4-2018

Statistical Properties of Population Stability Index

Bilal Yurdakul
Western Michigan University, bilalyurdakul@gmail.com

STATISTICAL PROPERTIES OF POPULATION STABILITY INDEX

by

Bilal Yurdakul

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Statistics
Western Michigan University
April 2018

Doctoral Committee:

Joshua Naranjo, Ph.D., Chair
Magdalena Niewiadomska-Bugaj, Ph.D.
Clifton E. Ealy, Ph.D.
Jeff Terpstra, Ph.D.

STATISTICAL PROPERTIES OF POPULATION STABILITY INDEX

Bilal Yurdakul, Ph.D.

Western Michigan University, 2018

Population stability is an important concept in model management. It is crucial to monitor whether the current population has changed from the population used during development of a model. For example, has the distribution of credit scores changed, and is the existing credit score model still valid? Population change may occur for many reasons–change in the economic environment, strategic change in the business, policy changes within the company, or changes in regulatory environment.

The population stability index (PSI) is a statistic that measures how much a variable has shifted over time, and is used to monitor applicability of a statistical model to the current population. In banking for example, a high PSI may result in an internal investigation of the reasons behind the change, or an audit by the Federal Reserve Bank. Since banks are heavily regulated by FRB, an unsuitable use of a model means additional risk.

There are not many studies about the statistical properties of PSI. Existing rules of thumb are: PSI < 0.10 means "little shift", $.10<\text{PSI}<.25$ means "moderate shift", and $\text{PSI}>0.25$ means "significant shift, action required". However, these benchmarks are being used without reference to Type I or Type II error rates. This thesis will try to fill the gap by providing statistical properties of PSI and some recommendations for the rules of thumb. —

Acknowledgements

I would like to express my gratitude for all the support I receive during my PhD. studies from my thesis advisor Dr. Joshua Naranjo. His positive comments and invaluable contributions were the main driver for writing this thesis. In addition, I am grateful for his patience as I was working on my dissertation along with my full-time job.

I also would like to thank my committee members Dr. Magdalena Niewiadomska-Bugaj, Dr. Clifton E. Ealy, Dr. Jeff Terpstra for their support and comments.

Lastly, I would like to thank my family for their endless support. In particular, I would like to thank my better half, Meral, for her unending support, encouragement and patience as she was taking care of our sons, Zeyd, Kays and me while I was working on my PhD.

Bilal Yurdakul

Table of Contents

Table of Contents—Continued

List of Tables

List of Figures

Chapter 1

Introduction

Population Stability Index (PSI) is one of the widely used model monitoring metrics which measure the difference between the model development sample and the current sample that the model is used for and therefore it is implemented in several statistical package as in Pruitt (2010). In practice, there is a general rule of thumb: if PSI is less than 10% the model is appropriate, and if PSI is between 10% and 25% then the current sample has to be investigated for reasons of the high PSI. If PSI is beyond 25%, it is highly advised to develop a new model on a more recent sample. "These are industry rules of thumb, and not definitive." is quoted in Siddiqi (2016). PSI mainly calculated for score distributions however it can also be applied to any other variables which exist on both current and development data sources. In this thesis, we will study the distribution of PSI and we will provide some guidelines around the rule of thumb. In addition, PSI will be compared to other statistical methods that are widely used for comparison of distributions.

The following quotes from FRB SR11-7 shows the importance of model monitoring;

> The use of models invariably presents model risk, which is the potential for adverse consequences from decisions based on incorrect or misused model

outputs and reports.

Validity: “The relevance of the data used to build the model should be evaluated to ensure that it is reasonably representative of the market conditions."

Monitoring: "Ongoing monitoring is essential to evaluate...necessary adjustment or replacement of the model."

Similar statements exists in ,(FED, 2011),(OCC, 2016) as well.

1.1 PSI

1.1.1 Notation and Calculation of PSI

PSI is a measure of population stability between two population samples. Consider the population of credit scores for a *base* year, say, 2007. Then consider the population of credit scores in a *target* year, say 2017. We are interested comparing base and target years. Did the distribution of credit scores remain the same, or did it change? PSI is calculated based on the multinomial classification of credit scores into *bins* or categories. For example, credit scores are classified into six bins in Table 1.1, with the highest scores in Grade A and lowest scores in Grade G. The table is based on data freely available from Lending Club. The cutoffs were determined by Lending Club. The population stability index looks at the difference between base and target proportions, and is computed as follows:

$$PSI(Y_b, Y; B) = \sum_{i=1}^{B} (y_i - y_{b_i})(ln(y_i) - ln(y_{b_i})) \tag{1.1}$$

$$= \sum_{i=1}^{B} (y_i - y_{b_i}) ln(y_i / y_{b_i}) \tag{1.2}$$

The $y_1, \ldots, y_B$ are the proportions of target year credit scores that fall in the i_{th} bin and $y_{b_1}, \ldots, y_{b_B}$ are the proportion of base target year credit scores that fall in the i_{th} bin where b stands for base and B represents number of bins. Table (1.1) illustrates the calculation of PSI. Using Equation (1.1), the calculation for the first row is:

$$(.253 - .177) * (ln(.253) - ln(.177) = 0.028 \tag{1.3}$$

and PSI is the sum over the rows.

Table 1.1: Grade Distribution of Credit Scores

Grade	Base	Target	Base - Target	ln(Base)-ln(Target)	Product
A	0.253	0.177	0.077	0.36	0.028
B	0.302	0.262	0.040	0.14	0.006
C	0.204	0.285	-0.081	-0.33	0.027
D	0.134	0.158	-0.024	-0.16	0.004
E	0.072	0.088	-0.016	-0.20	0.003
F	0.026	0.025	0.001	0.05	0.000
G	0.008	0.006	0.002	0.30	0.001
					PSI = 0.068

To establish notation for the theoretical results of Chapter 2, We will introduce a formal definition of PSI based on counts and sample sizes.

Definition 1.1.1. (PSI) Let N be the sample size for base population and M be the sample size for target population. Then PSI can be defined as;

$$PSI = \sum_{i=1}^{B} \left(\frac{n_i}{N} - \frac{m_i}{M}\right) \times \left(\ln \frac{n_i}{N} - \ln \frac{m_i}{M}\right) \tag{1.4}$$

$$= \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i) \times (\ln \hat{p}_i - \ln \hat{q}_i) \tag{1.5}$$

where n_i and m_i's are counts in the i^{th} bin, $\sum n_i = N$, $\sum m_i = M$, $\hat{p}_i = n_i/N$, and $\hat{q}_i = m_i/M$.

Calculation of PSI for *Grade A*

The following example is again based on the same dataset used in Table 1.1 and it will demonstrate how Definition 1.1.1 is used. PSI is calculated between the 2007-2011 dataset and 2015 snapshot for the rating. Since the model specifications are unknown and the development data was not provided, PSI shows that the rating distribution in 2015 is not very different from the loans that are originated between 2007-2011 based on their ratings. The following will exhibit the calculation steps for the first line of the table below.

Table 1.2: PSI calculation for Grade A

n_1	N	m_1	M	n_1/N	m_1/M	$n_1/N - m_1/M$	$ln(n_1/N) - ln(m_1/M)$
10085	39786	32000	181231	0.253	0.177	0.077	0.362

Using last two columns above, we can get to 0.028, PSI for grade A that is;

$$0.077 \times 0.362 = 0.028$$

Note the $181,231$ is the total number of observations in 2015 data and $39,786$ is the number of observations in the development dataset. In addition, number of bin in the table is 7 as in Table (1.1). According to the "rule of thumb" PSI indicates that both population are similar in terms of their rating distribution.

1.2 Literature Review

Most of the papers about population stability index are blog posts, studies about its use in the industry (Siddiqi, 2016),(Pruitt, 2010). There is a patent issued for a machine which does the calculation of PSI (Liu et al., 2009). In addition to these publications, there are also papers and manuals by governmental regulatory bodies. Although

these publications do not directly talk about PSI, they have mentions about measuring stability of population as an ongoing monitoring requirement (FDIC, 2007),(OCC, 2016),(FED, 2011),(FDIC, 2007).However, PSI is the main tool that is used to measure "population stability".

1.2.1 Kullback-Liebler Uncertainty

During literature review, we could not locate any academic work on statistical analysis of PSI as it was named and used in practice. Most of the references only talk about its practical use as mentioned above. In addition there were no discussion around why "rule of thumb" works or PSI's statistical properties. However, we noticed that PSI can be written as some form of so called Kullback-Liebler *divergence* defined in Kullback & Leibler (1951). KL *divergence* is well studied as it can be found in the following references (Wu & Olson, 2010), (Li et al., 2008), (Lin, 2017), (Yousefi et al., 2016), (Gottschalk, 2016). Let $p(x)$ and $q(x)$ be two distributions of a discrete random variable X.

Definition 1.2.1. The Kullback-Liebler divergence of $q(x)$ from $p(x)$ is

$$D_{KL}(q(x)|p(x)) = E_p\left(\ln\frac{p(X)}{q(X)}\right) = \sum_{i=1}^{B} p(x_i)\ln\frac{p(x_i)}{q(x_i)} \tag{1.6}$$

We might think of $p(x)$ as the *true* distribution, and $q(x)$ as a theory or model distribution, so that D_{KL} represents some sort of loss due to using the wrong distribution. Even though D_{KL} measures divergence of $q(x)$ from $p(x)$, it is technically not a distance measure because the definition is not symmetric, i.e. $D_{KL}(q(x)|p(x)) \neq D_{KL}(p(x)|q(x))$. However, we can easily obtain a symmetric measure of divergence by

defining

$$\begin{aligned} D^*(p,q) &= D_{KL}(q|p) + D_{KL}(p|q) \\ &= \sum p(x_i) \ln \tfrac{p(x_i)}{q(x_i)} + \sum q(x_i) \ln \tfrac{q(x_i)}{p(x_i)} \\ &= \sum p(x_i) \ln \tfrac{p(x_i)}{q(x_i)} - \sum q(x_i) \ln \tfrac{p(x_i)}{q(x_i)} \\ &= \sum (p(x_i) - q(x_i)) \ln \tfrac{p(x_i)}{q(x_i)} \\ &= \sum (p(x_i) - q(x_i))(\ln p(x_i) - \ln q(x_i)) \end{aligned}$$

which brings us to the formula for PSI.

Lemma 1.2.1.

$$PSI = D_{KL}(\hat{q}(x)|\hat{p}(x)) + D_{KL}(\hat{p}(x)|\hat{q}(x)) \tag{1.7}$$

Kullback-Liebler risk also has well-known application to entropy, cross-entropy and AIC (Akaike Information Criteria). We will not pursue this topic in this thesis. Although Kullback-Liebler Risk is not symmetric, PSI is symmetric under the assumption that cut-off values are predetermined

$$PSI(X, Y; B) = PSI(Y, X; B) \tag{1.8}$$

for any random variable X,Y binned into B bins. However, in practice that is not true, since the selection of $B-1$ cut-off points as percentiles of base population determines corresponding PSI. If one switches roles; i.e. make base population target population and v.s. then this changes cut-off points and consequently PSI. This is an additional layer of complexity of the use of PSI in practice.

1.2.2 Kullback's PSI Definition

Kullback tackle the problem in two parts for multinomial distributions where you compare cell percentages between two population. He defines the problem for "Single Sample" and "Two Sample" where he points out that the comparison can be done with

one sample against an assumed population percentages versus two sample against a pre-defined population percentages. Kullback defined $J = N \times PSI$ for "Single Sample" and $J = (1/N + 1/M)^{-1} \times PSI$ for two-sample case. His definition was in Chapter 6 of Kullback (1978) and his notation was $J(1, 2)$ which was defined as the sum of so called "information";

$$J(1, 2) = I(1, 2) + I(2, 1) \tag{1.9}$$

He called J, the divergence between H_1 and H_2, two simple statistical hypotheses. H_1 and H_2 were basically assuming different population percentages. Since the underlying distribution is Multinomial, he did not have to consider cut-off values for binning. However, per the current use of PSI, the cell counts or percentages are created from binning a baseline distribution in the development data set and then accordingly B bins are created. The bins are basically dependent on the base population's distribution. In addition, although he talked about PSI's distribution being χ^2, he did not calculate expectation or variance of PSI.

In his book, his main focus was around I, information, his tables were not for J, divergence. However, this thesis will focus on PSI and explore its properties.

Chapter 2

Theoretical Results

2.1 Properties of PSI

PSI compares two distributions $\mathcal{F}$, $\mathcal{G}$ based on a set of cut-off points, $\zeta_1, \zeta_2, ..., \zeta_{B-1}$. These cut-off points will form B intervals and the percentage of samples in corresponding intervals will be calculated from these distributions. Let $\mathcal{X}$ be a sample of size N from $\mathcal{F}$ and $\mathcal{Y}$ be a sample of size M from $\mathcal{G}$. Then we can define x_i and y_i to be the number of observations in the interval $(\zeta_{i-1}, \zeta_i]$ from the sample $\mathcal{X}$ and $\mathcal{Y}$ respectively. Obviously, x_0 and y_0 are the number of observation less than or equal to ζ_1 and x_B and y_B are the number of observation greater than ζ_{B-1}. Consequently,

$$\hat{p}_i = \frac{x_i}{N} \tag{2.1}$$

$$\hat{q}_i = \frac{y_i}{M} \tag{2.2}$$

are the population percentages in the i^{th} bin. Let $\mathcal{X}$ be the base sample and $\mathcal{Y}$ be the target sample for the rest of this chapter.

PSI is a non-negative index as proven in the next theorem. This is one of the direct

implications that PSI having logarithm in its formulation.

Theorem 2.1.1. $PSI = \sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i) \times (\ln(\hat{p}_i) - \ln(\hat{q}_i)) \in [0, \infty)$.

Proof. If $\hat{p}_i > \hat{q}_i$ then $ln(\hat{p}_i/\hat{q}_i) > 0$ and $\hat{p}_i - \hat{q}_i > 0$ therefore their multiplication is greater than 0. If $\hat{p}_i < \hat{q}_i$ then $ln(\hat{p}_i/\hat{q}_i) < 0$ and $\hat{p}_i - \hat{q}_i < 0$ therefore their multiplication is again greater than 0. Since each term is greater than or equal to 0, $PSI \geq 0$. If for any i, $\hat{p}_i = 0$ then $ln(\hat{p}_i/\hat{q}_i) = ln(0) = -\infty$, then PSI becomes ∞. □

Corollary 2.1.1.1. $PSI = 0 \iff \hat{p}_i = \hat{q}_i, \forall i$.

Proof. $\Rightarrow$: Since each term in the sum above is greater than equal to 0, if $PSI = 0$ then each term must be 0. which immediately implies that $\hat{p}_i = \hat{q}_i$ for all i. The other direction is obvious. □

2.2 Expectation of PSI

For the following results, we consider both population percentages as random. In practice, base population percentages considered to be fixed since sometimes PSI is used to measure the divergence of current population from the model development population. We will also provide results for the case when base percentages are fixed as a corollary.

Theorem 2.2.1. *Let* $(x_1, \ldots, x_B)$ *be a multinomial with parameters* N *and* $(p_1, \ldots, p_B)$ *and let* $(y_1, \ldots, y_B)$ *be a multinomial with parameters* M *and* $(q_1, \ldots, q_B)$. *Let* $\hat{p}_i = x_i/N$ *and* $\hat{q}_i = y_i/M$. *Then as* $min(N, M) \to \infty$,

$$\ln \hat{p}_i = \ln(p_i) + (\hat{p}_i - p_i) * \frac{1}{p_i} + \frac{(\hat{p}_i - p_i)^2}{2!} * \frac{-1}{p_i^2} + R_{i1} \tag{2.3}$$

$$\ln \hat{q}_i = \ln(q_i) + (\hat{q}_i - q_i) * \frac{1}{q_i} + \frac{(\hat{q}_i - q_i)^2}{2!} * \frac{-1}{q_i^2} + R_{i2} \tag{2.4}$$

where R_{i1} *is* $O_p(\frac{1}{N^{3/2}})$ *and* R_{i2} *is* $O_p(\frac{1}{M^{3/2}})$.

Proof. From the consistency of $\hat{p}$ and Taylor expansion of $\hat{p}$ around p, it follows that $|R_{i1}| = \left|\frac{(\hat{p}_i - p_i)^3}{3! * c^2}\right|$ for some c between p_i and $\hat{p}_i$. Therefore c is bounded from 0 and $1/c^2$ is bounded. Furthermore, $\sqrt{N}(\hat{p}_i - p_i)$ is $O_p(1)$ since it is asymptotic normal with mean 0 and bounded variance. Therefore $(\hat{p}_i - p_i)$ is $O_p(1/\sqrt{N})$. It follows that $(\hat{p}_i - p_i)^3$ is $O_p(1/N^{3/2})$ and R_{i1} is $O_p(1/N^{3/2})$. Similarly, $|R_{i2}| = \left|\frac{(\hat{q}_i - q_i)^3}{3! * c^2}\right|$ is $O(1/M^{3/2})$. □

Therefore we can approximate $\ln \hat{p}_i$ and $\ln \hat{q}_i$ by

$$\ln \hat{p}_i \sim \ln(p_i) + (\hat{p}_i - p_i) * \frac{1}{p_i} - \frac{(\hat{p}_i - p_i)^2}{2!} * \frac{1}{p_i^2} \tag{2.5}$$

$$\ln \hat{q}_i \sim \ln(q_i) + (\hat{q}_i - q_i) * \frac{1}{q_i} - \frac{(\hat{q}_i - q_i)^2}{2!} * \frac{1}{q_i^2} \tag{2.6}$$

If the above approximation is inserted into PSI then it can be written as the following theorem states:

Theorem 2.2.2. *Under the conditions of Theorem 2.2.1,*

$$PSI = PSI^* + O_p(\frac{1}{N^{3/2}}) + O_p(\frac{1}{M^{3/2}}) \tag{2.7}$$

where

$$PSI^* = \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i)\Big[(\ln p_i - \ln q_i) + \frac{\hat{p}_i - p_i}{p_i} - \frac{\hat{q}_i - q_i}{q_i} - \frac{(\hat{p}_i - p_i)^2}{2\, p_i^2} + \frac{(\hat{q}_i - q_i)^2}{2\, q_i^2}\Big]$$

Proof. By definition,

$$PSI = \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i) \times (\ln \hat{p}_i - \ln \hat{q}_i) \tag{2.8}$$

Use Theorem 2.2.1 to substitute for $\ln \hat{p}_i$ and $\ln \hat{q}_i$ in equation (2.8), we get

$$\begin{aligned} PSI &= \sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)\Big[(\ln p_i - \ln q_i) + \tfrac{\hat{p}_i - p_i}{p_i} - \tfrac{\hat{q}_i - q_i}{q_i} - \tfrac{(\hat{p}_i - p_i)^2}{2\,p_i^2} + \tfrac{(\hat{q}_i - q_i)^2}{2\,q_i^2} + R_{i1} - R_{i2}\Big] \\ &= PSI^* + \sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)R_{i1} + \sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)R_{i2} \end{aligned}$$

Since $(\hat{p}_i - \hat{q}_i) = O_p(1)$ and $R_{i1} = O_p(1/N^{3/2})$, then

$$\sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)R_{i1} = O_p\left(\frac{1}{N^{3/2}}\right)$$

Similarly, $\sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)R_{i2} = O_p(1/M^{3/2})$ and the result follows. □

We provide a simpler approximation for PSI under the null hypothesis.

Theorem 2.2.3. *Under the null hypothesis H_0: $p_i = q_i$, $i = 1...B$ and the conditions of Theorem 2.2.1, PSI can be written as*

$$PSI = PSI^{**} + O_p(\frac{1}{N^{3/2}}) + O_p(\frac{1}{M^{3/2}}) \tag{2.9}$$

where

$$PSI^{**} = \sum_{i=1}^{B}\frac{(\hat{p}_i - \hat{q}_i)^2}{p_i}$$

Proof. Ignoring terms that are $O_p(\frac{1}{N^{3/2}})$ and $O_p(\frac{1}{M^{3/2}})$, recall that

$$PSI \doteq \sum_{i=1}^{B}(\hat{p}_i - \hat{q}_i)\Big[(\ln p_i - \ln q_i) + \frac{\hat{p}_i - p_i}{p_i} - \frac{\hat{q}_i - q_i}{q_i} - \frac{(\hat{p}_i - p_i)^2}{2\,p_i^2} + \frac{(\hat{q}_i - q_i)^2}{2\,q_i^2}\Big]$$

Under the null, note that $(\ln p_i - \ln q_i) = 0$, and

$$\frac{\hat{p}_i - p_i}{p_i} - \frac{\hat{q}_i - q_i}{q_i} = \frac{\hat{p}_i - \hat{q}_i}{p_i}$$

Furthermore, $(\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)^2 = [(\hat{p}_i - p_i) - (\hat{q}_i - p_i)](\hat{p}_i - p_i)^2$ is $O_p(1/N^{3/2})$. Similarly, $(\hat{p}_i - \hat{q}_i)(\hat{q}_i - q_i)^2$ is $O_p(1/M^{3/2})$. The result follows. □

If $\hat{p}_i = 0$ or $\hat{q}_i = 0$ occur with nonzero probability, then PSI takes value infinity

with nonzero probability, and neither $E(PSI)$ nor $\mathrm{Var}(PSI)$ exist. In the remainder of this section, we will derive the mean, variance, and distribution of the more tractable approximations PSI^* and PSI^{**}. The finite sample simulations in Chapter 4 confirm that the distribution results based on the approximations reasonably describe the behavior of PSI.

Theorem 2.2.4. *Under the conditions of Theorem 2.2.1, PSI^* has expected value*

$$E(PSI^*) = \sum_{i=1}^{B}(p_i - q_i)\times(\ln p_i - \ln q_i) + (B-1)(\frac{1}{N}+\frac{1}{M}) + \frac{B - \sum_{i=1}^{B} q_i/p_i}{2N} + \frac{B - \sum_{i=1}^{B} p_i/q_i}{2M}$$

Proof.

$$\begin{aligned} E(PSI^*) = &\sum_{i=1}^{B}(p_i - q_i)(\ln p_i - \ln q_i) \\ &+ \sum_{i=1}^{B}\Big[\frac{E((\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i))}{p_i} - \frac{E((\hat{p}_i - \hat{q}_i)(\hat{q}_i - q_i))}{q_i}\Big] \\ &- \sum_{i=1}^{B}\Big[\frac{E((\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)^2)}{2\,p_i^2} - \frac{E((\hat{p}_i - \hat{q}_i)(\hat{q}_i - q_i)^2)}{2\,q_i^2}\Big] \end{aligned} \tag{2.10}$$

Now, by independence of $\hat{p}_i$ and $\hat{q}_i$, $E(p_i - \hat{q}_i)(\hat{p}_i - p_i) = 0$ so that

$$\frac{E((\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i))}{p_i} = \frac{E((\hat{p}_i - p_i + p_i - \hat{q}_i)(\hat{p}_i - p_i))}{p_i} \tag{2.11}$$

$$= \frac{1}{p_i}\Big[Var(p_i) + E(p_i - \hat{q}_i)(\hat{p}_i - p_i)\Big] \tag{2.12}$$

$$= \frac{1}{p_i}Var(p_i) \tag{2.13}$$

$$= \frac{1}{p_i}\Big[\frac{p_i(1-p_i)}{N}\Big] \tag{2.14}$$

$$= \frac{1-p_i}{N} \tag{2.15}$$

So we have

$$\sum_{i=1}^{B} \frac{E(\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)}{p_i} = \sum_{i=1}^{B} \frac{1 - p_i}{N} = \frac{B-1}{N}$$

Similarly

$$\sum_{i=1}^{B} \frac{E(\hat{p}_i - \hat{q}_i)(\hat{q}_i - q_i)}{q_i} = -\sum_{i=1}^{B} \frac{E(\hat{q}_i - \hat{p}_i)(\hat{q}_i - q_i)}{q_i} = -\frac{B-1}{M} \tag{2.16}$$

Furthermore,

$$\begin{aligned} \frac{E(\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)^2)}{2\ p_i^2} &= \frac{E((\hat{p}_i - p_i) + (\hat{q}_i - q_i) + (p_i - q_i))(\hat{p}_i - p_i)^2)}{2\ p_i^2} \\ &= \frac{E(\hat{p}_i - p_i)^3 + E(\hat{q}_i - q_i)(\hat{p}_i - p_i)^2 + E(p_i - q_i)(\hat{p}_i - p_i)^2}{2\ p_i^2} \end{aligned} \tag{2.17}$$

The third central moment of a binomial is $Np(1-p)(1-2p)$ so the first term in the numerator is $E(\hat{p}_i - p_i)^3 = Np(1-p)(1-2p)/N^3$ which is $O(1/N^2)$. The second term $E(\hat{q}_i - q_i)(\hat{p}_i - p_i)^2$ is 0 because of independence. The third term is $E(p_i - q_i)(\hat{p}_i - p_i)^2 = (p_i - q_i)Var(\hat{p}_i) = (p_i - q_i)p_i(1-p_i)/N$, so that ignoring the $O_p(N^{-2})$ terms we have

$$\frac{E(\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)^2)}{2\ p_i^2} = \frac{(p_i - q_i)p_i(1-p_i)}{2p_i^2 N} = \frac{(p_i - q_i)(1-p_i)}{2p_i N}$$

so that

$$\begin{aligned} \sum_{i=1}^{B} \frac{E(\hat{p}_i - \hat{q}_i)(\hat{p}_i - p_i)^2)}{2\ p_i^2} &= \sum_{i=1}^{B} \frac{(p_i - q_i)(1-p_i)}{2p_i N} = \sum_{i=1}^{B} \frac{(p_i - q_i - p_i^2 + q_i p_i)}{2p_i N} \\ &= \frac{B - \sum_{i=1}^{B} q_i/p_i - 1 + 1}{2N} = \frac{B - \sum_{i=1}^{B} q_i/p_i}{2N} \end{aligned}$$

Similarly,

$$\sum_{i=1}^{B} \frac{E(\hat{p}_i - \hat{q}_i)(\hat{q}_i - q_i)^2)}{2\, p_i^2} = -\sum_{i=1}^{B} \frac{E(\hat{q}_i - \hat{p}_i)(\hat{q}_i - q_i)^2)}{2\, p_i^2} = -\frac{B - \sum_{i=1}^{B} p_i/q_i}{2M}$$

Therefore, we have

$$E(PSI^*) = \sum_{i=1}^{B}(p_i - q_i)(\ln p_i - \ln q_i) + \frac{B-1}{N} + \frac{B-1}{M} + \frac{B - \sum_{i=1}^{B} q_i/p_i}{2N} + \frac{B - \sum_{i=1}^{B} p_i/q_i}{2M}$$

□

Corollary 2.2.4.1. *Under H_0: $p_i = q_i$, $i = 1...B$,*

$$E(PSI^*) = (B-1)\left(\frac{1}{N} + \frac{1}{M}\right) \tag{2.18}$$

Proof. Under H_0, the first term is 0. Also $B - \sum_{i=1}^{B} q_i/p_i = 0$. The result follows. □

2.3 Variance of PSI^{**}

In this section, we will extract the variance of PSI^{**} under H_0. We will start with a reminder of a general result.

Theorem 2.3.1. *Let $E(Y) = \mu$ and $Cov(Y) = \Sigma$ then $Var(Y'AY) = 2tr(A\Sigma A\Sigma) + 4\mu' A\Sigma A\mu$.*

Proof of this theorem can be found, for example, in Searle's 1971 book Linear Model, page 57.

Theorem 2.3.2. *Under the conditions of Theorem 2.2.1, and assuming the null hypothesis H_0: $p_i = q_i$, $i = 1...B$ is true, then*

$$Var(PSI^{**}) = 2(\frac{1}{N} + \frac{1}{M})^2(B-1)$$

Proof. Let

$$Y = \begin{bmatrix} \hat{p}_1 - \hat{q}_1 \\ \hat{p}_2 - \hat{q}_2 \\ \vdots \\ \hat{p}_B - \hat{q}_B \end{bmatrix} \tag{2.19}$$

then $\mu = E(Y) = 0$, therefore Theorem 2.3.1 implies that $Var(Y'AY) = 2tr(A\Sigma A\Sigma)$. Recall that

$$PSI^{**} = \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i)^2 \frac{1}{p_i} \tag{2.20}$$

$$= Y^T A Y \tag{2.21}$$

where

$$A = \begin{bmatrix} \frac{1}{p_1} & 0 & \dots & 0 \\ 0 & \frac{1}{p_2} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \frac{1}{p_B} \end{bmatrix} \tag{2.22}$$

To get Σ, we will calculate the following:

$$Var(\hat{p}_i - \hat{q}_i) = Var(\hat{p}_i) + Var(\hat{q}_i) \tag{2.23}$$

$$= \frac{p_i(1-p_i)}{N} + \frac{q_i(1-q_i)}{M} \tag{2.24}$$

$$= p_i(1-p_i)(\frac{1}{N} + \frac{1}{M}) \text{ under } H_0 : p_i = q_i \tag{2.25}$$

$$Cov(\hat{p}_i - \hat{q}_i, \hat{p}_j - \hat{q}_j) = Cov(\hat{p}_i, \hat{p}_j) - Cov(\hat{p}_i, \hat{q}_j) - Cov(\hat{p}_j, \hat{q}_i) + Cov(\hat{q}_i, \hat{q}_j) \tag{2.26}$$

$$= -\frac{p_i p_j}{N} + 0 + 0 - \frac{q_i q_j}{M} \tag{2.27}$$

$$= -p_i p_j \times [\frac{1}{N} + \frac{1}{M}] \text{ under } H_0 : p_i = q_i \tag{2.28}$$

$$A = \begin{bmatrix} \frac{1}{p_1} & 0 & \dots & 0 \\ 0 & \frac{1}{p_2} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \frac{1}{p_B} \end{bmatrix} \tag{2.29}$$

$$A\Sigma = (\frac{1}{N} + \frac{1}{M}) \begin{bmatrix} \frac{1}{p_1} & 0 & \dots & 0 \\ 0 & \frac{1}{p_2} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \frac{1}{p_B} \end{bmatrix} \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \dots & -p_1 p_B \\ -p_2 p_1 & p_2(1-p_2) & \dots & -p_2 p_B \\ \vdots & & \ddots & \\ -p_B p_1 & -p_B p_2 & \dots & p_B(1-p_B) \end{bmatrix} \tag{2.30}$$

$$= (\frac{1}{N} + \frac{1}{M}) \begin{bmatrix} 1-p_1 & -p_2 & \dots & -p_B \\ -p_1 & 1-p_2 & \dots & -p_B \\ \vdots & & \ddots & \\ -p_1 & -p_2 & \dots & 1-p_B \end{bmatrix} \tag{2.31}$$

Note that

$$row_i \bullet col_i = (1-p_i)^2 + p_i p_1 + .. + p_i p_{i-1} + p_i p_{i+1} + .. + p_i p_B \tag{2.32}$$

$$= (1-p_i)^2 + p_i(p_1 + .. + p_{i-1} + p_{i+1} .. + p_B) \tag{2.33}$$

$$= (1-p_i)^2 + p_i(1-p_i) \tag{2.34}$$

$$= 1 - p_i \tag{2.35}$$

Since the trace of a matrix is $\sum_{i=1}^{B} row_i \bullet col_i$ then

$$tr(A\Sigma A\Sigma) = (\frac{1}{N} + \frac{1}{M})^2 \sum_{i=1}^{B}(1 - p_i) \tag{2.36}$$

$$= (\frac{1}{N} + \frac{1}{M})^2 (B - 1) \tag{2.37}$$

Therefore,

$$Var(PSI^{**}) = 2(\frac{1}{N} + \frac{1}{M})^2 (B - 1) \tag{2.38}$$

□

2.4 PSI^{**} has an Approximate χ^2_{B-1} Distribution

Here we show that PSI^{**} is proportional to a chi square random variable. More specifically, it behaves like $(1/N + 1/M)\chi^2_{B-1}$.(Stapleton, 2009)

Theorem 2.4.1. *Under the conditions of Theorem 2.2.1, and assuming the null hypothesis H_0: $p_i = q_i$, $i = 1...B$ is true, then*

$$\left(\frac{1}{N} + \frac{1}{M}\right)^{-1} PSI^{**}$$

has an approximate χ^2 distribution with B-1 degrees of freedom.

Proof. Recall that

$$\begin{aligned} PSI &\sim \textstyle\sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i)^2 \frac{1}{p_i} \\ &= Y^T A Y \end{aligned}$$

where

$$Y = \begin{bmatrix} \hat{p}_1 - \hat{q}_1 \\ \hat{p}_2 - \hat{q}_2 \\ \vdots \\ \hat{p}_B - \hat{q}_B \end{bmatrix}$$

and

$$A = \begin{bmatrix} \frac{1}{p_1} & 0 & \dots & 0 \\ 0 & \frac{1}{p_2} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \frac{1}{p_B} \end{bmatrix}$$

Then $\left(\frac{1}{N} + \frac{1}{M}\right)^{-1} PSI^{**} = Y^T A^* Y$ where

$$A^* = \left(\frac{1}{N} + \frac{1}{M}\right)^{-1} \begin{bmatrix} \frac{1}{p_1} & 0 & \dots & 0 \\ 0 & \frac{1}{p_2} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \frac{1}{p_B} \end{bmatrix}$$

Since Y is (approximately) multivariate normal, then $Y' A^* Y$ is distributed as χ^2_{B-1} if $A^*\Sigma$ is idempotent (see for example Stapleton, 2009). Therefore we just need to show that $A^*\Sigma$ is idempotent, where

$$\Sigma = (\frac{1}{N} + \frac{1}{M}) \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \dots & -p_1 p_B \\ -p_2 p_1 & p_2(1-p_2) & \dots & -p_2 p_B \\ \vdots & & \ddots & \\ -p_B p_1 & -p_B p_2 & \dots & p_B(1-p_B) \end{bmatrix}$$

Now

$$A^*\Sigma = \begin{bmatrix} 1-p_1 & -p_2 & \dots & -p_B \\ -p_1 & 1-p_2 & \dots & -p_B \\ \vdots & & \ddots & \\ -p_1 & -p_2 & \dots & 1-p_B \end{bmatrix} \tag{2.39}$$

We need to show that $A^*\Sigma A^*\Sigma = A^*\Sigma$, or that the ij^{th} entry of both sides are the same. We have already shown in (2.32) that diagonal terms are the same. We will complete the proof by showing the off-diagonal terms are the same. The i^{th} row of $A^*\Sigma$ is

$$R_i = \begin{pmatrix} -p_1 & -p_2 & \dots (1-p_i) & \dots & -p_B \end{pmatrix} \tag{2.40}$$

and j^{th} column of $A^*\Sigma$ is

$$C_j = \begin{pmatrix} -p_j & -p_j & \dots (1-p_j) \cdots - p_j \end{pmatrix}^T$$

where $1-p_j$ is the j^{th} entry. By rewriting C_j as follows;

$$C_j^T = \begin{pmatrix} -p_j & -p_j & \dots (1-p_j) & \dots & -p_j \end{pmatrix} \tag{2.41}$$

$$= -p_j \begin{pmatrix} 1 & 1 & \dots 1 & \dots & 1 \end{pmatrix} \tag{2.42}$$

$$+ \begin{pmatrix} 0 & 0 & \dots & 1 & \dots & 0 \end{pmatrix} \tag{2.43}$$

then we will have the following;

$$R_i \bullet C_j = -p_j \sum_{j=1}^{B} R_{ij} + (-p_j) \tag{2.44}$$

$$= -p_j \tag{2.45}$$

since sum of each row of $A^*\Sigma$ is 0. So each entry of j^{th} column will be $-p_j$ except at the diagonal, and that is the same as ij^{th} entry of $A * \Sigma$. □

2.5 A Note for Expectation and Variance

In practice, the base year bin percentages are treated as population percentages to which target year percentages are compared, e.g. by some goodness-of-fit test. In other words $\{\hat{p}_i\}$ are treated as population parameters $\{p_i\}$ and only $\{\hat{q}_i\}$ are treated as random. We will refer to this in the succeeding discussions as *the One-Sample problem.* Then the results for this case is as follows:

Conjecture 2.5.1. *If base population percentages are fixed, then expectation of PSI^{**} is*

$$E(PSI^{**}) \approx \frac{B-1}{M} \tag{2.46}$$

and variance is

$$V(PSI^{**}) \approx \frac{2(B-1)}{M^2} \tag{2.47}$$

and consequently deviation is

$$SD(PSI^{**}) \approx \frac{\sqrt{2(B-1)}}{M} \tag{2.48}$$

where M is the sample size for the target population and B is the number of bins.

Heuristic Proof. The result is a direct implication of the previous theorems. If we follow the proof of theorems (2.2.4) and (2.3.2) and remove terms regarding $\hat{p}$ since $\hat{p}$ is not random and consequently the variance of $\hat{p}$ is 0. It implies that the terms with $1/N$ vanishes and results the variance and expectation above. □

Chapter 3

Proposed Tests and Alternative Methods

In this chapter, we will propose tests based on the results given in Chapter 2. We will compare PSI against some other well known statistical tests for distributional change. Statistical tests will briefly be explained and then simulation results will be presented in the next chapter.(Cochran, 1977)

3.1 Proposed Tests

Instead of using fixed critical values of PSI that are 10% and 25%, I propose to use 95^{th}, 99^{th} or 99.9^{th} theoretical percentiles of PSI based on either percentiles of normal or χ^2_{B-1} distribution.

3.1.1 Method I

In Theorems 2.2.4 and 2.3.2, both expectation and variance of PSI is given. Based on these results, we can construct the following test;

$$PSI > (1/N + 1/M)(B-1) + z_\alpha(1/N + 1/M) \times \sqrt{2(B-1)} \tag{3.1}$$

where RHS is the critical value for PSI** based on normal approximation and α can be 95^{th},99^{th} or 99.9^{th} percentiles.

3.1.2 Method II

In Chapter 2, we proved that $(1/N + 1/M)^{-1}PSI^{**}$ has a χ^2_{B-1} distribution. Based on theorem 2.4.1, we can do the the test;

$$PSI^{**} > \chi^2_{\alpha,B-1} \times (1/N + 1/M) \tag{3.2}$$

where RHS is the critical value for PSI** based on χ^2_{B-1} and α can be 95^{th},99^{th} or 99.9^{th} percentiles.

3.1.3 Tables of Method I and Method II

The tables below shows percentiles of proposed tests. Although 10% and 25% cut-offs are fixed, as sample sizes and B changes percentiles of the corresponding tests changes as well. So, it is important to take those parameters into consideration as we apply the test. There are more tables in Appendix B for quick reference however everything is reproducible using expressions in Method I and Method II.

Table 3.1: N:Size of Base Sample
M:Size of Target Sample

Method I: P_{95} of Normal Approximation of PSI, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	32.0%	24.0%	20.0%	18.6%	18.0%	17.6%
200	24.0%	16.0%	12.0%	10.7%	10.0%	9.6%
400	20.0%	12.0%	8.0%	6.7%	6.0%	5.6%
600	18.6%	10.7%	6.7%	5.3%	4.7%	4.3%
800	18.0%	10.0%	6.0%	4.7%	4.0%	3.6%
1000	17.6%	9.6%	5.6%	4.3%	3.6%	3.2%

Table 3.2: N:Size of Base Sample
M:Size of Target Sample

Method II: P_{95} of χ^2_{B-1}, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	33.8%	25.4%	21.1%	19.7%	19.0%	18.6%
200	25.4%	16.9%	12.7%	11.3%	10.6%	10.2%
400	21.1%	12.7%	8.5%	7.0%	6.3%	5.9%
600	19.7%	11.3%	7.0%	5.6%	4.9%	4.5%
800	19.0%	10.6%	6.3%	4.9%	4.2%	3.8%
1000	18.6%	10.2%	5.9%	4.5%	3.8%	3.4%

Table 3.3: N:Size of Base Sample
M:Size of Target Sample

Method I: P_{95} of Normal Approximation of PSI, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	58.3%	43.7%	36.4%	34.0%	32.8%	32.1%
200	43.7%	29.1%	21.9%	19.4%	18.2%	17.5%
400	36.4%	21.9%	14.6%	12.1%	10.9%	10.2%
600	34.0%	19.4%	12.1%	9.7%	8.5%	7.8%
800	32.8%	18.2%	10.9%	8.5%	7.3%	6.6%
1000	32.1%	17.5%	10.2%	7.8%	6.6%	5.8%

Table 3.4: N:Size of Base Sample
M:Size of Target Sample

Method II: P_{95} of χ^2_{B-1}, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	60.3%	45.2%	37.7%	35.2%	33.9%	33.2%
200	45.2%	30.1%	22.6%	20.1%	18.8%	18.1%
400	37.7%	22.6%	15.1%	12.6%	11.3%	10.6%
600	35.2%	20.1%	12.6%	10.0%	8.8%	8.0%
800	33.9%	18.8%	11.3%	8.8%	7.5%	6.8%
1000	33.2%	18.1%	10.6%	8.0%	6.8%	6.0%

3.2 χ^2 Goodness of Fit Test

Pearson's Goodness of fit (GOF) test is based on the statistic,

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_j - E_j)^2}{E_j} \tag{3.3}$$

where O_j are the observed bin counts for the target year and $E_j = Mp_j$, where p_j is the base year percentage of j^{th} bin. Base year cell frequencies $\{p_j\}$ are treated as fixed parameters. For the data in Table 3.5, the goodness-of-fit test statistic is

$$\chi^2_G = \frac{(18-24)^2}{24} + \frac{(26-18)^2}{18} + \cdots + \frac{(15-20)^2}{20} = 7.09 \tag{3.4}$$

Note that the expected frequencies depend only on base year frequencies, in contrast to the test of homogeneity that we discuss next.

Table 3.5: χ^2 Goodness of fit (expected counts depend on base year frequencies)

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Row Total
base	24	18	16	22	20	100
target	18	26	15	26	15	100
Total	42	44	31	48	35	200

3.3 χ^2 Test of Homogeneity (TOH)

Pearson's Test of Homogeneity is based on the statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{B}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where O_{ij} are the observed bin counts for the ij^{th} cell (see, e.g. Table 3.5). The expected values E_{ij} are calculated from a $B\times 2$ contingency table according to $E_{ij}=\frac{R_i * C_j}{T}$. Here, both base year and target year frequencies are considered random.

For the data in Table 3.5, the expected values are given in Table 3.6. The chi square statistic is calculated as

$$\begin{aligned}\chi^2_H &= \tfrac{(24-21)^2}{21}+\tfrac{(18-22)^2}{22}+\cdots+\tfrac{(20-17.5)^2}{17.5}\\ &\quad+\tfrac{(18-21)^2}{21}+\tfrac{(26-22)^2}{22}+\cdots+\tfrac{(15-17.5)^2}{17.5}\\ &= 3.39\end{aligned} \tag{3.5}$$

Table 3.6: Table of expected values E_{ij} forχ^2 test of homogeneity

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Row Total
base	21	22	15.5	24	17.5	100
target	21	22	15.5	24	17.5	100
Total	42	44	31	48	35	200

Chapter 4

Simulation Studies

In this chapter, we compare the various tests for $H_0 : p_i = q_i$, $i = 1, \ldots, B$ so we are testing whether base and target populations have the same bin probabilities.

4.1 Simulation Methodology

In order to create bins and corresponding cell counts, we have created samples of sizes 100, 400, and 1600 from multinomial, and normal distributions. To mimic credit score distributions, we have simulated from a normal distribution with mean 700 and deviation 100 (Dornhelm (2015)).

In addition to sample size, we examined PSI for number of bins 5, 10, 15, 20, and 25. Study of higher number of bins possible but it is worth to note that in practice, 10 or 20 bins are commonly used.

Table 4.1 shows an example. We generated a sample of 100 observations from $N(\mu = 700, \sigma = 100)$ for base year (column 1) and another 100 observations for target year (column 2). The bins are formed by taking 20^{th},40^{th},60^{th} and 80^{th} percentiles of the generating distribution $N(\mu = 700, \sigma = 100)$. The bin upper boundaries are given in column 3, so all observations less than 615.84 go into bin 1. The resulting bin

frequencies are presented in columns 4 and 5. The calculations for PSI are carried out in the remaining columns, resulting in PSI=8.07%. Since this is less than the traditional benchmark of 10%, then H_0 is not rejected. Base and Target year frequencies are not significantly different.

Here is a summary of simulation parameters:

- Number of bins: $10, 20$
- Sample size (N, M): $(100, 100), (400, 400), (1600, 1600)$
- Mean credit scores: $\mu_{base} = 700$, $\mu_{target} = (700, 695, 690, 685, 680)$

For binning the credit scores, we consider three simulation cases.

- Simulation A: One-sample case
    1. In this set up, we do not generate base year credit scores. Instead, we fix them at uniform frequencies. See Step 4.
    2. Generate N target year observations from $N(\mu, \sigma = 100)$.
    3. Create bin boundaries as percentiles of the true null distribution $N(700, 100)$.
    4. Create uniform base year bin frequencies. For example, when Bin=10, and N=400, each base year bin will have count equal to 40.
    5. Calculate target year bin frequencies according to bin boundaries in Step 3.
    6. Calculate PSI and chisquare tests.
- Simulation B: Two-sample with fixed bin boundaries
    1. Generate N base year observations from $N(700, 100)$.
    2. Generate N target year observations from $N(\mu, \sigma = 100)$.

3. Create bin boundaries as percentiles of the true null distribution $N(700, 100)$.
4. Calculate base year bin frequencies according to bin boundaries in Step 3.
5. Calculate target year bin frequencies according to bin boundaries in Step 3.
6. Calculate PSI and chi square tests.

- Simulation C: Two-sample with random bin boundaries

1. Generate N base year observations from $N(700, 100)$.
2. Generate N target year observations from $N(\mu, \sigma = 100)$.
3. Create bin boundaries as percentiles of the generated base sample. For example when Bin=10, the bin boundaries are the 10th, 20th, 30th,..., and 90th percentiles *of the generated base sample*.
4. Calculate base year bin frequencies according to bin boundaries in Step 3.
5. Calculate target year bin frequencies according to bin boundaries in Step 3.
6. Calculate PSI and chi square tests.

Table 4.1: Example of data generation, binning, and calculation of PSI

Base(B)[a]	Target(T)[a]	Bin boundaries[b]	B#	T#	$\hat{p}_i$	$\hat{q}_i$	$(\hat{p}_i - \hat{q}_i)(\ln\hat{p}_i - \ln\hat{q}_i)$
649.78	666.71	(000.00, 615.84)	18	11	0.18	0.11	0.0345
713.15	836.31	(615.84, 674.67)	20	28	0.20	0.28	0.0269
692.11	653.09	(674,67, 725.33)	28	27	0.28	0.27	0.0004
788.68	784.29	(725.33, 784.16)	15	19	0.15	0.19	0.0095
711.70	554.20	(784.16, ∞)	19	15	0.19	0.15	0.0095
731.86	659.97	Total	100	100	1.00	1.00	PSI=0.0807
641.82	622.36						
771.45	663.07						
617.47	824.01						
664.01	689.26						
708.99	717.26						
709.63	725.46						
679.84	638.55						
773.98	557.08						
712.34	666.90						
697.07	712.84						
661.11	801.81						
751.09	674.44						
608.62	669.75						
931.03	861.52						
$\vdots$	$\vdots$						
610.70	795.39						
584.24	673.40						
646.97	889.53						
944.57	657.00						
616.75	857.55						
741.35	716.19						
582.13	591.45						
582.60	757.69						

[a] Samples are generated from $N(\mu = 700, \sigma = 100)$ size 100.
[b] Bin boundaries are 20^{th},40^{th},60^{th}, and 80^{th} percentiles of $N(\mu = 700, \sigma = 100)$.

4.2 Simulation Results

Table 4.2 shows rejection rates of PSI, χ^2 Goodness of Fit and Homogeneity test for the setup of Simulation A. In this one-sample study, the base percentages are not generated at all, but fixed. For example, when Bin=10, then the p_i's are all set at 10%. The target population is generated from $N(\mu, \sigma = 100)$ and binned using percentiles of $N(\mu = 700, \sigma = 100)$, where the binning percentiles match the chosen base percentages p_i. Since this is a true goodness-of-fit application, the χ^2 GOF does very well here, with size close to 5% and increasing power as μ moves away from 700. The χ^2 homogeneity test is not applicable here because there is only one random sample. The traditional benchmarks PSI>10% and PSI>25% are either too liberal (when N=100) or too conservative (when N=400 or 1600). The theory in Chapter 2 explains this. According to equations (2.46) and (2.48), PSI has expected value $(B-1)/N$ and standard deviation $\sqrt{2(B-1)}/N$. For $B = 10$, the expected value and standard deviation of PSI are

N	E(PSI)	SD(PSI)
100	0.0900	0.0424
400	0.0225	0.0106
1600	0.0056	0.0026

which agrees with the simulations, i.e. PSI>.10 has 35% rejection rate for N=100, and virtually 0% rejection rate for N=400 or 1600.

As we have discussed in Chapter 2, PSI has an approximate $(1/N + 1/M)\chi^2_{B-1}$ or $\frac{1}{M}\chi^2_{B-1}$ depending on the application of PSI. Using this information, we also include rejection rate under H_0 on the 8^{th} column of Table 4.2. The rejection rates stay at around 50 per 1000, as expected.

However, on the 4^{th} row of the Table 4.2, when there are 20 bins and sample size is 100, one can notice that rejection rates are not correct when using 95^{th} percentile of

χ^2_{B-1}. When we looked into that closely, in many cases, PSI was ∞. This is due 0's in one of the bins either in base or target. When infinities were not counted, the number of rejections become 56, which is again the correct number of rejections. This suggests that sample size 100 is not enough for 20 bins. The table 4.3 shows rejection rates by number of bins. This table is created with sample size 100 where both populations are simulated from $N(\mu = 700, \sigma = 100)$. The 2^{nd} column shows rejection rate with infinities. After $Bin = 10$, the rejection rate deviates from 5% more and more; and also the number of infinities increases by number of bins. So, when the cell counts low, PSI is failing. We can conclude from this table that on average at least 10 observations are needed for a stable PSI calculation. The same is true for χ^2 tests as well.

Table 4.2 shows rejection rates of PSI, χ^2 GOF test, χ^2 Test of Homogeneity also compared to both 95^{th} percentile of $N\Big(\mu = \frac{B-1}{M}, \sigma^2 = \frac{2(B-1)}{M^2}\Big)$ and χ^2_{B-1} distributions. Base is simulated from $N(\mu = 700, \sigma = 100)$. The table shows results from simulating 1000 times. Since the p_i are fixed, only random quantity is q_i. This is also confirmed by the rejection rates from χ^2 test; whereas GOF tests give correct rejection rates, TOH fails, that is due to that TOH assumes two random population percentages. In practice, this distinction is very important as using incorrect version causes to use the incorrect distribution which will yield wrong rejections or non-rejections. So, if one calculates percentages of the base population and fixed that, then one needs to use the distribution above which are only depending on number of bin and sample size of the target population.

In Table 4.2, we included 95^{th} percentile of Normal distribution as well. Although, PSI has a χ^2_{B-1} distribution, normal distribution has a close rejection rate.

Table 4.2: Simulation A: One-sample (rejection rates per 1000 when base year frequencies are fixed[a] and uniform)

Bin	Sample Size (N)	Target Mean[1]	PSI>10%	PSI>25%	χ^2_G†	PSI>$\chi_{.95}$⋆	PSI>$Z_{.95}$∗
10	100	700	355	9	49	66	59
10	400	700	0	0	53	57	51
10	1600	700	0	0	60	62	53
20	100	700	971	306	41	166	145
20	400	700	6	0	53	68	56
20	1600	700	0	0	57	56	46
10	100	695	408	11	51	77	70
10	400	695	0	0	92	97	89
10	1600	695	0	0	215	218	204
20	100	695	969	313	58	185	178
20	400	695	6	0	75	94	73
20	1600	695	0	0	164	180	150
10	100	690	459	13	83	118	108
10	400	690	0	0	220	228	210
10	1600	690	0	0	792	797	779
20	100	690	978	350	66	214	196
20	400	690	41	0	146	177	145
20	1600	690	0	0	659	665	624
10	100	685	569	24	127	158	146
10	400	685	4	0	479	482	466
10	1600	685	0	0	995	996	994
20	100	685	984	404	90	252	232
20	400	685	101	0	352	385	348
20	1600	685	0	0	981	982	978
10	100	680	660	54	204	238	225
10	400	680	56	0	778	787	772
10	1600	680	0	0	1000	1000	1000
20	100	680	987	489	148	313	290
20	400	680	284	0	630	663	623
20	1600	680	1	0	1000	1000	1000

[a]For example when Bin=10, and N=400, each base year bin will have count equal to 40.

†Chi-square goodness-of-fit test

⋆ Chi-square approximation of PSI critical value

∗ Normal approximation of PSI critical value

Table 4.3: Rejection Rate by Bin and ∞ occurrences

Bin	Rejection Rate	Rejection without ∞
3	5.7%	5.7%
4	6.8%	6.8%
5	5.0%	5.0%
6	5.1%	5.1%
7	5.1%	5.1%
8	5.1%	5.1%
9	6.1%	6.1%
10	5.9%	5.9%
11	7.0%	7.0%
12	7.5%	7.2%
13	8.2%	7.8%
14	8.4%	7.5%
15	9.4%	7.3%
16	8.3%	6.1%
17	9.8%	5.9%
18	12.6%	6.4%
19	12.9%	4.6%
20	14.5%	3.5%
21	20.8%	4.4%
22	24.0%	3.0%
23	28.8%	3.0%
24	32.6%	1.7%
25	39.8%	1.4%

Third column is the rejection rate (# of Rejections without infinities)/1000 (infinities were not removed from the denominator)

However, if one calculates both base and target population's percentages dynamically by sampling from those, then it is appropriate to use the $(\frac{1}{N}+\frac{1}{M})\chi^2_{B-1}$ to compare with the calculated PSI. The assumption that the base population also random brings the additional $1/N$ term.

Table(4.4) and Table(4.5) shows the result of our simulation of 1000 runs. For the simulation, we used $N(\mu = 700, \sigma = 100)$ for base and for target we used μ 700,695,690,685, and 680. In this case, both GOF and TOH χ^2 tests rejects at around 5% for H_0. We observe similar issue when Bin=20 and sample size is 100, that 95%χ^2 rejects at a higher rate which we explained the reason earlier. As target mean goes further away, the rejection rates increase as expected. However, "the rule of thumb" rejection rates drop quickly as sample size increases. Table (4.4) and Table (4.5) has a slight difference in terms of sampling. For Table (4.4), we fixed the bins from the base distribution, whereas for Table (4.5), the bins are created from the base sample itself, not from the true base distribution. The chi-square goodness of fit test performs badly in the both cases, since GOF is a one sample test whereas we have two samples for both of these tables. In other words, for Table (4.4) we use true distribution percentiles so p_i's and q_i's keep changing so it is two-sample problem. For Table 4.5 p_i's are fixed and uniform but bin boundaries are random, then q_i's are determined based on those bin boundaries. So p_i's are q_i's are based on two random samples. Therefore, a one sample test GOF, fails to reject at much higher rate.

Table 4.4: Simulation B: Two-sample with fixed bin boundaries (rejection rates per 1000 when bins are based on percentiles of true distribution)

Bin	Sample Size	Target Mean[a]	PSI>.10	PSI>.25	χ^2_G†	χ^2_H‡	PSI>$\chi^2_{.95}$⋆	PSI>$Z_{.95}$∗
10	100	700	856	220	556	45	68	62
10	400	700	18	0	508	48	53	46
10	1600	700	0	0	509	50	51	44
20	100	700	1000	902	847	35	258	250
20	400	700	427	1	767	43	62	45
20	1600	700	0	0	735	47	49	42
10	100	695	869	225	559	54	79	75
10	400	695	26	0	550	57	67	56
10	1600	695	0	0	663	125	127	115
20	100	695	998	918	859	47	277	270
20	400	695	436	1	776	62	79	64
20	1600	695	0	0	821	101	102	92
10	100	690	867	265	597	50	87	76
10	400	690	58	0	664	122	132	117
10	1600	690	0	0	911	437	441	418
20	100	690	999	944	892	45	306	288
20	400	690	526	1	841	89	117	93
20	1600	690	0	0	946	320	332	296
10	100	685	891	302	634	71	103	92
10	400	685	148	0	804	240	246	236
10	1600	685	0	0	996	848	852	839
20	100	685	1000	946	904	66	332	320
20	400	685	647	5	888	169	195	173
20	1600	685	0	0	993	727	740	705
10	100	680	909	353	678	95	137	123
10	400	680	290	0	902	430	441	428
10	1600	680	2	0	1000	988	988	987
20	100	680	1000	939	893	77	359	343
20	400	680	819	10	947	288	340	304
20	1600	680	21	0	1000	952	958	944

[a] Base is $N(700, 100)$, target is $N(\mu, 100)$. Bin boundaries are percentiles of $N(700, 100)$.

† Chi-square goodness-of-fit test

‡ Chi-square test of homogeneity

⋆ Chi-square approximation of PSI critical value

∗ Normal approximation of PSI critical value

Table 4.5: Simulation C: Two-sample with random bin boundaries (rejection rates per 1000 when bins are based on percentiles of generated base sample)[a] (Simulation C)

Bin	Sample Size	Target Mean[b]	PSI>.10	PSI>.25	χ^2_G†	χ^2_H‡	PSI>$\chi^2_{.95}$⋆	PSI>$Z_{.95}$∗
10	100	700	826	232	440	40	70	62
10	400	700	29	0	477	61	66	61
10	1600	700	0	0	504	54	59	49
20	100	700	999	896	628	18	454	454
20	400	700	435	0	694	48	65	52
20	1600	700	0	0	721	46	54	37
10	100	695	833	235	452	44	72	68
10	400	695	30	0	504	70	77	70
10	1600	695	0	0	660	131	135	117
20	100	695	1000	906	616	30	432	431
20	400	695	430	1	715	57	73	62
20	1600	695	0	0	813	89	96	76
10	100	690	861	246	467	52	85	76
10	400	690	51	0	620	118	131	116
10	1600	690	0	0	912	442	443	415
20	100	690	999	900	658	25	449	449
20	400	690	520	3	780	98	120	100
20	1600	690	0	0	940	339	344	316
10	100	685	885	265	524	55	96	85
10	400	685	130	0	782	226	244	225
10	1600	685	0	0	996	853	854	842
20	100	685	1000	921	717	30	480	480
20	400	685	640	5	879	169	214	171
20	1600	685	3	0	995	720	730	694
10	100	680	896	329	605	92	137	129
10	400	680	277	2	886	405	421	404
10	1600	680	2	0	1000	990	991	989
20	100	680	1000	926	765	48	498	489
20	400	680	799	13	938	282	337	294
20	1600	680	19	0	1000	956	956	953

[a]Base is $N(700, 100)$. Bin boundaries are percentiles of generated sample.
[b]Target is $N(\mu, 100)$.
†Chi-square goodness-of-fit test
‡Chi-square test of homogeneity
⋆ Chi-square approximation of PSI critical value
∗ Normal approximation of PSI critical value

Table 4.6: Sample Percentiles of PSI, Sample Size=400 by Bin Size

Bin	Size	P_{25}	P_{50}	Mean PSI	P_{75}	P_{90}	P_{95}	P_{99}
3	400	0.29%	0.66%	0.96%	1.31%	2.13%	3.43%	4.47%
4	400	0.61%	1.18%	1.46%	1.96%	3.04%	4.45%	5.20%
5	400	1.01%	1.69%	2.05%	2.71%	4.05%	5.59%	6.78%
6	400	1.37%	2.16%	2.48%	3.27%	4.50%	6.36%	7.63%
7	400	1.73%	2.68%	3.00%	3.87%	5.26%	7.22%	8.66%
8	400	2.20%	3.35%	3.64%	4.69%	6.24%	8.23%	9.07%
9	400	2.53%	3.63%	3.96%	5.05%	6.73%	8.54%	9.71%
10	400	**2.90%**	**4.19%**	**4.57%**	**5.87%**	**7.57%**	**10.05%**	**11.57%**
11	400	3.48%	4.95%	5.23%	6.64%	8.20%	10.27%	12.12%
12	400	3.94%	5.25%	5.65%	6.99%	8.84%	11.30%	12.64%
13	400	4.43%	5.88%	6.25%	7.77%	9.52%	12.09%	13.54%
14	400	4.77%	6.23%	6.62%	8.05%	10.14%	12.88%	13.87%
15	400	4.99%	6.71%	7.02%	8.68%	10.58%	12.80%	14.20%
16	400	5.66%	7.31%	7.71%	9.53%	11.64%	14.15%	15.57%
17	400	6.05%	7.96%	8.23%	9.90%	12.04%	14.87%	16.59%
18	400	6.44%	8.32%	8.73%	10.63%	13.08%	15.69%	16.79%
19	400	7.07%	9.07%	9.32%	11.22%	13.37%	16.24%	17.67%
20	400	**7.49%**	**9.43%**	**9.77%**	**11.85%**	**13.86%**	**16.34%**	**17.64%**
21	400	7.84%	10.02%	10.30%	12.31%	14.58%	17.97%	19.56%
22	400	8.45%	10.30%	10.72%	12.58%	15.22%	18.35%	20.21%
23	400	8.87%	10.97%	11.33%	13.40%	15.69%	19.04%	21.21%
24	400	9.25%	11.28%	11.79%	13.74%	16.77%	19.90%	22.48%
25	400	9.98%	12.00%	12.46%	14.51%	17.18%	21.15%	23.56%
26	400	10.42%	12.59%	12.98%	15.21%	17.84%	20.91%	23.17%
27	400	10.86%	13.36%	13.53%	15.73%	18.31%	21.42%	23.18%
28	400	11.15%	13.79%	14.08%	16.44%	19.31%	23.06%	25.02%
29	400	12.10%	14.33%	14.77%	17.20%	20.12%	24.37%	25.72%
30	400	12.35%	14.95%	15.31%	17.82%	20.91%	24.45%	26.72%

Base distribution has $\mu = 700$ and $\sigma = 100$. Sample size is fixed at 400.
Number of bins are increased from 3 to 30.
Bin=10 and Bin=20 are widely used.

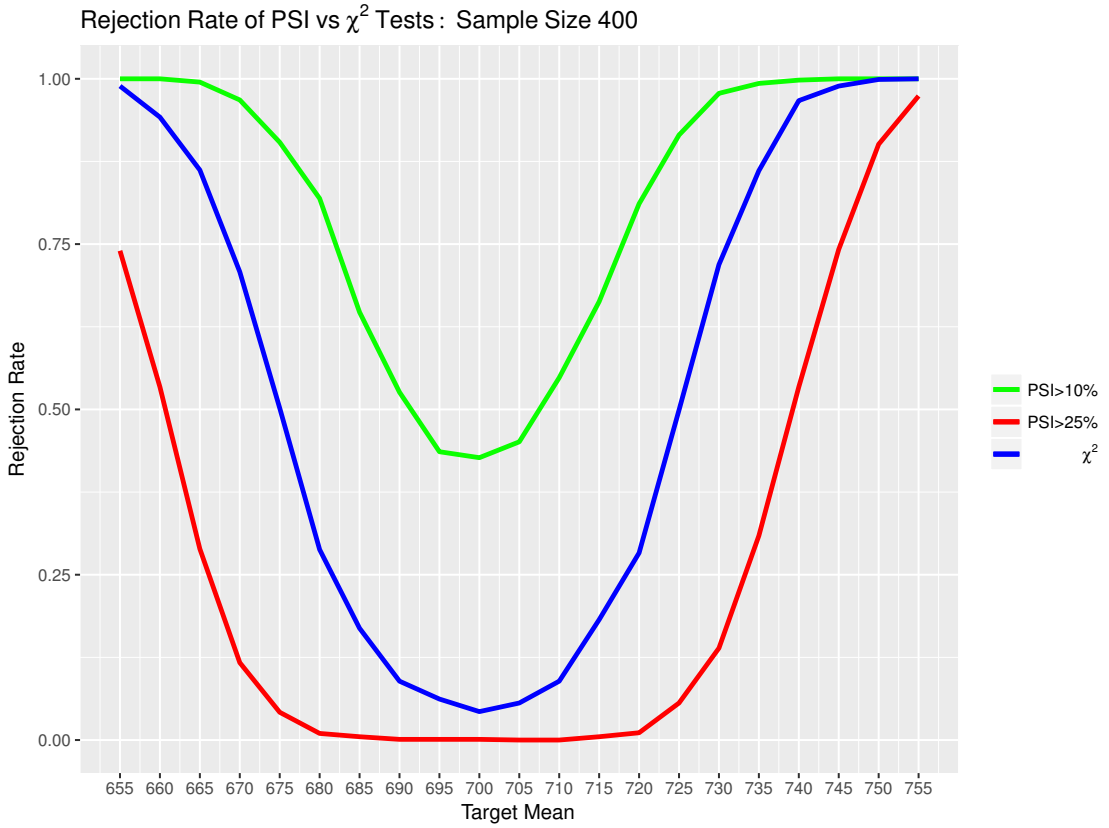

Figure 4.1: Rejection Rate of PSI and χ^2 by Target Mean for One-Sample

a. Rejection Rates for χ^2 test of homegeneity and PSI with critical values 0.10 and 0.25

b. Base population's mean is 700 and $B = 20$.

c. At 700, all tests are at their lowest rejection rate.

The figure (4.1) shows rejection rates for PSI "rule of thumb" and rejection rates based on χ^2-95^{th} percentile. Horizontal axis is for Target Mean, we systemically change Target population's mean to deviate the target population from base population. However, we kept variance same for both. The chart shows some similarity in shape between rejection rates based on "PSI>10%" and χ^2 95^{th} percentile.

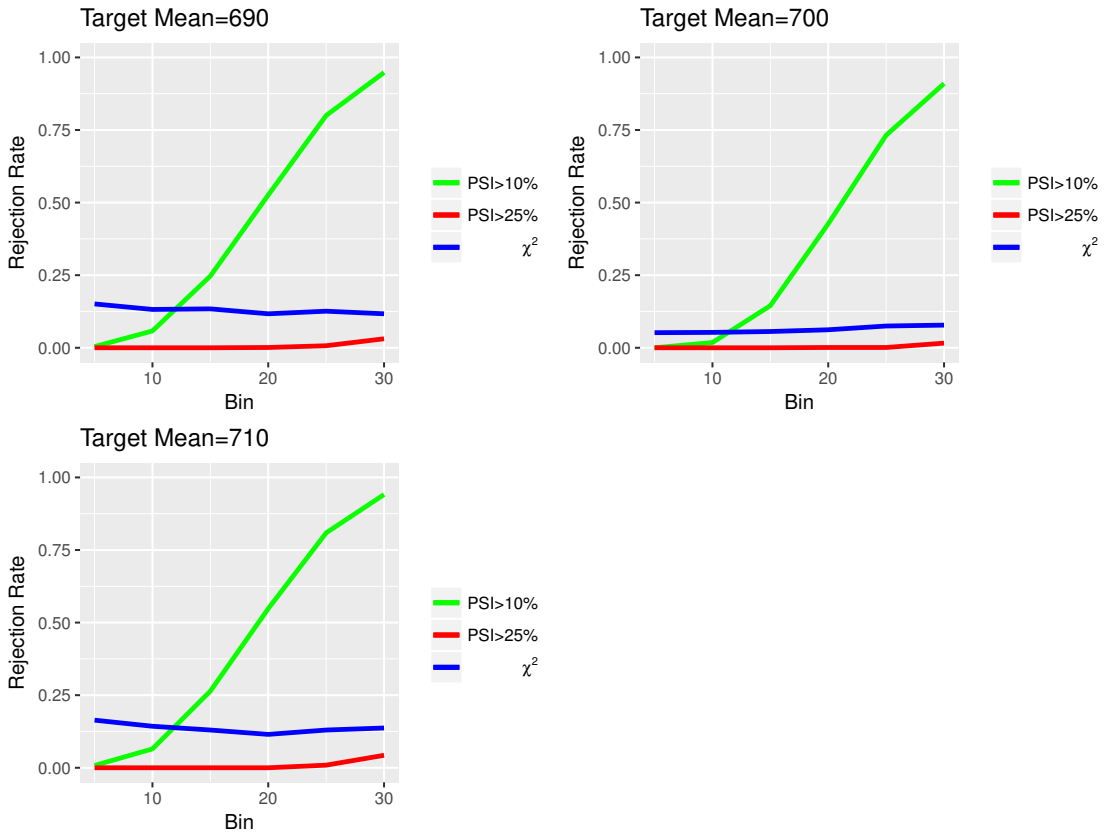

Figure 4.2: Rejection Rate of PSI and χ^2 Test of Homogeneity by Number of Bins

1. Rejection Rates for χ^2 test of homogeneity and PSI cut-off values 0.10 and 0.25 as Target population's mean is 690,700, and 710.
2. Base population's mean is 700.
3. Number of bins ranges from 3 to 30, and sample size is 400.

The figure (4.2) shows how the rejection rates are changing as we increase number of bins and as we deviate from base population. In this simulation, PSI is exceeding 25% for only higher bins so the red line is almost always on x-axis.

Chapter 5

Conclusion

As discussed in previous chapters, distribution of PSI* is only depend on B, N and M. PSI's distribution does not depend on the underlying distribution of the variable in consideration. This suggests that PSI can be used to measure divergence between two distribution of the same variable without considering variable's own distribution. In that sense, it is quite robust for it applicability to any kind of variable.

However, in practice, PSI is unfortunately blindly used in the industry. In this thesis, we provided some guidance by providing statistical properties of PSI and using its connection to χ^2 distribution, a table of recommended cut-off values are provided with respect to B, number of bins and N,M population sample size of base and target variables, respectively.

Instead of using fixed critical values 10% and 25%, I proposed to use 95^{th}, 99^{th} or 99.9^{th} theoretical percentiles of PSI based on either percentiles of normal or χ^2_{B-1} distribution in Chapter 3. The following critical values of PSI^{**} based on normal

approximation and χ^2_{B-1} distribution can be used for tests.

$$CV_1 = (1/N + 1/M)(B-1) + z_\alpha \times \sqrt{2(B-1)} \times (1/N + 1/M) \quad (5.1)$$

$$CV_2 = \chi^2_{\alpha,B-1} \times (1/N + 1/M) \quad (5.2)$$

The resulting critical values are listed as a table in Appendix B along with percentiles of simulated PSI. Critical values and simulation results agrees quite well. Since it is easier to imagine a percentile using normal approximation and corresponding expected value and variance, we suggest to use CV_1. Although, simulations shows that CV_1 is lower than CV_2 they are still close enough to each other especially for B=10.

Use of PSI is not bound to credit score or ratings only. In reality, the calculation can be applied any variable as long as the variable exist on both base dataset and target dataset. We study PSI up to bin size 30, this could be extended. It may also be theoretically interesting when B goes to infinity for future research.

In credit scoring or risk analysis, PSI is used to show a change in the population, the next step is to find the root of this change. This process is called "Characteristic Analysis" and this includes in depth analysis of model components to pin point the root cause of population change. PSI may still be valuable at this stage by calculating PSI for independent variables of the model.

Appendix A

A Note on Taylor Series

$$\ln(x) = \ln(x_0) + \sum_{k=1}^{\infty} \frac{1}{k} \times \frac{1}{x_0^k} \times (x - x_0)^k \times (-1)^{k-1} \tag{A.1}$$

In our case, the equation becomes;

$$\ln(\hat{p}) = \ln(p_0) + \sum_{k=1}^{\infty} \frac{1}{k} \times \frac{1}{p_0^k} \times (\hat{p} - p_0)^k \times (-1)^{k-1} \tag{A.2}$$

The radius of convergence is given by limit of the consecutive terms that is:

$$L = \frac{a_{n+1}}{a_n} \tag{A.3}$$

$$= \left| \frac{\frac{1}{k+1} \times \frac{1}{p_0^{k+1}} \times (\hat{p} - p_0)^{k+1} \times (-1)^k}{\frac{1}{k} \times \frac{1}{p_0^k} \times (\hat{p} - p_0)^k \times (-1)^{k-1}} \right| \tag{A.4}$$

$$= \left| \frac{k}{k+1} \times \frac{1}{p_0} \times (\hat{p} - p_0) \right| \tag{A.5}$$

From the convergence requirement we need to have;

$$\lim_{k\to\infty} L = \left|\frac{\hat{p} - p_0}{p_0}\right| \tag{A.6}$$

$$< 1 \tag{A.7}$$

which we can conclude that $0 < \hat{p} < 2p_0$. So Taylor series will converge only if the inequality holds.

Appendix B

PSI Tables for various N and M

Although most of our discussion for PSI was when both sample sizes are the same, this is not true in practice. To cover that gap, we include following tables which shows theoretical percentiles of PSI for various sizes of samples. These tables are based on Theorems 2.4.1, 2.2.4.1 and 2.3.2 of PSI given in Chapter 2 with a normality assumption of PSI.

This suggests that comparison of base and target with different sample sizes impacts PSI significantly. Accordingly one needs to consider the effect of sample size when deciding a rejection rule.

Theoretical P_{95} of Normal Approximation of PSI, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	32.0%	24.0%	20.0%	18.6%	18.0%	17.6%
200	24.0%	16.0%	12.0%	10.7%	10.0%	9.6%
400	20.0%	12.0%	8.0%	6.7%	6.0%	5.6%
600	18.6%	10.7%	6.7%	5.3%	4.7%	4.3%
800	18.0%	10.0%	6.0%	4.7%	4.0%	3.6%
1000	17.6%	9.6%	5.6%	4.3%	3.6%	3.2%

Table B.1: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{95} of χ^2_{B-1}, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	33.8%	25.4%	21.1%	19.7%	19.0%	18.6%
200	25.4%	16.9%	12.7%	11.3%	10.6%	10.2%
400	21.1%	12.7%	8.5%	7.0%	6.3%	5.9%
600	19.7%	11.3%	7.0%	5.6%	4.9%	4.5%
800	19.0%	10.6%	6.3%	4.9%	4.2%	3.8%
1000	18.6%	10.2%	5.9%	4.5%	3.8%	3.4%

Table B.2: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{99} of Normal Approximation of PSI, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	37.7%	28.3%	23.6%	22.0%	21.2%	20.8%
200	28.3%	18.9%	14.2%	12.6%	11.8%	11.3%
400	23.6%	14.2%	9.4%	7.9%	7.1%	6.6%
600	22.0%	12.6%	7.9%	6.3%	5.5%	5.0%
800	21.2%	11.8%	7.1%	5.5%	4.7%	4.2%
1000	20.8%	11.3%	6.6%	5.0%	4.2%	3.8%

Table B.3: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{99} of χ^2_{B-1}, B=10

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	43.3%	32.5%	27.1%	25.3%	24.4%	23.8%
200	32.5%	21.7%	16.2%	14.4%	13.5%	13.0%
400	27.1%	16.2%	10.8%	9.0%	8.1%	7.6%
600	25.3%	14.4%	9.0%	7.2%	6.3%	5.8%
800	24.4%	13.5%	8.1%	6.3%	5.4%	4.9%
1000	23.8%	13.0%	7.6%	5.8%	4.9%	4.3%

Table B.4: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{95} of Normal Approximation of PSI, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	58.3%	43.7%	36.4%	34.0%	32.8%	32.1%
200	43.7%	29.1%	21.9%	19.4%	18.2%	17.5%
400	36.4%	21.9%	14.6%	12.1%	10.9%	10.2%
600	34.0%	19.4%	12.1%	9.7%	8.5%	7.8%
800	32.8%	18.2%	10.9%	8.5%	7.3%	6.6%
1000	32.1%	17.5%	10.2%	7.8%	6.6%	5.8%

Table B.5: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{95} of χ^2_{B-1}, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	60.3%	45.2%	37.7%	35.2%	33.9%	33.2%
200	45.2%	30.1%	22.6%	20.1%	18.8%	18.1%
400	37.7%	22.6%	15.1%	12.6%	11.3%	10.6%
600	35.2%	20.1%	12.6%	10.0%	8.8%	8.0%
800	33.9%	18.8%	11.3%	8.8%	7.5%	6.8%
1000	33.2%	18.1%	10.6%	8.0%	6.8%	6.0%

Table B.6: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{99} of Normal Approximation of PSI, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	66.7%	50.0%	41.7%	38.9%	37.5%	36.7%
200	50.0%	33.3%	25.0%	22.2%	20.8%	20.0%
400	41.7%	25.0%	16.7%	13.9%	12.5%	11.7%
600	38.9%	22.2%	13.9%	11.1%	9.7%	8.9%
800	37.5%	20.8%	12.5%	9.7%	8.3%	7.5%
1000	36.7%	20.0%	11.7%	8.9%	7.5%	6.7%

Table B.7: N:Size of Base Sample
M:Size of Target Sample

Theoretical P_{99} of χ^2_{B-1}, B=20

N \ M	**100**	**200**	**400**	**600**	**800**	**1000**
100	72.4%	54.3%	45.2%	42.2%	40.7%	39.8%
200	54.3%	36.2%	27.1%	24.1%	22.6%	21.7%
400	45.2%	27.1%	18.1%	15.1%	13.6%	12.7%
600	42.2%	24.1%	15.1%	12.1%	10.6%	9.7%
800	40.7%	22.6%	13.6%	10.6%	9.0%	8.1%
1000	39.8%	21.7%	12.7%	9.7%	8.1%	7.2%

Table B.8: N:Size of Base Sample
M:Size of Target Sample

Bibliography

Cochran, W. G. (1977). *Sampling techniques*. Wiley.

Dornhelm, E. (2015). Us credit quality continues to climb – but will it level off?
URL http://www.fico.com/en/blogs/risk-compliance/us-credit-quality-continues-climb-will-level/

FDIC (2007). Scoring and Modeling VIII. SCORING AND MODELING.
URL http://www.fdic.gov/regulations/examinations/credit_card/pdf_version/ch8.pdf

FED (2011). Supervisory guidance on model risk management contents.
URL http://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

Gottschalk, S. (2016). Entropy and credit risk in highly correlated markets.
URL http://arxiv.org/pdf/1604.07042.pdf

Kullback, S. (1978). *Information theory and statistics*. Peter Smith.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.
URL http://projecteuclid.org/euclid.aoms/1177729694

Li, S., Kharidhi, S., & Kramer, M. (2008). Using PSI to Monitor Predictive Model Stability in the Database Marketing Industry.
URL http://analytics.ncsu.edu/sesug/2008/BI-007.pdf

Lin, A. Z. (2017). Examining Distributional Shifts by Using Population Stability Index (PSI) for Model Validation and Diagnosis.
URL http://www.lexjansen.com/wuss/2017/47_Final_Paper_PDF.pdf

Liu, H., Parise, G. F., & Ware, L. R. (2009). United States Patent: 8326575.
URL http://patentimages.storage.googleapis.com/pdfs/US8326575.pdf

OCC (2016). Sensitivity to Market Risk Installment Lending.
URL http://www.occ.gov/publications/publications-by-type/comptrollers-handbook/installment-lending/pub-ch-installment-lending.pdf

Pruitt, R. (2010). The Applied Use of Population Stability Index (PSI) in SAS® Enterprise Miner™ Posters SAS Global Forum 2010.
URL http://support.sas.com/resources/papers/proceedings10/288-2010.pdf

Siddiqi, N. (2016). *Intelligent Credit Scoring :Building and Implementing Better Credit Risk Scorecards*. John Wiley & Sons.

Stapleton, J. H. (2009). *Linear statistical models*, vol. 719. John Wiley & Sons.

Wu, D., & Olson, D. (2010). Enterprise risk management: coping with model risk in a large bank. *Journal of the Operational Research Society*, *61*(61), 179–190.
URL http://www.palgrave-journals.com/jors/

Yousefi, S., Se Supervisor, S., & Olsson, J. (2016). Credit Risk Management in Absence of Financial and Market Data.

URL http://www.math.kth.se/matstat/seminarier/reports/M-exjobb16/160617.pdf