12-1974

Distribution-Free Interval Estimation of the Largest α-Quantile

William E. Plouff
Western Michigan University

DISTRIBUTION-FREE INTERVAL ESTIMATION OF THE LARGEST α-QUANTILE

by

William E. Plouff

A Project Report
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Specialist in Arts Degree

Western Michigan University
Kalamazoo, Michigan
December 1974

ACKNOWLEDGEMENTS

I wish to express my appreciation to Professor Gerald L. Sievers for his patient guidance and refreshing encouragement during the preparation of this paper and throughout my graduate studies.

My gratitude also goes to Professor Michael R. Stoline and Professor Ronald L. Iman for their roles in my statistical training and for serving on my thesis committee.

I wish to also thank Western Michigan University for the financial support I received while pursuing the degree of Specialist in Arts in Mathematics.

William E. Plouff

DISTRIBUTION FREE INTERVAL ESTIMATION OF THE LARGEST α-QUANTILE

William E. Plouff, Sp.A.

Western Michigan University, 1974

Several procedures based on order statistics from unbalanced samples are given for the interval estimation of the largest α-quantile of several continuous distributions. Infimum of probability coverage is discussed for these procedures for one-sided and two-sided intervals.

Alternate procedures based on different types of order statistics are given for the interval estimation of the largest median of several continuous symmetric distributions. Infimum of probability coverage is also discussed for the intervals arising from these procedures.

MASTERS THESIS M-6688

PLOUFF, William E.
DISTRIBUTION-FREE INTERVAL ESTIMATION OF THE LARGEST α-QUANTILE.

Western Michigan University, Sp.A., 1974
Mathematics

Xerox University Microfilms, Ann Arbor, Michigan 48106

TABLE OF CONTENTS

I. INTRODUCTION

Statistical problems concerned with the ordering of several unknown parameters have been of wide interest. There are numerous types of problems. One of these is the problem of estimating the largest parameter. Both point and interval estimation problems have been considered by several authors. This paper will focus on a procedure based on order statistics in a distribution-free setting given by Rizvi and Saxena [1]. They considered the interval estimation of the largest α-quantile of several continuous distributions using equal size samples from each population. An extension of this procedure to the case of unbalanced sample sizes will be the first of two main discussions in this paper. The second will concern an improvement of the procedure for the specific case when the parameter of interest is the median and the distributions are assumed to be symmetric.

Alam [2], Bechofer [3], Blumenthal and Cohen [4], and Dudewicz [5] have considered point estimation of the largest parameter. Estimating the largest mean from k $(k \geq 2)$ populations was often the problem considered, particularly for k normal distributions.

Saxena and Tong [6] considered the construction of a fixed-width confidence interval for the largest mean of k $(k \geq 1)$ normal populations, with known variances, based on the largest sample mean. Their procedure used symmetric intervals of a fixed length about the maximum of a set of consistent estimators of the k unknown parameters. The infimum of probability coverage for these intervals was discussed for the special case when the parameter of interest is a location parameter. Many of the techniques used in the proofs of theorems in [6] were used later by Rizvi and Saxena [1], and will be used in Chapter 2 when we discuss the extension of [1] to the case of unbalanced sample sizes.

Dudewicz [7] and [8] developed a procedure for obtaining two-sided confidence intervals for ranked means by using the k $(k \geq 1)$ sample means in yet a different way. He argued his confidence intervals were better than those described by Saxena and Tong in [6]. Tong [15] later argued that a class of hybrid estimators should be considered when discussing the k $(k \geq 1)$ sample problem of estimating the largest population mean.

Saxena [9] also discussed random-width confidence intervals for the largest variance from k $(k \geq 1)$ normal distributions. The use of consistent estimators and several procedural methods were similar to those in [6].

Tong [10] extended the idea of estimating the largest of k $(k \geq 1)$ means of k normal distributions for the case of unknown common variances. A multi-stage procedure was used in which an estimate of the variance was found.

Tong [11] later discussed an improvement of the usage of symmetric intervals for the largest mean of k $(k \geq 1)$ normal distributions.

Alam, Saxena, and Tong [12] extended confidence interval ideas for location and scale parameters to a more general setting of k $(k \geq 1)$ univariate distributions, indexed by a real parameter θ. They used the maximum of individual observations from the k populations to construct a confidence interval for the maximum parameter from the set of k parameters. The authors' approach involved the finding of two strictly increasing continuous functions and their corresponding inverses, which would satisfy certain conditions. With appropriate restrictions on these functions and the original univariate distributions the authors obtained several general theorems. The proofs of these theorems contained techniques similar to those in [6,7].

It was stated in [12] that a family of distributions with the monotone likely ratio property would satisfy the necessary conditions for the theorems if the parameter of interest is a location or scale parameter. The largest mean of a family of normal distributions was considered by the authors in an example. Consistent estimators were used as estimates for the population means. If the X_i's and $F(\cdot,\theta_i)$'s in [12] are replaced by the sample means and their corresponding cdf's the results are precisely those stated in [6 and 11].

In the references cited thus far, the main emphasis has been on finding confidence intervals for the largest parameter from k univariate distributions, each indexed by a corresponding parameter. Location and scale parameters were usually the parameters of interest; the confidence intervals were often centered around consistent estimators, or some other single function, from each population sample. Order statistics were not used in the procedure. The theorems and other main results in the papers were not given in detail since it was the techniques in the proofs and the general procedures that were important, not the actual results. In Rizvi and Saxena [1] and in our extension, in Chapter 2, of their ideas the distribution-free setting and the usage of order statistics led to results which are often very different than those in the papers previously mentioned.

In a univariate setting, discussions comparing confidence intervals for the median of a symmetric distribution arising from the sign test statistic (which is equivalent to the usual order statistic approach) and the Wilcoxon test statistic have been given by many authors, including Gibbons [13]. For many types of distributions, it is well known that for the one sample case the confidence interval related to the Wilcoxon test statistic has shorter expected length than the one related to the sign test statistic.

The methods in [1] for finding a confidence interval for the largest α-quantile can be thought of in terms of confidence intervals related to the sign test statistic. It would seem natural when considering the largest of the medians of k $(k \geq 1)$ symmetric distributions to attempt to parallel the procedures in [1] using confidence intervals related to the Wilcoxon test. In Chapter 3 we will discuss difficulties that arise in such an attempt.

Noether [14] described a confidence interval for the median of a symmetric distribution which was better than the one related to the sign test statistic. His (g,h) intervals are related to both the sign test and Wilcoxon test. In Chapter 3 we will show how his procedure can be extended to the k-sample problem. The resulting confidence intervals should have shorter expected length than those of Rizvi and Saxena [1] for the largest median of k $(k \geq 1)$ symmetric population distributions.

II. DISTRIBUTION-FREE INTERVAL ESTIMATION OF THE LARGEST α-QUANTILE FOR k UNBALANCED SAMPLES

The purpose of this section is to extend the procedure of Rizvi and Saxena [1] to the case of k (≥ 1) unbalanced samples.

Consider k (≥ 1) distributions with unknown continuous cumulative distribution functions F_i , $i = 1,2,\ldots,k$. Let $x_\alpha(F_i)$ denote the unique α-quantile $(0 < \alpha < 1)$ of F_i . If $x_\alpha(F_i)$ is not unique it can be defined to be so in the usual manner. Define $\theta = \max_{1\leq i\leq k} x_\alpha(F_i)$. For a specified constant γ , a random interval I is desired such that

$$\text{(i)} \qquad \inf_{\Omega} \Pr\{\theta \in I\} \geq \gamma ,$$

where Ω denotes the set of all possible k-tuples $(F_1,F_2,\ldots,F_k)$. Let $X_{1,i},X_{2,i},\ldots,X_{n,i}$ denote the order statistics of a random sample of size n_i from the i^{th} distribution, $i = 1,2,\ldots,k$.

The procedure in [1] for determining an interval I is based on the order statistics of random samples of equal sizes from each F_i , $i = 1,2,\ldots,k$. Intervals of the form (Y_s,Y_t) are found where $Y_r = \max_{1\leq i\leq k} X_{r,i}$, $1 \leq r \leq n$. The statistic Y_r , $1 \leq r \leq n$, does not have a meaningful interpretation as an endpoint for a confidence interval for

θ if the k samples are unbalanced. If we let the sample sizes be arbitrary a procedure using "proportional" order statistics seems reasonable. Consider the following simple example of our meaning of the term "proportional."

If $n_1 = 10$, $n_2 = 15$, and $n_3 = 20$ we refer to $X_{2,1}$, $X_{3,2}$ and $X_{4,3}$ as the two-tenths "proportional" order statistics from the first three samples.

In general, let μ and ν represent fractions such that $0 \le \mu < \alpha < \nu \le 1$, where the α-quantile is the quantile of interest. Let the μ^{th} proportional order statistic from sample i be $Y_{\mu n_i,i} = X_{[\mu n_i],i}$ and the ν^{th} proportional order statistic from sample i be $Y_{\nu n_i,i} = X_{[\nu n_i]+1,i}$ where [ ] represents the greatest integer function. For simplicity, here and in the following discussion μn_i means $[\mu n_i]$ and νn_i means $[\nu n_i]+1$. Define $Y_\mu = \max\limits_{1\le i\le k} Y_{\mu n_i,i}$ and $Y_\nu = \max\limits_{1\le i\le k} Y_{\nu n_i,i}$. We will be able to develop a procedure for the unbalanced case in a manner similar to some extent to the procedure in [1] for the balanced case.

Define $Y_0 = -\infty$ and $Y_1 = \infty$. For $\mu < \alpha < \nu$, let $I = (Y_\mu, Y_\nu)$ be the random interval under consideration. With $G_{\mu n_i,i}(p)$ denoting the incomplete beta function

$$G_{\mu n_i,i}(p) = \mu n_i \binom{n_i}{\mu n_i} \int_0^p w^{\mu n_i - 1}(1-w)^{n_i - \mu n_i} dw ,$$

where μn_i means $[\mu n_i]$ as before and

$$G_{\nu n_i,i}(p) = \nu n_i \binom{n_i}{\nu n_i} \int_0^p w^{\nu n_i - 1}(1-w)^{n_i - \nu n_i} dw ,$$

where νn_i means $[\nu n_i] + 1$, the cdf of $Y_{\mu n_i,i}$ is $G_{\mu n_i,i}(F_i(y))$ and the cdf of $Y_{\nu n_i,i}$ is $G_{\nu n_i,i}(F_i(y))$. Using the convention $G_{0,i}(\cdot) \equiv 1$ and $G_{1,i}(\cdot) \equiv 0$ the probability of coverage of θ by I can be stated as

$$Pr(\theta \in I) = Pr(Y_\mu \le \theta) - Pr(Y_\nu \le \theta)$$

$$\text{(ii)} \qquad = Pr(Y_{\mu n_i,i} \le \theta, i=1,2,\ldots,k) - Pr(Y_{\nu n_i,i} \le \theta, i=1,2,\ldots,k)$$

$$= \prod_{i=1}^{k} G_{\mu n_i,i}(F_i(\theta)) - \prod_{i=1}^{k} G_{\nu n_i,i}(F_i(\theta)) .$$

Note the above procedure and statement concerning probability of coverage are precisely the same as those in [1] if the k sample sizes are equal.

Using the convention $G_{0,i}(\cdot) \equiv 1$ and $G_{1,i}(\cdot) \equiv 0$, and noting that $\alpha \le F_i(\theta) \le 1$, for all i, we can state the following.

Theorem 2.1 (one-sided intervals):

(a) For $\mu = 0$, $\nu < 1$, that is, with $I = (-\infty, Y_\nu)$,

$\inf_{\Omega} Pr\{\theta \in I\} = 1 - G_{\nu n_i,i}(\alpha)$ where

$i' \in \{1,2,\ldots,k\}$ is chosen so that

$G_{\nu n_{i'},i'}(\alpha) \geq G_{\nu n_i,i}(\alpha)$, for all i ; and

(b) for $\mu > 0$, $\nu = 1$, that is, with $I = (Y_\mu,\infty)$,

$$\inf_{\Omega} \Pr\{\theta \in I\} = \prod_{i=1}^{k} G_{\mu n_i,i}(\alpha) .$$

<u>Proof</u>: (a) and (b) follow readily from (ii), thus the details are omitted. #

Tables of the incomplete beta function or equivalently cumulative binomial tables can be used to evaluate the right hand side of (a) and (b). For (a) of Theorem 2.1 we choose ν to be the smallest fraction such that $1 - G_{\nu n_{i'},i'}(\alpha) \geq \gamma$. For a given α , γ must satisfy $0 < \gamma < 1 - \alpha^{n_{i'}}$. Note that for $\mu < \mu'$ that $G_{\mu n_i,i}(\cdot) \geq G_{\mu' n_i,i}(\cdot)$. Therefore in (b) of Theorem 2.1 we choose μ to be the largest fraction such that

$$\prod_{i=1}^{k} G_{\mu n_i,i}(\alpha) \geq \gamma , \quad \text{with} \quad 0 < \gamma^{1/k} < \{1 - (1-\alpha)^{n_i}\} ,$$

for all i . It is clear from the above upper bounds on γ that for fixed k and α , I can satisfy (i) for any value of γ between 0 and 1 , provided $\min\{n_1,n_2,\ldots,n_k\}$ is large enough.

The amount of effort to find μ or ν for a given set of values $k,\alpha,\gamma,n_1,n_2,\ldots,n_k$ is often not as short as for the balanced case considered in [1]. The expressions (a) and (b) are more complicated for unbalanced cases and usually more trial-and-error calculations are required before the desired μ or ν is arrived at. In searching for an appropriate μ and ν such that (i) holds, we can alternately express $G_{\mu n_i, i}$ in terms of the cumulative binomial distribution

$$G_{\mu n_i, i}(\alpha) = \sum_{j=\mu n_i}^{n_i} \binom{n_i}{j} \alpha^j (1-\alpha)^{n_i - j} .$$

$G_{\nu n_i, i}(\alpha)$ can be expressed in a similar manner. (Keep in mind μn_i is rounded down and νn_i is rounded up in our convention.)

Let's consider finding ν in Theorem 2.1 (a) for a specified k, γ, α and $n_1, n_2, \ldots, n_k$. We need to use a trial-and-error procedure of examining $\nu n_1, \nu n_2, \ldots, \nu n_k$ for various values of ν. This involves checking the binomial cdf tables to determine $G_{\nu n_i, i}$ for each n_i and ν, $i = 1, 2, \ldots, k$. We need to find the appropriate smallest ν and the particular n_i such that (a) of Theorem 2.1 is satisfied. For many ν's it will follow that $n_{i'}$ of (a) will be $\min\{n_1, n_2, \ldots, n_k\}$. This is not always the case, however, due to the effects of the rounding off procedure for the νn_i's and the discreteness of the binomial distribution. One can use $\min\{n_1, n_2, \ldots, n_k\}$ to get an initial value for ν, then check (a) and adjust ν slightly if necessary.

If the sample sizes are appropriately large to allow a normal approximation to the binomial cdf, the continuous property of the normal distribution allows the following simplification of the above trial-and-error procedure. Assuming

$$G_{\nu n_i, i}(\alpha) \approx N\left(-\frac{\nu n_i - n_i \alpha}{\sqrt{n_i \alpha (1-\alpha)}}\right) ,$$

for all i , it follows that

$$G_{\nu n_{i'}, i'}(\alpha) \geq G_{\nu n_i, i}(\alpha) \text{ , for all } i \text{ , } \Rightarrow n_{i'} \leq n_i \text{ for all } i \text{ .}$$

Thus for the normal approximation we can find the appropriate ν by considering

$$N\left(-\frac{\nu n_{i'} - n_{i'} \alpha}{\sqrt{n_{i'} \alpha (1-\alpha)}}\right) \leq 1 - \gamma \quad \text{with}$$

$n_{i'} = \min\{n_1, n_2, \ldots, n_k\}$.

Let's now consider finding μ in (b) of Theorem 2.1 for a specified k, γ, α and $n_1, n_2, \ldots, n_k$. A procedure somewhat similar to that for part (a) can be used. The trial-and-error procedure to find μ involves working with a product of k-terms, all dependent on μ . Finding the largest μ such that

$$\prod_{i=1}^{k} G_{\mu n_i, i}(\alpha) \geq \gamma$$

is somewhat tedious since the precise relationship between the individual factors varies for different μ's . This is again due to the discreteness of the binomial cdf and the rounding off of the μn_i's .

A number of methods have been attempted to find a simpler procedure to determine the largest μ such that (b) of Theorem 2.1 holds. If a single function of the n_i's can be used to determine μ , only a single binomial cdf value would need to be considered. The following methods seem reasonable but have not yielded exact results:

(1) Choose μ such that $G_{\mu n_{i'}, i'}(\alpha) \geq \gamma^{1/k}$ where $n_{i'} = \min_{1 \leq i \leq k} \{n_1, n_2, \ldots, n_k\}$. This procedure often leads to very conservative probability statements, especially if the sample sizes are small or vary from one another to any great extent.

(2) Choose μ such that $G_{\mu\frac{\Sigma n_i}{k}, \frac{\Sigma n_i}{k}}(\alpha) \geq \gamma^{1/k}$. The Generalized Hölders Inequality can be used to make infimum statements. The resulting statement for (b) of Theorem 2.1 may be good or poor depending again on the relationship of the sample sizes.

If the normal approximation to the binomial cdf is good (i.e. if the sample sizes are large enough), the trial-and-error procedure to find μ for a specified k, γ, α and $n_1, n_2, \ldots, n_k$ in (b) of Theorem 2.1 does become easier by using the following approach. Assume that $n_1 \le n_2 \le \ldots \le n_k$. We note that from our assumption of good normal approximations

$$\prod_{i=1}^{k} G_{\mu n_i, i}(\alpha) \approx \prod_{i=1}^{k} N\left(-\frac{\mu n_i - n_i\alpha}{\sqrt{n_i\alpha(1-\alpha)}}\right) = \prod_{i=1}^{k} N\left(-\frac{\mu-\alpha}{\sqrt{\alpha(1-\alpha)}}\, n_1 \cdot \sqrt{\frac{n_i}{n_1}}\right) .$$

Let $-\frac{\mu-\alpha}{\sqrt{\alpha(1-\alpha)}} \cdot \sqrt{n_1} = z_1$. Using this notation,

$$\prod_{i=1}^{k} G_{\mu n_i, i}(\alpha) \approx \prod_{i=1}^{k} N\left(z_1 \cdot \sqrt{\frac{n_i}{n_1}}\right) .$$

Finding the smallest z_1 (hence the largest μ) such that $\prod_{i=1}^{k} N\left(z_1 \cdot \sqrt{\frac{n_i}{n_1}}\right) \ge \gamma$ is an easier task than the procedure described using the binomial cdf tables. Note, if $\sqrt{\frac{n_i}{n_1}} \approx 1$, for all i , $z_1 \approx z^*$ where z^* is a value such that $N(z^*) = \gamma^{1/k}$. If $\sqrt{\frac{n_i}{n_1}} \not\approx 1$ for at least one value of i , $i \in \{1,2,\ldots,k\}$, we should try values for z_1 such that $z_1 < z^*$.

We have noted that the previously discussed method of obtaining one-sided confidence intervals using proportional order statistics can involve lengthy trial-and-error calculations. We will now describe another method, involving intervals defined in a different way, which requires simpler calculation for its execution. We'll see that often the resulting one-sided intervals will involve the same order statistics for the two procedures. We will indicate that for most sets of sample sizes neither of the two procedures will be preferable, prior to a particular sampling. We will use the term "preferable" as Wilks [16] and Rizvi and Saxena [1] do. We will want to choose μ and ν in Theorem 2.1 such that the rank difference $(\nu n_i - \mu n_i)$ is minimized for a preassigned γ for each i, $i = 1,2,\ldots,k$. In the following procedure we will choose m_i and ℓ_i such that the rank difference $(\ell_i - m_i)$ is minimized for each i, $i = 1,2,\ldots,k$, for a preassigned γ used in the next theorem.

Henceforth for notational purposes, we will refer to the previous procedure as the μ,ν proportional procedure. The following procedure will be called the m,ℓ procedure.

To describe the one-sided m,ℓ procedure consider the following. Let $Y_{m_i,i}$ denote the m_i^{th} largest order statistic for the i^{th} population (with sample size n_i). We choose m_i to be the largest rank such that $G_{m_i,i}(\alpha) \geq \gamma^{1/k}$ for each i. Let $Y_{\ell_i,i}$ denote the ℓ_i^{th} largest order statistic from the i^{th} population and choose ℓ_i to be the smallest rank such that $G_{\ell_i,i}(\alpha) \leq 1 - \gamma$ for each i. Define $Y_m = \max_{1\leq i\leq k} Y_{m_i,i}$ and $Y_\ell = \max_{1\leq i\leq k} Y_{\ell_i,i}$. Use the convention $Y_{0,i} = -\infty$ and $Y_{n_i+1,i} = \infty$, for all i.

The one-sided intervals arising from this procedure are of the type $I = (Y_m, \infty)$ or $I = (-\infty, Y_\ell)$. Noting that $G_{0,i}(\cdot) \equiv 1$, $G_{n_i+1,i}(\cdot) \equiv 0$ and $\alpha \le F_i(\theta) \le 1$ for each i we state

Theorem 2.2 (one-sided intervals):

(a) For $m_i = 0$, $\ell_i \le n_i$, for all i , that is, with $I = (-\infty, Y_\ell)$,

$$\inf_\Omega \Pr\{\theta \in I\} = 1 - G_{\ell_{i^*}, i^*}(\alpha) \ge \gamma \text{ where } i^* \in \{1,2,\ldots,k\}$$

is chosen so that

$$G_{\ell_{i^*}, i^*}(\alpha) \ge G_{\ell_i, i}(\alpha) \text{ , for all } i \text{ ; and}$$

(b) for $m_i \ge 1$, $\ell_i = n_i + 1$, for all i , that is, with $I = (Y_m, \infty)$

$$\inf_\Omega \Pr\{\theta \in I\} = \prod_{i=1}^{k} G_{m_i,i}(\alpha) \ge \gamma \text{ .}$$

Proof: (a) and (b) follow directly from the way we defined m_i and ℓ_i for each i , $i = 1,2,\ldots,k$. #

Note for (a) and (b) to be meaningful we need the requirement that for (a) $0 < \gamma^{1/k} < (1 - (1-\alpha)^{n_i})$, for all i , for a given α , and for (b) that $0 < \gamma < 1 - \alpha^{n_{i*}}$ for a given α . It is clear from upper bounds on γ that for fixed k and α (i) can hold for any value of γ between 0 and 1, provided $\min\{n_1,n_2,\ldots,n_k\}$ is large enough.

To show the confidence intervals (Y_m,∞) and $(-\infty,Y_\ell)$ from the m,ℓ procedure can often be very similar to (Y_μ,∞) and $(-\infty,Y_\nu)$, respectively, from the μ,ν proportional procedure we'll compare them when the sample sizes are sufficiently large to allow the use of normal approximations for the binomial cdf's involved. In doing this we will look at several numerical examples for various sets of sample sizes.

In comparing $(-\infty,Y_\ell)$ to $(-\infty,Y_\nu)$ we see the two procedures involved are similar since only one term is involved in the infimum statement. We find νn_i, is usually ℓ_{i*}. Comparing (Y_m,∞) to (Y_μ,∞) is more difficult. Using the notation

$$G_{\mu n_i,i}(\alpha) \approx N\left(z_1\cdot\sqrt{\frac{n_i}{n_1}}\right), \quad \text{for all } i, \text{ where}$$

$$z_1 = \frac{-\mu+\alpha}{\sqrt{\alpha(1-\alpha)}}\cdot\sqrt{n_1} \quad \text{and} \quad G_{m_i,i}(\alpha) \approx N\left(\frac{-m_i+n_i\alpha}{\sqrt{n_i\alpha(1-\alpha)}}\right),$$

for all i, it follows from the two procedures that

$$N(z_1)\cdot N\left(z_1\cdot\sqrt{\frac{n_2}{n_1}}\right)\ \ldots\ N\left(z_1\cdot\sqrt{\frac{n_k}{n_1}}\right) = \gamma \quad \text{and} \quad N(z^*) = \gamma^{1/k},$$

where $z^* = \dfrac{-m_i+n_i\alpha}{\sqrt{n_i\alpha(1-\alpha)}}$, for all i.

It is easily verified that $z_1 \le z^* \le z_k = z_1\sqrt{\frac{n_k}{n_1}}$. If the sample sizes satisfy $\sqrt{\frac{n_i}{n_1}} \approx 1$ for all i it follows that the confidence intervals for the two procedures are similar.

The following examples illustrate the relationships of the rank of the order statistics obtained by the two procedures for various applications of part (b) of Theorems 2.1 and 2.2. We will use the normal approximation procedures.

We use a correction factor of .5 when using normal approximations.

Ex1. Let $n_1 = 16$, $n_2 = 25$, $n_3 = 36$, $n_4 = 49$, $\alpha = 1/2$, $\gamma = .90$. Then $z_1 = 1.56$ and $\prod_{i=1}^{4} G_{\mu n_i, i}(\alpha) = .90$.

For $z_1 = 1.56$ we get $\mu = .30$. Then

$[\mu n_1] = 5$ $\qquad$ $m_1 = 4$

$[\mu n_2] = 8$ $\qquad$ $m_2 = 8$

$[\mu n_3] = 11$ $\qquad$ $m_3 = 12$

$[\mu n_4] = 15$ $\qquad$ $m_4 = 18$.

Ex2. Let $n_1 = 80$, $n_2 = 90$, $n_3 = 90$, $n_4 = 100$, $\alpha = .3$, $\gamma = .80$

Then $z_1 = 1.52$, $\mu = .22$ and

$[\mu n_1] = 18$ $\qquad$ $m_1 = 18$

$[\mu n_2] = 20$ $\qquad$ $m_2 = 20$

$[\mu n_3] = 20$ $\qquad$ $m_3 = 20$

$[\mu n_4] = 23$ $\qquad$ $m_4 = 23$.

Ex3. Let $n_1 = 160$, $n_2 = 180$, $n_3 = 200$, $n_4 = 220$, $\alpha = .3$

$\gamma = .955$.

Then $z_1 = 1.38$, $\mu = .25$ and

$[\mu n_1] = 40$	$m_1 = 39$
$[\mu n_2] = 45$	$m_2 = 45$
$[\mu n_3] = 50$	$m_3 = 50$
$[\mu n_4] = 55$	$m_4 = 56$.

Ex4. Let $n_1 = 80$, $n_2 = 90$, $n_3 = 90$, $n_4 = 100$, $\alpha = .75$,

$\gamma = .943$.

Then $z_1 = 2.065$, $\mu = .65$ and

$[\mu n_1] = 52$	$m_1 = 51$
$[\mu n_2] = 59$	$m_2 = 58$
$[\mu n_3] = 59$	$m_3 = 58$
$[\mu n_4] = 65$	$m_4 = 65$.

We've seen the μ,ν proportional procedure and the m,ℓ procedure yield similar intervals for (a) of Theorem 2.1 and (a) of Theorem 2.2. From the normal approximation discussion for intervals of the form (Y_μ,∞) from (b) of Theorem 2.1 and of the form (Y_m,∞) from (b) of Theorem 2.2 we know that in most cases neither procedure is preferable. In one of the four numerical examples we see that one method is preferable to the other. This occurs in the last example, the μ,ν proportional method is preferable. This follows since each rank of the individual order statistics for the proportional method is greater than or equal to the rank of the corresponding order statistic from the m,ℓ procedure. Theoretically, from the inequalities $z_1 \le z^* \le z_k$ it follows that either procedure might be preferable for a particular set of values k,α,γ and $n_1,n_2,\ldots,n_k$, but it seems reasonable that in most cases neither procedure is preferable. For this reason, the m,ℓ procedure for finding one-sided random intervals satisfying (i) is often desirable because of the simpler calculations for its execution.

We now discuss two-sided intervals. We introduce the following terminology. Let $Y_{s_i,i}$ be the s_i^{th} order statistic from sample i and define $Y_s = \max_{1\le i\le k} Y_{s_i,i}$. Let $Y_{t_i,i}$ be the t_i^{th} order statistic from sample i and define $Y_t = \max_{1\le i\le k} Y_{t_i,i}$. For interval $I = (Y_s,Y_t)$, it follows in a manner similar to (ii) that

$$\text{(iii)}\quad \Pr(\theta \in I) = \prod_{i=1}^{k} G_{s_i,i}(F_i(\theta)) - \prod_{i=1}^{k} G_{t_i,i}(F_i(\theta)) .$$

Consider minimization over Ω of $\Pr\{\theta \in I\}$ for two-sided random intervals.

<u>Theorem 2.3 (two-sided intervals)</u>: For $1 \le s_i < t_i \le n_i$, for all i, that is, with $I = (Y_s,Y_t)$,

$$\inf_{\Omega} \Pr\{\theta \in I\} \geq \min\{ \prod_{i=1}^{r} G_{s_{\beta_i},\beta_i}(\alpha) - \prod_{i=1}^{r} G_{t_{\beta_i},\beta_i}(\alpha)\}$$

where the minimum is over all choices of $r = 1,2,\ldots,k$ and $\{\beta_1,\ldots,\beta_r\} \subseteq \{1,2,\ldots,k\}$.

Proof: The proof follows a pattern similar to Theorem 2 in [1]. Adapting the discussion in [1] to our problem we get the following:

$$\text{(iv)}\quad \Pr\{\theta \in (Y_s,Y_t)\} = \prod_{i=1}^{k} G_{s_i,i}(F_i(\theta)) - \prod_{i=1}^{k} G_{t_i,i}(F_i(\theta)) \ .$$

We see $\Pr\{\theta \in I\}$ involves F_i's evaluated at the constant θ , and $F_i(\theta) \geq \alpha$ for $i = 1,2,\ldots,k$. We can write $F_i(\theta) = \alpha + \delta_i$; where $0 \leq \delta_i \leq 1 - \alpha$. This enables us to reparameterize (iv) as a function of the δ_i's . Without loss of generality we can assume $F_k(\theta) = \alpha$. Thus the problem of minimization of (iv) over $\Omega = \{(F_1,F_2,\ldots,F_k)\colon F_i$ is continuous for each $i\}$ is reduced to its minimization over $\{(\delta_1,\ldots,\delta_{k-1})\colon 0 \leq \delta_i \leq 1 - \alpha ,\ i = 1,2,\ldots,k-1\}$.

$$\begin{aligned} \Pr\{\theta \in I\} &= G_{s_k,k}(\alpha) \prod_{i=1}^{k-1} G_{s_i,i}(\alpha+\delta_i) - G_{t_k,k}(\alpha) \prod_{i=1}^{k-1} G_{t_i,i}(\alpha+\delta_i) \\ \text{(v)}\qquad & \\ &= J(\delta_1,\ldots,\delta_{k-1}) \ , \text{ say.} \end{aligned}$$

For some j , fix $\delta_1,\ldots,\delta_{j-1},\delta_{j+1},\ldots,\delta_{k-1}$ and consider $\partial J/\partial\delta_j$. Using $G_{r_i,i}(p)$ written in its form as the incomplete beta function we define

$$g_{r_i,i}(p) = \frac{d}{dp} G_{r_i,i}(p) = r_i\binom{n}{r_i}p^{r_i-1}(1-p)^{n-r_i}, \; 0 \le p \le 1.$$

We can observe that $g_{t_i,i}(p)|g_{s_i,i}(p)$ is increasing in p for $t_i > s_i$ since

$$\frac{t_i\binom{n_i}{t_i}p^{t_i-1}(1-p)^{n_i-t_i}}{\binom{n_i}{s_i}p^{s_i-1}(1-p)^{n_i-s_i}} = Cp^{t_i-s_i}(1-p)^{s_i-t_i}$$

$$= C\left[\frac{p}{1-p}\right]^{t_i-s_i}$$

which increases as p increases since p is positive, C , a constant, is positive and $t_i - s_i > 0$.

Let $A = \prod_{i=1, i\neq j}^{k} G_{s_i,i}(\alpha+\delta_i)$

and $B = \prod_{i=1, i\neq j}^{k} G_{t_i,i}(\alpha+\delta_i)$.

We can obtain from (v) that

$$\frac{\partial J}{\partial \delta_j} = A \cdot g_{s_i,i}(\alpha+\delta_j)\left[1 - \frac{Bg_{t_i,i}(\alpha+\delta_j)}{Ag_{s_i,i}(\alpha+\delta_j)}\right] .$$

Since $\frac{g_{t_i,i}(\alpha+\delta_j)}{g_{s_i,i}(\alpha+\delta_j)}$ is increasing it follows that the expression in the brackets is decreasing in δ_j . Hence we conclude that $\frac{\partial J}{\partial \delta_j}$ has either the same sign at every value of δ_j or at most one change of sign from positive to negative and consequently $\inf_{\delta_j} J$ is either at $\delta_j = 0$ or $\delta_j = 1 - \alpha$. This holds for every j . Therefore, the infimum of J is achieved when a certain number of δ_j's are zero and the rest are equal to $1 - \alpha$. The statement of Theorem 2.3 thus follows. #

Note that for fixed k and α , I can satisfy (i) provided γ lies between 0 and

$$\min_{1 \le r \le k} \left\{ \prod_{i=1}^{r} \left[(1 - (1-\alpha)^{n_i} \right] - \prod_{i=1}^{r} \alpha^{n_i} \right\} .$$

This minimum can be made arbitrarily close to 1 by taking the n_i's large enough.

Let s_i be μn_i and t_i be νn_i for each i in (iii), where μ and ν are from the μ,ν proportional procedure. It follows from (iii) and Theorem 2.3 that for $I = (Y_\mu, Y_\nu)$

$$\text{(vi)} \quad \inf_{\Omega} \Pr\{\theta \in I\} \ge \min \left\{ \prod_{i=1}^{r} G_{s_{\beta_i'},\beta_i}(\alpha) - \prod_{i=1}^{r} G_{t_{\beta_i'},\beta_i}(\alpha) \right\} .$$

There are many μ's and ν's that would yield a value greater than γ , for a specified γ , if used in the right hand side of (vi). At the end of this chapter we will discuss an algorithm for finding μ and ν in an optimal way, similar to Rizvi and Saxena's optimum discussion in [1].

Typically, the minimum of the right hand side of (vi) occurs when $r = k$. To find μ and ν such that the right hand side of (vi) is greater than or equal to some specified γ a convenient starting point is to choose a γ_1 and γ_2 such that $\gamma_1 - \gamma_2 = \gamma$. Then μ is determined such that $\prod_{i=1}^{k} G_{\mu n_i, i}(\alpha) \geq \gamma_1$ by the procedure described for finding μ for Theorem 2.1 (b)(let $\gamma_1 = \gamma$ in that discussion). In a similar manner ν is determined, making ν as large as possible, such that

$$\prod_{i=1}^{k} G_{\nu n_i, i}(\alpha) \leq \gamma_2 .$$

Using these values for μ and ν the right side of (vi) is evaluated to check if the resulting value is greater than or equal to the specified γ. If it is not or if the infimum statement that results from (vi) is too conservative a slight adjustment of μ and ν is necessary. Note the interval determined in this way is not necessarily optimal.

Let's now define a two-sided random interval for the m,ℓ procedure. For a given γ, we choose a γ_1 and γ_2 such that $\gamma_1 - \gamma_2 = \gamma$ and $\gamma_1^{1/k} - \gamma_2^{1/k} \geq \gamma$. For each i, $i = 1,2,\ldots,k$, choose m_i as the largest integer such that $G_{m_i, i}(\alpha) \geq \gamma_1^{1/k}$, and choose ℓ_i as the smallest integer such that $G_{\ell_i,i}(\alpha) \le \gamma_2^{1/k}$.

With this selection of m_i and ℓ_i for each i , we can state the following.

Theorem 2.4 (two-sided intervals):

For $1 \le m_i < \ell_i \le n_i$, for all i , that is, with $I = (Y_m, Y_\ell)$,

$$\inf \Pr\{\theta \in I\} \ge \gamma .$$

Proof: We substitute m_{β_i} for s_{β_i} and ℓ_{β_i} for t_{β_i} for each β_i in (vi). By the way we define m_{β_i} and ℓ_{β_i} for each β_i it follows that

$$\inf_{\Omega} \Pr\{\theta \in I\} \ge \min_{1\le r\le k} \{(\gamma_1^{1/k})^r - (\gamma_2^{1/k})^r\} .$$

In turn, $\inf_{\Omega} \Pr\{\theta \in I\} \ge \min\{\gamma_1^{1/k} - \gamma_2^{1/k}, \gamma_1 - \gamma_2\}$ by lines (3.5) and (3.6) in [1]. Thus by the selection of γ_1 and γ_2 , Theorem 2.4 follows. #

The two procedures discussed above for finding two-sided random intervals can be compared in a manner similar to the earlier comparison of the one-sided intervals in (b) of Theorem 2.1 and (b) of Theorem 2.2. Using normal approximation arguments, it follows that in most cases neither procedure is preferable, prior to a particular sampling. Each procedure can be preferable for certain cases by making suitable selections of k, γ, α and $n_1, n_2, \ldots, n_k$. In most cases, intervals found by the two procedures are very similar if both procedures use the same γ_1 and γ_2, such that $\gamma_1 - \gamma_2 = \gamma$. For this reason, we see the m, ℓ procedure for finding a two-sided interval satisfying (i) is often desirable because of the simpler calculation for its execution.

We conclude our discussion of two-sided random intervals by discussing an optimum two-sided random interval for the μ, ν proportional procedure.

For specified k, γ, α and $n_1, n_2, \ldots, n_k$ we can find the μ and ν that satisfies $\inf_{\Omega} \Pr\{\theta \in (Y_\mu, Y_\nu)\} \geq \gamma$ in an optimum way by using the same type of algorithm as Rizvi and Saxena did in [1]. The procedure will be exactly the same as theirs if the sample sizes are all equal.

Consider ϵ such that $0 \leq \mu \leq \mu + \epsilon \leq 1$. Denote by $Q(\mu, \epsilon)$ the infimum of $\Pr\{Y_\mu < \theta < Y_{\mu+\epsilon}\}$ over Ω as given by Theorem 2.3. For every fixed ϵ, let $\mu'(\epsilon)$ be the value of μ for which $Q(\mu'(\epsilon), \epsilon) = \max_{1 \leq \mu \leq 1-\epsilon} Q(\mu, \epsilon)$. Choose the smallest ϵ, call it ϵ', such that $Q(\mu'(\epsilon'), \epsilon') \geq \gamma$. Then the optimum choice of the random interval satisfying the $\inf_{\Omega} \Pr\{\theta \in (Y_\mu, Y_\nu)\} \geq \gamma$ is $(Y_{\mu'(\epsilon')}, Y_{\mu'(\epsilon') + \epsilon'})$.

III. DISTRIBUTION-FREE INTERVAL ESTIMATION OF THE LARGEST MEDIAN FOR k SYMMETRIC DISTRIBUTIONS

In chapter two we discussed an extension of Rizvi and Saxena's procedure in [1] for finding confidence intervals for the maximum α-quantile in a distribution-free setting. We considered the case of k unbalanced samples.

We now turn our attention to a different problem. We assume equal sample sizes, restrict the discussion to k $(k \geq 1)$ unknown symmetric distribution and focus on finding a better confidence interval (in terms of shorter expected length) for the maximum median.

For $k = 1$, confidence intervals for the median arising from order statistics related to the Wilcoxon signed rank test often have shorter expected length than those using order statistics related to the sign test. Therefore, it is natural to extend usage of Wilcoxon "order statistics to the k-sample problem to parallel Rizvi and Saxena's extension using sign test "order statistics."

We will discuss a "Wilcoxon procedure" which enables us to find better confidence intervals for the largest median of k $(k \geq 1)$ symmetric distributions when considering one-sided random intervals. However, due to the complexity of the Wilcoxon statistic distribution under alternative hypotheses an attempt to use maximum Wilcoxon "order statistics" in a method similar to Rizvi and Saxena's two-sided random intervals for the k sample problem does not prove to be successful. A procedure for finding two-sided random intervals based on a procedure of Noether [14] will be discussed. An indication of its improvement over the Rizvi and Saxena method, in terms of intervals of shorter expected length, will also be given.

Consider k $(k \geq 1)$ unknown symmetric distributions with continuous cdf's F_i, $i = 1,2,\ldots,k$. Let θ_i be the median for the i^{th} distribution, $1 \leq i \leq k$, and define $\theta = \max_{1 \leq i \leq k} \theta_i$. For a specified γ we want to find an interval I such that

$$\text{(i)} \qquad \inf_{\Omega} \Pr\{\theta \in I\} \geq \gamma$$

where Ω denotes the set of all k-tuples $(F_1,F_2,\ldots,F_k)$. Let $x_{1,i}, x_{2,i},\ldots,x_{n,i}$ represent a sample from population i, $1 \leq i \leq k$. Let $u_{j,\ell,i} = 1/2(x_{j,i} + x_{\ell,i})$, $1 \leq j \leq \ell \leq n$, $1 \leq i \leq k$. Suppose the ordered sums $u_{j,\ell,i}$ are denoted by

$$W_{1,i} < W_{2,i} < \ldots < W_{M,i} \text{ where } M = \frac{n(n+1)}{2}.$$

$W_{r,i}$ is the r^{th} order statistic from the $\frac{n(n+1)}{2}$ averages of the $x_{j,i}$'s, $j = 1,2,\ldots,n$ for a fixed i, $1 \leq i \leq k$. Let $W_r = \max_{1 \leq i \leq k} W_{r,i}$. Define $W_0 = -\infty$ and

$W_{M+1} = \infty$. For $s < t$ consider random interval $I = (W_s, W_t)$. We discuss a procedure for finding s and t such that (i) holds.

Define $D_{j,i} = x_{j,i} - \theta$, and define

$$Z_{j,i} = \begin{Bmatrix} 1 & \text{if} & D_{j,i} > 0 \\ 0 & \text{if} & D_{j,i} < 0 \end{Bmatrix}$$

(note since the distributions are continuous $Pr(D_{j,i} = 0) = 0)$. Let

$T_i^+ = \sum_{j=1}^{n} Z_{j,i}\, r(|D_{j,i}|)$, where $r(|D_{j,i}|)$ is the rank of the absolute value of $D_{j,i}$ if the absolute values of the $D_{j,i}$'s are ordered. From Gibbons [13], page 117, it follows that

(ii) $$Pr(\theta < W_{t,i}) = Pr(M - t < T_i^+)$$

and

(iii) $$Pr(W_{s,i} < \theta) = Pr(T_i^+ < M - s) .$$

When $\theta_i = \theta$, values of s and t can be found from the Wilcoxon tables so that

$$Pr(M - t < T_i^+ < M - s) \geq \gamma ,$$

for some specified γ .

The probability of coverage of θ by $I = (W_s, W_t)$ is given by

(iv)
$$\Pr(\theta \in I) = \Pr(W_s \le \theta) - \Pr(W_t \le \theta)$$
$$= \prod_{i=1}^{k} \Pr(W_{s,i} \le \theta) - \prod_{i=1}^{k} \Pr(W_{t,i} \le \theta) .$$

For one-sided random intervals the minimization of $\Pr\{\theta \in I\}$ over Ω is given by Theorem 3.1.

Theorem 3.1 (one-sided intervals):

(a) For $s > 0$, $t = M + 1$, that is, with $I = (W_s, \infty)$,

(v) $$\inf_{\Omega} \Pr\{\theta \in I\} = \{\Pr[T^+ < M - s]\}^k ; \text{ and}$$

(b) for $s = 0$, $t \le M$, that is, with $I = (-\infty, W_t)$,

(vi) $$\inf_{\Omega} \Pr\{\theta \in I\} = 1 - \Pr(T^+ < M - t) ,$$

where T^+ has the null hypothesis distribution of the Wilcoxon statistic based on n observations.

Proof: The statements of (a) and (b) follow immediately from (ii) and (iii) since $\theta_i \le \theta$, for all i, and we defined $\Pr\{W_0 \le \theta) = 1$ and $\Pr\{W_{M+1} \le \theta) = 0$. #

Note that for $s < t$, $\Pr(W_{s,i} \le \theta) \ge P(W_{t,i} \le \theta)$, for all i. Therefore, in (a) we choose s from the Wilcoxon Signed Rank Test Statistic tables to be the largest integer such that the right hand side of (v) exceeds the specified γ . This can be done for any γ such that $0 < \gamma < \{1 - (1-1/2)^M\}^k$. In (b) of Theorem 3.1 we choose t from the Wilcoxon Signed Ranks Test Statistic tables to be the smallest integer such that the right hand side of (vi) exceeds γ with $0 < \gamma < 1 - (1/2)^M$. From the upper bounds on γ we see that for fixed n (hence for fixed M) I can satisfy $\Pr\{\theta \in I\} \geq \gamma$ for any value of γ between 0 and 1 if n is taken large enough.

Let's consider the minimization of (iv) for two-sided random intervals. Let β_i be defined as the difference $\theta - \theta_i$, for $i = 1,2,\ldots,k$. If we assume $\theta = \theta_k$, we can write (iv) as

$$\Pr(W_s \leq \theta \leq W_t) = \Pr(W_{s,k} \leq \theta_k) \prod_{i=1}^{k-1} \Pr(W_{s,i} \leq \theta_i + \beta_i)$$

$$\text{(vii)} \qquad - \Pr(W_{t,k} \leq \theta_k) \prod_{i=1}^{k-1} \Pr(W_{t,i} \leq \theta_i + \beta_i)$$

$$\text{where } 0 \leq \beta_i \leq \theta,\ i = 1,2,\ldots,k-1 .$$

Since we do not have a convenient form for the distribution of $W_{r,i}$ when $\beta_i > 0$, $1 \leq r \leq M$ and $1 \leq i \leq k - 1$, statements involving (vii) are difficult to handle. For this reason, we will discuss a procedure based on ideas in a paper by Noether [14].

Noether discusses what he calls (g,h) intervals for the median of a symmetric distribution with continuous cumulative distribution function $F(x)$ and center of symmetry η. For notational purposes, x_i is introduced as the i^{th} ordered observation from a random sample of size n and η_0 as the hypothesized value of η. Noether begins his discussion with the sign test based on the two statistics $S_- = \sum_{j=1}^{n} t_j$ and $S_+ = \sum_{j=1}^{n} (1 - t_j)$, where $t_j = t_j(\eta_0) = 1$ or 0 depending on whether for the j^{th} largest absolute difference $|x_i - \eta_0|$, x_i is smaller or greater than η_0. The hypotheses $\eta = \eta_0$ is rejected in favor of the alternative $\eta \neq \eta_0$ if the smaller of S_- and S_+ is sufficiently small, say less than or equal to some value c. The interval of acceptable η-values is bounded by x_d and x_{n+1-d}, the d^{th} smallest and d^{th} largest sample observations, where $d = c + 1$. This is equivalent to the usual sign test procedure.

Noether's generalization of the sign test intervals to what he calls (g,h) intervals consists of the following. Let m be an integer such that $2 \leq m \leq n$. We look only at the m largest differences $|x_i - \eta_0|$ and define $T_- = \sum_{j=1}^{m} t_j$ and $T_+ = \sum_{j=1}^{m} (1 - t_j)$. The hypothesis $\eta = \eta_0$ is rejected in the smaller of the two statistics is less than or equal to a specified value c. The lower bound for acceptable η_0 values when testing $\eta = \eta_0$ is shown to be the sample average $1/2(x_g + x_{n+1-h})$.

where $g = c + 1$ and $h = m - c$; the upper bound is $1/2(x_h + x_{n+1-g})$. The resulting confidence interval is

$$1/2(x_g + x_{n+1-h}) < \eta < 1/2(x_h + x_{n+1-g}) .$$

Note when $m = n$ this interval corresponds to the interval associated with the sign test. The significance level for the test with critical value c is $\sum_{s=0}^{c} b(s,m,1/2)$, where $b(s,m,1/2)$ is the binomial probability of s successes in m independent trials with success probability 1/2.

In terminology somewhat different from Noether, let's consider a random sample of size n, call the observations $z_1, z_2, \ldots, z_n$, from a symmetric population with continuous cumulative distribution function $F(x)$ and center of symmetry some value, say 0. Pick $\eta > 0$ (η could be chosen less than 0 and the following discussion could be handled in a similar way). Define $t_j(\eta) = 1$ or 0 depending on whether for the j^{th} largest absolute difference $|z_i - \eta|$ z_i is smaller or greater than η. We consider only the m largest absolute differences $|z_i - \eta|$ and define

$$T^- = \sum_{j=1}^{m} t_j \quad \text{and} \quad T^+ = \sum_{j=1}^{m} (1 - t_j) .$$

We note T^- and T^+ would be of the same form as T_- and T_+ in Noether if the t_j's were defined in terms of ordered observations, say $x_1, x_2, \ldots, x_n$, instead of the unordered observations $z_1, z_2, \ldots, z_n$.

The notation $T^+ = k$ indicates k of the t_j's are zero among the t_j's that correspond to the m largest $|z_i - \eta|$'s . Since the t_j's are not known to be independent, as they are in Noether's discussion under the null hypothesis, our evaluation of $\Pr(T^+ = k)$ will involve conditional probabilities. Let Y^* be the m^{th} largest $|z_i - \eta|$, $i = 1,2,\ldots,n$. Then $\Pr(T^+ = k|Y^* = y)$ is the probability that of the m observations from the conditional distribution of z given z is outside of $(\eta - y, \eta + y)$ we have k of these in $[\eta + y, \infty)$. In terms of the cdf $F(z)$ we have

$$\Pr(T^+=k|Y^*=y) = \binom{m}{k}\left[\frac{1-F(\eta+y)}{1-F(\eta+y)+F(\eta-y)}\right]^k\left[\frac{F(\eta-y)}{1-F(\eta+y)+F(\eta-y)}\right]^{m-k}$$

Let $h(y)$ denote the density function of Y^* . We could write $h(y)$ in explicit form as an order statistic probability density function. We won't, however, since it will not be used in the following discussion. We will merely state that the range of y is $(0,\infty)$. With the above notation it follows

$$\Pr(T^+ = k) = \int_0^\infty \binom{m}{k}\left[\frac{1-F(\eta+y)}{1-F(\eta+y)+F(\eta-y)}\right]^k \cdot \left[\frac{F(\eta-y)}{1-F(\eta+y)+F(\eta-y)}\right]^{m-k} h(y)\,dy \ .$$

From this we see

$$\Pr(T^{+} \le c) = \int_0^{\infty} \sum_{k=0}^{c} \binom{m}{k} \left[\frac{1-F(\eta+y)}{1-F(\eta+y)+F(\eta-y)}\right]^{k} \cdot$$

$$\left[\frac{F(\eta-y)}{1-F(\eta+y)+F(\eta-y)}\right]^{m-k} \cdot h(y)\,dy \;.$$

The binomial expression in the square brackets has a supremum over y expressible as

$$\sum_{k=0}^{c} \binom{m}{k} (p^*)^{k} (1-p^*)^{m-k} \quad \text{for some} \quad p^* < 1/2 \;.$$

Therefore,

$$\Pr(T^{+} \le c) \le \int_0^{\infty} \left[\sum_{k=0}^{c} \binom{m}{k} (p^*)^{k} (1-p^*)^{m-k}\right] h(y)\,dy$$

$$= \sum_{c=0}^{c} \binom{m}{k} (p^*)^{k} (1-p^*)^{m-k} \int_0^{\infty} h(y)\,dy$$

$$= \sum_{k=0}^{c} \binom{m}{k} (p^*)^{k} (1-p^*)^{m-k} \;.$$

If we write p^* as $1/2 - \beta$, for some β such that $0 \le \beta \le 1/2$, we note

$$\Pr(T^{+} \le c) \le \sum_{k=0}^{c} \binom{m}{k} (1/2 - \beta)^{k} (1/2 + \beta)^{m-k} \;.$$

In a similar manner we can show

$$\Pr(T^-=k)\,|Y^*=y) = \binom{m}{k}\left[\frac{F(\eta-y)}{1-F(\eta+y)+F(\eta-y)}\right]^k\left[\frac{1-F(\eta+y)}{1-F(\eta+y)+F(\eta-y)}\right]^{m-k} .$$

It follows that

$$\Pr(T^- \le c) \le \sum_{k=0}^{c} \binom{m}{k}(1-p^*)^k (p^*)^{m-k}$$

$$= \sum_{k=0}^{c} \binom{m}{k}(1/2 + \beta)^k (1/2 - \beta)^{m-k} ,$$

where p^* and β are defined as before.

Note we elected to discuss $\Pr(T^+ \le c)$ and $\Pr(T^- \le c)$ using t_j's defined in terms of unordered observations from the symmetric population distribution. As previously mentioned, defining the t_j's in terms of ordered or unordered observations doesn't alter T^+ or T^-.

As in Noether's notation, let $x_1, x_2, \ldots, x_n$ be the ordered observations, $g = c + 1$ and $h = m - c$. The statement $1/2(x_g + x_{n+1-h}) > \eta$ indicates η is closer to the $(c + 1)$st smallest observation than to the $(m-c)^{th}$ largest observation in the sample. Thus at least $(m-c)$ of the m largest $|z_i - \eta|$'s have t_j's with value 0, so $T^- \le c$. Therefore,

$$\Pr(1/2(x_g + x_{n+1-h}) > \eta) = \Pr(T^- \le c)$$

(viii)

$$\le \sum_{k=0}^{g-1} \binom{m}{k} (1/2 + \beta)^k (1/2 - \beta)^{m-k} .$$

Similarly, $1/2(x_n + x_{n+1-g}) < \eta$ indicates η is closer to the g^{th} largest observation than to the h^{th} smallest observation. Thus at least $m - c$ of the m largest $|z_i - \eta|$'s have corresponding t_j's with value 1, so $T^+ \le c$. Therefore,

$$\Pr(1/2(x_h + x_{n+1-g}) < \eta) = \Pr(T^+ \le c)$$

$$\le \sum_{k=0}^{g-1} \binom{m}{k} (1/2 - \beta)^k (1/2 + \beta)^{m-k} .$$

From this it follows

$$\Pr(1/2(x_h + x_{n+1-g}) > \eta) \ge 1 - \sum_{k=0}^{g-1} \binom{m}{k} (1/2 - \beta)^k (1/2 + \beta)^{m-k}$$

$$= 1 - \sum_{k=m-g-1}^{m} \binom{m}{k} (1/2+\beta)^k (1/2-\beta)^{m-k}$$

(ix)

$$= \sum_{k=0}^{m-g} \binom{m}{k} (1/2+\beta)^k (1/2-\beta)^{m-k}$$

$$= \sum_{k=0}^{h-1} \binom{m}{k} (1/2+\beta)^k (1/2-\beta)^{m-k} .$$

Note if we had selected $\eta < 0$ in the above discussion the expressions for $Pr(T^- \le c)$ and $Pr(T^+ \le c)$ would remain the same except that p^* would be greater than 1/2. Thus the roles of $(1/2 - \beta)$ and $(1/2 + \beta)$ would be interchanged.

For the case $k = 1$, we see from (viii) and (ix) that for η greater than the population median

$$Pr(1/2(x_g + x_{n+1-h}) < \eta < 1/2(x_h + x_{n+1-g}))$$

$$\ge \sum_{k=0}^{h-1} \binom{m}{k} (1/2 + \beta)^k (1/2 - \beta)^{m-k}$$

$$- \sum_{k=0}^{g-1} \binom{m}{k} (1/2 + \beta)^k (1/2 - \beta)^{m-k} .$$

Let's now consider extending the above ideas to determine a two-sided random interval for the maximum median of k $(k \ge 1)$ symmetric distributions with continuous cdf's. Fix m, $2 \le m \le n$, and note $g + h = m + 1$. Define $\theta = \max_{1 \le i \le k} \theta_i$, where θ_i is the population median of the i^{th} distribution. Let $W_{g,i} = 1/2(x_{g,i} + x_{n+1-h,i})$ and $W_{h,i} = 1/2(x_{h,i} + x_{n+1-g,i})$ for each i, $i \le 1 \le k$, where $x_{1,i}, x_{2,i}, \ldots, x_{n,i}$ denote the ordered observations from sample i. Define $W_g = \max_{1 \le i \le k} W_{g,i}$ and $W_h = \max_{1 \le i \le k} W_{h,i}$. Using this notation we can state the following.

Theorem 3.2 (two-sided intervals): For $1 \le g < h \le m$ where $g = c + 1$ and $h = m - c$, that is, with $I = (W_g, W_h)$,

$$\inf_{\Omega} \Pr\{\theta \in I\} \ge \min\{(1/2)^m \sum_{s=0}^{h-1} \binom{m}{s} - (1/2)^m \sum_{s=0}^{g-1} \binom{m}{s} ,$$

$$\left[(1/2)^m \sum_{s=0}^{h-1} \binom{m}{s}\right]^k - \left[(1/2)^m \sum_{s=0}^{g-1} \binom{m}{s}\right]^k \} .$$

Proof:

$$\Pr\{\theta \in I\} = \Pr(W_{g,i} \le \theta,\ i = 1,2,\ldots,k)$$

$$- \Pr(W_{h,i} \le \theta,\ i = 1,2,\ldots,k) .$$

Using (viii) and (ix) with $\beta = \beta_j$ for all j such that $\theta_j \neq \theta$, and using Noether's discussion when $\theta_j = \theta$ (thus $\beta_j = 0$), it follows

$$\Pr(W_{g,j} \le \theta) \ge \sum_{s=0}^{h-1} \binom{m}{s} (1/2 + \beta_j)^s (1/2 - \beta_j)^{m-s}$$

and $$\Pr(W_{h,j} \le \theta) \le \sum_{s=0}^{g-1} \binom{m}{s} (1/2 + \beta_j)^s (1/2 - \beta_j)^{m-s}$$

for each j. In turn

$$\Pr(W_{g,i} \le \theta, i=1,2,\ldots,k) \ge \prod_{i=1}^{k} \sum_{s=0}^{h-1} \binom{m}{s} (1/2+\beta_i)^s (1/2-\beta_i)^{m-s}$$

and

$$\Pr(W_{h,i} \le \theta, i=1,2,\ldots,k) \le \prod_{i=1}^{k} \sum_{s=0}^{g-1} \binom{m}{s} (1/2+\beta_i)^s (1/2-\beta_i)^{m-s} .$$

Therefore,

$$\Pr\{\theta \in I\} \ge \prod_{i=1}^{k} \sum_{s=0}^{h-1} \binom{m}{s} (1/2+\beta_i)^s (1/2-\beta_i)^{m-s}$$

$$- \prod_{i=1}^{k} \sum_{s=0}^{h-1} \binom{m}{s} (1/2+\beta_i)^s (1/2-\beta_i)^{m-s} .$$

To find the infimum of the right hand side of this expression we use the results of Theorem 2 in Rizvi and Saxena directly. The statement of Theorem 3.2 thus follows. #

Note if η was chosen less than 0 in the above discussion that Theorem 3.2 would be altered since the values $(1/2 + \beta_i)$ and $(1/2 - \beta_i)$ would be interchanged in the discussion. To use Rizvi and Saxena's argument we would interchange the values by making the lower sums into appropriate upper sums. Resultingly, Theorem 3.2 would be changed only to the extent that the products in the infimum probability statement would involve upper sums.

The above procedure for two-sided intervals is equivalent to Rizvi and Saxena's for the case $m = n$.

For the case $k = 1$, Noether discusses a method of choosing g and h. Based on this method, he gives an efficiency discussion of his (g,h) intervals relative to the sign test interval. When $m < n$, he shows for many types of continuous distributions his method yields intervals with shorter expected length than the corresponding sign test intervals. It is natural to expect this property to carry over to the k sample problem, thus yielding random intervals of shorter expected length for the maximum median using our procedure than for Rizvi and Saxena's procedure in [1].

BIBLIOGRAPHY

1. Rizvi, M. H. and Saxena, K. M. L.,"Distribution-Free Interval Estimation of the Largest α-Quantile," Journal of the American Statistical Association, 67 (1972), 196-8.

2. Alam, K., "A Two Sample Estimate of the Largest Mean," Annals of the Institute of Statistical Mathematics, 19, No. 2 (1967), 271-83.

3. Bechofer, R. E., "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances," Annals of Mathematical Statistics, 25 (1954), 16-39.

4. Blumenthal, S. and Cohen, A., "Estimation of the Larger of Two Normal Means," Journal of the American Statistical Association, 63 (1968), 861-76.

5. Dudewicz, E. J., "Estimation of Ordered Parameters," Technical Report, No. 66, Department of Operations Research, Cornell University.

6. Saxena, K. M. L., and Tong, Y. L., "Interval Estimation of the Largest Mean of k Normal Populations with Known Variances," Journal of the American Statistical Association, 64 (1969), 296-9.

7. Dudewicz, E. J., "Two-Sided Confidence Intervals for Ranked Means," Journal of the American Statistical Association, 67 (1972), 462-4.

8. Dudewicz, E. J., "Confidence Intervals for Ranked Means," Naval Research Logistics Quarterly, 17 (1970), 69-78.

9. Saxena, K. M. L., "Interval Estimation of the Largest Variance of k Normal Populations," Journal of the American Statistical Association, 66 (1971), 408-10.

10. Tong, Y. L., "Multi-stage Interval Estimations of the Largest Mean of k Normal Populations," Journal of the Royal Statistical Society, Series B, 32, No. 1 (1970), 272-7.

11. Tong, Y. L., "An Asymptotically Optimal Sequential Procedure for the Estimation of the Largest Mean," Annals of Statistics, 1, No. 1 (1973), 175-9.

12. Alam, K., Saxena, K. M. L. and Tong, Y. L., "Optimal Confidence Interval for a Ranked Parameter," Journal of the American Statistical Association, 68 (1973), 720-4.

13. Gibbons, J. D., Nonparametric Statistical Inference, New York: McGraw-Hill, Inc., 1971. Pp. 106-19.

14. Noether, G. E., "Some Simple Distribution-Free Confidence Intervals for the Center of a Symmetric Distribution," Journal of the American Statistical Association, 68 (1973), 716-9.

15. Tong, Y. L., "Interval estimation of the ordered means of two normal populations based on hybrid estimators," Biometrika, 58 (1971), 605-13.

16. Wilks, S. S., Mathematical Statistics, New York: John Wiley and Sons, Inc., 1962.

17. Walsh, J. E., "Some Bounded Significance Level Properties of the Equal-Tail Sign Test," Annals of Mathematical Statistics, 22 (1951), 408-17.