a similar fashion, \( \text{Var}[\mathbb{E}(X'|X)] = \rho_{XX'}^2 \sigma_X^2 = \rho_{XX'}^2 \sigma_X^2 \) since the variance of parallel measures is equal, \( \sigma_X^2 = \sigma_X^2 \).
corresponding independent variables. Note, the fpc is only employed when sample estimates are used in place of population parameters.

As a simple illustration of the importance of accounting for measurement error, suppose a sample of size \(n=4,000\) is randomly drawn using SRS from a population of size \(N=10,000\). If the observed variance is 225, then the fpc adjusted population variance and standard error are \((1−0.4)*225=135\) and \(\sqrt{(135/4000)≈0.1837}\), respectively. Now, suppose the parallel test reliability for the random variable is 0.7, then the fpc and reliability adjusted variance and standard error are \(0.7^2*135=66.15\) and \(\sqrt{[0.7^2(135/4000)]≈0.1286}\), respectively. A Monte Carlo analysis was conducted to confirm these theoretical expectations wherein 10,000 samples of the predicted true scores were drawn from a population of predicted true scores. This analysis yielded a fpc and reliability adjusted variance and standard error of 65.938 and 0.1284, respectively, with a coverage probability of 94.8% for \(\alpha=0.05\). That is, the reliability unadjusted CI was 30% larger than the reliability adjusted CI. As a rule, reliability adjusted standard errors produce \((1−\rho_{XX'})\)% smaller CIs than their unadjusted counterparts. Although this example employed a single variable, it is clear the property holds for multiple independent variables since the variances are only summed and divided by a constant.

As this analysis illustrates, it is possible to remove the error variance from the observed variance and thereby, shrink the size of a CI. However, in order to expand the error variance property \(\text{Var}(\hat{w}|\mathbf{X})\) to dependent variables, one must first derive the impact of the constituent reliabilities on the conditional covariance. Based on these derivations (see Appendix J), the reliability adjusted covariance, for the purpose of constructing a CI to capture the mean true score of a composite or the mean parallel observed score, is given by \(\text{Cov}[\mathbf{E}(T_1|X),\mathbf{E}(T_2|Y)]=\text{Cov}[\mathbf{E}(X'|X),\mathbf{E}(Y'|Y)]=\rho_{XX'}\rho_{Y'Y}\sigma_X\sigma_Y\). Armed with this, one can derive the final Summative Confidence variance estimator by replacing the variance and
covariance estimators in $\text{Var}(u\bar{x})$ with their true score counterparts. For $k$ dependent variables, the master formula for Summative Confidence is equal to the weighted sum of the predicted true score error variances and covariances for each random variable, which is

$$
\text{Var}(u\bar{x}) = \frac{1}{k^2} \left[ \sum_{i=1}^{k} w_i^2 \rho_{XX_i}^2 \sigma_i^2 + 2 \sum_{i<j} w_i w_j \rho_{XX_i} \rho_{XX_j} \rho_{\bar{x} \bar{x}} \right] \frac{1}{\sqrt{n_i n_j}}
$$

$$
= \frac{1}{k^2} \left[ \sum_{i=1}^{k} w_i^2 R_{XX_i}^2 \left( 1 - \frac{n_i}{N_i} \right) \frac{S^2}{n_i} + 2 \sum_{i<j} w_i w_j \left( 1 - \frac{n_i}{N_i} \right) \frac{n_j}{N_j} \frac{R_{XX_i} R_{XX_j} R_{SS_i} R_{SS_j}}{\sqrt{n_i n_j}} \right]
$$

where $\sqrt{n_i n_j} = n_i$ when $n_i = n_j$ and the $i^{th}$ and $j^{th}$ variables are drawn from the same population (i.e., the fpcs are equal). While it is likely that only the last expression will be employed on a regular basis, one will need to use the middle expression whenever the correlation between two variables is known (or can be reasonably approximated) but due to the study design the average covariance term cannot be computed for all the data due to missing responses. Note, the fpc is only employed when sample estimates are used in place of population parameters.

Notice then that this variance estimator accounts for sampling error and measurement error. While it may be employed to construct a CI for an evaluative conclusion, another mechanism by which one can gage the quality of a conclusion is reliability. Based on the definition of parallel reliability, one can define a (fpc-adjusted) Summative Confidence reliability to equal the ratio of the true score error variance to the observed score error variance, $\rho_{\text{SC}} = \frac{\text{Var}(u\bar{x})}{\text{Var}(u\bar{x})}$. Notice then that both $\text{Var}(u\bar{x})$ and $\rho_{\text{SC}}$ unify the universe of statistical and measurement theory to produce two statistics: one that represents the variability in predicted true scores and one that represents the reliability of a summative conclusion.
To illustrate the method, a Monte Carlo simulation in which 100,000 samples of size \( n = 3,000 \) were repeatedly drawn from a population of size 10,000 using SRS. A composite mean was generated from four random variables, where \( X_1 \sim N(50,15) \), \( X_2 \sim N(75,20) \), \( X_3 \sim N(100,30) \), \( X_4 \sim N(125,50) \), \( \rho_{XX_1} = 0.3 \), \( \rho_{XX_2} = 0.5 \), \( \rho_{XX_3} = 0.7 \), \( \rho_{XX_4} = 0.9 \), and the following population correlations existed among the four variables:

\[
\Sigma_{4 \times 4} = \begin{bmatrix}
1.00 & 0.30 & 0.20 & 0.30 \\
0.30 & 1.00 & 0.45 & 0.25 \\
0.20 & 0.45 & 1.00 & 0.60 \\
0.30 & 0.25 & 0.60 & 1.00
\end{bmatrix}
\]

Examination of the results confirmed the simulation met these specifications within reasonable allowances: \( X_1 \sim N(49.98,14.811) \), \( X_2 \sim N(74.99,20.217) \), \( X_3 \sim N(100.01,30.472) \), \( X_4 \sim N(124.96,50.377) \), and a maximum \( |R - \rho| < 0.02 \). Therefore, the sampling error variance of the composite variable computed from the average of the four predicted true scores is

\[
\text{Var}(\bar{x}) = \frac{1}{16} \left[ 1 - \frac{3000}{10000} \right] \times \frac{1}{3000} \left[ \sum_{i=1}^{4} R_{XX_i}^2 s_i^2 + 2 \sum_{i<j} R_{XX_i} R_{XX_j} \Sigma_{XX_{ij}} \right]
\]

\[
= \frac{0.7}{16(3000)} \left[ (0.3)^2 (15) + (0.5)^2 (20) + (0.7)^2 (30) + (0.9)^2 (50) \right]
\]

\[
= \frac{0.7}{16(3000)} \left[ 0.15*0.3*0.5*\sqrt{15*20} + 0.20*0.3*0.7*\sqrt{15*30} + 0.30*0.3*0.9*\sqrt{15*50} + 0.45*0.5*0.7*\sqrt{20*30} + 0.25*0.5*0.9*\sqrt{20*50} + 0.60*0.7*0.9*\sqrt{30*50} \right]
\]

\[
\approx \frac{0.7}{16(3000)} (112.6587) = 0.00164.
\]

Hence, the standard error of the sample composite is equal to approximately 0.0405. Notice, the variance estimate is considerably smaller than the unadjusted for reliability variance estimate of 0.00335 \( = 0.7*229.4457/(16*3000) \). In fact, the ratio of these two variances estimates equals the Summative Confidence reliability, \( \rho_{SC} \approx 0.00164/0.00335 = 0.49 \). The analysis also revealed that the population variance and standard error of the composite were
approximately equal to 7.0412 (=112.6587/16) and 0.0265 [=\sqrt{7.0412/10000}]]. Inspection of the simulation results revealed that the population variance and error variance of the composite were 7.0641 and 0.0266, while the standard error of the sample composite is equal 0.0406. Moreover, the probability coverage was equal to 95.1% at \( \alpha=0.05 \). The nearly identical results between the predicted and observed values clearly illustrate the ability of \( \text{Var}(\mu_{\mathbf{x}}) \) to correctly compute the variance of the composite. Moreover, when compared to the observed score population variance (14.3363), these results further illustrate the importance of accounting for both sampling and measurement error.

A perhaps not so obvious issue is weighting. A surprising result is obtained if one were to sum the amount of variability each variable contributes to the total variance in the previous problem. In this case, variability is composed of both variance and covariance. Since there are two covariance terms for each pair of random variables, the total variability a random variable contributes to the composite variance is equal to the sum of the variance of the variable and each of its corresponding covariance divided by the square of the number of random variables in the composite. Hence, in the previous example, \( X_1 \) contributed nearly 0.3031 \( \frac{[(0.3)^2(15)+0.15*0.3*0.5*\sqrt{15*20}+0.2*0.3*0.7*\sqrt{15*30}+0.3*0.3*0.9*\sqrt{15*50}]/4^2}{\text{variance units, which only accounts for about 4.3\% of the total composite variance (7.0412)}. In contrast, \( X_4 \) contributed approximately 3.8072 variance units, which accounts for 54.1\% of the total composite variance. Such an imbalance in the amount of information contributed by each variable may be cause for concern.

One solution may be to find the \textit{a posteriori} weights that when applied to each predicted true score would adjust the amount of variability contributed by each random variable to the composite so that it is in-line with the desired weighting scheme. One method for accomplishing this task may be to set up a series of nonlinear equations that calculate the
amount of variance contributed by each variable to the composite variance and then use the Newton-Raphson method to compute the weights necessary to produce the variance in-line with a specified weighting scheme. The present author set about this task and was able to generate such a weight matrix ($w_1=3.6656$, $w_2=1.7837$, $w_3=0.8942$, and $w_4=0.5604$) that when applied to the predicted true scores, resulted in a composite for which all four random variables made approximately equal variance contributions. Unfortunately, the weights also resulted in a composite mean (119.1) that was significantly larger than the expected composite mean (87.5). Given that this discrepancy, further work is needed. Consequently, the SAS code for this method was not included in the appendix.

Classical Reliability

It is safe to say that all measures are imperfect since no instrument is capable of measuring with perfect accuracy or precision. From the perspective of research and evaluation, in general, and Summative Confidence, in specific, this implies that results sans measurement error are more precise than their counterparts. The removal of measurement error, however, requires knowledge of the parallel test reliability. A multitude of such estimators exist. For example, a test is considered reliable if it yields similar results for a person measured at different times. Hence, such reliability estimators require two administrations of a test in order to measure the stability of a construct over time. Alternatively, a test is considered reliable if all its items have large (positive) covariances (McDonald, 1999). In such instances, only a single administration of a test is required to measure the degree to which its items are homogeneous.

Due to the large number of potential estimators, the type of reliability coefficient used to calculate the true score variance can affect the magnitude of the variance and its semantic meaning. This section will cover the methods employed most frequently. Since the
majority, but not all, of these methods are well-known to researchers, they will be treated as common knowledge and presented with limited citations. Details will be provided for newer methods, such as bootstrap reliability. Since these methods only serve as inputs to the Summative Confidence algorithm, they will be presented without proof. Interested readers are directed to the seminal works of the measurement theorists presented in the previous section. SAS code for select methods is provided in Appendix K.

Stability and Equivalence

Stability refers to the ability of a test or method to yield consistent results over time. Hence, there is an assumption the construct being measured is constant over time. Naturally, this implies that two tests must be administered at two different periods in time in order to measure stability. Differences between parallel measures then enable one to estimate the distribution of measurement error across the group of respondents. Although the period can vary from a few minutes (e.g., multiples measures of blood pressure) to several years (e.g., IQ), it must be shorter than the amount of time for change to naturally occur in a person’s true score. For example, one expects observe blood pressure to change based upon diet, mood, and stress, to name a few factors. Hence, a long period confounds the interpretability of the reliability coefficient since one would not expect it to remain stable over extended periods of time. In contrast, since intelligence is believed to be relatively stable across one’s lifetime, one would expect IQ scores to remain fairly consistent even across decades.

Three methods are generally employed to measure stability, all of which depend upon the correlation coefficient. The most commonly used method, referred to as the test-retest method, entails the administration of a test (instrument) at different points in time to the same respondents. A correlation, known as the coefficient of stability, is then computed between the two tests. Another popular method used to estimate reliability entails the administration
of two alternative forms of a test or instrument to the same respondents. A correlation, known as the coefficient of equivalence, is then computed between the two tests. Unlike the test-retest method, the alternative form method is administered at the same time, allowing for a break between tests, if needed, to avoid fatigue. Hence, while the test-retest method is primarily concerned with temporal changes, the alternate form method is concerned with differences in item sampling. Finally, these two methods may be combined by administering one form of the test at one point in time and an alternative form of the test at a subsequent point in time, known as test-retest with alternate forms. Naturally, the correlation coefficient, known as the coefficient of stability and equivalence, is affected by both content sampling and temporal changes.

A large number of correlation coefficients have been developed since Karl Pearson first proposed the product-moment correlation coefficient, the majority of which are special cases of his correlation coefficient. The type of correlation coefficient one should compute depends on the level of measurement of and the relationship between the two variables (i.e., parallel tests). If the two variables are measured with an interval or ratio scale and their relationship is reasonably linear then one can compute the Pearson correlation coefficient. If the variables are measured with an ordinal scale then one can compute the Spearman-rank correlation coefficient. However, if one can further assume that the distribution underlying each ordinal variable is normal then one can compute a polychoric correlation coefficient. If the variables are dichotomous and nominal then one can compute either the Phi correlation coefficient or Cohen’s Kappa (Stemler, 2007). Finally, if the variables are nominal with more than two categories, then one can compute Fleiss’ Kappa (Fleiss, 1981). However, since from the perspective of Summative Confidence neither the mean nor the median of a set of
nominal variables is interpretable, parallel reliability is restricted to variables with at least an ordinal level of measurement.

A critical question in reliability theory is, what is an adequate level of reliability? Frequently, one encounters the recommendation that a reliability coefficient of 0.7 is often sufficient for establishing reliability (Nunnally, 1978). However, the lower the reliability, the more measurement error creeps into one's conclusions, which results in a lower degree of validity. Hence, the acceptableness of a reliability coefficient depends upon its use.

Internal Consistency

Since the reliability of a test is a function of its item covariances (McDonald, 1999), it stands to reason that a test constructed from homogenous items (e.g., items sampled from the same content domain) will exhibit greater reliability than one composed of heterogeneous items. Item homogeneity is determined by the size of the content domain and the quality of the items. The smaller the content domain, the better the items are written\(^{102}\), the more the item difficulty matches the ability of respondents\(^{103}\), and the less subjective the scoring\(^{104}\), then the greater the reliability will be. That is, one can have greater confidence the respondent’s score will generalize to other items from the same content domain. However, measurement errors due to random variables continue to affect reliability. Thus, a coefficient of internal consistency is an index of both content homogeneity and random noise.

---

\(^{102}\) An item is free of technical flaws if it is representative of the content domain (as opposed to an item that examines a superfluous detail contained within the content domain) and there is a low risk that respondents will misinterpret the item or be able to guess the “correct” response based on information other than their knowledge. With regard to the latter point, social desirability must not be a critical factor in answering the item.

\(^{103}\) Items that are too difficult, either because respondent lacks the ability or knowledge to answer the item, are likely to result in guessing, which detracts from reliability. Likewise, items that are too easy, because the vast majority of respondents report the same response, also detract from reliability due to their impact on variance.

\(^{104}\) Although oral and written questions date back hundreds of years, the reliability of such tests has been shown to be considerably lower than objectively scored tests (e.g., multiple choice, true-false) (Hopkins, 1998).
Two general methods can be used to estimate reliability based upon a single test administration. The first method is an extension of the alternate form method, wherein the period between the two administrations is eliminated by administering both forms at the same time. Known as the *split-half method*, a single test is divided into two parts (subtests) in such a way that the subtests are as nearly parallel as possible in terms of length, content, difficulty of items, and so on. Several strategies have been devised for dividing the test into half, including assigning odd numbered items to form 1 and even numbered items to form 2, randomly assigning items to the two forms, and matching items for content then assigning one to form 1 and the other to form 2. The two subtests are then scored separately and a correlation coefficient is computed between them. However, since this correlation is likely to underestimate the true parallel test reliability because longer tests are more reliable than shorter tests, one can employ the Spearman-Brown prophecy formula (Lord & Novick, 1968; Crocker & Algina, 1986) to estimate the reliability for the full-length test. Accordingly, the projected full-length reliability is given by $\rho_{XX'} = \frac{k\rho_{YY'}}{1+(k-1)\rho_{YY'}}$, where $\rho_{XX'}$ represents the projected reliability, $\rho_{YY'}$ represents the subtest reliability, and $k$ represents the ratio of the length of the full-length test to the subtest. Hence, if the split-half correlation was 0.7 then the projected full-length reliability would be approximately $2(0.7)/[1+0.7(2-1)] \approx 0.824$. It is important to note that the Spearman-Brown prophecy formula requires that the split-halves are parallel, additional items function in the same way as those of the subtests, and the length of the new test is such that fatigue, boredom, or other factors do not affect responses. The greater the violation of these assumptions, the greater the deviation between the projected and observed scores.

Owing to its ease, correcting (stepping-up) the split-half reliability by the Spearman-Brown prophecy formula has great appeal and dates back nearly a century. However, as
Brownell (1933) aptly demonstrated, the parallel reliability estimated on the basis of the split-half method suffers from a significant drawback. Namely, there are many ways of dividing a test into halves. In fact, if a test has an even number of questions, there are \( k! / \{2[(k/2)]!\} \) different ways of dividing \( k \) items into halves (Crocker & Algina, 1986), whereas if a test has an odd number of questions, there are \( k! / \{2[(k/2)]!\} \) different ways of dividing the \( k \) items into \((k/2-0.5)\) and \((k/2+0.5)\) halves. For example, a test comprised of 10 items can be split into \( 10! / \{2[(5)!\} = 126 \) possible distinct halves. However, if a test was comprised of 49 items, one could be split into \( 49! / (24!*25!) = 63,205,303,218,876 \) possible distinct halves.

Given the large number of potential split-halves a test may be divided into and the fact that these possibilities are likely to yield a range of reliability estimates. Ideally, one would like to compute the exact parallel test reliability such that it reflects the central tendency of all the possible split-half tests. However, even with the impressive amount of computing power available to people today, such an endeavor would consume massive amounts of resources for little gain. Fortunately, an alternative method arises out of sampling theory. As was previously proven, one can estimate a population parameter with great precision given a modest sample size that is randomly sampled from the population. Hence, if one were to repeatedly sample with replacement (i.e., bootstrap) the stepped-up reliability computed from the population of all possible split-half tests, a sampling distribution would emerge from which one could estimate the mean (or median) and construct a nonparametric CI based on appropriate lower and upper bound percentiles.

To demonstrate this method, a Monte Carlo was conducted in which 60 random variables (representing items) with a population size of 1,000,000 were simulated so as to have a random (unstructured) correlation matrix with an average correlation of 0.1. Random means and variances were assigned to each variable such that the expected value and the
variance for the 60 items were approximately equal to \( N(100,1) \) and the folded normal distribution \( |N(25,2.25)| \), respectively. Furthermore, since simulating an unstructured correlation matrix often results in a negative definite matrix, the matrix was corrected by setting the negative eigen values to zero and standardizing the reconstituted correlation matrix. This yielded a positive definite matrix whose values were close to those of the original matrix. Unfortunately, the correction also produced a small negative bias, relative to the 0.1 correlation, resulting in an average population correlation of 0.0764 and subsequently a sample correlation of 0.0766.\(^{105}\)

A sampling distribution was created by computing 10,000 random stepped-up split-half reliability estimates from the 59,132,290,782,430,712 possible distinct halves. Give the use of random assignment, the equivalence of the distinct halves (Rodriguez, 2007) was tested to determine whether the method violated the three assumptions underlying parallel reliability. These tests revealed that the average properties of each subtest were nearly identical: means (99.8806 versus 99.8802), variances (2.6893 versus 2.6920)\(^{106}\), and correlations (0.0766 versus 0.0766). These results clearly demonstrate that it is safe to treat the bootstrap reliability method as a measure of parallel test reliability.

The Spearman-Brown prophecy formula was used to estimate the expected reliability of the full test (60 items) based on the average correlation for the simulated sampling distribution. Not surprisingly, this estimated reliability, \( 60 \times 0.0764/(1+59 \times 0.0764) \approx 0.8323 \), was very close to the bootstrap reliability of 0.8160. However, unlike the parallel test reliability estimate, a CI for the expected reliability cannot be estimated from knowledge of

\( ^{105} \) Note, the negative bias that was observed was a function of the correction performed on the simulated correlation matrix. Since in the real-world correlation matrices are generally positive definite, or at the very least do not possess as many negative eigen values as are generally found in simulated correlation matrices, the bootstrap reliability method described herein would not be affected by such a negative bias.

\( ^{106} \) Since the average variance for each of the 30 items was 25 then the total variance of the composite is equal to the sum of all the variance and covariances divided by 30\(^2\), which is \( [30 \times 25 + 30 \times 29 \times 0.0766 \times 25]/30^2 = 2.6845 \).
the average sample correlation. It is in this regard that the true power of the bootstrap reliability method is realized. By obtaining the 2.5 and 97.5 percentiles from the sampling distribution, one can construct a 95% CI, which in this example was (0.7074, 0.8969).

It is worth noting, although the Pearson correlation between split-half subtests was used in this example, one could just as easily have utilized the Spearman-rank correlation and computed the median of the sampling distribution to attain a completely nonparametric bootstrap reliability estimate. Moreover, because the CI was constructed using percentiles rather than the standard error, it was not symmetric about the bootstrap reliability estimate, which is consistent with expectation. As a point of comparison, confidence intervals about a correlation are never symmetric, unless the correlation is zero. Furthermore, this method can be expanded to enable one to generalize to the population reliability parameter (rho) by performing a second bootstrap on the sampling unit. That is, the double-bootstrap reliability method randomly samples both the sampling unit and the sampling distribution of all possible split-half subtests, thereby accounting for sampling error and the content domain.

The ability to construct a CI for the reliability estimate places this method at a premium to all those that came before it and even to those yet to be discussed. Furthermore, this ability has clear implications for Summative Confidence. As was previously argued, it would behoove evaluators to remove measurement error from the variance estimates used to construct a CI on a summative conclusion because doing so would result in a tighter CI. However, low reliability also connotes low validity. Hence, it is unwise to use unreliable measures in order to shrink the CI. Moreover, since the primary purpose of Summative Confidence is to inform evaluators and decision-makers as to the quality of an evaluative conclusion, a degree of conservatism is warranted. Hence, rather than employ the reliability estimate or its lower bound, one should utilize the upper bound estimate when computing
reliability adjusted variance estimates. Although this may seem counterintuitive on the surface, it is consistent with the fear expressed by notable psychometricians (Guttman, 1945; McDonald, 1999) who warned that if reliability is underestimated, it may lead to an overestimation of reliability that is corrected for attenuation (i.e., removal of measurement error). Along this line of thought, employing the upper bound reliability will ensure that a conservative CI is produced to aid decision-making. That is, if the CI for the summative conclusion is informative despite being computed under conservative assumptions, then the CI computed under liberal assumptions will also be informative. Similarly, using the upper bound reliability estimate will produce a higher Summative Confidence reliability $\rho_{SC}$ which would indicate that greater faith can be placed in the reliability of the summative conclusion.

Due to the tremendous number of ways in which a test can be divided into unique subtests, the split-half reliability method fell out of favor with the scientific community once Cronbach (1951) published his seminal paper on coefficient alpha ($\alpha$). Without doubt, coefficient alpha is currently the most popular method for measuring internal consistency. Unlike the previous methods, which utilize the correlation coefficient, alpha relies on an analysis of the variance-covariance item structure. Under a very restricted set of assumptions, alpha can be shown to equal $\rho_{XX'}\geq[k/(k-1)]*(1-\sum_{i=1}^{k}\sigma_i^2/\sigma_X^2)\equiv\alpha$, where $\sigma_i^2$ and $\sigma_X^2$ represent the individual item and total composite variances, respectively. Alternatively, the standardized coefficient alpha can be shown to equal $\alpha\equiv k\bar{p}/[1+(k-1)\bar{p}]$, where $\bar{p}$ represents the average correlation coefficient between items. One cannot help but notice the similarity of this formula to the Spearman-Brown prophecy formula, which is not surprising given that both arise from CTT. Another way of expressing alpha occurs when items are dichotomously scored. Known more commonly as the Kuder-Richardson 20 or KR$_{20}$, after the psychometricians (Kuder & Richardson, 1937) who introduced it, coefficient alpha
reduces to $KR_{20} \equiv \frac{k}{(k-1)} \left( 1 - \sum_{i=1}^{k} p_i q_i / \sigma_i^2 \right)$, where $p_i$ represents the proportion of people who answered the $i^{th}$ item correctly, $q_i$ represents the proportions who answered the item incorrectly, and $pq_i$ represents the Bernoulli variance for the item.

As was the case with the aforementioned split-half reliability methods, coefficient alpha and $KR_{20}$ are unrelated to stability of scores over time or the equivalence of scores across alternate forms of a test. Instead, they are characterized as *coefficients of precision*. Examination of both statistics and their use leads one to several revelations and concerns. First, unless all of the items are perfectly parallel, coefficient alpha and $KR_{20}$ are lower bound estimates for the parallel reliability of a composite (i.e., the full-length test). Second, the greater the covariance, the higher the value of the coefficient of precision will be. That is, a positive association exists between the average correlation between items and the total test reliability, where a higher average correlation begets a higher reliability coefficient. Third, although higher coefficients of precision are frequently interpreted to connote that the test is unidimensional (i.e., the items represent a single common factor), this is not so because high covariance can be produced by more than a single common factor. Therefore, the coefficient of precision for a multidimensional construct is not easily interpreted, providing it can be interpreted at all. Fourth, although by definition the range space for reliability is zero to one, neither of the coefficients of precision have a known lower bound. Chen and Krauss (2004) provide an example in which coefficient alpha was -1.12. Examination of their data by the present author revealed a negative average item correlation of -0.0802. This in fact is not surprising since the only time the coefficient of precision can be negative is when the composite variance is smaller than the sum of the item variances, which only occurs when the covariance is negative. Lastly, although not intuitively obvious, both coefficients of precision underestimate the true reliability when a continuous latent construct is skewed or is...
measured by less than a 6-point ordinal scale (Zumbo, Gadermann, & Zeisser, 2007). While the first three issues are widely known and included in many contemporary measurement theory textbooks, knowledge of the latter two issues has yet to be widely disseminated. Moreover, the fact that the coefficient of precision is a lower bound estimate of the parallel reliability, has the potential to yield negative estimates, and consistently underestimates the reliability of ordinal variables calls into question its value, particularly with regard to which reliability estimator one should utilize in a Summative Confidence analysis.

More recently, McDonald (1999) has proposed an alternative reliability coefficient, known as coefficient omega (ω), based upon the common factor model. Like alpha, omega estimates the reliability of a set of items. However, omega relies on the common factor loadings produced by a factor analysis. A factor loading represents the correlation between an item and the latent construct. Hence, analogous to r-squared, the square of a factor loading (λ) (i.e., λ²) represents the shared variance between the item and the latent construct. That is, lambda-squared represents the true score variance of the item. Therefore, following from the definition of classical reliability, omega can be expressed by the ratio of the square of the sum of all the factor loadings (i.e., the shared variance of the latent construct with all the items or true score variance for the test) to the composite variance. Mathematically, omega is expressed as ω=(∑ₖλᵢ)²/σ² = (∑ₖλᵢ)²/[(∑ₖλᵢ)²+(∑ₖΨᵢ²)], where Ψ represents the error variance. It is important to note that if all of the items that comprise a test are parallel measures, then α=ω; otherwise, α<ω, providing the test is unidimensional.107 Hence, alpha is

107 Zinbarg, Revelle, Yovel, and Li (2005) proved that for multidimensional tests, coefficient alpha may be less than, equal to, or greater than coefficient omega when the observed scores of a test can be decomposed into four parts: a general factor (i.e., a factor common to all the items), group factor (i.e., a factor that is only common to some of the factors), specific factor (i.e., a factor unique to each item), and random error. They noted that “increases in the variance of either the general factor loadings and/or the group factor loadings will tend to depress the value of [alpha] to [omega]...[whereas] when there is little variability in the general factor loadings and relatively strong group factors, alpha will tend to be greater than [omega]” (pp. 128-129).
a special case of omega. However, like alpha, omega assumes that observed scores for the random variables are continuous.

Unfortunately, the need to employ ordinal measures in evaluation far surpasses the ability to employ continuous measures. One of the methods currently utilized to account for the non-continuous nature of ordinal variables is via the use of polychoric correlations—a correlation between the two latent normal variables that underlie the observed ordinal variables. Fortunately, several software platforms exist for computing a polychoric correlation. Furthermore, Zumbo, Gadermann, and Zeisser (2007) proposed ordinal $\alpha$ and $\omega$ wherein the previous equations for $\alpha$ and $\omega$ are applied to the polychoric correlation matrix rather than the raw scores. Monte Carlo analysis of ordinal $\alpha$ by Zumbo et al. revealed that it consistently estimated the true reliability “regardless of the magnitude of the theoretical reliability, the number of scale points, and the skewness of the scale point distributions” (p. 21). Therefore, three methods exist for handling ordinal data: ordinal $\alpha$ and $\omega$, and the nonparametric bootstrap reliability method.
Evaluators work at the center of all disciplines. Scarcely a field can be found to which evaluation has not already made a significant contribution or has the potential to make a significant contribution. Evaluation is best classified as a transdiscipline—such as statistics, logic, management, information science—in that its subject matter is the study and improvement of other disciplines (Scriven, 1991). According to Scriven (1991), evaluation is also an emerging science in that it involves the production of knowledge. However, he points out, it would be extremely misleading if evaluation was defined as the provision of practical information since in addition to the collection and summary of data, evaluation entails the identification and utilization of relevant values and standards for the purpose of synthesizing the information collected into evaluative conclusions.

Yet, in the opinion of this author, evaluation, in part, falls short of being able to claim the mantle of a full-fledged scientific discipline because it has not been able to develop a method for evaluating the credibility of its conclusions. It has long been hoped that metaevaluation would fill this void. However, Wingate (2009) calls into question the use of metaevaluation as an *ex post facto* methodology. This revelation should come as no surprise to experienced evaluators who have known for a long time that “the reliability of evaluations is a largely unknown quantity…The few data on [the replications of evaluations]…make clear that reliability, once you factor out spurious effects such as shared bias, is *not high* [emphasis added]” (Scriven, 1991, p. 310).
One of the objectives of Summative Confidence is to align evaluation more closely with other scientific disciplines. Specifically, it seeks to provide evaluators with a mechanism for mathematically computing the degree of precision or reliability associated with their evaluative conclusions. This is an ambitious goal and perhaps the only way in which the subjectivity of the evaluator may be held in check. Yet, in return for the considerable effort necessary to conduct the analysis, evaluators may benefit from the protection the analytical result provides against the charge that their conclusions are subjective or amount to nothing more than professional opinions—assuming, of course, that the analysis indicates that their summative conclusion is likely to replicate in a future study.

This chapter will briefly summarize the highlights of the method and how it should be implemented. The second section will discuss its implications for planning better evaluation studies. Next, a reflection on the contributions to evaluation, statistics, and measurement theory will be presented. Lastly, the chapter will close with some final thoughts for evaluators, researchers, and students who wish to pursue the study and application of Summative Confidence farther, yet find the computations forbidding.

The Summative Confidence Algorithm

This section presents the master formula for Summative Confidence, highlights how the terms of the formula are linked to the 11 factors of interest discussed in the introduction, outlines the algorithm that should be followed to properly conduct the Summative Confidence analysis, and discusses the implications of Summative Confidence for evaluation studies. At this stage, it is hoped that readers will have a deeper appreciation for the issues involved in conducting an evaluation. Evaluation is not synonymous with simply rendering a professional opinion following the examination of the data collected, at least it does not have to be. It has the potential to evolve into a scientific discipline in which its conclusions are
accountable to the quantity and quality of the data collected and evaluators are held accountable for the quality of their conclusions and professional opinions.

While the derivation of the Summative Confidence theory has been laborious, the final formula is remarkably simple. For \( k \) dependent variables, the sampling error variance once measurement error is removed is equal to the weighted sum of the predicted true score error variances and covariances for each random variable, which is given by

\[
\text{Var}(\bar{w}x) = \frac{1}{k^2} \sum_{i=1}^{k} w_i R_{XX}^2 \left[ 1 - \frac{n_i}{N_i} \right] S_i^2 + 2 \sum_{i<j} w_i w_j \left[ 1 - \frac{n_i}{N_i} \right] R_{XY} R_{XX} R_{YY} s_{ij}^2
\]

Notice that this formula is remarkably similar to the one that was logically deduced in Chapter 2. The only noteworthy differences is the incorporation of the reliability terms (the parallel test reliability \( \rho_{XX}^2 \), and the reliability indices \( \rho_{XX} \) and \( \rho_{YY} \)). However, since the sample size for two variables must be the same for one to compute a covariance estimate, \( \sqrt{n_i n_j} \) can simply be replaced with \( n \). In situations in which \( n_i \neq n_j \), one can either (1) set the covariance to zero for cases where the units of analysis disagree or (2) extrapolate the covariance based upon the observed data. Furthermore, the formula presented above is restricted to the condition in which \( N_i = N_j \). When the population sizes differ, the full formula for \( \text{Var}(\bar{w}x) \). Note, although the remainder of this section focuses on constructing a Summative Confidence interval, a perfectly viable alternative is to compute the Summative Confidence reliability \( \rho_{SC} \), which is equal to the ratio of the true score error variance to the observed score error variance, \( \rho_{SC} = \frac{\text{Var}(\bar{w}x)}{\text{Var}(\bar{w}x)} \).

Comparison of the terms found in the formula to the list of factors that this study set out to investigate reveals that it accounts for 8 of the 11 factors, with the final factor accounted for by alpha in the CI construction process. The number of variables included in an evaluation is accounted for by the term \( k^2 \). The variance associated with each variable is
accounted for by the term $S^2_i$. The correlation among the $i^{th}$ and $j^{th}$ variables is accounted for by the term $R_{ij}$, which is part of the covariance denoted by the term $R_{i,j}S_j$. The sample size of each variable is accounted for by the term $n_i$. The sampling error is accounted for by the term $1-n_i/N_i$. The weights assigned to each variable are accounted for by the terms $w_i^2$, $w_i$, and $w_j$. And finally, the measurement error associated with a variable as well as the impact of a standard are accounted for by the terms $R^{2}_{XX}$ and $R^{2}_{YY}$. 

Application of the method is straight-forward when the normality of the composite variable can be assumed. Given a continuous or a discrete random variable with more than 15 ordinal categories (Jöreskog & Sörbom, 1999), the results of the Monte Carlo simulations conducted in this study revealed that the sampling distribution of the composite variable is asymptotically normally distributed, even when the composite is nonnormal. With respect to the composite mean across all the sampling units (i.e., the grand mean), this means that a $100(1-\alpha)\%$ CI is given by $\bar{X} \pm z_{1-\alpha} \sqrt{\text{Var}(\bar{w}X)}$. Note, in the case of the grand mean, one does not need to compute the expected value of all the predicted scores since it is equal the expected value of the observed scores. The expected value of the predicted score is only needed when one is interested in constructing a CI for an individual observation found in the vector associated with the composite predicted true score.

Unfortunately, many of the variables evaluators collect are unlikely to be continuous or have more than 15 ordinal categories. In fact, given the effort required to compute the inverse cdf of each threshold boundary, one is not likely to elect to transform all the dissimilar distributions into a common distribution with more than 15 categories. One solution for reducing the workload may be to apply the cut-scores to the original distribution and utilize a software package to compute the class probabilities from the sample data. Unfortunately, this will undoubtedly introduce error into the estimated probabilities. Given a
large enough sample size or sampling frame, or absent knowledge of the underlying probability distribution and its parameters, estimating the class probabilities from the sample data may be a reasonable course of action.\textsuperscript{108}

Regardless of the transformation selected, one will need to ensure that the new random variables all share a common distribution. The simplest course of action is to utilize dichotomization. Clearly then all of the transformed variables would conform to a Bernoulli distribution and so a composite can be generated by summing all the variables. However, such a transformation would discard a great deal of information contained in the original distributions. Hence, polychotomization may be a better option, albeit one that is analytically more challenging, particularly when one or more of the observed variables have less classes than the desired common distribution. Whenever the latter scenario is encountered, evaluators will be faced with two choices. The first is simply to polychotomize all the variables to the variable with the lowest number of classes. Hence, if one of the observed random variables is binary, this option would require one to dichotomize all of the variables, which could result in the loss of a lot of information. Alternatively, one could recode the lowest observed value to the lowest scale point of the transformed variable (assuming a direct transformation is appropriate) and the highest observed value to another scale point (perhaps the highest value) of the transformed variable. The variance of the transformed variable is equal to the variance of the observed variable times the square of the difference between the lowest scale point and the highest value assigned to the transformed variable.

\textsuperscript{108} Naturally, one may wonder, given the relative computational simplicity of setting cut-scores on the original distribution and then computing the class probabilities via a software package, why would one ever need to employ the inverse cdf method? One example of an instance in which this route is necessary occurs when one needs to compute a Summative Confidence on the summary data provided in a published evaluation report. In such cases, one is not likely to have access to the original data. However, one can use the reported sample statistics as estimates of the parameters of the original distribution, whose family can generally be deduced based on historical evidence. Hence, given the summary sample statistics and an assumed distribution, one can compute the inverse cdf for each threshold to obtain the class probabilities of a transformed variable.
Going from a $k$-class observed variable to an $m$-class transformed variable where $2<k<m$ requires computing a weighted variance estimate based on the difference in values between the transformed and observed classes. Specifically, the variance of the transformed variable will equal $(n-1)^{-1}\left\{\sum_{i=1}^{n}(x_i+c_i)^2-\left[\sum_{i=1}^{n}(x_i+c_i)\right]^2/n\right\}$, where $c_i$ denotes the difference between the observed value $x_i$ and the transformed value $y_i=x_i+c_i$.

Given the choice of using polychotomous transformations, one could assume that equidistance exists between the values of each transformed variable (although this is not indicated), which would connote that the variables are measured on an interval scale. Such an assumption would greatly simplify the remaining analyses because one can now compute the mean and variance of the transformed constituent variables and proceed as usual—i.e., compute the composite mean score $\bar{X}$ and error variance $\text{Var}(\nu\bar{X})$. From the perspective of this author, however, it is difficult to see the conditions that would justify making this assumption. Thus, the equidistance assumption should be clearly stated when reporting the results of a Summative Confidence analysis.

Alternatively, one can construct a nonparametric discrete CI on the transformed random variables. As a result, the composite grand mean would need to be replaced with the composite grand median, which can be computed by finding the median of all the medians of the transformed variables. A series of $100(1-\alpha)\%$ nonparametric discrete CIs must then be computed—one for each transformed random variable—as described in Chapter 4. Although this analysis is somewhat complex to perform manually, it can be easily performed by the SAS code provided in Appendix H. Once a nonparametric CI is obtained for each transformed variable, one can estimate the error variance for each variable using $\sigma^2_X=[(UB-LB)/2\bar{z}_{\alpha/2}]^2$, where UB and LB denote the upper and lower boundaries of the CI. By substituting each $\sigma^2_X$ term for its corresponding $\sigma^2_i/n_i$ term, one can then estimate
Note, when estimating the nonparametric \( \text{Var}(\bar{w} \bar{x}) \), one should always use nonparametric reliability and correlation estimates. Lastly, it is advisable to compare the nonparametric CI on the grand median with the parametric CI on the grand mean (i.e., the one for which equidistance is assumed). If the two CIs are very similar then one can use the parametric CI since most readers will likely find it easier to understand. If the CIs are different, particularly if they lead to different conclusions, then one should use the nonparametric CI. Fortunately, based on the results of the Monte Carlo simulations conducted in this study, chances are that the nonparametric CI will be tighter than its parametric counterpart.

A number of correlation coefficients can be used to estimate the population correlation between two variables. If the transformed variables are continuous then one should use the Pearson correlation to estimate \( R_{ij} \); if the transformed variables are discrete with more than two classes then one should use the Spearman-rank correlation; and if the transformed variables are dichotomous then one should use the Phi correlation. However, if it is reasonable to assume that the latent distributions of two dichotomous variables are bivariate normal then one should use the tetrachoric correlation, whereas if the latent distributions of two polychotomous variables are bivariate normal then one should use the polytropic correlation.

Similarly, a number of reliability coefficients can be used to estimate the parallel test reliability. When a test (generically speaking) is administered on more than one occasion then one should use the test-retest correlation, making sure, of course, to select an appropriate correlation coefficient. There are times, however, when the alternate form and test-retest with alternate form may be more suitable. However, since Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001) is able to compute parallel
coefficients whilst removing unwanted source of variability, it is better to employ the
generalizability coefficient rather than the aforementioned coefficients. Moreover,
Generalizability Theory enables one to compute the index of dependability (Brennan &
Kane, 1977), which should be used whenever an absolute decision needs to be made for a
critically-important variable with a fixed standard. That is, the index of dependability reflects
the reliability of correctly classifying a score with respect to a cutscore (standard).

When a test is only administered once, undoubtedly, many will choose to compute
coefficient alpha. However, this statistic is only appropriate when all the constituent
variables of the test (e.g., test items) are continuous and (homogenous) unidimensional.
However, coefficient omega is generally a better choice when the constituent variables of a
test are continuous. When the constituent variables of the test are dichotomous then one
should use KR20, whereas if the constituent variables of the test have more than two classes
and if, and only if, the latent distributions of the constituent variables of the test can be
assumed to be normal, then one should compute ordinal alpha or omega. However, this
author prefers to use the bootstrap parallel test coefficient because it enables one to
compute a CI on the reliability estimate. Moreover, in the case of discrete variables, one can
utilize the Spearman-rank correlation to compute a nonparametric reliability estimate and CI.
Given the fact that many evaluators will choose to use polychotomization in order to
preserve as much of the original information as possible, this method enables one to
compute the parallel reliability of ordinal variables.

Lastly, a few words are warranted on a few of the factors. Undoubtedly, there may be
instances in which the reliability of a factor is unknown and cannot be computed from the
available data. For example, a single indicator may be used to measure a factor at one point
in time. In such instances, one can either use a historical estimate or one found in the
literature or simply set the reliability to unity, which will result in a more conservative CI. Likewise, if the population size for a factor is unknown then one should assume that it is infinite (or extremely large) in which case the finite population correlation will also equal unity. The selection of weights, however, requires more care. This author prefers to conduct evaluations within a collaborative framework. Hence, weights, values, and standards are either solicited from key stakeholders or they are provided with the opportunity to accept, reject, or refine the estimates recommended by the evaluator or the literature.

Hypothetical Example of Summative Confidence

To illustrate Summative Confidence, let us return one last time to the hypothetical tenure review example. As was illustrated in Figure 1, the tenure review committee based its recommendation for tenure on 20 values: 7 microvalues organized under the Research macrovalue, 3 microvalues organized under the Teaching macrovalue, 1 microvalue representing the Accolades macrovalue, 6 microvalues organized under the Service macrovalue, 2 microvalues organized under the Academic Interests macrovalue, and 1 microvalue representing the Collegiality macrovalue. Furthermore, the macrovalues were divided into critically important (Research and Teaching) and non-critically important (Accolades, Service, Academic Interests, and Collegiality), where the critically important macrovalues accounted for 60% of the decision-space and the remaining macrovalues accounted for 40% of the decision-space.

In the example to follow, several simplifying assumptions were made. First, a hypothetical pdf was assigned to each of the microvalues. In a real-world evaluation, the pdf of each variable would be determined via induction from historical or actual data based on how well the data meets the properties of the hypothesized pdf. Herein, the distributions were determined via deduction based upon the definition of the hypothesized pdf and scale
used to measure each variable. Second, all the microvalues were assumed to be independent and sampled from an infinitely large population even though both assumptions are certain to be incorrect given that all of the microvalues are dependent upon the same candidate and the tenure review committee could not have possibly been sampled from a large population. Third, since the summative conclusion is a binary variable, dichotomous transformations for the 20 microvalues was deemed to be an appropriate method for establishing a common distribution. Moreover, the \( i^{\text{th}} \) transformed random variable \( Y_i \) was computed from the hypothesized pdf and \( \tau \) for the corresponding microvalue \( X_i \), wherein \( Y_i = 1 \) if \( X_i > \tau \) and zero otherwise. Fourth, a consensus was reached by the tenure review committee for the cut-score \( \tau_i \) for each microvalue, wherein values lower than \( \tau_i \) demark poor performance and vice versa. Fifth, the composite variable was assumed to be asymptotically normally distributed. Sixth, the majority of the data was assumed to have been abstracted from the candidate’s application package. As such, in the majority of instances, the sample size was equal to unity or the size of the tenure review committee (herein, assumed to equal 10). And lastly, with a few noted exceptions, measurement error was assumed to equal zero.

Table 16 presents the values used in this analysis. The mean and variance for the \( i^{\text{th}} \) transformed random variable \( Y_i \) was equal to \( \mathbb{E}(Y_i) = 1 - F_{X_i}(\tau_i) \) and \( \text{Var}(Y_i) = F_{X_i}(\tau_i)[1 - F_{X_i}(\tau_i)] \). A detailed illustration of how the means and variances were computed for the Research macrovalue may be found in the Transformation section (Chapter 4). In light of the values presented in the table, the expected value of the composite can be computed by \( \mathbb{E}(\rho Y) = \sum_{i=1}^{k} w_i \mathbb{E}(Y_i) \), where \( w_i \) and \( \mathbb{E}(Y_i) \) denote the weight and expected value for the \( i^{\text{th}} \) macrovalue. Therefore, \( \mathbb{E}(Y) = 0.3(0.7149+0.6672+0.7181+0.2500+0.7500+0.200+0.1694)/7 + 0.3(0.9192+0.9502+0.9004)/3+0.1(0.6321)+0.1(0.8488+0.9522+0.8088+0.5940+0.9084+0.6570)/6+0.1(0.7500+0.6300)/2+0.1(0.7716) = 0.7145. \)
### Hypothetical Case Based on the Tenure Review Example

<table>
<thead>
<tr>
<th>Research ($w_7$=0.3)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1. Number of publications</td>
<td>$Poi(6)$</td>
<td>4</td>
<td>1</td>
<td>0.7149</td>
<td>0.2038</td>
</tr>
<tr>
<td>X2. Number of presentations</td>
<td>$Poi(10)$</td>
<td>8</td>
<td>1</td>
<td>0.6672</td>
<td>0.2221</td>
</tr>
<tr>
<td>X3. Quality of publications</td>
<td>$N(2,0.75)$</td>
<td>1.5</td>
<td>6</td>
<td>0.7181</td>
<td>0.2024</td>
</tr>
<tr>
<td>X4. Quality of conference presentations</td>
<td>$U(0,100)$</td>
<td>75%</td>
<td>10</td>
<td>0.2500</td>
<td>0.1875</td>
</tr>
<tr>
<td>X5. Impact of publications</td>
<td>$Mult(1;“H”=0.35, “M”=0.40, “L”=0.25)$</td>
<td>L</td>
<td>10</td>
<td>0.7500</td>
<td>0.1875</td>
</tr>
<tr>
<td>X6. Impact of conference presentations</td>
<td>$Mult(1;“5”=.05,“4”=.15,“3”=.25,“2”=.35,“1”=.20)$</td>
<td>3</td>
<td>10</td>
<td>0.2000</td>
<td>0.1600</td>
</tr>
<tr>
<td>X7. Change in the number/quality of publications/presentations</td>
<td>$Bin(6,0.25)$</td>
<td>2</td>
<td>10</td>
<td>0.1694</td>
<td>0.1407</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teaching ($w_2$=0.3)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X8. Quality of student evaluations</td>
<td>$N(4,6,1.0)$</td>
<td>3.2</td>
<td>35</td>
<td>0.9192</td>
<td>0.0742</td>
</tr>
<tr>
<td>X9. Teaching awards received</td>
<td>$Poi(3)$</td>
<td>0</td>
<td>1</td>
<td>0.9502</td>
<td>0.0473</td>
</tr>
<tr>
<td>X10. Amount of teaching workload</td>
<td>$Poi(8)$</td>
<td>4</td>
<td>1</td>
<td>0.9004</td>
<td>0.0897</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accolades ($w_3$=0.1)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X11. Number of non-teaching awards</td>
<td>$Poi(2)$</td>
<td>0</td>
<td>1</td>
<td>0.6321</td>
<td>0.2325</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service ($w_4$=0.1)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X12. Number of graduate advisees</td>
<td>$Poi(6)$</td>
<td>3</td>
<td>1</td>
<td>0.8488</td>
<td>0.1283</td>
</tr>
<tr>
<td>X13. Amount of external funding</td>
<td>$N(25,000,9,000,000)$</td>
<td>$20,000$</td>
<td>1</td>
<td>0.9522</td>
<td>0.0455</td>
</tr>
<tr>
<td>X14. Number of dissertation/theses committees served on</td>
<td>$Poi(8)$</td>
<td>5</td>
<td>1</td>
<td>0.8088</td>
<td>0.1547</td>
</tr>
<tr>
<td>X15. Number of university committees served on</td>
<td>$Poi(2)$</td>
<td>1</td>
<td>1</td>
<td>0.5940</td>
<td>0.2412</td>
</tr>
<tr>
<td>X16. Number of workshops conducted</td>
<td>$Poi(4)$</td>
<td>1</td>
<td>1</td>
<td>0.9084</td>
<td>0.0832</td>
</tr>
<tr>
<td>X17. Leadership roles within professional organizations</td>
<td>$Bin(3,0.30)$</td>
<td>0</td>
<td>1</td>
<td>0.6570</td>
<td>0.2254</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Academic Interests ($w_5$=0.1)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X18. No shift in interests</td>
<td>$Bin(1,0.75)$</td>
<td>0</td>
<td>10</td>
<td>0.7500</td>
<td>0.1875</td>
</tr>
<tr>
<td>X19. Importance of interests to profession</td>
<td>$Mult(1;“5”=.15,“4”=.25,“3”=.23,“2”=.20,“1”=.17)$</td>
<td>2</td>
<td>10</td>
<td>0.6300</td>
<td>0.2331</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collegiality ($w_6$=0.1)</th>
<th>Hypothesized Pdf</th>
<th>$\tau^1$</th>
<th>$n^2$</th>
<th>Mean$^3$</th>
<th>Variance$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>X20. Collegiality with staff, students, and faculty</td>
<td>$N(4,1,0.65)$</td>
<td>3.5</td>
<td>50</td>
<td>0.7716</td>
<td>0.1762</td>
</tr>
</tbody>
</table>

---

1 The $\tau^i$ cut-score for the X random variable that is used to generate a dichotomous transformation, wherein the $i^{th}$ transformation is equal to $Y_i ~ Bin[1,1−F_X(\tau_i)]$.

2 Although clearly incorrect, for the sake of simplicity, assume that the sample size $n$ is drawn from an infinite population $N$.

3 The expected value of the transformed variable is equal to $E(Y) = 1−F_X(\tau_i)$.

4 The variance of the transformed variable is equal to $Var(Y) = F_X(\tau_i)[1−F_X(\tau_i)]$. 


Similarly, suppose all the variables where \( n=1 \) were measured with no measurement error—not an entirely unreasonable assumption given that the data were abstracted from the candidate’s curriculum vita—whereas the reliability of the other variables was 0.8. The sampling error variance is then given by \( \text{Var}(\bar{wY})=k^{-2}\sum_{i=1}^{k}w_{i}^{2}R_{XX}^{2}/n \), where \( R_{XX}^{2} \) denotes the square of the reliability coefficient for the \( i^{th} \) macrovalue. Since, organizationally, microvalues are located within their corresponding macrovalue, the error variance must be accumulated in two steps: first within each macrovalue and then across macrovalues. Thus, for the Research macrovalue, the predicted true score error variance is equal to \( \text{Var}(\bar{wY})=[0.2038+0.2221+0.8^{2}(0.2024/6+0.1875/10+0.1600/10+0.1407/10)]/7^{2} \approx 0.01 \).

In a similar fashion, the predicted true score error variances for the other macrovalues can be computed to equal 0.0154 for Teaching, 0.2325 for Accolades, 0.0244 for Service, 0.0067 for Academic Interests, and 0.0023 for Collegiality. The predicted true score error variance for the evaluand then is equal to the sum of the weighted error variances for the 10 macrovalues divided by 100 (i.e., the square of the number of macrovalues), i.e., \( \text{Var}(\bar{wY})=[0.3^{2}(0.01)+0.3^{2}(0.0154)+0.1^{2}(0.2325)+0.1^{2}(0.0244)+0.1^{2}(0.0067)+0.1^{2}(0.0023)]/10^{2} \approx 0.00005. \)

Finally, a two-sided \( 100(1-\alpha)\% \) CI is given by \( \bar{E}(wY) \pm z_{\frac{1-\alpha}{2}} \sqrt{\text{Var}(\bar{wY})} \). Therefore, a 95% CI for the composite mean is given by \( 0.7145 \pm 1.96*\sqrt{0.00005}=(0.7006,0.7284) \). However, since the tenure review committee is likely to only be interested in the lower bound estimate, in this example, it is more appropriate to compute \( \bar{E}(wY) - z_{\frac{1-\alpha}{2}} \sqrt{\text{Var}(\bar{wY})} \), which here is equal to \( 0.7145-1.645*\sqrt{0.00005}=(0.7029,1) \). Hence, if the policy for recommending candidates for tenure stipulated that only candidates with a composite score greater than 0.7 can be recommended, the tenure review committee can feel reassured that their decision to recommend the candidate for tenure would be reaffirmed if an independent review committee replicated their analysis under identical conditions.
The Implications of Summative Confidence

Several direct implications follow readily from the Summative Confidence master formula. First, reliability coefficients should always be used to shrink the size of the CI interval. Although the temptation may be to use lower bound reliability estimates in order to obtain a tighter CI, this practice can be misleading if the true reliability estimate is higher. Hence, good practice dictates that one should employ a more conservative reliability estimate (e.g., the upper bound) so as to protect against shrinking the CI unreasonably. The Monte Carlo simulations performed in this study clearly showed that the finite population correction should be applied whenever information is available on the population size. Hence, since the population size encountered in many evaluations are not always large, it would behoove evaluators to collect such information whenever there is reason to suspect that the sampling fraction will be greater than 5%. A more subtle implication arises from the covariance term. The lower the covariance is, then the tighter the CI will be. Therefore, the more independent variables one employs, the smaller the size of the CI interval. Similarly, the smaller the weight assigned to variables with high variance, the smaller the CI interval will be. To protect against the potential temptation of selecting weights so as to influence the outcome of the Summative Confidence analysis, however, it is best to collect such data prior to conducting the final analysis. That is, weights must be \textit{a priori}.

Not surprisingly, sample size, variance, and the number of variables in the evaluation contribute the most to the size of the CI interval. Clearly, the larger the sample size of each variable, the smaller the CI interval will be. In contrast, the lower the variance of each variable, particularly critically-important or highly weighted variables, the smaller the CI interval will be. However, the factor that may be of greatest importance is the number of constituent variables that comprise the composite. Given that the cumulative error variance
is divided by \( k^2 \), it is easy to see how adding more variables reduces the size of the CI interval. Naturally, this makes sense because the more information one has about the merit, worth, or significance of the evaluand, then the more precise one's summative conclusion should be. However, a note of caution has been raised by some evaluators, who are worried that some evaluators or clients will include trivial factors into an evaluation as a means of influencing the overall conclusion—see “numerical weight and sum” (Scriven, 1991; Davidson, 2005). This problem can be readily resolved via the use of a weighting system applied at the macrovalue level. That is, if a critically-important value is designated to account for only 25% of the variance of a composite score then it does not matter whether there is only a single microvalue that corresponds to the macrovalue or an infinite number of microvalues. No matter how many microvalues are introduced, the weight of the macrovalue will continue to be 25%. Hence, trivial variables can never over power other variables so as to impact an evaluative conclusion.

This remedy highlights the importance of organizational structure, which is not a factor that appears in the Summative Confidence formula. However, its impact on the precision of results should be evident. Increasing the number of microvalues that are organized underneath a macrovalue (recall Figure 1), clearly will improve the reliability with which that macrovalue is measured. Another means by which the “numerical weight and sum” dilemma can be avoided is by determining the contribution of each microvalue to a macrovalue. That is, trivial microvalues should have smaller correlations with unrelated macrovalues. Hence, if one weighed each microvalue by its correlation with the latent construct, like the lambda values produced by a confirmatory factor analysis (Brown, 2006), then one could further control the impact of trivial factors. However, the specific coefficient that would need to be incorporated into the master formula has yet to be derived.
A more subtle implication of Summative Confidence is that it can and should be used to plan future evaluations. That is, given the results of an evaluation or pilot study, one can examine questions regarding how to best plan a future evaluation so as to attain a specific level of confidence or reliability. The ability to compute the true score sampling error variance \( \text{Var}(\bar{w}_x) \) for the summative conclusion implies that one knows the amount of variability that each macrovalue and microvalue contributed to the estimate. Therefore, an evaluator can collect more data (i.e., increase the sample size) of microvalues whose contribution was higher than desired. In addition to influencing the error variance of the total composite by manipulating the sample size of constituent variables, an evaluator controls the number of microvalues per macrovalue and even the number of macrovalues. Clearly, the more microvalues nested within a macrovalue the lower the error variance for the macrovalue composite score, all other factors held equal. The same is also true of the number of macrovalues that comprise the composite score of the evaluand.

Contributions to Evaluation, Statistics, and Psychometrics

Among psychometricians, it is well-known that \textit{without reliability there can be no validity}, whilst among statisticians, it is well-known that \textit{without variance there can be no inference}. Likewise, among evaluators, it is well-known that \textit{without values and standards there can be no evaluation}. Clearly then the single greatest contribution of this study is the integration of statistical theory and psychometric theory for the purpose of application to the discipline of evaluation. While wide-scale adoption of Summative Confidence does not, in and of itself, have the ability to lift the emerging evaluation profession to the level of a scientific discipline—since evaluation has yet to develop and empirically validate theories independent of other disciplines—it puts the discipline on the road to this goal. Furthermore, it has the potential to raise the profile of evaluation among people who may equate it with nothing
more than a professional opinion, which is not to suggest that there is no value in professional opinions—there is. However, just as the medical profession has shifted away from experienced based practice to evidence-based medicine over the past 30 years (Gugiu & Gugiu, 2010), so too must evaluation begin to hold itself to a higher standard.

Such a shift will undoubtedly require the profession to hold evaluators accountable for the quality of their conclusion. Furthermore, it may even require clarification of the terms by which the profession defines itself. Evaluation is a transformative process whereby the underlying probability distributions of a series of pertinent random variables are transformed into a common distribution based on prespecified rules designed to evaluate the merit, worth, or significance of each variable prior to their synthesis into an evaluative conclusion or recommendation. Notice, this definition departs slightly from the one commonly accepted by evaluators, which was first introduced by Dr. Michael Scriven. Namely, “evaluation refers to the process [emphasis added] of determining the merit, worth, or value of something, or the product of that process” (Scriven, 1991, p. 139). Thus, another important contribution of this study is the explication of how dissimilar distributions, which are typical of the variables collected in evaluation practice, can be transformed into a common distribution that enables one to add apples and oranges in a meaningful way.

A nonparametric CI was derived specifically for ordinal variables, one based on discrete order statistics. Based on the Monte Carlo simulations, the method holds great promise for evaluators and researchers alike. Moreover, the ability to construct a CI signifies the ability to perform inferential tests. Therefore, in the future, the nonparametric CI method may usher in new statistical tests. Important contributions were also made to measurement theory. A bootstrap split-half reliability method was developed. Moreover, this method can be adopted so as to be able to compute a nonparametric reliability coefficient.
and CI, thereby enabling it to be used with ordinal data without having to assume an underlying normal distribution.

Besides these key contributions, a few minor, albeit still important, ones were also made. SAS code was written for the purpose of simulating correlated nonnormal data using the GLD. The only other programs that serve this purpose were written in programming languages that are not generally known to researchers outside of computer science. This method was used on a number of occasions. Lastly, a SAS function was programmed that enables one to compute the cdf of the multinomial distribution.

Future Research and Final Remarks

The theoretical ideas explicated in this document have evolved over the long period during which they were synthesized into a coherent method and approach. In fact, Summative Confidence today departs remarkably from this author’s original formulation from four years ago. Its evolution has been guided by an ever-growing appreciation for statistical and psychometric theory, with the solution to old problems and the emergence of new problems often tumbling out simultaneously. It is possible that inconsistencies still remain in the present method. Yet, even if it stands up to criticism, Summative Confidence will continue to evolve as evaluators begin to apply the method, statisticians propose more sophisticated statistical solutions to those presented herein, and psychometricians begin to integrate the numerous other measurement theories that have not called themselves to this author’s attention. The first place to begin, however, is to apply Summative Confidence to a real-world evaluation. It is only by putting theory into practice that one can truly verify the merits of a novel method. No doubt, evaluators will discover along the way that numerous issues and challenges require further development and refinement. How can one determine whether an evaluation is suitable for a Summative Confidence analysis? Can such an analysis
be conducted from the data generally published in evaluation reports? How can one account for variability in values or standards? Can Summative Confidence be employed to maximize precision given budgetary restrictions?

Today’s reader, coming to the new emerging methodology of Summative Confidence for the first time, no doubt finds it forbidding. As evaluators and methodologists become accustomed to its ideas and ways of treating data, this strangeness will pass as has been the case throughout history when a new idea or method was introduced. As the theory is put in different words by successive writers, it will gain further clarity and traction. As other statistical and psychometric theoreticians integrate new factors and theories, it will begin to cover the span of evaluation study designs. Finally, as software programs are developed to free evaluators from the complex analyses that are required, it will become inseparable from the evaluation theory and practice of the next generation.
REFERENCES


Healy, M., & Perry, C. (2000). Comprehensive criteria to judge validity and reliability of qualitative research within the realism paradigm. 3 (3).


253


APPENDIX A

Glossary of Key Terms
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>The degree of discrepancy between an estimated value and the actual value. This is not synonymous with precision because a result can be accurate but measured with low precision.</td>
</tr>
<tr>
<td><strong>Alpha</strong></td>
<td>The level of confidence set for a Summative Confidence analysis. Alpha is the Family Type I Error and refers to the probability of incorrectly concluding the Summative Confidence interval contains the true evaluative conclusion when in fact it does not.</td>
</tr>
<tr>
<td><strong>Alternative-form method</strong></td>
<td>The administration of two similar forms of a test to the same group of examinees within a very short period of time.</td>
</tr>
<tr>
<td><strong>Bar</strong></td>
<td>A performance standard in which failure to meet the benchmark results in the failure of the entire evaluand.</td>
</tr>
<tr>
<td><strong>Bernoulli distribution</strong></td>
<td>Describes a discrete probability distribution produced by an experiment, known as a Bernoulli trial, that results in one of two possible outcomes.</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>A sample statistics is said to be biased if the expected value of the statistic averaged over all the samples is not equal to the population parameter.</td>
</tr>
<tr>
<td><strong>Binomial distribution</strong></td>
<td>Describes a discrete probability distribution produced by an experiment in which the successful outcomes resulting from a set of Bernoulli trials are summed.</td>
</tr>
<tr>
<td><strong>Central Limit Theorem</strong></td>
<td>According to this theorem, if the sum of the variables has a finite variance, then it will be approximately normally distributed.</td>
</tr>
<tr>
<td><strong>Coefficient alpha</strong></td>
<td>The correlation used to estimate the internal consistency of items continuously scored. It represents an estimate of the lower bound estimate of the parallel test reliability.</td>
</tr>
<tr>
<td><strong>Coefficient of equivalence</strong></td>
<td>The correlation obtained from an alternative-form method (see Alternative-form method).</td>
</tr>
<tr>
<td><strong>Coefficient of precision</strong></td>
<td>The correlation between test scores when examinees respond to the same test items repeatedly and there are no changes in examinees over time.</td>
</tr>
<tr>
<td><strong>Coefficient of stability</strong></td>
<td>The correlation obtained from the test-retest method (see Test-retest method).</td>
</tr>
<tr>
<td><strong>Coefficient of stability and equivalence</strong></td>
<td>The correlation coefficient obtained from a test-retest with alternate forms method (see Test-retest with alternate forms method).</td>
</tr>
<tr>
<td><strong>Coefficient omega</strong></td>
<td>A coefficient of precision derived from the factor loadings produced by a factor analysis.</td>
</tr>
<tr>
<td><strong>Complex variable</strong></td>
<td>A composite variable that synthesizes multidimensional data collected from multiple data streams into a single vector whose elements denote the evaluand’s performance on (or their quality with respect to) the criteria of merit or worth (as judged by some method of measurement) that are included in the evaluation. Moreover, since evaluation data often are collected from multiple data streams, the units of analysis (e.g., case identification numbers, people, respondents) for these datasets are different thereby</td>
</tr>
</tbody>
</table>
preventing one from meaningfully merging the data and analyzing it in a single step.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composite variance</td>
<td>The amount of variability in the scores of a complex variable. This variability is a function of the variability of each constituent variable and the covariability between each pair of these variables.</td>
</tr>
<tr>
<td>Confidence interval</td>
<td>A range of values that has a specific probability (i.e., confidence level) of containing the true value of a parameter.</td>
</tr>
<tr>
<td>Confidence level</td>
<td>The probability that the true value of a variable would be contained within a specific confidence interval if the evaluation were repeated ad infinitum under parallel conditions.</td>
</tr>
<tr>
<td>Constituent variables</td>
<td>The variables that are summed (or averaged) to form the composite variable (i.e., evaluative conclusion).</td>
</tr>
<tr>
<td>Construct</td>
<td>A theoretical concept or idea that is composed of multiple attributes and measured by an instrument or method.</td>
</tr>
<tr>
<td>Construct validity</td>
<td>The degree to which the measure of a construct is cohesive (internally consistent), is associated with the measures of related constructs (a.k.a. convergent validity), and is unassociated with the measures of unrelated constructs (a.k.a. divergent validity).</td>
</tr>
<tr>
<td>Correlation</td>
<td>The degree to which two or more variables are related with each other. The range of possible values is between -1 and +1 with negative values representing indirect relationships (i.e., when the values of one variable increase, the values of the other variable decrease) and positive values representing direct relationships (i.e., when the values of one variable increase or decrease so do the values of the other variable). Moreover, -1 means a perfect negative correlation, +1 means a perfect positive correlation, and 0 means no correlation at all.</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>The magnitude of correlation between two variables.</td>
</tr>
<tr>
<td>Covariance</td>
<td>A measure of how much two random variables vary together. If the variables are positively related, then their covariance will also be positive, and vice versa. If the two variables are independent, then their covariance will be zero.</td>
</tr>
<tr>
<td>Coverage probability</td>
<td>The proportion of sample confidence intervals (CIs) that contain the population value, if the study was replicated a large number of times under parallel conditions.</td>
</tr>
<tr>
<td>Criteria of merit</td>
<td>See Value.</td>
</tr>
<tr>
<td>Cumulative distribution function (cdf)</td>
<td>A function that specifies, for all real values x, the probability that the random variable is less than or equal to x.</td>
</tr>
<tr>
<td>Decision-makers</td>
<td>The individuals who are principally responsible for setting the goals of the evaluand, managing its operations, or controlling the resources necessary to operate it. Generally, they are also the individuals who will utilize the results of the evaluation to implement decisions that affect the evaluand.</td>
</tr>
<tr>
<td>Dependent variables</td>
<td>Two variables are dependent if knowledge of the value of one variable provides information about the value of another variable.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Dichotomization</td>
<td>A procedure used to divide the distribution of a random variable into two mutually exclusive classes or groups based upon a single cut-score.</td>
</tr>
<tr>
<td>Error variance</td>
<td>The amount of unexplained variability in the scores associated with a measure, method, or variable. The larger the error variance, the poorer the precision with which the variable was measured.</td>
</tr>
<tr>
<td>Evaluand</td>
<td>A generic term for the entity under investigation.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>The process of determining the merit, worth, or meaningful significance of something, or the product of the process.</td>
</tr>
<tr>
<td>Evaluative conclusion</td>
<td>A decision made or an opinion formed that expresses the merit, worth, or meaningful significance of the evaluand and is typically formulated after weighing all of the relevant factual information; performance on relevant indicators of merit; and comparison of performance to recognized, widely used, or valid standards.</td>
</tr>
<tr>
<td>Event</td>
<td>A subset of the sample space.</td>
</tr>
<tr>
<td>Expected mean square</td>
<td>A measure of variability expressed algebraically by the treatment effects and estimated by the observed mean squares.</td>
</tr>
<tr>
<td>Expected value</td>
<td>Also known as a weighted mean, it is the sum of the probability of each possible outcome of a study multiplied by the outcome value.</td>
</tr>
<tr>
<td>Experiment</td>
<td>A controlled trial wherein the magnitude of a set of dependent variables is controlled by the experimenter to uncover the relation between these variables and a set of independent variables.</td>
</tr>
<tr>
<td>Factor loading</td>
<td>The correlation between an item and the latent construct.</td>
</tr>
<tr>
<td>Finite population correction (fpc)</td>
<td>A correction factor employed to produce an unbiased estimator of the population variance and covariance.</td>
</tr>
<tr>
<td>Hard-hurdle</td>
<td>A performance standard in which failure to meet the benchmark results in the failure of the macrovalue.</td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>Individual differences between the stakeholders from whom data are being collected.</td>
</tr>
<tr>
<td>Impactees</td>
<td>A generic term used to refer to people or organizations that are affected, either directly or indirectly, by the evaluand.</td>
</tr>
<tr>
<td>Independent variables</td>
<td>Two variables are independent if knowledge of the value of one variable provides no information about the value of another variable. That is, there is a zero correlation between the variables.</td>
</tr>
<tr>
<td>Infimum</td>
<td>The largest lower bound of a set.</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>The degree of consistency with which a group of examinees performed across items or subsets of items on a single test form.</td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>The degree of consistency with which different raters rate the same entity the same on a given scale of measurement.</td>
</tr>
<tr>
<td>Interval scale</td>
<td>A measurement scale with equidistance between points but no true zero. Virtually all statistical analyses are permissible for such data.</td>
</tr>
<tr>
<td>Law of Large Numbers</td>
<td>As the sample size increases, sample estimators tend to converge to the population parameters.</td>
</tr>
<tr>
<td>Macrovalue</td>
<td>A composite variable that is constructed from several related microvalues.</td>
</tr>
<tr>
<td><strong>Measurement error</strong></td>
<td>The unexplainable discrepancy between a measurement and the entity which the measurement instrument is intended to measure.</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Metaevaluation</strong></td>
<td>An evaluation of one or more evaluations for the purpose of determining the merit and worth of the original evaluation(s).</td>
</tr>
<tr>
<td><strong>Method of moments</strong></td>
<td>A method of estimation of population parameters such as mean, variance, median, and so forth, by equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated.</td>
</tr>
<tr>
<td><strong>Microvalue</strong></td>
<td>See Value.</td>
</tr>
<tr>
<td><strong>Monotone</strong></td>
<td>Either an increasing or a decreasing function.</td>
</tr>
<tr>
<td><strong>Multinominal distribution</strong></td>
<td>A generalization of the binomial distribution wherein the trials that are aggregated may have two or more outcomes.</td>
</tr>
<tr>
<td><strong>Nested model</strong></td>
<td>A design in which factor B is nested in factor A. That is, unlike a fully-crossed design, every level of factor A does not co-occur with every level of factor B.</td>
</tr>
<tr>
<td><strong>Nominal coverage probability</strong></td>
<td>It refers to the desired probability for a CI—i.e., $100(1-\alpha)%$.</td>
</tr>
<tr>
<td><strong>Normal distribution</strong></td>
<td>Describes a continuous probability distribution that conforms to the function discovered by Abraham DeMoivre. It is considered by many the most important distribution in statistics.</td>
</tr>
<tr>
<td><strong>Nominal scale</strong></td>
<td>A measurement scale that only attempts to classify data into classes or groups that lack a natural order or units of measure. Therefore, an entity either belongs or does not belong to a group. The only permissible mathematical operations for such data are medians, percentiles, and ranks.</td>
</tr>
<tr>
<td><strong>Object of measurement</strong></td>
<td>The sampling unit of primary interest, usually people.</td>
</tr>
<tr>
<td><strong>Ordinal scale</strong></td>
<td>A measurement scale that has a natural order (i.e., classes or scale points that can be sorted from low to high on some attribute) but lacks units of measure and equidistance between classes. The only permissible mathematical operations for such data are medians, percentiles, and ranks.</td>
</tr>
<tr>
<td><strong>Parameter</strong></td>
<td>A quantifiable characteristic of feature of a population (e.g., mean, variance). It can also refer to the true value of a population on some dimension of interest (e.g., conclusion, performance).</td>
</tr>
<tr>
<td><strong>Partially nested model</strong></td>
<td>A three-factor ANOVA model in which nested and crossed factors occur.</td>
</tr>
<tr>
<td><strong>Poisson distribution</strong></td>
<td>Describes a phenomenon that consists of counting the number of Bernoulli trials that occur within a continuous interval of measurement, such as time, length, area, or volume.</td>
</tr>
<tr>
<td><strong>Polychotomization</strong></td>
<td>It is a generalization of dichotomization in which the probability distribution of the raw random variable is partitioned into $k$ classes based on the designation of $k-1$ cut-scores.</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td>The total number of objects or entities that share a common characteristic, feature, or quality.</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>The degree of error with which a variable is estimated. Highly precise estimates have less error (i.e., smaller confidence intervals).</td>
</tr>
</tbody>
</table>
whereas imprecise estimates contain more error. This is not synonymous with accuracy because a result can be measured with a high degree of precision but produce an inaccurate result.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability distribution</td>
<td>A mathematical relation by which probabilities are assigned to all the events that constitute the sample space of a random variable.</td>
</tr>
<tr>
<td>Probability distribution</td>
<td>Describes the probability that the value of a random variable $P(x)$ can attain such that $0 ≤ P(x) ≤ 1$ and the sum of the probabilities over the entire range of values equals 1.</td>
</tr>
<tr>
<td>Professional evaluation</td>
<td>The systematic investigation of an evaluand in order to determine its merit, worth, and significance. Furthermore, the central conclusions reached by this investigation, must be based on the synthesis of performance data across the key values/needs that are relevant for determining the merit of the evaluand, and any defensible standards for those values/needs that are deemed critically important to the overall performance of the evaluand.</td>
</tr>
<tr>
<td>Random variable</td>
<td>A real-valued function of the events of a sample space.</td>
</tr>
<tr>
<td>Ratio scale</td>
<td>A measurement scale with equidistance between points and a true zero. All statistical analyses are permissible for such data.</td>
</tr>
<tr>
<td>Reliability</td>
<td>The consistency of a measure or method.</td>
</tr>
<tr>
<td>Reliability index</td>
<td>The correlation between the true scores and observed scores.</td>
</tr>
<tr>
<td>Research</td>
<td>An active, diligent, and systematic process of inquiry aimed at discovering, interpreting, and revising facts—objective and verifiable observations.</td>
</tr>
<tr>
<td>Robust</td>
<td>A statistical technique is called robust if it leads to basically correct conclusions and probability estimates, even when the statistical assumptions of the technique are violated.</td>
</tr>
<tr>
<td>Sample</td>
<td>An amount of a group of objects or entities that share a common characteristic, feature, or quality.</td>
</tr>
<tr>
<td>Sample size</td>
<td>The size of the sample taken from the population of sampling units.</td>
</tr>
<tr>
<td>Sampling distribution</td>
<td>The frequency or probability distribution of a statistic obtained from an extremely large number of random samples drawn from a specified population. Thus, given a score, one can determine the probability of observing scores of lower, higher, or the same value.</td>
</tr>
<tr>
<td>Sampling error</td>
<td>The amount of error caused by observing a sample rather than the population. The smaller the sample size is, the poorer will be one’s ability to generalize the results of a sample to the population.</td>
</tr>
<tr>
<td>Sample space</td>
<td>The set of all possible outcomes that occur as a result of an experiment.</td>
</tr>
<tr>
<td>Soft-hurdle</td>
<td>A performance standard in which failure to meet the benchmark results in the failure of the microvalue.</td>
</tr>
<tr>
<td>Split-half method</td>
<td>The method of estimating the stepped-up (by the Spearman-Brown prophecy formula) correlation coefficient from the scores of two subtests (half-tests) obtained upon the administration of a single test to a group of examinees.</td>
</tr>
<tr>
<td>Stability</td>
<td>The quality, state, or degree of being of not changing over repeated test administrations across time.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Standard</td>
<td>A performance level associated with a particular criterion or dimension of merit that is of great importance to the performance of the evaluand. For example, a standard may demark acceptable from unacceptable performance or excellent from less than excellent performance on a dimension of merit that is pertinent to the overall performance of the evaluand.</td>
</tr>
<tr>
<td>Standard error of a mean</td>
<td>The standard deviation of the distribution of sample means used to estimate the population mean.</td>
</tr>
<tr>
<td>Standard error of estimate</td>
<td>The standard deviation of the discrepancy between the evaluand’s observed performance on a measure and its predicted performance on a parallel measure.</td>
</tr>
<tr>
<td>Standard error of measurement</td>
<td>The standard deviation of the discrepancy between the evaluand’s observed performance on a measure and the true value of the construct being measured.</td>
</tr>
<tr>
<td>Standard error of prediction</td>
<td>The average discrepancy between an observed score and the observed score on a parallel measure.</td>
</tr>
<tr>
<td>Statistic</td>
<td>A quantifiable characteristic of feature of a sample (e.g., mean, variance) used to estimate its corresponding population parameter.</td>
</tr>
<tr>
<td>Statistical moments</td>
<td>The first four statistical moments of a distribution are mean, variance, skewness, and kurtosis.</td>
</tr>
<tr>
<td>Summative conclusion</td>
<td>An overall conclusion that is reached after weighing all of the evaluative conclusions and adjusting for poor or excellent performance on criteria of merit deemed critically important to the proper functioning of the evaluand.</td>
</tr>
<tr>
<td>Summative confidence</td>
<td>A mathematical statement that expresses the degree of confidence one may place on an evaluative conclusion that was formulated by synthesizing performance of the evaluand on critically important and non-critically important dimensions.</td>
</tr>
<tr>
<td>Summative evaluation</td>
<td>The synthesis of multiple values (variables)—typically composed of different distributions and thereby requiring transformation into a common distribution—and standards into an evaluative conclusion.</td>
</tr>
<tr>
<td>Supremum</td>
<td>The smallest upper bound of a set.</td>
</tr>
<tr>
<td>Synthesis</td>
<td>Either the integration of facts with values in order to formulate an evaluative statement or the integration of multiple evaluative conclusions into a summative conclusion.</td>
</tr>
<tr>
<td>Test-retest method</td>
<td>The administration of two identical tests or instruments at different points in time to the same respondents.</td>
</tr>
<tr>
<td>Test-retest with alternate form method</td>
<td>The administration of one form of the test at one point in time and an alternative form of the test at a subsequent point in time.</td>
</tr>
<tr>
<td>Triangulation</td>
<td>A method of establishing the accuracy of a conclusion by comparing information from multiple data sources and methods.</td>
</tr>
<tr>
<td>True score (value)</td>
<td>The mean of the sampling distribution that would form if an individual was repeatedly tested and they did not experience fatigue or recall.</td>
</tr>
<tr>
<td>Type I error</td>
<td>Also known as a false positive, this error refers to the probability of concluding that a relationship exists between two variables (like an</td>
</tr>
</tbody>
</table>
independent and dependent variable) when in fact no relationship exists between them.

<table>
<thead>
<tr>
<th>Uniform distribution</th>
<th>Describes either a continuous or discrete probability distribution in which the events found in the sample space are all equally probable.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>Criteria or dimensions of merit that are used to formulate an evaluative conclusion about the performance of an evaluand.</td>
</tr>
<tr>
<td>Variance</td>
<td>A measure of statistical dispersion—equal to the squared discrepancy between the observed and expected value—of a variable. The larger the variance, the greater the average discrepancy.</td>
</tr>
<tr>
<td>Vector</td>
<td>A single column (or row) of data.</td>
</tr>
<tr>
<td>Weighting scheme</td>
<td>The amount of mathematical importance that is accorded to each value that is used to formulate the summative conclusion.</td>
</tr>
<tr>
<td>z-score</td>
<td>The number of standard deviations above or below the mean of a standard normal distribution.</td>
</tr>
</tbody>
</table>
APPENDIX B

SAS Code for Generating GLD Lambda Values
%Let L3 = .0000008; /* INPUT SEED FOR LAMBDA 3 */
%Let L4 = &L3; /* INPUT SEED FOR LAMBDA 4 */
%Let CV = .00000001; /* DEFAULT CONVERGENCE CRITERION */
%Let MAXITER = 500; /* DEFAULT MAXIMUM ITERATIONS */
%Let Rep = 2000; %Let N = 400;
%Macro FindLambda(a1=, a2=, a3=, a4=);
PROC IML;
START Moments;
MEAN = &a1;
VAR = &a2;
SKEW = &a3;
%Let S = &a3;
KURT = &a4;
IF SKEW < 0 THEN DO;
SKEW = abs(SKEW);
%Let S = %sysfunc(abs(%sysevalf(&a3))); END;
FINISH;
START Valid;
IF %sysfunc(min(&a4,%sysevalf(&S**2-2)))=&a4 THEN DO;
minvalue = %sysevalf(&S**2-2);
PRINT "INVALID REGION: Skew=" &S " Kurt=" &a4;
PRINT "You must increase Kurtosis to more than: " minvalue;
ABORT;
END;
FINISH;
START NEWTON;
RUN FUN; /* EVALUATE FUNCTION AT STARTING VALUES */
DO ITER=1 TO &MAXITER /* ITERATE UNTIL MAXITER ITERATIONS */
WHILE(MAX(ABS(F))>&CV); /* OR CONVERGENCE */
RUN DERIV; /* EVALUATE DERIVATIVES IN J */
DELTA = SOLVE(J,F); /* SOLVE FOR CORRECTION VECTOR */
Lambda = Lambda + DELTA; /* THE NEW APPROXIMATION */
RUN FUN; /* EVALUATE THE FUNCTION */
END;
FINISH;
START FUN;
L3 = Lambda[1]; L4 = Lambda[2]; /* EXTRACT THE VALUES */
A = 1/(1+L3) - 1/(1+L4);
B = 1/(1+2*L3) + 1/(1+2*L4) - 2*BETA(1+L3,1+L4);
C = 1/(1+3*L3) - 1/(1+3*L4) - 3*BETA(1+2*L3,1+L4) + 3*BETA(1+L3,1+2*L4);
D = 1/(1+4*L3) + 1/(1+4*L4) - 4*BETA(1+3*L3,1+L4) + 6*BETA(1+2*L3,1+2*L4) -
   4*BETA(1+L3,1+3*L4);
L2 = SQRT((B - A**2)/&a2);
L1 = &a1-A/L2;
/* The first part of F estimates SKEWNESS the second part estimates KURTOSIS */
F = ((B-A**2)**(-1.5))*(C-3*A*B+2*A**3) - SKEW; / /* normalize kurtosis */
END;
FINISH;
START DERIV;
dA_L3 = -(1+L3)**(-2);
dB_L3 = -2*(1+2*L3)**(-2) - 2*BETA(1+L3,1+L4)*(DIGAMMA(1+L3)-DIGAMMA(2+L3+L4));
dC_L3 = -3*(1+3*L3)**(-2) - 6*BETA(1+2*L3,1+L4)*(DIGAMMA(1+2*L3)-
   DIGAMMA(2+2*L3+L4))+3*BETA(1+L3,1+2*L4)*(DIGAMMA(1+L3)-
   DIGAMMA(2+L3+2*L4));
dD_L3 = -4*(1+4*L3)**(-2) - 12*BETA(1+3*L3,1+L4)*(DIGAMMA(1+3*L3)-
   DIGAMMA(2+3*L3+L4))+12*BETA(1+2*L3,1+2*L4)*(DIGAMMA(1+2*L3)-
   DIGAMMA(1+2*L3)-
DIGAMMA(2+2*L3+2*L4)-4*BETA(1+L3,1+3*L4)*(DIGAMMA(1+L3)-DIGAMMA(2+L3+3*L4));

dA_L4 = (1+L4)**(-2);

dB_L4 = -2*(1+2*L4)**(-2) - 2*BETA(1+L3,1+L4)*(DIGAMMA(1+L4)-DIGAMMA(2+L3+L4));

dC_L4 = 3*(1+3*L4)**(-2) - 3*BETA(1+2*L3,1+L4)*(DIGAMMA(1+L3)-DIGAMMA(2+2*L3+L4))+6*BETA(1+L3,1+2*L4)*(DIGAMMA(1+2*L4)-DIGAMMA(2+L3+2*L4));

dD_L4 = -4*(1+4*L4)**(-2) - 4*BETA(1+3*L3,1+L4)*(DIGAMMA(1+L4)-DIGAMMA(2+3*L3+L4))+12*BETA(1+2*L3,1+2*L4)*(DIGAMMA(1+2*L4)-DIGAMMA(2+2*L3+2*L4))-12*BETA(1+L3,1+3*L4)*(DIGAMMA(1+3*L4)-DIGAMMA(2+L3+3*L4));

da3L3 = (B-A**2)**(-1.5)*(dC_L3-3*(A*dB_L3+B*dA_L3) + 6*(A**2)*dA_L3) + (C-3*A*B+2*A**3)*(-1.5)*(B-A**2)**(-2.5)*(dB_L3-2*A*dA_L3);;

da3L4 = (B-A**2)**(-1.5)*(dC_L4-3*(A*dB_L4+B*dA_L4) + 6*(A**2)*dA_L4) + (C-3*A*B+2*A**3)*(-1.5)*(B-A**2)**(-2.5)*(dB_L4-2*A*dA_L4));

da4L3 = (B-A**2)**(-2)*(dD_L3-4*(A*dC_L3+C*dA_L3) + 6*(A**2)*dB_L3+2*A*B*dA_L3) - 12*(A**3)*dA_L3) + (D-4*A*C+6*(A**2)*B-3*A**4)*(-2*(B-A**2)**(-3))*(dB_L3-2*A*dA_L3));

da4L4 = (B-A**2)**(-2)*(dD_L4-4*(A*dC_L4+C*dA_L4) + 6*(A**2)*dB_L4+2*A*B*dA_L4) - 12*(A**3)*dA_L4) + (D-4*A*C+6*(A**2)*B-3*A**4)*(-2*(B-A**2)**(-3))*(dB_L4-2*A*dA_L4));

J = (da3L3) || (da3L4) || (da4L3) || (da4L4); /* EVALUATE JACOBIAN */

FINISH;

START NEGATIVE;

If L2>0 & L3<0 & L4<0 then Do;

L2 = -L2;
T = L3;
L3 = L4;
L4 = T;
end;

If &a3 < 0 then DO;

SKEW = -1*SKEW;
T = L3;
L3 = L4;
L4 = T;
L1 = L1 + 2*A/L2;
end;

FINISH;

DO;

Lambda={&L3, &L4}; /* starting values */

RUN Moments ;
RUN Valid ;
RUN NEWTON;
RUN NEGATIVE;
PRINT MOMENTS, MEAN VAR SKEW KURT ;
PRINT "Estimates of Lambdas 1 to 4." ;
PRINT L1 L2 L3 L4 ;
CREATE Lambdas var {MEAN VAR SKEW KURT L1 L2 L3 L4} ;
APPEND;
CLOSE Lambdas ;

END ;QUIT;%Mend;
APPENDIX C

SAS Code for Generating Positive Definite Correlation Matrices
%Macro ValidCorrMatrix;
  O=M;
  Iteration=0;
  do while (min(E)<=0);
    Iteration=Iteration+1;
    i=loc(E<0);
    /* Following 2 lines generate a truly unstructured corr matrix. */
    %if %Upcase(&Cov)=UN %then %do;
      max=max(E);
      E[i]=max*ranuni(0);
    %end;
    %else %do;
      /* Use the following line if you want the new corr matrix to closely mirror
       the input corr matrix. */
      E[i]=0;
    %end;
    E=diag(E);
    O=V*E*V`; /* Adjusted (new) correlation matrix */
  end;
%Macro Standardize;
  d=M-O;
  T=abs(trace(d));
  max=max(abs(d));
  sum=sum(abs(d))-T;
  ave=sum/(ncol(M)**2 - ncol(M));
%mend ValidCorrMatrix;
%Macro Standardize;
  Sd=O;
  do col=1 to ncol(Sd);
    do row=1 to ncol(Sd);
    end;
  end;
  O=Sd;
%mend Standardize;
%Macro CheckMatrix;
  proc iml;
  use Corr;
  READ all var(“V1”:”V&v”) INTO M;
  CLOSE Corr;
  call eigen(E,V,M);
  if min(E)<0 then do;
    %ValidCorrMatrix;
    names=(“V1”:”V&v”);
    create Corr from O[c=(names)];
    append from O;
  end;
quit;
%mend CheckMatrix;
quit;
APPENDIX D

SAS Code for Illustrating the Central Limit Theorem (Model 1)
OPTIONS nonotes nosource errors=0; * Abbreviate size of log file;
proc datasets Kill nolist; quit;
/* */
%Let Model= Model1a;
%Let ModNo= 1;
%Let Sims = 250; /* Number of simulations */
%Let Cov = 1N ; /* In=Independent, CS=Compound symmetry, AR=autoregressive, UN=unstructured*/
%Let Rho  = 0;
%Let Alpha= 0.05;
%Let Crit = 0.15; /* Standard used by Shapiro-Wilks test for normality. Set to 0.15. */
%Let minV = 2 ; /* Minimum # variables */
%Let maxV = 300; /* Maximum # variables */
%Let byV  = 0 ; /* By # variables */
%Let minN = 10 ; /* Minimum sample size */
%Let maxN = 1000; /* Maximum sample size */
%Let byN  = 10 ; /* By sample size */
/* */
%Let Loc = C:\Users\Cristian\Desktop\My files\SAS Files\Dissertation\Phase I\Test Models\Support files; Libname Lam "&Loc .. .. .. .. .."]; Libname Loc "C:\Users\Cristian\Desktop\My files\SAS Files\Dissertation\Phase I\Test Models"];
/* */
/* Generate a correlation matrix */
/* */
%macro CorMatrix;
data Corr(keep= V1-V&v);
%let v2=%eval(&v*&v);
format V1-V&v2 5.3;
length _numeric_ 4;
array a{&v,&v} V1-V&v2;
if &ModNo=2 then rho=&rho/10;
else rho=ρ;
do c=1 to &v;
r=0;
do while (r<c);
  r=r+1;
  if r=c then a(r,c)=1;
  else do;
    /* Generate only + numbers */
    if UPCASE("&Cov") = "UN" then cor = ranuni(0);
    else if UPCASE("&Cov") = "IN" then cor=0;
    else if UPCASE("&Cov") = "AR" then cor=abs(rho**(c-r));
    else if UPCASE("&Cov") = "CS" then cor=rho;
    a(r,c)=cor;
    a(c,r)=cor;
  end;
end;
do r=1 to &v;
  %do c=1 %to &v;
    V&c=a(r,c);
  %end;
output;
end;
%mend CorMatrix;
/* */
/*----------------------------------------------------------------------------------------*/
/* Check to make sure correlation matrix is positive */
/*----------------------------------------------------------------------------------------*/
%%Macro ValidCorrMatrix;
  O=M;
  Iteration=0;
  do while (min(E)<=0);
    Iteration=Iteration+1;
    i=loc(E<0);
    /* Following 2 lines generate a truly unstructured corr matrix. */
    %if %Upcase(&Cov)=UN %then %do;
      max=max(E);
      E[i]=max*ranuni(0);
    %end;
    %else %do;
      /* Use the following line if you want the new corr matrix to closely mirror the input corr matrix. */
      E[i]=0;
    %end;
    E=diag(E);
    O=V*E*V`; /* Adjusted (new) correlation matrix */
  end;
%%Standardize;
  d=M-O;
  T=abs(trace(d));
  max=max(abs(d));
  sum=sum(abs(d))-T;
  ave=sum/(ncol(M)**2 - ncol(M));
%%end ValidCorrMatrix;

%%Macro Standardize;
  Sd=O;
  do col=1 to ncol(Sd);
    do row=1 to ncol(Sd);
    end;
  end;
  O=Sd;
%%end Standardize;

%%Macro CheckMatrix;
  proc iml;
    use Corr;
    READ all var("V1","V&v") INTO M;
    CLOSE Corr ;
    call eigen(E,V,M);
    if min(E)<0 then do;
      %ValidCorrMatrix;
      names=("V1","V&v");
      create Corr from O[c=(names)];
      append from O;
    end;
  quit;
  data Corr;
set Corr;
length _numeric_ 4;
run;
%Mend CheckMatrix;
/*-----------------------------------------------*/
/*-----------------------------------------------*/
/* Generate correlated nonnormal data */
/*-----------------------------------------------*/
%macro Cronbach;
ods listing close;
ods results off;
od$ output CronbachAlpha=CronbachAlpha;
proc corr data = V alpha nocorr nomiss noprobt nosimple ;
   var V1-V&v;
   by Dist vars n Rho sim;
run;
ods results on;
ods listing;
data CronbachAlpha;
   set CronbachAlpha;
   length _numeric_ 4;
run;
proc append base=Cronbach data=CronbachAlpha force;
run;
%mend Cronbach;

%Macro GenerateData(v=, N=);
%do Rho=%sysvalf(10*&minR) %to %sysvalf(10*&maxR) ;
%CorrMatrix;  /* Move these two lines below the DO LOOP */
/*%CheckMatrix; below if cov=Unstructured */
%do I=1 %to &Sims;
   Proc IML;
      Use Corr;
      READ all var("V1":"V&v") INTO Corr;
      CLOSE Corr ;

      Use Lam.Lambdas;
      READ all var("L1":"L4") INTO L ;
      READ all var{_Dist_} INTO Dist ;
      CLOSE Lam.Lambdas ;

      START CORR;
      Cholesky=root(Corr);
      data=rannor([Dn,&v,0]);
      Z=DATA*Cholesky;
      U=CDF('NORMAL',Z);
      FINISH;

      START D;
      X=(min(Dist):max(Dist))~;
      do i=1 to nrow(L);
         R=R//repeat(X[i],&N,1);
         Y=Y//repeat(Y[i],&N,1);
      end;
L=Y;
FINISH;

START VARS;
V=J(Dn,&v,);
do i=1 to Dn;
    X = L[i,1] + (U[i,1]##L[i,3] - (1-U[i,1]##L[i,4]) / L[i,2];
    V[i,]=X;
end;
name2=("V1"."V&maxV");
FINISH;

START RHO; /* average correlation */
RHO=abs(Corr);
RHO=(RHO[+]-nrow(Corr))/(nrow(Corr)**2-nrow(Corr));
FINISH;

START KEYS;
I=J(Dn,7,1);
I[,1]=R;  * Assign distribution;
I[,2]=J(Dn,1,&i);  * Assign simulation ;
I[,3]=J(Dn,1,&N);  * Assign sample size ;
I[,4]=J(Dn,1,&V);  * Assign # variables ;
I[,5]=repeat((1:&N),Ln,1);  * Assign order;
I[,6]=V[,];  * Average of all V ;
I[,7]=J(Dn,1,RHO);  * Average correlation;
name1=("Dist" | "Sim" | "N" | "Vars" | "Order" | "Ave" | "Rho");
FINISH;

START MAIN;
Ln=nrow(L);
Dn=&N*nrow(L);
RUN CORR;
RUN D;
RUN VARS;
RUN RHO;
RUN KEYS;
if &v=&maxV then V=I||V;
else do;
    X=J(Dn,&maxV-&v,);
    V=I||V||X;
end;
CREATE V FROM V[c=(name1 | name2)];
APPEND FROM V;
FINISH;
RUN MAIN;
quit;

data V;
set V;
length _numeric_ 4;
run;
proc append base=&Model data=V force; run;
%Cronbach;
%Macro Model1;
%Let cnt = 0;
%do v=&minV %to &maxV;
  %Let cnt = %eval(&cnt + 1);
  %if &cnt=10 %then %Let v = &maxV;
  %Put %Str( )Subrun &cnt of 10: Vars = &v%Str( )%Sysfunc(Time(),timeampm.);
  %do N=&minN %to &maxN;
    %GenerateData(v=&v, N=&N);
    %Let N=%eval(&N + 10* &byN + 39);
    %Let byN = %eval(&byN + 1);
  %end;
  %Let byN = 10;
  %Let v=%eval(&v + 5* &byV + 7);
  %if &cnt<4 %then %Let byV = %eval(&byV);
    %else %Let byV = %eval(&byV + 3);
proc univariate data=&Model noprint;
  output out=Interim/*(drop=abc)*/ PROBN=PROBN ;*MEAN=abc RHO;
  var ave ;*RHO;
  by vars n sim dist;
run;
proc append base=Normal data=Interim force;
proc datasets nolist;
  delete &Model Interim V;
quit;
%end;

Data Loc.&model%cmpres(Data);
  set Normal;
run;
%mend Model1;  
%Model1;

proc sort data=Cronbach; by Dist Vars N Sim; run;
proc transpose data=Cronbach out=Cronbach(rename=(Col1=Raw Col2=Standardized));
  by Dist Vars N Sim;
run;
Title1 "Probability Analysis";
Title2;
proc IML;
  use Normal;
  read all var{Dist Vars N Sim /*Rho*/ PROBN} into N;
  close Normal;
  use Cronbach;
  read all var{Standardized} into Cronbach;
  close Cronbach;
START MAIN;
   C=J(nrow(N),1,&Crit);
   NT=J(nrow(N),3,);
   NT[,1]=(N[,5] >= C[,1]); /* 1=Passed the Normal test, 0=Failed the Normal test */
   NT[,2]=&ModNo; /* Model number */
   N=N || NT || Cronbach;
   names=("Dist" || "Vars" || "N" || "Sim" || "Rho" || "Prob" || "Normal" || "Model" ||
   "Cronbach");

CREATE Loc.&Model FROM N[c=names];
APPEND FROM N;
FINISH;
run

quit;

Title 1;
proc sort data=Loc.&Model;
   by dist vars n ;
run;
Proc Means data=Loc.&Model noprint;
   var Normal;
   output out=Normal(drop=_Type_ _Freq_) Sum=count;
   by dist vars n ;
run;
proc iml;
use Normal;
read all var{Dist Vars N Count} into N;
CLOSE Normal ;

START MAIN;
   Prob = N[,4]/&Sims;
CREATE Normal FROM N[c=("Dist" || "Vars" || "N" || "Prob")];
APPEND FROM N;
FINISH;
run

quit;
*/ --------------------------------------------------------------- */
*/ Merge the Model 1 mini-datasets into one dataset */
*/ --------------------------------------------------------------- */
%let loc = C:\Users\Cristian\Desktop\My files\Dissertation\Data\Model 1;
libname lib "&loc";
libname Lam "C:\Users\Cristian\Desktop\My files\SAS Files\Dissertation\Phase I";

%let Sims = 1000; /* Number of simulations */
data Model1;
   set lib.Model1a lib.Model1b lib.Model1c lib.Model1d;
run;
proc sort data=Model1;
   by dist vars n sim;
run;
Proc Means data=Model1 noprint;
  var Normal;
  output out=Normal(drop=_Type_ _Freq_) Sum=count;
  by dist vars n;
run;

proc iml;
  use Normal;
  read all var{Dist Vars N Count} into N;
  close Normal;

  start MAIN;
    Prob = N[,4]/&Sims;
    create Normal from N[c=("Dist" | "Vars" | "N" | "Prob")];
    append from N;
  finish;
run MAIN;
quit;

data Normal;
  length _numeric_ 4;
  merge Normal
    lam.lambdas(keep=skew kurt _dist_ rename=(_dist_=dist));
  by dist;
run;

/*-----------------------------------------------------------------------------------------*/
/* Likehood of attaining a normal composite based on interpolated skewness and kurtosis data*/
/*-----------------------------------------------------------------------------------------*/

proc sort data=Normal; by skew kurt; run;
proc means data=Normal noprint;
  output out=BySkewKurt(drop=_type_ _freq_) mean=prob;
  var prob;
  by skew kurt;
run;
proc g3grid data=BySkewKurt out=BySkewKurt;
  grid skew*kurt=prob /
    axis1=0 to 2 by .1 /* Skewness */
    axis2=0 to 10 by .5; /* Kurtosis */
run;
quit;

data BySkewKurt;
  set BySkewKurt;
  if kurt>=round(1.8*(skew**2 + 1)-3,0001) ;
run;

goptions reset=all gunit=pt ftext='Garamond/bo' ctext=black htitle=14 htext=10
  norotate vsize=4.0in hsize=8.0in;
goptions device=EMF gsfmode=replace gsfname=gout TRANSPARENCY;
filename gout "&Loc\Figures\SkewKurt.emf";
proc g3d data=BySkewKurt gout = work.threeWay;
  plot skew*kurt=prob/rotate=150 tilt=80 grid name="skewKurt"
    caxis=black ctext=black ctop=bib cbottom=blue
zticknum=10 zmin=0 zmax=1 yticknum=5 xticknum=5 xytype=3;
label Skew="Skewness" Kurt="Kurtosis" Prob="Probability";
run; quit; Title1;
/* --------------------------------------------------------------- */
/* Likehood of attaining a normal composite based on interpolated N and # of variables aggregated */
/* --------------------------------------------------------------- */
proc sort data=Normal; by N Vars; run;
proc means data=Normal noprint;
  output out=ByNVars(drop=_Type_ _Freq_) Mean=Prob;
  var Prob;
  by N Vars;
run;
proc g3grid data=ByNVars out=ByNVars;
  grid N*Vars=Prob /
    axis1=10 to 1000 by 25 /* Sample size */
    axis2=2 to 300 by 3 ;/* # variables aggregated */
run;
quit;
goptions reset=all gunit=pt ftext='Garamond/bo' ctext=BLACK htitle=14 htext=10
  norotate vsize=4.0 in hsize=8.0 in ;
goptions device=EMF gsfmode=replace gsfname=gout TRANSPARENCY;
filename gout "&Loc\Figures\SkewKurt.emf";
GOPTIONS DEVICE=activex;
ODS HTML path="C:\Users\Cristian\Desktop\My files\Dissertation\Data\Model 1\figures\"
  body="NVars.htm";
goptions border;
proc g3d data=ByNVars gout = work.ThreeWay;
  plot N*Vars=Prob/rotate=150 tilt=80 grid name="NVars"
    caxis=black ctext=black cttop=bib cboottom=blue
    zticknum=10 zmin=0 zmax=1 yticknum=10 xticknum=10 xytype=3;
label N="Sample size" Vars="No. variables" Prob="Probability";
run; quit; Title1;
ODS HTML CLOSE;
/* --------------------------------------------------------------- */
/* Generate Figures 7 and 8 investigating the impact of sample size on passing the normality test */
/* --------------------------------------------------------------- */
/* Run this block of code first. Change the sample size below for each run. */
/* In order to produce the panel figure, you need to run analysis for four different sample sizes, e.g., 50, 150, 250, 500. */
%Let N=250;
proc iml;
  Use Lam.Lambdas;
  READ all var("L1","L4") INTO L where(skew=2 & kurt=10) ;
  CLOSE Lam.Lambdas ;

  START GLD;
    U=ranuni(j(N,V,0)) ;
    X=j(N,V,0);
  END GLD;
do i=1 to V;
end;
X=X[,:,]; /* composite mean */
call sort(X,1);
id=j(N,2,i); /* sim number */
id[,2]=j(1:N); /* order number */
X=id | X;
CV=CV//X;
FINISH GLD;
START MAIN;
sim=3000; /* # simulations */
N=&N; /* sample size */
V=150; /* # variables aggregated */
do j=1 to sim;
    RUN GLD;
end;
CREATE CLT FROM CV [c=('sim' | 'order' | 'X1')];
APPEND FROM CV;
FINISH MAIN;
run; quit;
proc sort data=CLT;
    by order;
run;
Proc Means data=CLT noprint;
    var X1;
    output out=CLT2(drop=_Type_ _Freq_) Mean=X1;
    by order;
run;

goptions reset=all  gunit=pt device=EMF ftext='Garamond' HBY=0
    htitle=12 htext=10 vsize=5in hsize=7in;
proc univariate DATA=CLT2 normal noprint ;
    histogram / normal NMIDPOINTS=20
        outhistogram = MyHist ;
    inset N ProbN="S-W (p-value)" normal(KSDPVAL="  K-S")
        / pos = nw format = 15.2 cfill = ywh height=3;
    VAR X1;
run;
/ * Run this block of code second. You will need to update the
        Shapiro-Wilks statistic below with the p-value reported by
        the Proc Univariate analysis produced by the previous code */
%annomac;
data anno;
    length function  color $ 8 Text $ 30;
    retain xsys ysys hsys '1' function 'label'
        color 'black' size 3.5 x 19 y 82;
    Text="N=&N"
        output;
    Text="Shapiro-Wilks = 1.00"; y=78;
    output;
%line(8,84,30,84,blue,21,.25); output;
%line(8,75,30,75,blue,21,.25); output;
run;

axis1 value=none order=(0 to 30) label=none major=none minor=none style=0;
axis2 value=none label=none major=none minor=none;
symbol1 c=red ci=navy v=none i=sm1 l=1 w=1;
symbol2 c=Navy ci=red v=none i=sm1 l=21 w=1;
legend1 across=1
  origin=(390,270)
  mode=share frame label=none
  shape=line(0.5) in value=(height=12 'Composite' 'Normal');

proc gplot data=MyHist(where=(-5<=_MIDPT_<=5)) gout = work.Hist /*uniform*/;
  plot _OBSPCT_*_MIDPT_ _EXPPCT_*_MIDPT_/overlay annotate=ano legend=legend1
    vaxis=axis1 haxis=axis2 noframe name="N_&N";
run;quit;
/* Run this block of code once all four gplots are produced */
goptions device=EMF gsfmode=replace gsfname=gout TRANSPARENCY;
filename gout "C:\Users\Cristian\Desktop\My files\Dissertation\Data\Model 1\Figures\Shapiro2..emf";*
file destination;
proc greplay igout=work.Hist tc=sashelp.templt template=l2r2 nofs;
  treplay 1:N_250 2:N_500 3:N_1000 4:N_2000 name='Shapiro2' des='Shapiro-Wilks test';
run;quit;
/* -------------------------------------------------------------------------- */
APPENDIX E

SAS Code for Illustrating the Central Limit Theorem (Model 2)
goptions reset = all;
OPTIONS nonotes nosource errors=0; * Abbreviate size of log file ;
proc datasets Kill nolist; quit;

/* Let Model= Model2n; */
%Let Model = Model2n;
%Let ModNo = 2;
%Let Sims = 100; /* Number of simulations */
%Let Cov = CS; /* In=Independent, CS=Compound symmetry, AR=Autoregressive, UN=Unstructured*/
%Let minR = 0.10;
%Let maxR = 1.00;
%Let byR = 0.10;
%Let Alpha = 0.05; /* Type I Error*/
%Let Cnt = 0.15; /* Standard used by Shapiro-Wilks test for normality. Set to 0.15. */
%Let minV = 2; /* Minimum # variables */
%Let maxV = 300; /* Maximum # variables */
%Let byV = 0; /* By # variables */
%Let minN = 10; /* Minimum sample size */
%Let maxN = 1000; /* Maximum sample size */
%Let byN = 10; /* By sample size */

%Let Loc = C:\SAS Files\Dissertation\Phase I\Test Models\Support files ;
Libname Lam "&Loc\..\..\..\..\..\..\..";
Libname Loc "C:\SAS Files\Dissertation\Phase I\Test Models\";

/* Generate a correlation matrix */
APPEND CODE FROM APPENDIX E HERE

/* Check to make sure correlation matrix is positive */
APPEND CODE FROM APPENDIX E HERE

/* Generate correlated nonnormal data */
APPEND CODE FROM APPENDIX E HERE

/* MODEL 2 */
%Macro Model2;
%Let cnt = 0;
%do v=&minV %to &maxV;
%Let cnt = %eval(&cnt + 1);
%if &cnt=10 %then %Let v = &maxV;
%put %str( )Subrun &cnt of 10: Vars = &v %str( )%sysfunc(Time(),timeampm.);
%do N=&minN %to &maxN;
%GenerateData(v=&v, N=&N);
%Let N=%eval(&N + 10*&byN + 39);
%end;
%end;

}%Macro Model2;
%Let byN = %eval(&byN + 1);
%end;
%Let byN = 10;
%Let v=%eval(&v + 5*byV + 7);
%if &cnt<4 %then %Let byV = %eval(&byV);
%else %Let byV = %eval(&byV + 3);

proc univariate data=&Model noprint;
   output out=Interim/*(drop=abc)*/ PROBN=PROBN ;*MEAN=abc RHO;
   var ave /*RHO*/;
   by vars n Rho sim dist;
run;

data Interim;
set Interim;
   length _numeric_ 4;
run;

proc append base=Normal data=Interim force; run;
proc datasets nolist;
delete &Model Interim V;quit;
%end;
/* Data Loc.&model%cmpres(Data);*/
/* set Normal;*/
/* run;*/
%mend Model2;

proc sort data=Cronbach; by Dist Vars N Rho Sim; run;
proc transpose data=Cronbach out=Cronbach(rename=(Col1=Raw Col2=Standardized));
   by Dist Vars N Rho Sim;
run;
%StepTime(Step='Generate Model 2 data:')  ; * Append interim time in log. ;

Title1 "Probability Analysis";
Title2;
proc IML;
   use Normal;
   read all var{Dist Vars N Sim Rho PROBN} into N;
   CLOSE Normal ;

   use Cronbach;
   read all var{Standardized} into Cronbach;
   CLOSE Cronbach;

START MAIN;
   i=loc(Cronbach<0);
   Cronbach[i]=0;

   C=J(nrow(N),1,&Crit);
   NT=J(nrow(N),2,); NT[,1]=(N[,6] >= C[,1]) ; /* 1=Passed the Normal test, 0=Failed the Normal test */
   NT[,2]=&ModNo; /* Model number */
   N=N || NT || Cronbach;
names=('Dist' || 'Vars' || 'N' || 'Sim' || 'Rho' || 'Prob' || 'Normal' || 'Model' || 'Cronbach');
CREATE Loc.&Model FROM N[c=names];
APPEND FROM N;
FINISH;
runtime;
quit;
Title1;
data Loc.&Model;
set Loc.&Model;
length _numeric_ 4;
run;
proc sort data=Loc.&Model;
by dist vars n rho;
run;
Proc Means data=Loc.&Model noprint;
var Normal;
output out=Normal(drop=_Type_ _Freq_) Sum=count;
by dist vars n rho;
run;
proc iml;
use Normal;
read all var{Dist Vars N Rho Count} into N;
CLOSE Normal;
START MAIN;
Prob = N[,5]/&Sims;
CREATE Normal FROM N[c=('Dist' || 'Vars' || 'N' || 'Rho' || 'Prob')];
APPEND FROM N;
FINISH;
runtime;
quit;
/*StepTime(Step='Calculate probability:') ; * Append interim time in log. ;
*/
/*/ ---------------------------------------------------------- */
/*/ Merge the Model 2 mini-datasets into one dataset */
/*/ ---------------------------------------------------------- */
%let loc = C:\Users\Cristian\Desktop\My files\Dissertation\Data\Model 2;
libname lib "&loc";

***** NOTE, you will need to reset the sim counter for each file so that it goes from 0-1000;
/*data lib.Model2n;*/ *** change file name accordingly;
/* set lib.Model2n;*/ *** change file name accordingly;
/* sim=sim+900;*/ *** increment counter for each file by appropriate amount;
/*run;*/
data Model2;
set Model1 /* Rho = 0 */
rho=round(rho,1);
run;

proc sort data=Model2;
   by dist vars n rho sim;
run;

Proc Means data=Model2 noprint;
   var Normal;
   output out=Normal2(drop=_Type_ _Freq_) Sum=count;
      by dist vars n rho;
run;

proc iml;
    use Normal2;
    read all var{Dist Vars N Rho Count} into N;
    close Normal2;
    start MAIN;
        Prob = N[,5]/&Sims;
        create Normal2 from N[c=("Dist" || "Vars" || "N" || "Rho" || "Prob")];
        append from N;
    finish;
run MAIN;
quit;

data Normal2;
    length _numeric_ 4;
    merge Normal2
       Lam.Lambdas(keep=Skew Kurt _Dist_ rename=(_Dist_=Dist));
    by Dist;
run;

/*-----------------------------------------------------------------------------------------*/
/* Likehood of attaining a normal composite based on interpolated skewness and kurtosis data */
/*-----------------------------------------------------------------------------------------*/

proc sort data=Normal2; by Skew Kurt; run;
proc means data=Normal2 noprint;
   output out=BySkewKurt(drop=_Type_ _Freq_) Mean=Prob;
      var Prob;
      by Skew Kurt;
run;

proc g3grid data=BySkewKurt out=BySkewKurt;
   grid Skew*Kurt=Prob / axis1=0 to 2 by .1 /* Skewness */
      axis2=0 to 10 by .5 /* Kurtosis */
run;
quit;

Data BySkewKurt;
    set BySkewKurt;
    if Kurt>=round(1.8*(Skew**2 + 1)-3,0001) ;
run;

goptions reset=all gunit=pt ftext='Garamond/bo' ctext=BLACK htitle=14 htext=10
norotate vsize=4.0in hsize=8.0in;
goptions device=EMF gsffmode=replace gsfname=gout TRANSPARENCY;
filename gout "&Loc\Figures\SkewKurt.emf";
proc g3d data=BySkewKurt gout = work.ThreeWay;
   plot Skew*Kurt=Prob/rotate=200 tilt=80 grid name="SkewKurt"
      caxis=black ctext=black ctop=blue cbottom=blue
      zticknum=10 zmin=0 zmax=1 yticknum=5 xticknum=5 xtype=3;
   label Skew="Skewness" Kurt="Kurtosis" Prob="Probability";
run; quit;
Title1;
/* ------------------------------------------------------------------------ */
/* Likehood of attaining a normal composite based on interpolated N and # of variables aggregated */
/* ------------------------------------------------------------------------ */
proc sort data=Normal2; by N Vars; run;
proc means data=Normal2 noprint;
   output out=ByNVars(drop= _Type_ _Freq_) Mean=Prob;
   var Prob;
   by N Vars;
run;
proc g3grid data=ByNVars out=ByNVars;
   grid N*Vars=Prob /
      axis1=10 to 1000 by 25 /* Sample size */
      axis2=2 to 300 by 3 ;/* # variables aggregated */
run;
quit;

goptions reset=all gunit=pt ftext='Garamond/bo' ctext=BLACK htitle=14 htext=10
norotate vsize=4.0in hsize=8.0in;
GOPTIONS DEVICE=activex;
ODS HTML path="&Loc\Figures\" body="NVars2.htm";
goptions border;
proc g3d data=ByNVars gout = work.ThreeWay;
   plot N*Vars=Prob/rotate=150 tilt=80 grid name="NVars"
      caxis=black ctext=black ctop=blue cbottom=blue
      zticknum=10 zmin=0 zmax=1 yticknum=10 xticknum=10 xtype=3;
   label N="Sample size" Vars="No. variables" Prob="Probability";
run; quit;
Title1;
ODS HTML CLOSE;
/* ------------------------------------------------------------------------ */
/* Likehood of attaining a normal composite based on interpolated sample size and rho*/
/* ------------------------------------------------------------------------ */
proc sort data=Normal2; by N Rho; run;
proc means data=Normal2 noprint;
   output out=ByNRho(drop= _Type_ _Freq_) Mean=Prob;
   var Prob;
   by N Rho;
run;
proc g3grid data=ByNRho out=ByNRho;
   grid N*Rho=Prob /
      axis1=10 to 1000 by 25 /* Sample size */
/* correlation b/w variables */

run;
quit;

data ByNRho;
set ByNRho;
if Prob<0 then Prob=0;
run;

/* Likehood of attaining a normal composite based on interpolated # variables aggregated and rho */
proc sort data=Normal2; by Vars Rho; run;
proc means data=Normal2 noprint;
output out=ByVarsRho(drop=_Type_ _Freq_) Mean=Prob;
by Vars Rho;
run;
proc g3grid data=ByVarsRho out=ByVarsRho;
grid Vars*Rho=Prob / axis1=2 to 300 by 3 /* # variables aggregated */
axis2=0 to 1 by .01 /* correlation b/w variables */
run; quit; Title1;
ODS HTML CLOSE;
*/
APPENDIX F

SAS Code for Illustrating Transformation Theory
/* Dichotomous Transformations */
proc IML;
    START SIM;
        call randseed(123456); /* seed */
        X=J(N,7,.);
        xx=J(N,1,1);
        call randgen(xx,'POISSON',6);
        X[,1]=xx;
        call randgen(xx,'POISSON',10);
        X[,2]=xx;
        call randgen(xx,'NORMAL');
        X[,3]=2+sqrt(0.75)*xx;
        call randgen(xx,'UNIFORM');
        X[,4]=100*xx;
        m5 = RANDMULTINOMIAL(N,T,p_X5); /* Random values from multinomial */
        m6 = RANDMULTINOMIAL(N,T,p_X6); /* Random values from multinomial */
        do i=1 to N;
            X[i,5]=(ncol(m5)+1)-loc(m5[i,]); /* reverse order values */
            X[i,6]=(ncol(m6)+1)-loc(m6[i,]); /* reverse order values */
        end;
        call randgen(xx,'BINOMIAL',6,.25);
        X[,7]=xx;
    FINISH SIM;
    START DICHOTOMIZE;
        Y=J(N,7,.);
        Y[,1]=X[,1]>4;
        Y[,2]=X[,2]>8;
        Y[,3]=X[,3]>1.5;
        Y[,4]=X[,4]>75;
        Y[,5]=X[,5]>1; /* 1=High, 2=Moderate, 3=Low */
        Y[,6]=X[,6]>3; /* 5=Very likely, 4=Likely, 3=Possibly, 2=Probably not, 1=Definitely not */
        Y[,7]=X[,7]>2;
    FINISH DICHOTOMIZE;
    START Mean;
        Mean_X = X[:,]'; /* Sample mean for each class */
        Mean_Y = Y[:,]'; /* Sample mean for each class */
    FINISH Mean;
    START Variance;
        Var_X=J(7,1,1); Var_Y=Var_X;
        ss_x=J(7,1,1); ss_y=ss_x;
        sum_x=J(7,1,1); sum_y=sum_x;
        ss_x=X[##,]';
        sum_x=X[+,]';
        do i=1 to 7;
    FINISH Variance;
Var_X[i]=(ss_x[i]-(sum_x[i]**2)/N)/(N-1);
end;

ss_y=Y[#];
sum_y=Y[+];
do i=1 to 7;
Var_Y[i]=(ss_y[i]-(sum_y[i]**2)/N)/(N-1);
end;
FINISH Variance;

START MAIN;
N=100000;
p_X5 = {0.35,0.40,0.25}; /* Probability of each class */
p_X6 = {0.05,0.15,0.25,0.35,0.20}; /* Probability of each class */
T = 1; /* number of trials */
RUN SIM;
RUN DICHOTOMIZE;
RUN MEAN;
RUN VARIANCE;
Transform=X||Y;
print Mean_X[f=10.4 l="Raw random variables" c="Mean"] Var_X[f=10.4 c="Variance"];
print Mean_Y[f=10.4 l="Transformed random variables" c="Mean"] Var_Y[f=10.4 c="Variance"];
CREATE Transform FROM Transform [c=("X1":"X7","Y1":"Y7")];
APPEND FROM Transform;
FINISH MAIN;
runquit;
data Transform;
  set transform;
  mean=mean(of Y1-Y7);
  sum=sum(of Y1-Y7);
run;
goptions reset=all gunit=pt ftext='Garamond' ctext=BLACK htitle=12 htext=8
  norotate vsize=4in hsize=5in noborder;
%macro pdf(var=,href=,S=,E=,by=,bin=);
  proc univariate data=Transform CIQUANTDF(alpha=.05);*noprint;
  var &var ;
  title1 "Probability mass function of raw variable";
  histogram / href=&href chref=red endpoints=&S to &E by &by MAXNBIN=&bin
    name='pdf' NOHLABEL ;
  inset mean median var="Variance" / pos = ne format = 6.3;
run;
%mend pdf;

%pdf(var=x1,href=4,S=0,E=16,by=1,bin=17);
%pdf(var=x2,href=8,S=1,E=22,by=1,bin=23);
%pdf(var=x3,href=1.5,S=-1.0,E=4.8,by=1,bin=10);
%pdf(var=x4,href=75,S=0,E=100,by=5,bin=11);
%pdf(var=x5,href=2,S=1,E=3,by=1,bin=3);
%pdf(var=x6,href=4,S=1,E=5,by=1,bin=5);
%pdf(var=x7,href=2,S=0,E=7,by=1,bin=8);

290
data Mult;
  do i=1 to 1000000;
    x1=RANPOI(123456,1);
    x2=RANPOI(987654,4);
    if x1=0 then y1=1;
    else if x1=1 then y1=2;
    else if x1=2 then y1=3;
    else y1=4;
    if x2<=2 then y2=1;
    else if x2<=4 then y2=2;
    else if x2<=6 then y2=3;
    else y2=4;
    output;
  end;
run;

proc freq data=Mult;
  table y1 y2;
run;

* Do NOT make changes to these macros ;
%macro Openloop(j=);
  %let k=k0;
  %do i=1 %to &j;
    %if %eval(&i)>1 %then %let k = &k-%eval(&i-1);
    /* %sysfunc(byte(59) generates a semicolon so the code can run */
    do k&i=0 to T-%eval(byte(59)) ;
      %if %eval(&i)=%eval(&j) %then k%eval(&j+1) = T-%k%eval(&i) %sysfunc(byte(59)) ;
    %end;
%end;
%mend Openloop;

%macro Closeloop(j=);
  %do i=1 %to &j;
    end %sysfunc(byte(59)) ;
  %end;
%mend Closeloop;

%macro MultinomialProb;
  %Let Prob=fact(T); %Let Fact=1;
  %do i=1 %to %eval(&c+1);
    %Let Prob = &Prob*(prob[&i]##k&i);
    %Let Fact = &Fact*fact(k&i);
  %end;
  PMF = (&Prob)/(&Fact);
%end;
%mend MultinomialProb;

%macro SaveMult(j=) ;
  %Let Class=;
  %do i=1 %to &j;
    %Let Class = &Class k&i;
    Mult[r,&i]=k&i;
  %end;
  Mult[r,&i]=PMF; Mult[r,%eval(&i+1)]=CDF;
%end;
%mend SaveMult;

* Only make changes to prob and T ;
proc IML;
proc = {0.20, 0.25, 0.35, 0.20}; /* Probability of each class */
T = 4; /* number of trials */
prob = {
0.20,
0.25,
0.35,
0.20,
};
/* Probability of each class */
k = nrow(Prob); /* Number of classes */
call symputx('c', k-1); /* Generate Macro variable C= k-1 */

START SS;
n=1;
do i=1 to k-1;
   n=n*(T+i);
end;
n=n/fact(k-1);
Mult=J(n, k+2, 0);
FINISH SS;

START CDF;
%MultinomialProb;
CDF=CDF+PMF;
%SaveMult(j=%eval(&c+1));
FINISH CDF;

START Main;
RUN SS; /* Sample space for multinomial distribution */
CDF=0; k0=0; r=0; /* Initialize values */
%Openloop(j=&c) ;
r=r+1; /* Row counter */
RUN CDF;
%Closeloop(j=&c) ;
call symputx('maxPMF', max(Mult[,k+1]));
CREATE Mult FROM Mult [c=("k1","k%eval(&c+1)" || "PMF" || "CDF")];
APPEND FROM Mult;
FINISH Main;
run; quit;

data Mult;
set Mult;
n=&c+1;
SD=k1/n; D=k2/n; A=k3/n; SA=k4/n;
if SA >= .5 then y3=4;
else if sum(A,SA) >= .5 then y3=3;
else if sum(D,A,SA) >= .5 then y3=2;
else y3=1;
run;

proc tabulate data=Mult format=10.4;
class y3;
var pmf;
table y3*pmf;
run;
/* Continuous Order Statistics */
/* */
proc iml;
START MAIN;
r=1000000;
n=5;
k=3;
LB=-.5;
UB=.5;
x=j(n,1,.)

m=j(r,2,0);
do j=1 to r;
    do i=1 to n;
        x[i]=rannor(1);
    end;
call sort(x, {1});
m[j,1] = x[k];
    if x[k]>=LB & x[k]<=UB then m[j,2]=1;
end;
p=m[,2][+] / r;
print p[f=8.6];
FINISH MAIN;
RUN; QUIT;

/* */
data _null_;  
p1=probnorm(0.5);
p2=probnorm(-0.5);
prob=0;
do i=3 to 5;
    prob=prob+comb(5,i)*((p1**i)*((1-p1)**(5-i))-(p2**i)*((1-p2)**(5-i)));
end;
put prob=;
prob=0;
do i=0 to 2;
    prob=prob+comb(5,i)*((p2**i)*((1-p2)**(5-i))-(p1**i)*((1-p1)**(5-i)));
end;
put prob=;
run;
/* Discrete Order Statistics */
data _null_;  
n=15;
k=8;
LOS=7;
UOS=9;
P1=.3;
P2=.3;
P3=.4;
C=0;
do i=0 to n-LOS;
    do j=0 to UOS-1;
        if i+j<=n then C=C+(fact(n)/(fact(i)*fact(n-i)*fact(j)))*(Pi1**i)*
            (Pi2**(n-i-j))*(Pi3**j);
    end;
end;
put C=;

C=0;
do j=0 to UOS-1;
    do i=0 to n-LOS;
        if i+j<=n then C=C+(fact(n)/(fact(i)*fact(n-i-j)*fact(j)))*(Pi1**i)*
            (Pi2**(n-i-j))*(Pi3**j);
    end;
end;
put C=;

C=cdf('binom',UOS-1,Pi3,n)-cdf('binom',LOS-1,Pi2+Pi3,n);
put C=;
run;

%Let N = 15; /* Sample size */
%Let sim = 1000000; /* Number of simulations */
%Let prob = {0.1,0.2,0.3,0.25,0.15}; /* Probability of each class */
%macro CheckCoverage(median=, LR=, UR=);
proc IML;
    START SIM;
        do j=1 to &sim;
            Y = j(&N,
                1,
                .);
            /* categorical variables */
            X = RANDMULTINOMIAL(&N,1,
                p[j,]);/* Random values from multinomial */
            do i=1 to &N;
                Y[i,1]=loc(X[i,]);
            end;
            call sort(Y, {1});
            if Y[&LR] <= &median & Y[&UR] >= &median then C[j]=1;
        end;
    FINISH SIM;

    START MAIN;
        seed= 1;
        p=j(&sim,nrow(&prob),.);/* probability distribution of each variable */
        C=j(&sim,1,0); /* probability distribution of each variable */
        do i=1 to &sim;
            p[i,]=&prob;
        end;
        RUN SIM;
        Coverage=C[+]/&sim;
        print Coverage[f=percent12.2];
    FINISH MAIN;
run;quit;
%mend CheckCoverage;
%CheckCoverage(median=3, LR=7, UR=9);
/* */
data _null_;
    n=10;
    Pi1=.45;
\[ \begin{align*}
Pi_2 &= 0.3; \\
Pi_3 &= 0.25; \\
LOS &= 5; \\
UOS &= 6;
\end{align*} \]

\[ C = \text{cdf('binom',UOS,1,\Pi_3,n)} - \text{cdf('binom',LOS-1,\Pi_2+\Pi_3,n)}; \]

\[ \text{put LOS=} \quad \text{UOS=} \quad \text{C=}; \]

\[ \begin{align*}
LOS &= 4; \\
UOS &= 7;
\end{align*} \]

\[ C = \text{cdf('binom',UOS-1,\Pi_3,n)} - \text{cdf('binom',LOS-1,\Pi_2+\Pi_3,n)}; \]

\[ \text{put LOS=} \quad \text{UOS=} \quad \text{C=}; \]

\[ \begin{align*}
LOS &= 3; \\
UOS &= 8;
\end{align*} \]

\[ C = \text{cdf('binom',UOS-1,\Pi_3,n)} - \text{cdf('binom',LOS-1,\Pi_2+\Pi_3,n)}; \]

\[ \text{put LOS=} \quad \text{UOS=} \quad \text{C=}; \]

\[ \text{run;} \]

/*  *----------------------------------------------------------------------------------------------------------------------------*/

%Let N = 25; /* Sample size */
%Let V = 10;
%Let sim = 1000000; /* Number of simulations */
%Let prob = {0,20,0,25,0,30,0,25}; /* Probability of each class */

%Macro CheckCoverage(median=, LR=, UR=);
proc IML;
START SIM;
START MAIN;
run;
/*  *----------------------------------------------------------------------------------------------------------------------------*/

%Let N = 25; /* Sample size */
%Let V = 10;
%Let sim = 1000000; /* Number of simulations */
%Let prob = {0,20,0,25,0,30,0,25}; /* Probability of each class */

%Macro CheckCoverage(median=, LR=, UR=);
proc IML;
START SIM;
START MAIN;
run;
/*  *----------------------------------------------------------------------------------------------------------------------------*/
p[i,]=&prob;
end;

RUN SIM;
Cov=Cov[:];
CI=median(CI);
print LR[c="Lower rank" f=15.] UR[c="Upper rank" f=15.] CI[c="Confidence interval"
c=\{"LCI" "UCI"\}] Cov[c="Coverage" f=percent12.2];
FINISH MAIN;
orunquit;
%mend CheckCoverage;
%CheckCoverage(median=3, LR=5, UR=6);
%CheckCoverage(median=3, LR=4, UR=7);
%CheckCoverage(median=3, LR=3, UR=8);
/* --------------------------------------------- */
APPENDIX G

SAS Code for Illustrating Simple Random Sampling
/* ---------------------- */
/* Simple Random Sampling with replacement */
/* ---------------------- */
data Pop;
input X @@;
datalines;
1 2 2 3
;
run;

proc iml;
use Pop;
read all var _num_ into Pop;
close Pop;
START FREQ;
scale=unique(tot)';
counts=j(nrow(scale),3,0);
counts[,1]=scale;
do i=1 to nrow(scale);
counts[i,2]=sum(tot=scale[i]);
counts[i,3]=counts[i,2]/sim;
end;
FINISH FREQ;

START BIAS;
bias=j(sim,1,0);
do i=1 to sim;
bias[i]=2*tot[i]-T;
end;
bias=abs(bias[+])/sim;
FINISH BIAS;

START SRSWR;
samp=j(2,1,0);
do i=1 to 2;
select=ceil(nrow(Pop)*ranuni(123));
samp[i]=Pop[select];
end;
FINISH SRSWR;

START MAIN;
Pop=Pop';
Print Pop;
Pop=Pop';
T=8;
sim=100000;
tot=j(sim,1,0);
do j=1 to sim;
RUN SRSWR;
tot[j]=samp[+];
end;
RUN FREQ;
RUN BIAS;
Print sim bias;
Print counts[c="Total","Count","Percentage"] f=10.4];
FINISH MAIN;
RUN; QUIT;
/* Illustrate impact of finite population correction factor on variance */

%macro SingleVar(fpc=);
Proc IML;
START STATS(pop,fpc,ds,n) global (mean,var,se);
    mean=ds[:];                /* mean */
    if pop=1 then denom=n;    /* Run when sample = population */
    else denom=n-1;           /* Run when sample ^= population */
    Y=ds##2;                  /* Estimate of Y */
    if "&fpc"="Y" then
        var=fpc*(Y[+]n*mean##2)/denom; /* Correct variance estimate */
    else var=(Y[+]n*mean##2)/denom; /* Incorrect variance estimate */
    se=sqrt(var/denom);
    FINISH STATS;

START POPULATION;            /* Generate a population */
    Pop=j(popN,2,.) ;
    Pop[,1]=(1:popN);        /* Generate popN random values from Normal distribution */
    Pop[,2]=sqrt(var)*normal(j(popN,1,seed))+mu;
    fpc=1;                  /* fpc=1 since the sample and population are the same */
    RUN STATS(1,fpc,Pop[,2],popN);
    mu_Pop=mean;
    total=Pop[+];           /* population total */
    print "Population parameters";
    print mean[f=15.8] var[f=15.8] se[f=15.8] total[f=20.8];,
    FINISH POPULATION;

START SRS;                    /* Generate simple random sample without replacement */
    fpc=1-popN/popN;        /* finite population correction factor */
    select=uniform(j(popN,1,seed))<1-fpc; /* randomly select cases from population*/
    sum=select[+];          /* number of randomly selected cases */
    if sum>sampN then do;   /* oversampled the population (N too large) */
        Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */
        Samp=Samp[1:sampN];  /* keep the first n randomly selected values */
    end;
    else if sum<sampN then do; /* undersampled the population (N too small) */
        do while (sum<sampN); /* add more randomly selected case to reach sampN*/
            if sum>0 then Y=j(sum,2,.) ;
            Y=Pop[loc(select=0),];
            id=Y[ceil(nrow(Y)*uniform(seed)),1];
            select[id]=1;
            sum=select[+];       /* number of randomly selected cases */
        end;
        Samp=Pop[loc(select=1),2];
    end;
    else Samp=Pop[loc(select=1),2];
    total=popN*mean;         /* estimated population total */
    FINISH SRS;

START CHECKCI;
    z=probit(1-alpha/2);
    FINISH CHECKCI;
LB = mean - z*se;
UB = mean + z*se;
if LB < mu_Pop & mu_Pop < UB then inCI = 1;
else inCI = 0;
FINISH CHECKCI;

START SAMPDIST; /* Generate sampling distribution */
dist = j(sim, 5, .); stat = j(1, 5, .);
do i = 1 to sim;
    RUN SRS;
    RUN STATS(0, fpc, Samp, sampN);
    RUN CHECKCI;
    dist[i, 1] = mean;
    dist[i, 2] = var;
    dist[i, 3] = se;
    dist[i, 4] = total;
    dist[i, 5] = inCI;
end;
do i = 1 to 5;
    stat[i] = dist[, i];
end;
print "Sampling distribution (fpc=&fpc)";
print stat[l="f=17.8 c=("Mean" || "Variance" || "SE" || "Total" || "Coverage")];
FINISH SAMPDIST;

START MAIN;
    seed = 123;
    alpha = .05;
    sim = 1000;
    popN = 10000; /* population size */
    sampN = 5000; /* sample size */
    mu = 100; /* population mean */
    var = 25; /* population var */
RUN POPULATION;
RUN SAMPDIST;
Pop = Pop[, 2];
CREATE Population FROM Pop [c="X1"]; APPEND FROM Pop;
CREATE SRS FROM dist [c=("Mean" || "Variance" || "SE" || "Total" || "inCI")];
APPEND FROM dist;
FINISH MAIN;
RUN; QUIT;
%mend SingleVar;
title1 "Sample statistics without the population correction factor";
%SingleVar(fpc=N);
title1 "Sample statistics with the population correction factor";
%SingleVar(fpc=Y);
title1;
%macro MultiVar(vars=,fpc=); /* number of variables */
Proc IML;
START SIMCORR;
    Cholesky = root(Corr);
    N = rannor(J(popN,&vars,seed)) ; /* Generate 1,500 cases */
    Z = N*Cholesky;
    U = CDF('NORMAL',Z);
    /* GLD values to produce a normal distribution with skew=0 and kurt=0 */
    L11 = 7.534508E-14;
    L12 = 0.1974513695;
    L13 = 0.1349124547;
    L14 = 0.1349124547;
    Pop = j(popN,&vars,.,);
    do i=1 to &vars;
        Pop[,i] = L11 + (U[,i]**L13 - (1-U[,i])**L14)/L12;
        Pop[,i] = sqrt(var)*Pop[,i] + mu;
    end;
FINISH SIMCORR;
START STATS(pop,fpc,ds,n) global (mean,var,se);
    if pop=1 then ds=ds || ds[.,.]; /* mean */
    mean = ds[.,.];
    if pop=1 then denom = n; /* Run when sample = population */
    else denom = n-1; /* Run when sample ^= population */
    Y = ds##2;
    var = (fpc*(Y[+,.] - n*mean##2)/denom); /* GLD values to produce a normal distribution with skew=0 and kurt=0 */
    mean = mean;
    se = sqrt(var/denom);
FINISH STATS;
START POPULATION; /* Generate a population */
    fpc = 1; /* fpc=1 since the sample and population are the same */
    id = (1:popN); /* case id */
    RUN SIMCORR;
    RUN STATS(1,fpc,Pop,popN);
    total = Pop[+,.]; /* population total */
    Pop = id || Pop;
    mu_Pop = mean;
    print "Population parameters";
    print mean[f=15.8 r=("X1"||"X&vars" || "Mean") var[f=15.8] se[f=15.8] total[f=20.8]...; /* number of variables */
    Population = Pop[,2:ncol(Pop)];
    CREATE Population FROM Population [c=("X1"||"X&vars" || "Mean")];
    APPEND FROM Population;
    free Population;
FINISH POPULATION;
START SRS; /* Generate simple random sample without replacement */
    if "&fpc"="Y" then
        fpc = 1-sampN/popN; /* finite population correction factor */
    else fpc = 1;
select=uniform(j(popN,1,seed))<sampN/popN; /* randomly select cases from population */
sum=select[+]; /* number of randomly selected cases */
Samp=j(sampN,&vars,:);
if sum>sampN then do; /* oversampled the population (N too large)
  x=Pop[loc(select=1),2:ncol(Pop)];
  Samp=x[1:sampN,]; /* keep the first sampN randomly selected values */
end;
else if sum<sampN then
  do while (sum<sampN); /* add more randomly selected case to reach sampN */
    if sum>0 then 
      Y=j(sum,&vars+1,1);
      id=Y[ceil(nrow(Y)*uniform(seed),1)];
      select[id]=1;
      sum=sum[+]; /* number of randomly selected cases */
  end;
  Samp=Pop[loc(select=1),2:ncol(Pop)];
end;
else Samp=Pop[loc(select=1),2:ncol(Pop)];
FINISH SRS;

START CHECKCI;
z=probit(1-alpha/2);
LB=mean-z#se;
UB=mean+z#se;
inCI=j(&vars+1,1,0);
do i=1 to &vars+1;
  if LB[i]<mu_Pop[i] & mu_Pop[i]<UB[i] then inCI[i]=1;
end;
FINISH CHECKCI;

START SAMPDIST; /* Generate sampling distribution */
stat=j(sim,5,);
do j=1 to sim;
  RUN SRS;
  RUN STATS(0,fpc,Samp,sampN);
  total=popN*mean; /* estimated population total */
  RUN CHECKCI;
  stat[j,1]=mean[&vars+1];
  stat[j,2]=var[&vars+1];
  stat[j,3]=se[&vars+1];
  stat[j,4]=total[&vars+1];
  stat[j,5]=inCI[&vars+1];
  end;
  stat=stat[:,];
  print "Sampling distribution (fpc=&fpc)";
  print stat[f=17.8 c=("Mean" "Variance" "SE" "Total" "Coverage")];
FINISH SAMPDIST;

START MAIN;
  seed=12345;
  alpha=.05;
  sim=1000;
  popN=100000; /* population size */
sampN=20000; /* sample size */
mu=100; /* population mean */
var=25; /* population var */
Corr=&rho;
RUN POPULATION;
RUN SAMPDIST;

CREATE SRS FROM Samp [c=('X1':"X&vars" || "Mean")];
APPEND FROM Samp;
FINISH MAIN;
RUN; QUIT;

proc corr data=Population;
var X:;
run;
%mend MultiVar;
%LET Rho=
{1.0 0.0 0.0 0.0,
 0.0 1.0 0.0 0.0,
 0.0 0.0 1.0 0.0,
 0.0 0.0 0.0 1.0} ;
title1 "Sample statistics without the population correction factor";
%MultiVar(vars=4,fpc=N);
title1 "Sample statistics with the population correction factor";
%MultiVar(vars=4,fpc=Y);

%LET Rho=
{1.0 0.1 0.2 0.3,
 0.1 1.0 0.4 0.5,
 0.2 0.4 1.0 0.6,
 0.3 0.5 0.6 1.0} ;
title1 "Sample statistics without the population correction factor";
%MultiVar(vars=4,fpc=N);
title1 "Sample statistics with the population correction factor";
%MultiVar(vars=4,fpc=Y);
title1;

/-------------------------------------------------------------------------*/
/* Produce Figure 13: Illustration of Sampling Distribution */
/-------------------------------------------------------------------------*/

%macro SD(sim=);
Proc IML;
START POPULATION; /* Generate a population */
  Pop=j(popN,2,.);
  Pop[1]=(1:popN); /* Generate popN random values from uniform distribution */
  Pop[2]=uniform(j(popN,1,seed));
FINISH POPULATION;

START SRS; /* Generate simple random sample without replacement */
  select=uniform(j(popN,1,seed))<sampN/popN; /* randomly select cases from population */
  sum=select[+]; /* number of randomly selected cases */
  if sum>sampN then do; /* oversampled the population (N too large) */
    Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */
    Samp=Samp[1:sampN]; /* keep the first nn randomly selected values */
  endif;
%mend SD;
end;

else if sum<sampN then
    do; /* undersampled the population (N too small) */
        do while (sum<sampN); /* add more randomly selected case to reach sampN*/
            if sum>0 then Y=(sum,2,);
            Y=Pop[loc(select=0),];
            id=Y[ceil(nrow(Y)*uniform(seed)),1];
            select[id]=1;
            sum=sum[+]; /* number of randomly selected cases*/
        end;
        Samp=Pop[loc(select=1),2];
    end;
else Samp=Pop[loc(select=1),2];
FINISH SRS;

START SAMPDIST; /* Generate sampling distribution */
    ave=j(sim,1,);
    do i=1 to sim;
        RUN SRS;
        ave[i]=samp[;];
    end;
FINISH SAMPDIST;

START MAIN;
    seed=123;
    sim=&sim;
    popN=50; /* population size */
    sampN=10; /* sample size */

RUN POPULATION;
RUN SAMPDIST;
CREATE Population FROM Pop [c=('Pop')];
APPEND FROM Pop;
CREATE SampDist FROM ave [c=('Samp')];
APPEND FROM ave;
FINISH MAIN;
RUN; QUIT;
%mend SD;
%
SD(sim=1000);

goptions reset=all gunit=pt ftext='Garamond' ctext=BLACK htitle=14 htext=12 norotate vsize=3.5in hsize=5in noborder;

proc univariate data=Population ;
    title "Population distribution of a uniformly distributed variable";
    label Pop="Population (N=50)";
    var Pop;
    histogram Pop/endpoints=1 to 50 vaxis=0 to 2.5 by .5 name="Pop50";
run;

%macro SampDist;
%do i=1 %to 3;
    %if &i=1 %then %let N=1000;
    %else %if &i=2 %then %let N=10000;
    %else %let N=100000;
%end;


%SD(sim=&N);
proc univariate data=SampDist ;
   title "Sampling distribution of the mean of &N uniform variables";
       label Samp="Sample (n=10)";
   var Samp;
       histogram Samp/normal name="SD&N";
run;
%mend SampDist;

%options reset=all device=EMF gsfmode=replace gsfname=gout ;
filename gout 'C:\Users\Cristian\Desktop\Fig17.emf';

proc greplay igout=work.GSEG tc=sashelp.templt
       template=l2r2 nofs;
       treplay 1:Pop50 3:SD1000 2:SD10000 4:SD100000 ;
run; quit;
APPENDIX H

SAS Code for Constructing Parametric and Nonparametric Confidence Intervals
/* Check of coverage probability of CIs constructed from different probability distributions */

Proc IML;
   START POPULATION; /* Generate a population */
      Pop=j(popN,2,1:popN);
      if Type="Uniform" then Pop[,2]=Pop[,1]; /* Generate ranks 1-50 */
      else if Type="Bernouli" then
         Pop[,2]=ranbin(j(popN,1,seed1),1,p); /* (n,# vars,seed),# Bernouli trials, prob */
      else if Type="Binomial" then
         Pop[,2]=ranbin(j(popN,1,seed1),sampN,p);
      else Pop[,2]=RANPOI(j(popN,1,seed1),lam);
      mu=Pop[,2];
   FINISH POPULATION;

   START SRS; /* Generate simple random sample without replacement */
      select=uniform(j(popN,1,seed2))<sampN/popN; /* randomly select cases from pop. */
      sum=select[+]; /* number of randomly selected cases */
      if sum>sampN then do; /* oversampled the population (N too large) */
         Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */
         Samp=Samp[1:sampN]; /* keep the first nn randomly selected values */
      end;
      else if sum<sampN then do; /* undersampled the population (N too small) */
         /* add more randomly selected cases to reach sampN */
         do while (sum<sampN);
            if sum>0 then Y=j(sum,2,1:popN);
            Y=Pop[loc(select=0),];
            id=Y[ceil(nrow(Y)*uniform(seed2)),1];
            select[id]=1;
            sum=select[+]; /* number of randomly selected cases */
         end;
         Samp=Pop[loc(select=1),2];
      end;
   end SRS;

   START CI;
      mean=samp[1];
      Y=samp##2;
      var=fpc*(Y[+]-sampN*mean##2)/(sampN-1);
      se=sqrt(var/sampN);
      t=TINV(1-alpha/2,sampN-1);
      LCI=mean-t*se;
      UCI=mean+t*se;
   FINISH CI;

   START SAMPDIST; /* Generate sampling distribution */
      fpc=1-sampN/popN; /* finite population correction factor */
      Dist=j(sim,6,0);
      do i=1 to sim;
         RUN SRS;
   end SAMPDIST;

307
RUN CI;
Dist[,1]=mean;
Dist[,2]=var;
Dist[,3]=se;
Dist[,4]=LCI;
Dist[,5]=UCI;
if LCI<mu & UCI>mu then Dist[,6]=1;
end;
FINISH SAMPDIST;

START MAIN;
seed1=123456;
seed2=67891;
p=0.1;
lam=10;
alpha=.05;
sim=100000;
popN=100;        /* population size */
sampN=40;        /* sample size */
*Type="Uniform";
*Type="Bernouli";
*Type="Binomial";
Type="Poisson";
RUN POPULATION;
RUN SAMPDIST;
Pop=Pop[,2];

CREATE Population FROM Pop [c=("Pop")];
APPEND FROM Pop;
CREATE SampDist FROM Dist
[c=("Mean" || "VAR" || "SE" || "LCI" || "UCI" || "Coverage")];
APPEND FROM Dist;
Dist=Dist[,];
print mu[f=6.2] t[f=6.2] Dist
[c=("Mean" || "VAR" || "SE" || "LCI" || "UCI" || "Coverage") f=6.3];
FINISH MAIN;
RUN; QUIT;

/*--------------------------------------------------------------------------------*/
/* Produce Figure 18: Illustration of the distribution of CIs about the unknown population parameter */
/*--------------------------------------------------------------------------------*/

Proc IML;
START POPULATION;    /* Generate a population */
Pop=j(popN,2,1);
Pop[,1]=1:popN;     /* Generate popN random values from Poisson distribution */
Pop[,2]=RANPOI(j(popN,1,seed1),lam);
mu=Pop[,2];
FINISH POPULATION;

START SRS;        /* Generate simple random sample without replacement */
fpc=1-sampN/popN;    /* finite population correction factor */
select=uniform(j(popN,1,seed2))<sampN/popN;  /* randomly select cases from pop. */
sum=select[+];       /* number of randomly selected cases */
if sum>sampN then do; /* oversampled the population (N too large) */  
   Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */  
   Samp=Samp[1:sampN]; /* keep the first nn randomly selected values */  
end;

else if sum<sampN then  
   do; /* undersampled the population (N too small) */  
      /* add more randomly selected cases to reach sampN */  
      do while (sum<sampN);  
         if sum>0 then Y=j(sum,2,1);  
         Y=Pop[loc(select=0),];  
         id=Y[ceil(nrow(Y)*uniform(seed2)),1];  
         select[id]=1; /* number of randomly selected cases */  
         sum=sum+1;  
      end;  
   end;

   Samp=Pop[loc(select=1),2];
end;
else Samp=Pop[loc(select=1),2];

FINISH SRS;

START CI;
   Y=samp##2;  
   mean=samp[1,];  
   var=fpc*(Y[+]-sampN*mean##2)/(sampN-1);  
   se=sqrt(var/sampN);  
   t=TINV(1-alpha/2,sampN-1);  
   LCI=mean-t*se;  
   UCI=mean+t*se;
FINISH CI;

START SAMPDIST;  /* Generate sampling distribution */  
   Dist=j(sim,4,0);  
   do i=1 to sim;  
      RUN SRS;  
      RUN CI;  
      Dist[i,1]=mean;  
      Dist[i,2]=LCI;  
      Dist[i,3]=UCI;  
      if LCI<mu & UCI>mu then Dist[i,4]=1;  
   end;
FINISH SAMPDIST;

START MAIN;
   seed1=12345;  
   seed2=678910;  
   alpha=.05;  
   sim=100;  
   lam=10;  
   popN=1000;  /* population size */  
   sampN=200;  /* sample size */
RUN POPULATION;
RUN SAMPDIST;
p=Dist[,4];  
Pop=Pop[,2];  
Dist=(1:sim)’ | | Dist;

309
CREATE Population FROM Pop [c="Pop"];  
APPEND FROM Pop;  
CREATE SampDist FROM Dist [c="ID" | "Close" | "Low" | "High" | "InCI"];  
APPEND FROM Dist;  
FINISH MAIN;  
RUN; QUIT;

data SampDist2;  
set SampDist;  
drop Low High Close ;  
if InCI=1 then do;  
      CI=High; output;  
      CI=Low; output;  
      CI=Close; output;  
end;  
else do;  
      nCI=High; output;  
      nCI=Low; output;  
      nCI=Close; output;  
end;  
run;

goptions reset=all gunit=pt ftext='Garamond' ctext=BLACK htitle=12 htext=8  
norotate vsize=3.5in hsize=5in noborder;  
symbol1 interpol=hilot  
   cv=black  
   width=1.2;  
symbol2 interpol=hilot  
   cv=red  
   width=1.2;  
axis1 order=(0 to 100 by 5)  
   offset=(3,3)  
   color=black  
   label="Simple random samples drawn from a population")  
   major=(height=5 width=0.1)  
   minor=(number=4 color=gray height=3 width=1)  
   width=2;  
axis2 color=black  
   label=none  
   major=(height=5)  
   minor=(number=3 color=gray height=0.5)  
   offset=(2,2);  

goptions device=EMF gsfmode=replace gsfname=gout;  
filename gout 'C:\Users\Cristian\Desktop\Fig18.emf';  
proc gplot data=SampDist2;  
   plot CI*ID nCI*ID/overlay haxis=axis1 vaxis=axis2 vref=10.174 cvref=blue name="CI";  
run;quit;

/*  *-------------------------------------------------------------------------------------------------------------------------------*/  
/* Constructing a CI on the sample median of continuous random variable  *-----------------------------------------------------------------------*/

310
data Continuous;
    do i=1 to 500;
        X=2*rannor(98765)+10;
        X=X**2.5;
        output;
    end;
    drop i;
run;

data_null_;
    a=240;
    b=260;
    p=0.5;
    n=500;
    c=0;
    do i=a to b-1;
        c=c+comb(n,i)*(p**i)*((1-p)**(n-i));
    end;
    put "CI coverage: " c;
run;

data_null_;
    alpha=0.05;
    p=0.5; /* quantile (median=0.5) */
    n=500; /* sample size */
    prob=0;
    j=1;
    LOS=1; /* Lower bound order statistic */
    do until(prob>=1-alpha/2 | j>=n); /* start search algorithm */
        prob=cdf('BINOM',j-1,p,n);
        if prob<=alpha/2 then LOS=j; /* Upper bound order statistic */
        j=j+1;
    end;
    if p<=.5 then UOS=j-1;
    else UOS=j;
    put "Lower bound order statistic: " LOS;
    put "Upper bound order statistic: " UOS;
    c=0;
    do i=LOS to UOS-1;
        c=c+comb(n,i)*(p**i)*((1-p)**(n-i));
    end;
    put "CI coverage: " c;
run;

/ * Produce Tables 11 & 12: Unadjusted and FPC adjusted coverage prob. and mean CIs for 3 Distributions */
/ */
/ */ Produce Tables 11 & 12: Unadjusted and FPC adjusted coverage prob. and mean CIs for 3 Distributions */
/ */
/* Run this code to generate N(50,25,0,0) */
data continuous;
    do i=1 to 150;
        x=5*rannor(12321)+50;
    end;
output;
end;
drop i;
run;

/*/ -------------------------------------- */

/*/ Run this code to generate the two skewed distributions */
DATA Continuous;
* Generate a Normal: mu=50, var=25, skew=2.0, kurt=10 ;
*L1=-0.5154104205;
*L2=-0.1783773597;
*L3=-0.0387687344;
*L4=-0.1168183637;

* Generate a Normal: mu=50, var=25, skew=2.0, kurt=10 ;
L1=0.5154104205;
L2=-0.1783773597;
L4=-0.0387687344;
L3=-0.1168183637;

DO i = 1 TO 150;
   U = RANUNI (111);
   X = L1 + (U**L3 - (1 - U)**L4) / L2 ;
   X = 5*X+50;
   OUTPUT;
END;
KEEP X ;
RUN;

/*/ -------------------------------------- */

/*/ This code will generate the values for Tables 11 & 12 one value at a time */
Proc IML;
use Continuous;
read all var _num_ into Pop;
close Continuous;

START POPULATION;
   Pop=Pop;
   popN=nrow(Pop);
   R=rank(Pop);
   Pop=R || Pop;
   if p=0.5 then do;
      k=(popN+1)/2;
      OS=median(Pop[,2]);
   end;
   else do;
      k=ceil(p*(popN+1));
      if k>popN then k=popN;
      OS=Pop[ loc(Pop[,1]=k),2];
   end;
   fpc=1-sampN/popN;
   print "Population parameters";
   print popN[c="Pop size"] sampN[c="Samp size"] f=12.0]
      fpc[c="FPC" f=percent8.2] p[c="Percentile" f=percent15.2]
      OS[c="Order statistic" f=18.2] alpha[c="Alpha" f=percent10.2]
      nominal[c="Nominal" f=percent10.2];
FINISH POPULATION;
START SRS; /* Generate simple random sample without replacement */
/* randomly select cases from population */
select=uniform((popN,1,seed1))<sampN/popN;
sum=select[+]; /* number of randomly selected cases */

if sum=sampN then Samp=Pop[loc(select=1),2];
else if sum>sampN then do; /* oversampled the population (N too large) */
    Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */
    Samp=Samp[1:sampN]; /* keep the first sampN randomly selected values */
end;
else do; /* undersampled the population (N too small) */
    Pop=Pop[| select; 
do until(sum >= sampN);
        Y=Pop[loc(Pop[3]=0),];
        select=j(nrow(Y),1,0);
        do until(select[+]>0);
            select=uniform(j(nrow(Y),1,seed1))<(sampN-sum)/(popN-sum);
        end;
        Y[3]=select;
        id=Y[loc(Y[3]=1),1];
        if sum+nrow(id)>sampN then id=id[1:(sampN-sum)];
        do zz=1 to nrow(id);
            Pop[loc(Pop[1]=id[zz],3]=1;
        end;
        sum=Pop[+,3];
    end;
    Samp=Pop[loc(Pop[3]=1),2];
    Pop=Pop[1:2];
end;

FINISH SRS;

START FindQuantile;
prob=0;
if p<=0.5 then do;
    j=1;
    LOS=1;
    do until((cdf('BINOM',j-1,p,sampN)
        -cdf('BINOM',LOS-1,p,sampN))>=1-alpha | j=sampN);
        if cdf('BINOM',j-1,p,sampN)<=alpha/2 then LOS=j;
    j=j+1;
end;
    UOS=j;
end;
else do;
    j=sampN; UOS=sampN+1;
    do until((cdf('BINOM',UOS-1,p,sampN)
        -cdf('BINOM',j-1,p,sampN))>=1-alpha | j=0);
        if cdf('BINOM',j-1,p,sampN)>=1-alpha/2 then UOS=j;
    j=j-1;
end;
    if UOS>sampN then UOS=sampN;
    if UOS<sampN then LOS=j-1;
        else LOS=j;
end;
if p=0.5 then k=(sampN+1)/2;
   else k=ceil(p*(sampN+1));
if k>sampN then k=sampN;
FINISH FindQuantile;

START Asymmetric;
R=rank(Samp);
Samp=R | | Samp;
if p=0.5 then SOS=median(Samp[2]);
   else SOS=Samp[loc(Samp[1]=k),2];
   LCI=Samp[loc(Samp[1]=LOS),2];   /* Lower bound observation */
   UCI=Samp[loc(Samp[1]=UOS),2];   /* Upper bound observation */
FINISH Asymmetric;

START FPC;    /* Adjust CI for Finite Population Correction */
LOS1=k-\sqrt{\text{fpc}}*(k-LOS +0.5);
UOS1=k+\sqrt{\text{fpc}}*(UOS-k +0.5);

LOS1=ceil(LOS1);
UOS1=floor(UOS1);
if LOS1<=0 then LOS1=1;
if UOS1>sampN then UOS1=sampN;

LCI2=Samp[loc(Samp[1]=LOS1),2];
UCI2=Samp[loc(Samp[1]=UOS1),2];
FINISH FPC;

START SAMPDIST;    /* Generate sampling distribution */
dist=j(sim,7,0);
do i=1 to sim;
   RUN SRS;
   RUN Asymmetric;
   RUN FPC;
   dist[i,1]=SOS;
   dist[i,2]=LCI;
   dist[i,3]=UCI;
   if LCI<=OS & UCI>=OS then dist[i,4]=1;
   dist[i,5]=LCI2;
   dist[i,6]=UCI2;
   if LCI2<=OS & UCI2>=OS then dist[i,7]=1;
   end;
dist=dist(:,
print "Sample statistics"
print k["k"] LOS1["LOS"] UOS1["UOS"]
   dist[="Order statistic" | | "LCI" | | "UCI" | | "Coverage" | | "LCI (adj)" | |
   "UCI (adj)" | | "Coverage (adj)""); f=10.3;
FINISH SAMPDIST;

START MAIN;    *seed1=123456;
   seed1=0;
   alpha=0.05;
   nominal=1-alpha;
   sim=100000;
   sampN=75;    /* Sample size */
p=0.90;    /* percentile */
/* Constructing a CI on the sample median of discrete random variable */
/* This code illustrates that the lower and upper bound CI estimates should be allowed to cross the median */
data _null_;<br>alpha=0.05;<br>p1=.35;<br>p2=.30;<br>p3=.35;<br>n=100;<br>m=(n+1)/2;<br><br>do i=n to 0 by -1;<br>    p=cdf('BINOM',i-1,Pi3,n);<br>    if p>=1-alpha/2 then UOS=i;<br>end;<br>put;<br><br>do i=0 to n;<br>    p=cdf('BINOM',i-1,Pi2+Pi3,n);<br>    if p<=alpha/2 then LOS=i;<br>end;<br>put;<br><br>C=cdf('BINOM',UOS-1,Pi3,n)-cdf('BINOM',LOS-1,Pi2+Pi3,n);<br>put LOS= UOS= C=;<br>put;<br><br>p1=cdf('BINOM',56-1,Pi2+Pi3,n);<br>put "When a=56 then p1= " p1;<br><br>p2=cdf('BINOM',45-1,Pi3,n);<br>put "When b=45 then p2= " p2;<br>diff=p2-p1;<br>put "...and the probability coverage is p2-p1= " diff;<br>put;<br><br>p3=cdf('BINOM',50-1,Pi2+Pi3,n);<br>p4=cdf('BINOM',51-1,Pi3,n);<br>diff=p4-p3;<br>put "if a=50 and b=51 then probability coverage is " diff;<br><br>maxLOS=quantile('BINOM',alpha,Pi2+Pi3,n);<br>minUOS=quantile('BINOM',1-alpha,Pi3,n)+1;<br>if maxLOS<1 then maxLOS=1;<br>if minUOS>n then minUOS=n;<br>put maxLOS= minUOS=;
min=1;
do j=minUOS to n;
   exit=0;
   if maxLOS=1 then do;
      LOS=1;
      UOS=ceil(m);
   end;
   else if minUOS=n then do;
      LOS=floor(m);
      UOS=n;
   end;
   else if cdf('BINOM',j-1,Pi3,n)-cdf('BINOM',maxLOS-1,Pi2+Pi3,n)<min then
      do i=maxLOS to 1 by -1;
         InCI=((i<=m) and (j>=m)) or ((i>=m) and (j<=m));
         if (InCI=1 and exit=0) then do;
            C=cdf('BINOM',j-1,Pi3,n)-cdf('BINOM',i-1,Pi2+Pi3,n);
            if C>=1-alpha and C<min then do;
               min=C;
               LOS=i;
               UOS=j;
               exit=1;
            end;
            else if C>min then i=0;
         end;
   end;
end;
C=cdf('BINOM',UOS-1,Pi3,n)-cdf('BINOM',LOS-1,Pi2+Pi3,n);
if LOS>UOS then do;
a=LOS;
   LOS=UOS;
   UOS=a;
end;
put
"Absolute tightest CI occurs when " LOS= UOS= C=;
run;
/* -------------------------------------------------------------------------- */
/* Produce Table 13: Coverage probabilities and mean CIs for X~Mult(1;0.15,0.2,0.3,0.22,0.13) */
/* -------------------------------------------------------------------------- */
Proc IML;
START SIM;
m=(n+1)/2; /* Order statistic of the median */
Cov=j(sim,1,0); /* Probability coverage */
C=j(sim,1,0); /* Probability coverage */
Cl=j(sim,2,0); /* Probability coverage */
Pr=j(sim,3,0); /* Probability coverage */
W=j(sim,1,0); /* Width of CI as function of max */
Y = j(n,sim,0); /* categorical variables */
R=j(n,sim,0); /* categorical variables */
do j=1 to sim;
   Cat = RANDMULTINOMIAL(n,1,prob); /* Random values from multinomial */
do i=1 to n;
Y[i,j]=loc(Cat[i,]);
end;
R[i,j]=rank(Y[i,j]);
end;
FINISH SIM;

START FREQ;

/* Value of the median */
X=J(1,2,);
if mod(n,2)=1 then do;
X[1]=Y[loc(R[,s]=m)];
X[2]=X[1];
end;
else do;
X[1]=Y[loc(R[,s]=floor(m)),s];
X[2]=Y[loc(R[,s]=ceil(m)),s];
end;

Scale=unique(Y[,s])';
Counts=J(nrow(Scale),3,);
Pi=J(3,1,0);
Counts[,1]=Scale;
do i=1 to nrow(Scale);

Counts[i,2]=sum(Y[,s]=Scale[i]);
Counts[i,3]=Counts[i,2]/n;
if Scale[i]<X[1] then Pi[1]=Pi[1]+Counts[i,3];
end;
Pr[s,1]=Pi[1];
Pr[s,2]=Pi[2];
Pr[s,3]=Pi[3];
FINISH FREQ;

START FindQuantile;
maxLOS=quantile('BINOM',alpha,Pi[2]+Pi[3],n);
minUOS=quantile('BINOM',1-alpha,Pi[3],n)+1;
if maxLOS<1 then maxLOS=1;
if minUOS>n then minUOS=n;

min=1;
do j=minUOS to n;
exit=0;
if maxLOS=1 then do;
LOS=1;
UOS=ceil(m);
end;
else if minUOS=n then do;
LOS=floor(m);
UOS=n;
end;
else if cdf('BINOM',j-1,Pi[3],n)-cdf('BINOM',maxLOS-1,Pi[2]+Pi[3],n)<min then do i=maxLOS to 1 by -1;
inCI=((i<=m) & (j>=m)) | ((i>=m) & (j<=m));
if (InCI=1 & exit=0) then do;
cc=cdf('BINOM',j-1,Pi[3],n);
end;
end;
END.
-cdf('BINOM',i-1,Pi[2]+Pi[3],n);
if cc>=1-alpha & cc<min then do;
    min=cc;
    LOS=i;
    UOS=j;
    exit=1;
end;
else if cc>min then i=0;
end;
end;
cc=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
diff=abs(cc-(1-alpha));
if Pi[1]>Pi[3] then do;
    cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    if diff>abs(cc2-(1-alpha)) then LOS=LOS+1;
end;
else if Pi[1]<Pi[3] then do;
    cc2=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
    if diff>abs(cc2-(1-alpha)) then UOS=UOS-1;
end;
else do;
    cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
    cc3=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    cc4=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    diff2=abs(cc2-(1-alpha));
    diff3=abs(cc3-(1-alpha));
    if diff>abs(cc4-(1-alpha)) then do;
        LOS=LOS+1;
        UOS=UOS-1;
    end;
    else if diff2>diff3 & diff>diff2 then LOS=LOS+1;
    else if diff2>diff3 & diff>diff3 then UOS=UOS-1;
    else if diff2<diff3 & diff>diff3 then UOS=UOS-1;
    else if diff2<diff3 & diff>diff2 then LOS=LOS+1;
end;
FINISH FindQuantile;

START COVERAGE;
if LOS<=UOS then do;
    LCI=Y[loc(R[,s]=LOS),s];
    UCI=Y[loc(R[,s]=UOS),s];
end;
else if UOS<LOS then do;
    LCI=Y[loc(R[,s]=UOS),s];
    UCI=Y[loc(R[,s]=LOS),s];
end;
if LCI <= med & UCI >= med then Cov[s]=1;/ * Checks whether median is in CI */
W[s]=UCI-LCI;
CI[s,1]=LCI;
CI[s,2]=UCI;
C[s]=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
FINISH COVERAGE;
START MAIN;
prob = {0.15,0.20,0.30,0.22,0.13}; /* Probability of each class */
med=3;
n=50;
sim=100000;
alpha=.01;
nominal=1-alpha;
call randseed(12345); /* seed used by RANDMULTINOMIAL */

RUN SIM;
do s=1 to sim;
    RUN FREQ;
    RUN FindQuantile;
    RUN COVERAGE;
end;
Pr=Pr[:];
Cov=Cov[:];
C=C[:];
CI=CI[:];
W=W[:]/4;

reset NOAUTONAME NONAME ;
print n[c="N"] Cov[c="Coverage" f=percent10.2] C[c="Theoretical" f=percent10.2]
    nominal[c="Nominal" f=percent10.2]
    W[c="CI width" f=percent10.1] CI[c=('LCI' || 'UCI') f=5.2]
    Pr[c=('Pi1' || 'Pi2' || 'Pi3') f=5.3];

FINISH MAIN;
rerunquit;
/* --------------------------------------------------------------- */
*/ Produce Table 14: Coverage probabilities for random categorical distributions with varying no. of classes*/
/* ----------------------------------------------- */
/* Note: There is a small bug in the program. It occasionally crashes when the number of classes is high (>7).
   However, if one reruns the program, it eventually converges upon a solution. */
Proc IML;
/* This module designated the rank order in which a probability will be assigned to a class.
   It's purpose it to prevent the first classes from gobbling up the majority of the probability. */
START ORDER;
    ORDER=j(sim,k,.)
    do j=1 to sim;
        ID=(1:k)
        do i=1 to k;
            X=ceil((k+1-i)*ranuni(seed));
            ORDER[j,i]=ID[X];
            if i<k then ID=ID[loc(ID^=ID[X])];
        end;
    end;
FINISH ORDER;

/* This module assigns a random prob. to each class in the order designated by the previous mod. */
START PROB;
    do i=1 to sim;
        pp=0;
        do j=1 to k-1;
            r=1;
            /* set min prob so prob in a single category does not become too large */
            do until (r>minp);
r=ranuni(seed);
end;
p[i,ORDER[i,j]]=(1-pp)*r;
pp=pp+p[i,ORDER[i,j]];
end;
p[i,ORDER[i,k]]=1-pp;
pp=pp+p[i,ORDER[i,k]];
end;
FINISH PROB;

/* This module generates random categorical variables based on the probabilities designated by the previous module. */
START SIM;
Y = j(n,sim,.) ; /* categorical variables */
R=j(n,sim,.) ; /* rank order of categorical variable */
do j=1 to sim;
    Cat = RANDMULTINOMIAL(n,1,p[j,]) ; /* Random values from multinomial */
do i=1 to n;
        Y[i,j]=loc(Cat[i,]);
    end;
    R[j,]=rank(Y[j,]);
end;
FINISH SIM;

/* This module calculates the frequencies and probabilities of the previously generate random variable. */
START FREQ;
/* Value of the median */
i=0;
do until(p[s,1;i][+]>=.5); /* Determines the population median */
i=i+1;
    med=i;
end;
pop_med[s]=med;

m=(n+1)/2; /* Order statistic of the sample median */
X=J(1,2,.) ;
if mod(n,2)=1 then do;
    X[1]=Y[loc(R[s]=m)] ;
    X[2]=X[1];
end;
else do;
    X[1]=Y[loc(R[s]=floor(m)),s] ; /* lower bound estimate for even N */
    X[2]=Y[loc(R[s]=ceil(m)),s] ; /* upper bound estimate for even N */
end;
/* --------------------------- */
Scale=unique(Y[s,]) ;
Counts=J(nrow(Scale),3,.) ;
Pi=j(3,1,0) ;
Counts[,1]=Scale;
do i=1 to nrow(Scale);
    Counts[i,2]=sum(Y[s]=Scale[i]);
    Counts[i,3]=Counts[i,2]/n;
    if Scale[i]<X[1] then Pi[1]=Pi[1]+Counts[i,3];
*/ This module locates the order statistics for the lower and upper bounds of the CI on the median*/
START FindQuantile;
maxLOS=quantile('BINOM',alpha,Pi[2]+Pi[3],n);
minUOS=quantile('BINOM',1-alpha,Pi[3],n)+1;
if maxLOS<1 then maxLOS=1;
if minUOS>n then minUOS=n;
min=1;
do j=minUOS to n ;
exit=0;
if maxLOS=1 then do;
    LOS=1;
    UOS=ceil(m);
end;
else if minUOS=n then do;
    LOS=floor(m);
    UOS=n;
end;
else if cdf('BINOM',j-1,Pi[3],n)-cdf('BINOM',maxLOS-1,Pi[2]+Pi[3],n)<min then 
do i=maxLOS to 1 by -1;
    inCI=((i<=m) & (j>=m)) | ((i>=m) & (j<=m));
    if (InCI=1 & exit=0) then do;
        cc=cdf('BINOM',j-1,Pi[3],n)
        -cdf('BINOM',j-1,Pi[2]+Pi[3],n);
        if cc>=1-alpha & cc<min then do;
            min=cc;
            LOS=i;
            UOS=j;
            exit=1;
        end;
    else if cc<min then i=0;
end;
end;
cc=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
diff=abs(cc-(1-alpha));
if Pi[1]>Pi[3] then do;
    cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    if diff>abs(cc2-(1-alpha)) then LOS=LOS+1;
end;
else if Pi[1]<Pi[3] then do;
    cc2=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    if diff>abs(cc2-(1-alpha)) then UOS=UOS-1;
end;
else do;
    cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    cc3=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
    cc4=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
    diff2=abs(cc2-(1-alpha));
    diff3=abs(cc3-(1-alpha));
    if diff>abs(cc4-(1-alpha)) then do;
        LOS=LOS+1;
        UOS=UOS-1;
end;
end;
else if diff2>diff3 & diff>diff2 then LOS=LOS+1;
else if diff2>diff3 & diff>diff3 then UOS=UOS-1;
else if diff2<diff3 & diff>diff3 then UOS=UOS-1;
else if diff2<diff3 & diff>diff2 then LOS=LOS+1;
end;
FINISH FindQuantile;

START COVERAGE;
if LOS<=UOS then do;
    LCI=Y[loc(R[,s]=LOS),s];
    UCI=Y[loc(R[,s]=UOS),s];
end;
else if UOS<LOS then do;
    LCI=Y[loc(R[,s]=UOS),s];
    UCI=Y[loc(R[,s]=LOS),s];
end;
if LCI <= med & UCI >= med then Cov[s]=1; /* Checks whether GM is in CI */
FINISH COVERAGE;

START MAIN;
seed= 0;
alpha=0.05;
n=20;
minp=.01; /* Sets minimum probability (0-1) generated in k-1 classes. This reduces
the likelihood of generating too many classes with a zero probability. */
k = 3; /* Number of classes */
sim=100000; /* Number of variables simulated */
p=j(sim,k,0); /* probability distribution of each variable */
cov=j(sim,1,0);
pop_med=j(sim,1,.);
call randseed(seed); /* seed used by RANDMULTINOMIAL */

RUN ORDER;
RUN PROB;
RUN SIM;
do s=1 to sim;
    RUN FREQ;
    RUN FindQuantile;
    RUN COVERAGE;
end;
Cov=Cov[:]; /* seed used by RANDMULTINOMIAL */

RUN ORDER;
RUN PROB;
RUN SIM;
do s=1 to sim;
    RUN FREQ;
    RUN FindQuantile;
    RUN COVERAGE;
end;
Cov=Cov[:]; /* seed used by RANDMULTINOMIAL */

print k[v="No. Classes"] n[v="N"] Cov[v="Coverage" f=percent8.2];
X = 2 * rannor(98765) + 10;
X = X ** 1.5;
output;
end;
drop i;
run;

/* This method produces results identical to the continuous method for constructing CI on the median.
I am using it to replace my code in order to reduce on the redundancy of pasting the same code. */
proc univariate data=discrete CIQUANTDF(alpha=.05);
var x;
histogram;
run;

/* This code will need to be run for each of the cells of Table 15. No changes are needed. */
proc iml;
use discrete;
read all var _num_ into Mult;
close discrete;

START FREQ;
m = (n + 1) / 2;          /* Order statistic of the median */
ub = ceil(m);
lb = floor(m);
/* Value of the median */
X = J(1, 2, .);
if mod(n, 2) = 1 then do;
    X[1] = Mult[ loc(R = m) ];
    X[2] = X[1];
end;
else do;
    X[1] = Mult[ loc(R = floor(m)) ];
    X[2] = Mult[ loc(R = ceil(m)) ];
end;
/* ----------------------------- */
Scale = unique(Mult); 
Counts = J(nrow(Scale), 3, .);
Pi = J(3, 1, 0);
Counts[, 1] = Scale;
do i = 1 to nrow(Scale);
    Counts[i, 2] = sum( Mult = Scale[i] );
    Counts[i, 3] = Counts[i, 2] / n;
    if Scale[i] < X[1] then Pi[1] = Pi[1] + Counts[i, 3];
end;
FINISH FREQ;

START FindQuantile;
maxLOS = quantile('BINOM', alpha, Pi[2] + Pi[3], n);
minUOS = quantile('BINOM', 1 - alpha, Pi[3], n) + 1;
if maxLOS < 1 then maxLOS = 1;
if minUOS > n then minUOS = n;
min = 1;
do j=minUOS to n;
  exit=0;
  if maxLOS=1 then do;
    LOS=1;
    UOS=ceil(m);
  end;
  else if minUOS=n then do;
    LOS=floor(m);
    UOS=n;
  end;
  else if cdf('BINOM',j-1,Pi[3],n)-cdf('BINOM',maxLOS-1,Pi[2]+Pi[3],n)<min then
    do i=maxLOS to 1 by -1;
      inCI=((i<=m) & (j>=m)) | ((i>=m) & (j<=m));
      if (InCI=1 & exit=0) then do;
        cc=cdf('BINOM',j-1,Pi[3],n)
          -cdf('BINOM',j-1,Pi[2]+Pi[3],n);
        if cc>=1-alpha & cc<min then do;
          min=cc;
          LOS=i;
          UOS=j;
          exit=1;
        end;
      else if cc>min then i=0;
    end;
  end;
cc=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
diff=abs(cc-(1-alpha));
if Pi[1]>Pi[3] then do;
  cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
  if diff>abs(cc2-(1-alpha)) then LOS=LOS+1;
end;
else if Pi[1]<Pi[3] then do;
  cc2=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
  if diff>abs(cc2-(1-alpha)) then UOS=UOS-1;
end;
else do;
  cc2=cdf('BINOM',UOS-1,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
  cc3=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS-1,Pi[2]+Pi[3],n);
  cc4=cdf('BINOM',UOS-2,Pi[3],n)-cdf('BINOM',LOS,Pi[2]+Pi[3],n);
  diff2=abs(cc2-(1-alpha));
  diff3=abs(cc3-(1-alpha));
  if diff>abs(cc4-(1-alpha)) then do;
    LOS=LOS+1;
    UOS=UOS-1;
  end;
  else if diff2>diff3 & diff>diff2 then LOS=LOS+1;
  else if diff2>diff3 & diff>diff3 then UOS=UOS-1;
  else if diff2<diff3 & diff>diff3 then UOS=UOS-1;
  else if diff2<diff3 & diff>diff2 then LOS=LOS+1;
end;
FINISH FindQuantile;

START ASYMMETRIC;
RUN FindQuantile;
C=round(cdf('binom',UOS-1,Pi[3],n)-cdf('binom',LOS-1,Pi[2]+Pi[3],n),.0001);
L=Mult[loc(R=LOS)]; /* Lower bound observation */
U=Mult[loc(R=UOS)]; /* Upper bound observation */
sym="N'';
FINISH ASYMMETRIC;

START SYMMETRIC;

LOS=floor(m); UOS=ceil(m);
C=round(cdf('binom',UOS-1,Pl[3],n)-cdf('binom',LOS-1,Pl[2]+Pl[3],n),0.001);
do until ((C>1-alpha) | (LOS=1 & UOS=n));
    LOS=LOS-1;
    UOS=UOS+1;
    C=round(cdf('binom',UOS-1,Pl[3],n)-cdf('binom',LOS-1,Pl[2]+Pl[3],n),0.001);
end;
L=Mult[loc(R=LOS)]; /* Lower bound observation */
U=Mult[loc(R=UOS)]; /* Upper bound observation */
sym="Y'';
FINISH SYMMETRIC;

START PRINT;
if LOS>UOS then do;
a=LOS;
    LOS=UOS;
    UOS=a;
end;
X=X[:];
if sym="N" then do;
Print alpha[f=4.2 l='\Alpha'\]
DC[l="Desired coverage" f=percent25.2],;
Print "Asymmetric confidence interval";
end;
else Print "Symmetric confidence interval";
Print n[l="N"] X[l="median" f=5.1] m[l="order"] LOS UOS L[l="LCI" f=5.1]
U[l="UCI" f=5.1] C[l="Coverage" f=percent12.2];
FINISH PRINT;

START MAIN;
alpha=0.05;
DC=1-alpha; /* Desired coverage */
n=nrow(Mult);
R=rank(Mult);

RUN FREQ;
RUN ASYMMETRIC;
RUN PRINT;
RUN SYMMETRIC;
RUN PRINT;
FINISH MAIN;
runquit;
/* --------------------------------------------------------------- */
APPENDIX I

SAS Code for Testing Classical Test Theory
/* WARNING: This program takes approximately 2 hours to run. There are two commented out sections in the code. The first is a line in the CTT module, which refers to module ERROR. This module can be used to generate nonnormal errors. The second commented out section, also located in the CTT module, contains code that can be used to assign unique variance estimates to each case. */

START ERROR;
   U1=RANUNI(j(N,V,seed2));
   * Generate a Normal: mu=100, var=2.0, skew=2.0, kurt=10 ;
   L1=-0.5154104205;
   L2=0.1783773597;
   L4=-0.0387687344;
   L3=-0.1168183637;
   NN = L1 + (U1##L3 - (1 - U1)##L4) / L2 ;
FINISH ERROR;

START CTT;
   SEM=sqrt(var*(1-rho)); /* Theoretical standard error of measurement */
   SEE=sqrt(var*rho*(1-rho)); /* Theoretical standard error of estimation */
   SEP=sqrt(var*(1-rho**2)); /* Theoretical standard error of prediction */
   ST=sqrt(var*rho);
   T=ST*rannor(j(N,1,seed1))+mu;
   E=SEM*rannor(j(N,V,seed2)); /* Replace with the line below for nonnormal error */
   X=T+E;
   /* The following line can be used to assign unique variance estimates to each case */
   /* T=ST*rannor(j(N,1,seed1))+mu;
   SEM2=abs(S*rannor(j(N,1,seed2))+sqrt(SEM**2-S**2));
   ratio=max(SEM2)/min(SEM2);
   E=SEM2#rannor(j(N,V,seed3));
   X=T+E; */
FINISH CTT;

START MEANVAR;
   var_X=(((X##2)[+,-N*X[:,]])/(N-1));
   var_X2=var_X;
   var_T=((T##2)[+,-N*T[:,]])/(N-1);
   var_E=((E##2)[+,-N*E[:,]])/(N-1);
   E_X=X[:,];
   E_T=T[:,];
   E_E=E[:,];
FINISH MEANVAR;

START ASSUMPTIONS;
   Cov_TE=j(V,1,);
   Corr_TE=j(V,1,);
FINISH ASSUMPTIONS;
$\text{Corr}_X T = \text{corr}(X, T)$

$$\text{do } i = 1 \text{ to } V;$$

$$\text{TE} = T \| E[i];$$

$$\text{Corr} = \text{corr}(\text{TE});$$

$$\text{Corr}_{TE}[i] = \text{corr}[2, 1];$$

$$\text{Cov}_{TE}[i] = \text{corr}_{TE}[i] \times \sqrt{\text{var}_T \times \text{var}_E[i]};$$

$$\text{XT} = X[j] \| T;$$

$$\text{Corr} = \text{corr}(\text{XT});$$

$$\text{Corr}_{XT}[i] = \text{corr}[2, 1];$$

$$\text{end;}$$

$$\text{Cov}_{TE} = \text{corr}(\text{TE});$$

$$\text{Corr}_{TE} = \text{corr}(\text{TE});$$

$$\text{Cov}_{XT} = \text{corr}(\text{XT});$$

$$\text{Cov}_{EE} = \text{j}(V, V, .);$$

$$\text{Cov}_{XX} = \text{j}(V, V, .);$$

$$\text{Corr}_{EE} = \text{j}(V, V, .);$$

$$\text{Corr}_{XX} = \text{j}(V, V, .);$$

$$\text{do } i = 1 \text{ to } V;$$

$$\text{do } j = 1 \text{ to } V;$$

$$\text{if } j < i \text{ then do};$$

$$\text{EE} = E[i] \| E[j];$$

$$\text{Corr} = \text{corr}(\text{EE});$$

$$\text{Corr}_{EE}[i, j] = \text{corr}[2, 1];$$

$$\text{Cov}_{EE}[i, j] = \text{corr}_{EE}[i, j] \times \sqrt{\text{var}_E[i] \times \text{var}_E[j]};$$

$$\text{XX} = X[i] \| X[j];$$

$$\text{Corr} = \text{corr}(\text{XX});$$

$$\text{Corr}_{XX}[i, j] = \text{corr}[2, 1];$$

$$\text{Cov}_{XX}[i, j] = \text{corr}_{XX}[i, j] \times \sqrt{\text{var}_X[i] \times \text{var}_X[j]};$$

$$\text{end};$$

$$\text{end};$$

$$\text{var}_E = \text{var}_E[i];$$

$$\text{var}_X = \text{var}_X[i];$$

$$\text{Cov}_{EE} = \text{corr}(\text{EE});$$

$$\text{Cov}_{XX} = \text{corr}(\text{XX});$$

$$\text{sqrtCorr}_{XX} = \sqrt{\text{corr}(\text{XX})};$$

$$\text{Corr}_{XT} \times \text{var}_X = (\text{Corr}_{XT})^2 \times \text{var}_X;$$

$$\text{Corr}_{XX} \times \text{var}_X = (1 - \text{Corr}_{XT})^2 \times \text{var}_X;$$

$$\text{RhoVar} = \text{Rho} \times \text{var};$$

$$\text{RhoVar} = (1 - \text{Rho}) \times \text{var};$$

$$\text{sqrtRho} = \sqrt{\text{Rho}};$$

$$\text{SX} = \sqrt{\text{var}_X};$$

print "Property 1: E(X) = E(T)" E_X, "E(T)" E_T, "f = 10.4";

"Property 2: Cov(T, E) = 0" Cov_TE, "f = 10.4";

"Property 3: Cov(X, X') = Cov(X, X')" Cov_XX, "f = 12.4";

"Property 4: Corr(X, X') = Corr(X, X')" Corr_XX, "f = 15.4";

"Property 5: Var(X) = Var(T) + Var(E)" var_X, "f = 17.4";
"Property 10: \( \text{Var}(T) = [\text{Corr}(X,T)^2] \text{Var}(X) = \rho \text{Var}(X) \)"
\( \text{Var}(T) = [\text{Corr}(X,T)^2] \text{Var}(X) = 13.4 \),
"Property 11: \( \text{Var}(E) = [1-\text{Corr}(X,T)^2] \text{Var}(X) = (1-\rho) \text{Var}(X) \)"
\( \text{Var}(E) = 5.4 \),
\( \text{Corr}(X,T)^2 \text{Var}(X) = 15.4 \).

FINISH ASSUMPTIONS;

START COVERAGE;

\( z = \text{probit}(1-\alpha/2) \);
\( \text{print / "Theoretical values";} \);
\( \text{print SEM}[] = "\text{Standard error of measurement} f=30.4"; \)
\( \text{SEE}[][ = "\text{Standard error of estimation} f=30.4"; \)
\( \text{SEP}[][ = "\text{Standard error of prediction} f=30.4"; \)
\( \text{SEM}[] = j(v,1,\cdot); /\* \text{Observed standard error of measurement} */ \)
\( \text{SEE}[] = j(v,1,\cdot); /\* \text{Observed standard error of estimation} */ \)
\( \text{SEP}[] = j(v,1,\cdot); /\* \text{Observed standard error of prediction} */ \)
\( \text{Coverage1} = j(N,V,0); /\* \text{Coverage probability for CI on individual true scores} */ \)
\( \text{Coverage2} = j(N,V,0); /\* \text{Coverage probability for CI on individual predicted true scores} */ \)
\( \text{Coverage3} = j(N,V,0); /\* \text{Coverage probability for CI on individual parallel test scores} */ \)
\( T_{\text{hat}} = \rho X + (1-\rho) E_X; /\* \text{Predicted true score/predicted parallel test score} */ \)
\( \text{do} \ i = 1 \text{ to} \ V; \)
\( \text{SEM}[] = \sqrt{\text{var}(X)2[i] * (1-\rho)}; \)
\( \text{SEE}[] = \sqrt{\text{var}(X)2[i] * \rho * (1-\rho)}; \)
\( \text{SEP}[] = \sqrt{\text{var}(X)2[i] * (1-\rho^2)}; \)
\( \text{Coverage1}[] = \text{loc}(X[i] \geq T_{\text{hat}} - z*\text{SEM}[i] \& X[i] \leq T_{\text{hat}} + z*\text{SEM}[i],i] = 1; \)
\( \text{Coverage2}[] = \text{loc}(T \geq T_{\text{hat}}[i] - z*\text{SEE}[i] \& T \leq T_{\text{hat}}[i] + z*\text{SEE}[i],i] = 1; \)
\( \text{cnt}[] = j(N,V-1,0); \)
\( \text{if} \ i = 1 \text{ then} XX = X[2:V]; \)
\( \text{else if} \ i = V \text{ then} XX = X[1:V-1]; \)
\( \text{else} \ XX = X[1:i-1] || X[i+1:V]; \)
\( \text{do} \ j = 1 \text{ to} \ V; \)
\( \text{cnt}[] = \text{loc}(XX[j] \geq T_{\text{hat}}[i] - z*\text{SEP}[i] \& XX[j] \leq T_{\text{hat}}[i] + z*\text{SEP}[i],j] = 1; \)
\( \text{end}; \)
\( \text{Coverage3}[] = \text{cnt}[+]; \)
\( \text{end}; \)
\( \text{/* Probability for predicting the parallel score based on the observed score */} \)
\( \text{Coverage3}[] = (\text{Coverage3}[] / (V-1))[i]; \)
\( \text{SEM}[]; \)
\( \text{SEE}[]; \)
\( \text{SEP}[]; \)
\( \text{print "Observed values";} \)
\( \text{print SEM}[] = "\text{Standard error of measurement} f=30.4"; \)
\( \text{SEE}[] = "\text{Standard error of estimation} f=30.4"; \)
\( \text{SEP}[] = "\text{Standard error of prediction} f=30.4"; \)
\( \text{plusminus} = \text{byte}(177); \)
\( \text{Coverage1}[] = \text{Coverage1}[]; \)
\( \text{Coverage2}[] = \text{Coverage2}[]; \)
\( \text{Coverage3}[] = \text{Coverage3}[]; \)
\( \text{reset nocenter;} \)
\( \text{print "CI on true score (T) to predict an observed score (X): \( T + z*\text{SEM} \) Coverage1 f=10.4",} \)
\( \text{"CI on predicted true score (T_hat) to predict T: \( T_{\text{hat}} + z*\text{SEE} \) Coverage2 f=12.4",} \)
"CI on $T_{\hat{}}$ to predict parallel score ($X'$):
$T_{\hat{}} \pm z*\text{SEP}$ Coverage3="Coverage" $f=17.4$;"

FINISH COVERAGE;

START MAIN;
seed1=123;
seed2=98765;
seed3=9101112;
alpha=.05;
N=100000;
V=300;
mu=100;
Var=225;
rho=.35;
RUN ERROR;
RUN CTT;
RUN MEANVAR;
RUN ASSUMPTIONS;
RUN COVERAGE;
FINISH MAIN;
RUN; QUIT;

/*  --------------------------------------------------------------------------------------------------------
——
—
—
——
------------------------------------------------------------------------------------------------------------
*/

/*  ILLUSTRATE IMPACT OF FINITE POPULATION CORRECTION FACTOR ON VARIANCE
------------------------------------------------------------------------------------------------------------
*/

Proc IML;
START CTT;
ST=sqrt(var*rho);  /* Standard deviation of the true score */
SEM=sqrt(var*(1-rho));  /* Standard error of measurement */
T=ST*rannor(j(popN,1,seed))+mu;  /* True score */
E=SEM*rannor(j(popN,1,seed));  /* Measurement error, where error is normal */
X=T+E;  /* Observed score */
$T_{\hat{}}=\rho*X+(1-\rho)*X[:]$;  /* Predicted true score/predicted parallel test score */
FINISH CTT;

START STATS(pop,fpc,ds,n) global (mean,var,se);
mean=ds[:];  /* mean */
if pop=1 then denom=n;  /* Run when sample = population */
else denom=n-1;  /* Run when sample ^= population */
var=fpc*((ds##2)[+]n*mean##2)/denom;  /* Correct variance estimate */
se=sqrt(var/denom);
FINISH STATS;

START PRINT;
fpc=1-sampN/popN;  /* finite population correction factor */
nominal=1-alpha;
reset NOAUTONAME NONAME ;
print "Simulation parameters";
print sim[c="# Simulations" $f=10.0]$ sampN[c="Sample size" $f=15.0$]
   popN[c="Population size" $f=20.0$] fpc[c="FPC" $f=percent10.3$
   rho[c="Rho" $f=8.3$] alpha[c="Alpha" $f=8.2$] nominal[c="#Nominal" $f=8.2$],;
print "Population parameters";
print mean[c="Mean" $f=15.4$] var[c="Variance" $f=15.4$] se[c="SE" $f=15.4$],;
FINISH PRINT;
START POPULATION; /* Generate a population of predicted true/parallel scores */
  RUN CTT;
  Pop=j(popN,2,seeds);
  Pop[,1]=j(1:popN);
  Pop[,2]=T_hat; /* Predicted true/parallel score */
  fpc=1; /* fpc=1 since the sample and population are the same */
  RUN STATS(1,fpc,Pop[,2],popN);
  mu_Pop=mean;
  RUN PRINT;
FINISH POPULATION;

START SRS; /* Generate simple random sample without replacement */
  fpc=1-sampN/popN; /* finite population correction factor */
  select=uniform(j(popN,1,seed))<1-fpc; /* randomly select cases from population */
  sum=select[+]; /* number of randomly selected cases */
  if sum>sampN then do; /* oversampled the population (N too large) */
    Samp=Pop[loc(select=1),2]; /* identify value of randomly selected cases */
    Samp=Samp[1:sampN]; /* keep the first nn randomly selected values */
  end;
  else if sum<sampN then do;
    /* undersampled the population (N too small) */
    do while (sum<sampN); /* add more randomly selected case to reach sampN*/
      if sum>0 then Y=j(sum,2,seeds);
      Y=Pop[loc(select=0),];
      id=Y[ceil(nrow(Y)*uniform(seed)),1];
      select[id]=1;
      sum=sum[+]; /* number of randomly selected cases*/
    end;
    Samp=Pop[loc(select=1),2];
  end;
else Samp=Pop[loc(select=1),2];
FINISH SRS;

START CHECKCI; /* Generate sampling distribution */
  z=probit(1-alpha/2);
  LB=mean-z*se;
  UB=mean+z*se;
  if LB<=mu_Pop & mu_Pop<=UB then inCI=1;
else inCI=0;
FINISH CHECKCI;

START SAMPDIST; /* Generate sampling distribution */
  dist=j(sim,4,seeds);
  stat=j(1,4,seeds);
  do i=1 to sim;
    RUN SRS;
    RUN STATS(0,fpc,Samp,sampN);
    RUN CHECKCI;
    dist[i,1]=mean;
    dist[i,2]=var;
    dist[i,3]=se;
    dist[i,4]=inCI;
  end;
  do i=1 to 4;
    stat[i]=dist[i,];
  end;
print "Sampling distribution";
print stat[l=" f=13.4 e=("Mean" | "Variance" | "SE" | "Coverage")];
FINISH SAMPDIST;

START MAIN;
seed=123;
alpha=.05;
sim=10000;
popN=10000; /* population size */
sampN=4000; /* sample size */
mu=100; /* population mean */
var=225; /* population var */
rho=0.7;
RUN POPULATION;
RUN SAMPDIST;
Pop=X || T || E;
CREATE Population FROM Pop [c=('X' || 'T' || 'E')];
APPEND FROM Pop;
CREATE SRS FROM dist [c=('Mean' || 'Variance' || 'SE' || 'inCI')];
APPEND FROM dist;
FINISH MAIN;
RUN; QUIT;
/* QUIT; */
/* */
/* */
/* SUMMATIVE CONFIDENCE */
/* */
/* */

Proc IML;
START SIMCORR;
V=ncol(mu);
intRho=j(V,V,.);
do i=1 to V;
do j=1 to V;
    intRho[i,j]=rho[i,j]/sqrt(rel[i]*rel[j]);
    if i=j then intRho[i,j]=1;
end;
end;
Cholesky=root(intRho);
FINISH SIMCORR;
START CTT;
ST=sqrt(var#rel); /* Standard deviation of the true score */
SEM=sqrt(var#(1-rel)); /* Standard error of measurement */
T=rannor(j(popN,V,seed1)); /* Generate true scores */
T=T*Cholesky; /* Correlate true scores */
T=ST#T+mu; /* Add variance and means to correlated true scores */
E=SEM#rannor(j(popN,V,seed2)); /* Error is assumed to be normal */
X=T+E; /* Observed score */
T_hat=rel#X+(1-rel)#X[;]; /* Predicted true score/predicted parallel test score */
T_hat=wgt#T_hat;
var_E=(((E##2)+.popN*E[..])##2)/(popN-1));
var_X=(((X##2)+.popN*X[..])##2)/(popN-1));
var_T=(((T##2)+.popN*T[..])##2)/(popN-1));
var_T_hat=(((T_hat##2)+.popN*T_hat[..])##2)/(popN-1));
T=T[..];
mu_X=X[..];
T_hat=T_hat[..]; /* Create composite true score by averaging the four values */
corr=corr(X); /* Observed inner-item population correlations */
diff=rho-corr;
diff=max(diff);
X=X[:,];
var_X2=((X#2)+.popN*X[:,]#2)/(popN-1));
SE_X2=sqrt(var_X2/popN);
FINISH CTT;

START STATS(pop,fpc,ds,n) global (mean,var,se);
mean=ds[:]; /* population/sample mean */
if pop=1 then denom=n; /* Run when sample = population */
else denom=n-1; /* Run when sample ^= population */
var=fpc*((ds##2)+n*mean##2)/denom; /* Correct variance estimate */
se=sqrt(var/denom); /* Standard error of the mean */
FINISH STATS;

START PRINT;
fpc=1-sampN/popN; /* finite population correction factor */
nominal=1-alpha;
print "Simulation parameters";
print sim[="# Simulations" f=10.0] sampN[="Sample size" f=15.0]
popN[="Population size" f=20.0] fpc[="FPC" f=percent10.3]
alpha[="Alpha" f=8.2] nominal[="Nominal" f=8.2];
print "Simulated means",mu[="X1","X4"] f=10.4],"Simulated variances"
var_X[="X1","X4"] f=10.4],"Simulated reliability", rel[="X1","X4"] f=10.4];
print "Observed means",mu_X[="X1","X4"] f=10.4],"Observed variances"
var_X[="X1","X4"] f=10.4];
print rho[="Expected correlations" f=10.4] corr[="Observed correlations" f=10.4];
print "Maximum correlation difference:" diff[=""];
print "Unadjusted variance and standard error";
print var_X2[="Var(X)" f=10.4] SE_X2[="SE(X)" f=10.4];
print /"Expected composite population values";
print grandmean[="E(X)" f=15.4] SC[="Var(T_hat)" | "SE(T_hat)"]
print "Observed population parameters";
print mean[="Mean" f=15.4] var[="Variance" f=15.4] se[="SE" f=15.4];
FINISH PRINT;

START POPULATION; /* Generate a population of predicted true/parallel scores */
RUN CTT;
Pop=j(popN,2,);
Pop[,1]=(1:popN);
Pop[,2]=T_hat; /* Predicted true/parallel score */
fpc=1; /* fpc=1 since the sample and population are the same */
RUN STATS(1,fpc,Pop[,2],popN);
mu_Pop=mean;
RUN PRINT;
FINISH POPULATION;

START SRS; /* Generate simple random sample without replacement */
fpc=1-sampN/popN; /* finite population correction factor */
select=uniform(j(popN,1,seed3)<1-fpc); /* randomly select cases from population */
sum=select[+]; /* number of randomly selected cases */
if sum>sampN then do; /* oversampled the population (N too large) */
Samp=Pop[loc(select=1,2)]; /* identify value of randomly selected cases */
else if sum<sampN then
end;

333
do; /* undersampled the population (N too small) */
do while (sum<sampN); /* add more randomly selected case to reach sampN*/
if sum>0 then Y[j(sum,2,);
Y=Pop[loc(select=0),];
id=Y[ceil(nrow(Y)*uniform(seed3)),1];
select[id]=1;
sum=sum+[; /* number of randomly selected cases */
end;
Samp=Pop[loc(select=1),2];
end;
else Samp=Pop[loc(select=1),2];
FINISH SRS;

START CHECKCI;
z=probit(1-alpha/2);
LB=mean-z*se; /* Lower bound CI */
UB=mean+z*se; /* Upper bound CI */
if LB<=mu_Pop & mu_Pop<=UB then inCI=1; /* Probability coverage */
else inCI=0;
FINISH CHECKCI;

START SAMPDIST; /* Generate sampling distribution */
dist=[(sim,4,); stat=[(1,4,); 
do i=1 to sim;
RUN SRS;
RUN STATS(0,fpc,Samp,sampN);
RUN CHECKCI;
dist[i,1]=mean;
dist[i,2]=var;
dist[i,3]=se;
dist[i,4]=inCI;
end;
do i=1 to 4;
stat[i]=dist[i,i];
end;
print "Sampling distribution";
print stat[="f=13.4 c=("Mean" | "Variance" | "SE" | "Coverage")];
FINISH SAMPDIST;

START SUMCONF;
grandmean=mu[:];
varX=var;
SC=[(1,2,0,); 
do i=1 to V;
do j=i to V;
if i<=j then do;
   if i=j then do;
      /* Error variance of estimation/prediction */
      VEE=(wgt[i]*wgt[j]*rel[i]*rel[j]*var[i]);
   end;
else do; /* Error covariance of estimation/prediction */
      CEE=2*(wgt[i]*wgt[j]*rho[i,j]*rel[i]*rel[j]*sqrt(var[i]*var[j]));
   end;
end;
end;
SC=SC/V**2; SC[2]=sqrt(SC[2]);
FINISH SUMCONF;

START MAIN;
  seed1=3354;                     seed2=12189;                     seed3=923447;
  alpha=.05; /* Type I error rate */
  sim=100000; /* # of simulations */
  popN=10000; /* population size */
  sampN=3000; /* sample size */
  mu={50 75 100 125}; /* population means */
  var={15 20 30 50}; /* population variances */
  rel={.3 .5 .7 .9}; /* parallel reliability */
  wgt={1 1 1 1}; /* weighting scheme */
  rho={1.00 0.15 0.20 0.30, 0.15 1.00 0.45 0.25, 0.20 0.45 1.00 0.60, 0.30 0.25 0.60 1.00}; /* inner-item correlations */
RUN SIMCORR;
RUN SUMCONF;
RUN POPULATION;
RUN SAMPDIST;
FINISH MAIN;
RUN; QUIT;
/*--------------------------------------------------------------------------------*/
APPENDIX J

Extension of the Law of Total Covariance
Related to the law of total variance, which was used to decompose the variance of a random variable into the model (true score) variance and error variance, the law of total covariance can be used to decompose covariance into two components. This law states that if $X$, $Y$, and $Z$ are random variables, then given a specific value of $Z$, the covariance of $X$ and $Y$ can be decomposed into the sum of the covariance of the conditional expectation (model covariance) and the average conditional error covariance. That is, \( \text{Cov}(X,Y) = \text{Cov}[\text{E}(X,Y|Z)] + \text{E}[\text{Cov}(X,Y|Z)] \). However, since one is often interested in a pair of conditional covariates, a more complex law is needed to account for two covariates. It is simple to show that \( \text{Cov}(X,Y) = \text{Cov}[\text{E}(X|U),\text{E}(Y|V)] + \text{E}[\text{Cov}(X|U,Y|V)]. \)

\[
\text{Cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) = \text{E}\left[\text{E}(XY|U,V) - \text{E}(X|U)\text{E}(Y|V)\right]
\]

since

\[
\begin{align*}
\text{Cov}(X|U,Y|V) &= \text{E}(XY|U,V) - \text{E}(X|U)\text{E}(Y|V) \\
&= \text{E}\left[\text{Cov}(X|U,Y|V) + \text{E}(X|U)\text{E}(Y|V) - \text{E}(X|U)\text{E}(Y|V)\right] \\
&= \text{E}\left[\text{Cov}(X|U,Y|V) + \text{E}(X|U)\text{E}(Y|V) - \text{E}(X|U)\text{E}(Y|V)\right]
\end{align*}
\]

as was to be proven. Therefore, knowing $U$ and $V$ reduces the covariance of $X$ and $Y$ if $U$ is related to $X$ or $V$ is related to $Y$, since $\text{Cov}(X,Y) \geq \text{E}[\text{Cov}(X|U,Y|V)]$. Furthermore, if the relationships between $X$ and $U$ and $Y$ and $V$ are linear (Property 8) then this modified law of total covariance can be expressed as

\[
\begin{align*}
\text{Cov}[\text{E}(X|U),\text{E}(Y|V)] &= \text{Cov}\left[\text{E}(X) + \rho_{XY}\frac{\sigma_X}{\sigma_U}\left[U - \text{E}(U)\right],\text{E}(Y) + \rho_{YV}\frac{\sigma_Y}{\sigma_V}\left[V - \text{E}(V)\right]\right] \\
&= \text{E}\left[\text{E}(X) + \rho_{XY}\frac{\sigma_X}{\sigma_U}\left[U - \text{E}(U)\right],\text{E}(Y) + \rho_{YV}\frac{\sigma_Y}{\sigma_V}\left[V - \text{E}(V)\right]\right] \\
&= \text{E}\left[\rho_{XY}\frac{\sigma_X}{\sigma_U}\text{E}(X)[V - \text{E}(V)] + \rho_{XY}\frac{\sigma_X}{\sigma_U}\text{E}(Y)[U - \text{E}(U)] \right] - \text{E}(X)\text{E}(Y) \\
&= \rho_{XY}\rho_{YV}\frac{\sigma_X\sigma_Y}{\sigma_U\sigma_V}\text{Cov}(U,V) = \rho_{XY}\frac{\sigma_X\sigma_Y}{\sigma_U\sigma_V}\rho_{UV}\frac{\sigma_U}{\sigma_U}\frac{\sigma_V}{\sigma_V}
\end{align*}
\]

and

\[
\begin{align*}
\text{E}[\text{Cov}(X|U,Y|V)] &= \text{E}\left[\text{E}(XY|U,V) - \text{E}(X|U)\text{E}(Y|V)\right] \\
&= \text{E}\left[\text{E}(XY|U,V) - \text{E}(X|U)\text{E}(Y|V)\right] \\
&= \text{E}(XY) - \text{E}\left[\text{E}(X) + \rho_{XY}\frac{\sigma_X}{\sigma_U}\left[U - \text{E}(U)\right],\text{E}(Y) + \rho_{YV}\frac{\sigma_Y}{\sigma_V}\left[V - \text{E}(V)\right]\right]
\end{align*}
\]
Here is the text of the page with the document's content from the image:

\[ \begin{align*}
&= \mathbb{E}(XY) - \mathbb{E}\left[ \frac{\sigma_{XY}}{\sigma_U} \mathbb{E}(Y) + \frac{\sigma_{XY}}{\sigma_U} \mathbb{E}(U) \right] V - \mathbb{E}(V) \\
&= \mathbb{E}(XY) - \mathbb{E}(X) \mathbb{E}(Y) - \mathbb{E}(X) \mathbb{E}(U) - \mathbb{E}(Y) \mathbb{E}(V) + \mathbb{E}(X) \mathbb{E}(U) + \mathbb{E}(Y) \mathbb{E}(V) \\
&= \text{Cov}(X,Y) - \rho_{XY} \text{Cov}(U,V) \\
&= \rho_{XY} \sigma_X \sigma_Y - \rho_{XY} \sigma_X \sigma_Y = \rho_{XY} \sigma_X \sigma_Y - \rho_{XY} \sigma_X \sigma_Y = \sigma_{XY} \rho_{XY} = \rho_{XY} \sigma_X \sigma_Y \\
&= \sigma_{XY} \rho_{XY} = \mathbb{E}(X,Y), \text{ as expected.}
\end{align*} \]

Note that \( \rho_{UT} \rho_{UX} \sigma_{UT} \sigma_{UX} \rho_{TY} \sigma_{TY} = \sigma_{XY} \rho_{XY} = \mathbb{E}(X,Y) \), which reduces to the variance estimator for parallel tests, which one is more likely to construct in a CI for the purpose of capturing the mean true score of a composite or the mean parallel observed score of a composite, the latter of which would be important if one wished to determine the likelihood a result would replicate if studied under parallel conditions. The reliability adjusted covariance estimator is given by \( \text{Cov}(\mathbb{E}(X|U, Y|V)) = \mathbb{E}(X,Y) \text{Cov}(\mathbb{E}(X|U), \mathbb{E}(Y|V)) = \rho_{XY} \rho_{YY} \sigma_X \sigma_Y \), which, of course, reduces to the variance estimator \( \rho_{XY}^2 \sigma_X^2 \) when \( X = Y \).
APPENDIX K

SAS Code for Single-Bootstrap Reliability
/* Bootstrap Reliability */
/* WARNING: This program takes approximately 20:35 hours to run. */
proc iml;
START SIMCORR;
    r= J(V, V, 1);
    corr= J(V, V, 1);
    do j= 1 to V;
        do i= 1 to V;
            if i>j then r[i,j] = ranuni(seed1);
            r[j,i]= r[i,j];
            if i^=j then corr[i,j]= rho+(r[i,j]-0.5)* (1-rho);
        end;
    end;
    z= corr;
    z[loc(diag(z))]=.;
    meanCorr1= z[:];
FINISH SIMCORR;
START CHECK;
    call eigen(Eig, Vec, Corr);
    if min(Eig)<0 then do;
        RUN VALIDATE;
        Corr= NewCorr;
    end;
    z= corr;
    z[loc(diag(z))]=.;
    meanCorr2= z[:];
    free z NewCorr;
FINISH CHECK;
START VALIDATE;
    NewCorr= Corr;
    negEig= Eig[loc(Eig<0)];
    Eig[loc(Eig<0)]= 1E-10;  /* Do not set to zero */
    /* This code slightly reduces the negative bias of the correction of negative eigen values. */
    k= nrow(negEig);
    do i= 1 to k;
        Eig[i]= Eig[k+1:V][+]*Rho/((1-Rho)**(k+1-i));
    end;
    Eig= diag(Eig);
    NewCorr= Vec* Eig* Vec`;  /* Adjusted (new) correlation matrix */
    call eigen(Eig, Vec, NewCorr);
    RUN Standardize;
FINISH VALIDATE;
START STANDARDIZE;
    Sd= NewCorr;
    do col= 1 to ncol(Sd);
        do row= 1 to ncol(Sd);
        end;
    end;
    NewCorr= Sd;
FINISH STANDARDIZE;
START SIM;
    Cholesky=\text{root}(\text{Corr});
    X=\text{normal}(j(N,V,seed1));
    X=X*\text{Cholesky};

    Mu=\text{normal}(j(1,V,seed3))+100;
    SD=\text{abs}(1.5*\text{normal}(j(1,V,seed3))+5);
    X=SD\#X+Mu;

    corrX=corr(X);
    corrX[loc(diag(corrX))]=.;
    meanCorrX=corrX[:];
FINISH SIM;

START SPLITHALF; /* Generate simple random sample without replacement */
    select=\text{ranbin}(j(1,V,seed2),1,0.5); /* randomly select items from full-length test */
    if select[+]\!/V<0.5 then select=1-select;
    sum=select[+]; /* number of randomly selected items */

    if sum\!/V>0.5 then do;
        do while (sum\!/V>0.5);
            Y=loc(select=1);
            id=Y[1,ceil(ncol(Y)*\text{uniform}(0))];
            select[id]=0;
            sum=select[+]; /* number of randomly selected cases */
        end;
    end;
    Half1=loc(select=1);
    Half2=loc(select=0);
FINISH SPLITHALF;

START ASSUMPTIONS;
    Test=j(N,2,2);
    Test[,1]=X[,Half1][,]; /* Subtest 1 */
    Test[,2]=X[,Half2][,]; /* Subtest 2 */

    EqMu[i,1]=Test[,1]; /* Mean of subtest 1 */
    EqMu[i,2]=Test[,2]; /* Mean of subtest 2 */

    EqVar[i,1]=((Test[,1]##2)[+]\text{-}N*EqMu[i,1]##2)/(N-1); /* Variance of subtest 1 */
    EqVar[i,2]=((Test[,2]##2)[+]\text{-}N*EqMu[i,2]##2)/(N-1); /* Variance of subtest 2 */

    corr1=corr(X[,Half1]);
    corr1[loc(diag(corr1))]=.;
    EqCorr[i,1]=corr1[1]; /* Average correlation of subtest 1 items */

    corr2=corr(X[,Half2]);
    corr2[loc(diag(corr2))]=.;
    EqCorr[i,2]=corr2[1]; /* Average correlation of subtest 2 items */

    PTC=corr(Test)[1,2]; /* Parallel test correlation */
    Rel[i]=2*PTC/(1+PTC); /* Stepped-up with Spearman-Brown prophecy formula */
FINISH ASSUMPTIONS;

START PARALLEL;
    Rel=[sim,1,1];
    EqMu=[sim,2,2];
EqVar=j(sim,2,);  
EqCorr=j(sim,2,);  
do i=1 to sim;  
   RUN SPLITHALF;  
   RUN ASSUMPTIONS;  
end;  
call sort(Rel, {1});  
LCI=Rel[floor(sim*alpha/2)];  
UCI=Rel[ceil((1-alpha/2)*sim)];  
  diff=UCI-LCI;  
Rel=Rel[:];  
EqMu=EqMu[:];  
EqVar=EqVar[:];  
EqCorr=EqCorr[:];  
reset NOAUTONAME NONAME ;  
print sim[c="No. of simulations" f=18.0] V[c="No. of items" f=18.0]  
   N["Sample size" f=18.0] Rho["Desired correlation" f=23.4]  
   alpha["Alpha" f=10.2],  
   meanCorr1["Expected correlation" f=20.4]  
   meanCorr2["Expected correlation (pos. def.)" f=30.4]  
   meanCorrX["Observed average correlation" f=30.4],  
   Rel["Parallel reliability" f=10.4] LCI["LCI" f=10.4] UCI["UCI" f=10.4]  
   diff["Diff" f=10.4];  
print "Assumptions Tests";  
reset nocenter;  
print "Means of parallel tests are equivalent" EqMu["Test1" "Test2" f=10.4];  
print "Variance of parallel tests are equivalent" EqVar["Test1" "Test2" f=10.4];  
print "Inter-item correlations of parallel tests are eq." EqCorr["Test1" "Test2" f=10.4];  
reset center;  
FINISH PARALLEL;  
START MAIN;  
   seed1=56161;  
   seed2=236181;  
   seed3=1165556;  
   alpha=0.05;  
   sim=10000;  /* Number of simulations */  
   N=100000;  /* Sample size */  
   V=60;  /* Number of items on full-length test */  
   rho=0.1;  /* Expected value for correlation matrix */  
RUN SIMCORR;  /* Simulate correlation matrix with desired average Rho */  
RUN CHECK;  /* Check to make sure matrix is positive definite */  
RUN SIM;  /* Generate random variables based on correlation matrix */  
RUN PARALLEL;  
FINISH;  
runquit;  /* */