2019-01-01

Confidence Intervals for the Expected P-value

Emmanuel Kofi Abrefa
University of Texas at El Paso

CONFIDENCE INTERVALS FOR THE EXPECTED P-VALUE

EMMANUEL KOFI ABREFA

Master's Program in Statistics

APPROVED:

Xiaogang Su, Ph.D., Chair

Sangjin Kim, Ph.D.

Thompson Sarkodie-Gyan, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

to my

FATHER James, SISTERS Patricia and Frederickah and my ADVISOR Professor Su

with love

CONFIDENCE INTERVALS FOR THE EXPECTED P-VALUE

by

EMMANUEL KOFI ABREFA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2019

Acknowledgements

I would like to express my deep-felt gratitude to my advisor, Professor Xiaogang Su of the Mathematical Science Department at The University of Texas at El Paso, for his advice, encouragement, enduring patience and constant support. He was never ceasing in his belief in me although I was often doubting in my own abilities. He was always ready to provide clear explanations when I was hopelessly lost, constantly driving me with energy when I was tired and always giving me his time, in spite of his busy schedule.

I also wish to thank the other members of my committee, Dr. Thompson Sarkodie-Gyan of the Electrical and Computer Engineering Department and Dr. Sangjin Kim of the Mathematical Science Department, both at The University of Texas at El Paso. Their suggestions, comments and additional guidance were invaluable to the completion of this work.

Abstract

The p-value is widely used in many application fields. In common practice, a scientific finding is deemed statistically significant if its resultant p-value is less than a pre-specified significance level, for example $\alpha = 0.05$, albeit many statistically significant results are not reproducible in new studies. Mixed reasons including misuses, abuses, misunderstanding and misinterpretation arouse intensive debates and conservatives around the p-value from time to time over the years. Yet no reasonable solutions have been proposed. In this research, we make efforts to close the gap by advocating the use of confidence level for the expected p-value p_0. This allows us to perform a second-level testing problem ($H_0' : p_0 \geq 0.05 vs. H_a' : p_0 < 0.05$.) for each hypothesis testing problem (H_0 vs. H_a). Bootstrap-based confidence intervals are put forward. In particular, we investigate an infinitesimal jackknife (IJ) approach that possibly reduces variance in certain scenarios. The proposed method is empirically assessed and illustrated via both simulation studies and analyses of real data sets.

Table of Contents

Chapter 1

Introduction

1.1 History of p-values

The noticeable quality of the p-value in scientific literature is credited to Fisher, who did not imagine this likelihood measure but rather popularized its broad use for all types of factual research strategies. As indicated by Fisher, the right meaning of the p-value is "the probability of the observed result, plus more extreme results, if the null hypothesis were true." Fisher's motivation was not to utilize the p-value as a basic decision-making instrument but to give researchers an adaptable measure of statistical inference within the mind-boggling procedure of logical derivations. Additionally, there are imperative assumptions related to the appropriate utilization of the p-value. To start with, there is no connection between the causal factor being researched and the result of intrigue (ie, the null hypothesis is valid). Second, the study design and analyses which provide the effect size, confidence intervals, and p-value for the specific study project are completely free of systemic error (ie, there are no misclassification, selection, or confounding biases). Third, the appropriate statistical test is selected for the analysis (eg, the χ^2 test for a comparison of proportions). Within the different philosophies of statistical inference, both the Fisherian and the Neyman-Pearsonian approaches depend on the 'frequentist' translation of likelihood, which indicates that an experiment is hypothetically viewed as one of an infinite number of precisely repeated trials that yield measurably autonomous outcomes. Frequentist methods are the premise of practically all biomedical statistical techniques taught for clinical preliminaries and epidemiologic examinations. Although both the Fisherian and Neyman-Pearsonian approaches have numerous similitudes, they have imperative

philosophical and down to earth contrasts.

1.2 Controversies about the p-value

Recently, there have been doubts and controversies over its usage and over the possible relationship regarding its interpretation in research studies. It is imperative to emphasize that the p-value itself is unpretentious but the problem has to do with its misuse and misinterpretation in the field which leads authors or researchers to inappropriate conclusions. The way that p-values have been casually characterized and translated seems to have prompted colossal disarray and discussion with respect to their place in statistical analysis.

1.3 Proposed Compromise

This research topic is motivated by the doubts and controversies at present regarding the usage of p-values in application fields. In this project, we try to come up with a novel compromise by proposing confidence intervals for the expected p-value to demonstrate that there is, in fact, no ambiguity about what the p-value is, as opposed to what has been claimed in recent public debates in the statistics field. In other words, we seek to achieve that using the p-value is a valid and acceptable way to test the null and the alternative hypotheses in many application fields. To this end, we consider four types of bootstrap-based confidence intervals, including an infinitesimal jackknife (IJ) approach recently revisited by Efron (2015, JASA) for bagging estimators.

Chapter 2

Literature Review And Background

2.1 Controversies about P-values

In earlier part of 2015, a journal from the field of psychology which sought to ban p-values was published. In fact, the editors of Basic and Applied Social Psychology (BASP) announced that the journal would no longer publish papers containing P values, because the values were too often used to support lower-quality research. However, the editors encouraged authors to feel free to submit papers to BASP with P values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication according to Woolston (2015)

According to Baker (2016), p-values are widely used in science to test null hypotheses. For example, in a medical study looking at smoking and cancer, the null hypothesis could be that there is no link between the two. Many researchers interpret a lower P value as stronger evidence that the null hypothesis is false. Many also accept findings as 'significant' if the P value comes in at less than 0.05. But P values are slippery, and sometimes, significant P values vanish when experiments and statistical analyses are repeated.

The editor, David Trafimow and associate editor Michael Marks, who are psychologists at New Mexico State University in Las Cruces, in explaining the new development, said that P values have become a crutch for scientists dealing with weak data. "We believe that the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research", as stated by Woolston (2015)

On March 7, 2016, the American Statistical Association released a "Statement on Statistical Significance and P-Values" with six principles underlying the proper use and inter-

pretation of the p-value. In its statement, the society advises researchers to avoid drawing scientific conclusions or making policy decisions based on P values alone. Additionally, it emphasized that researchers should describe not only the data analyses that produced statistically significant results but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust, as stated by Wasserstein et al. (2016).

The statement's six principles, many of which address misconceptions and misuse of the p-value, are the following:

1. "P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis." See Wasserstein et al. (2016)

2.2 Properties of P-values

A p-value is a statistic that is evaluated from a random sample, it therefore has a distribution in the same way that a sample mean has a distribution, as stated by Sackrowitz and Samuel-Cahn (1999). This distribution also has features that are common to all hypothesis tests. Comprehending the distribution of p-values is the key to seeing how they are

interpreted. In testing any hypothesis, when the null hypothesis, H_o, is true, all p-values between 0 and 1 are equally likely. In other words, the p-value has a uniform distribution between 0 and 1, in this case, regardless of the sample size of the experiment.

On the contrary, if the alternative hypothesis, H_a is true, then the distribution of the p-value under the alternative hypothesis depends on both sample size and the true value or range of true values of the tested parameter. See Boos and Stefanski (2011). Besides other factors, the p-values have a distribution for which p-values near zero are more likely than p-values near 1. In fact, the exact distribution under the alternative hypothesis relies upon the particular hypotheses being tested and the true value of the parameter, but it generally favors values close to 0.

2.3 Ad hoc Strategies

P-values are normally used to test (and expel) a 'null hypothesis', which for example states that there is no difference between two groups, or that there is no relationship between a pair of characteristics. Therefore, they usually provide numerical summary of the evidence against H_o. The smaller the p-value, the more uncertain an observed set of values would occur by chance — under the assumption that the null hypothesis is true. See Su et al. (2016).

A p-value of 0.05 or less is commonly interpreted as meaning that a finding is statistically significant and warrants publication. However, according to Nuzzo (2015), the American Statistical Association has stated that this assertion is not necessarily true.

Baker (2016) establishes that a p-value of 0.05 does not mean that there is a 95 % chance that a given hypothesis is correct. Instead, it signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed. "Additionally, a p-value cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect."

The New England Journal of medicine also made a publication on "Significance of Positive Crossmatch Test in Kidney Transplantation" in which they proposed that the p-value should be compared to a significance level of 0.01 in making statistically decision instead of 0.05. See Patel and Terasaki (1969). However, this does not help solve the problem because the p-value itself is a statistic and depends on the sample data. With a large enough sample size, the p-value could be lower than any preset significance level, regardless how weak the signal strength is as long as it exists.

Chapter 3

Methodology

3.1 Motivation For The Study

As it has already been established, the p-value is a statistic that is evaluated from a random sample, it therefore has a distribution in the same way that a sample mean has a distribution.

Let $X_1, X_2, \ldots, X_n$ be sequence of independent identically distributed random variables, from $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. Suppose we want to test the following hypothesis:

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

With the one-sample t-test approach, the test statistic is given by

$$Z_{obs} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

The resultant p-value is

$$\begin{aligned} \hat{P} &= P\left\{ Z > \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\} \\ &= 1 - \Phi\left\{ \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\} \end{aligned}$$

$$\begin{aligned} E_D\left(\hat{P}\right) &= E_{\bar{X}}\left\{ 1 - \Phi\left\{ \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\} \right\} \\ &= 1 - E_{\bar{X}}\Phi\left\{ \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\} \end{aligned}$$

Consider $\Phi\left\{\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\}$

We have $\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$.

Let $y = \Phi\left\{\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\}$ with $w = \left\{\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right\}$; $w \sim \mathcal{N}\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$.
Now,
$w = \Phi^{-1}(y)$
$\frac{dw}{dy} = \frac{d}{dy}(\Phi^{-1}(y)) = \frac{1}{\phi(\Phi^{-1}(y))}$

To carry out the transformation, we make use of the formula below:
$f_y(y) = f_x(g^{-1}(y))|\frac{d}{dy}(g^{-1}(y))|$

$$
\begin{aligned}
f_y(y) &= \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-\left(\Phi^{-1}(y) - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2}{2}\right\} \cdot \frac{1}{\phi(\Phi^{-1}(y))} \\
&= \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-\left(\Phi^{-1}(y) - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2}{2}\right\} \cdot \frac{1}{\frac{1}{\sqrt{2\pi}} exp\left(\frac{-(\Phi^{-1}(y))^2}{2}\right)} \\
&= exp\left\{\frac{-(\Phi^{-1}(y))^2}{2} + \Phi^{-1}\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) - \frac{1}{2}\left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)^2 + \frac{(\Phi^{-1}(y))^2}{2}\right\} \\
&= exp\left\{\left(\frac{-2\Phi^{-1}(y) - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}}{2}\right) \cdot \left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)\right\} \\
&= exp\left\{\frac{(\mu-\mu_0) - 2\frac{\sigma}{\sqrt{n}}\Phi^{-1}(y)}{2\sigma^2} n(\mu-\mu_0)\right\}
\end{aligned}
$$

We want 95% Confidence Interval for y, so we proceed as follows:

$$P(L \leq w \leq U) = 95\%$$

$$P(\Phi(L) \leq \Phi(w) \leq \Phi(U)) = 95\%$$

$$\left\{ L = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + Z_{0.025} \qquad , \qquad U = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + Z_{0.975} \right\}$$

Finally, $(1-\alpha)100\%$ at Confidence Interval for p-value can be explicitly given as:

$$\left[1 - \Phi\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} - Z_{1-\frac{\alpha}{2}} \right) \qquad , \qquad 1 - \Phi\left(\frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + Z_{1-\frac{\alpha}{2}} \right) \right]$$

Suppose now that σ^2 is unknown, we test the same hypothesis:

Consider $F_t\left(\frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}\right)$

Let $y' = F_t\left(\frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}\right)$ with $w' = \left(\frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}\right)$.

This w' has a non-central t distribution. To show this, we proceed as follows:

We have $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Standardizing the above result yield

$\bar{X} - \mu_0 \sim \mathcal{N}\left(\mu - \mu_0, \frac{\sigma^2}{n}\right)$

$\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \left(\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$

The non-central t test statistic is given by:

$$T = \frac{Z + \mu}{\sqrt{\frac{V}{\nu}}} = \frac{\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} + \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t\left(n-1, \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

Let $w' = F_t^{-1}(y')$

$$\frac{dw'}{dy'} = \frac{dF_t^{-1}(y')}{dy'} = \frac{1}{f_t(F_t^{-1}(y'))}$$

The pdf of non-central t distribution is given by :

$$f(t) = \frac{\nu^{\frac{\nu}{2}} exp(-\frac{\nu\mu^2}{2(t^2+\nu)})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})2^{\frac{\nu-1}{2}}(t^2+\nu)^{\frac{\nu+1}{2}}} \int_0^\infty y^\nu exp\left(-\frac{1}{2}\left(y - \frac{\mu t}{\sqrt{t^2+\nu}}\right)^2\right) dy$$

We carry out the transformation as follows:

$f_{y'}(y') = f_t(g^{-1}(y'))|\frac{d}{dy'}(g^{-1}(y'))|$

$$f(t) = \frac{\nu^{\frac{\nu}{2}} exp(-\frac{\nu\mu^2}{2(t^2+\nu)})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})2^{\frac{\nu-1}{2}}(t^2+\nu)^{\frac{\nu+1}{2}}} \int_0^\infty y^\nu exp\left(-\frac{1}{2}\left(y - \frac{\mu t}{\sqrt{t^2+\nu}}\right)^2\right) dy \cdot \frac{1}{f_t(F_t^{-1}(y'))}$$

The explicit form of this distribution is overwhelming and cannot be obtained easily.

3.2 Four Methods Proposed For This Study

The experimentally determined $\hat{p}$ relies upon the sample data D, making $\hat{p}$ a statistic with a sampling distribution in the same way that a sample mean has a distribution. Since $\hat{p}$ assumes a crucial role in settling on significant practical decisions, the need to deduce properties of the underlying distribution is highly desirable. However, directly studying the sampling distribution of $\hat{p}$ is generally difficult due to the complexity of the procedures in obtaining $\hat{p}$. It is therefore rational to resort to the bootstrap resampling method. To this end, we consider four types of bootstrap-based confidence intervals:

1. Bootstrap Confidence Interval
2. Percentile Interval
3. Infinitesimal Jackknife (IJ) SE for $\tilde{p}$
4. Bias-Corrected IJ SE for $\tilde{p}$

3.3 Confidence Interval for $\hat{p}$

The first two methods are known as the traditional bootstrap methods which make use of $\hat{p}$ in constructing the confidence interval for the expected p-value. It is imperative to understand that these methods are not so reliable because they usually provide wide intervals due to the fact that the variances associated with such computations are sometimes large as stated by Efron (2014) The methodology employed here is simply based on bootstrap distribution of $\hat{p}$. The procedure comprises taking B bootstrap samples $D_b : b = 1, \ldots B$. With each bootstrap sample D_b, exactly the same procedure for obtaining $\hat{p}$ with the original sample data D is repeated to obtain a bootstrap estimate $\hat{p}_b$. This gives rise to a bootstrap sample of the estimates, $\{\hat{p}_b : 1, \ldots, B\}$, from which their bootstrap distribution could be investigated. Nonetheless, the methods employed here also encompasses a novel way of obtaining inference on p from the empirical bootstrap distribution of $\hat{p}$. Since the first two methods are known as the traditional bootstrap methods, it is essential to elaborate on traditional ways of making bootstrap inference. The sample variance of $\hat{p}_b$'s, known as the bootstrap estimate of the variance of $\hat{p}$ is:

$$V_B(\hat{p}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{p}_b - \tilde{p})^2$$

where

$$\tilde{p} = \frac{1}{B}\sum_{b=1}^{B}\hat{p}_b \tag{3.1}$$

is the sample average of $\hat{p}_b$'s. Accordingly, a $(1-\alpha)\times 100\%$ bootstrap confidence interval for p is given by:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} SE_B(\hat{p})$$

with $SE_B(\hat{p}) = \sqrt{\hat{V}_B(\hat{p})}$ where $z_{1-\frac{\alpha}{2}}$ denotes $(1-\frac{\alpha}{2})-$th percentile of the standard normal $\mathcal{N}(0,1)$ distribution.

The percentile confidence interval is another common method employed in this study. Let $\hat{p}_{\frac{\alpha}{2}}$ and $\hat{p}_{1-\frac{\alpha}{2}}$ denote the $\frac{\alpha}{2}-$th and the $(1-\frac{\alpha}{2})-$th percentile of $\hat{p}_b$'s respectively.
A $(1-\alpha)\times 100\%$ bootstrap percentile confidence interval for p is given by:

$$\left(\hat{p}_{\frac{\alpha}{2}}, \hat{p}_{1-\frac{\alpha}{2}}\right)$$

3.4 Confidence Interval for $\tilde{p}$

It is crucial to expatiate on the sample average $\tilde{p}$ because of its variance reduction properties. According to Efron (1982), $\tilde{p}$ is a form of model averaging which facilitates a bagging estimator for p and it is sometimes called bootstrap smoothing. In comparison to the original sample estimate $\hat{p}$, $\tilde{p}$ is more favorable because it may have reduced variation, making it a more dependable tool for setting standard errors and confidence intervals. The last two methods utilize this bagging estimator in its interval estimation.
Moreover, Efron (1982) has proposed a general method of computing standard errors for bagging estimators based on the infinitesimal jackknife (IJ) approach. One paramount term in the IJ formula is a $B\times n$ incidence matrix $\mathbf{N}=(N_{bi})$ associated with B bootstrap samples, where N_{ni} counts how many times the $i-$th observation shows up in the $b-$th bootstrap sample for $i=1,\ldots,n$ and $b=1,\ldots B$. There is therefore no equivocation that $\sum_{i=1}^{n} N_{bi}=n$ or $\mathbf{N}^T\mathbf{j}_B=n\cdot\mathbf{j}_B$ where $\mathbf{j}_B$ is the $B\times 1$ vector with all element being 1. Additionally, let $\hat{\mathbf{p}}=(\hat{p_b})\in\mathbb{R}^B$ represent the vector containing bootstrap estimates. Zhang et al. (2019) has established that, the proposed **IJ** formula for $SE_{IJ}(\tilde{p})$ is given by the following:

Theorem 1: The IJ estimate of variance of the bagged estimator $\tilde{p}$ is given by

$$\hat{V}_{IJ}=\sum_{i=1}^{n}\left[\frac{1}{B}\sum_{b=1}^{B}N_{bi}(\hat{p}_b-\tilde{p})\right]^2$$

$$=\sum_{i=1}^{n}\left[\hat{\text{cov}}(N_{bi},\hat{p}_b)\right]^2$$

where $\hat{\text{cov}}(N_{bi}, \hat{p_b}) = \sum_{b=1}^{B} N_{bi}(\hat{p_b} - \tilde{p})/B$ is the sample covariance between quantities N_{bi} and $\hat{p_b}$ over the B bootstrap samples. In matrix form,

$$\hat{V}_{IJ} = \mathbf{j}_n^T \mathbf{N^T P_B^{\perp} \hat{p}/B^2}$$

where $\mathbf{j}_n = (1) \in \mathbb{R}^n$, $\mathbf{P}_{\frac{1}{B}} = \mathbf{I}_B - \mathbf{j}_B(\mathbf{j}_B^T\mathbf{j}_B)^{-1}\mathbf{j}_B^T$ with $\mathbf{I}_B$ denoting the $B \times B$ identity matrix and $\mathbf{P}_B^{\perp}\mathbf{N}$ centers matrix $\mathbf{N}$ with its column averages.

Applying theorem 1, a $(1-\alpha) \times 100\%$ **IJ** confidence interval for p can be given by

$$\tilde{p} \pm z_{1-\frac{\alpha}{2}} SE_{IJ}$$

with

$$SE_{IJ} = \sqrt{\hat{V}_{IJ}(\tilde{p})} = \frac{1}{B}\sqrt{\mathbf{j}_B^T \mathbf{P}_B^{\perp} \hat{\mathbf{p}}}$$

The next theorem justifies the significance of the bagging estimator $\tilde{p}$ over the sample estimator $\hat{p}$.

Theorem 2: The **IJ** variance estimate for $\tilde{p}$ is smaller than the bootstrap variance estimate for $\hat{p}$, *i.e.*, $\hat{V}_{\mathbf{IJ}} \leq \hat{V}_B(\hat{p})$ implying that $\tilde{p}$ is more precise than $\hat{p}$ as an estimator of p. It is interesting to know that there are two sources of variation associated with the variance estimate $\hat{V}_{\mathbf{IJ}}$. One of these variation is known as the sampling error which stems from randomness with data collection. In practice, only a finite number of B bootstrap samples are taken and this therefore give rise to another variation called the Monte Carlo variation. With small or moderate B, the latter source i.e. the Monte Carlo noise, often dominates the error of $\hat{V}_{\mathbf{IJ}}$ which makes $\hat{V}_{\mathbf{IJ}}$ biased upwards. As a result, Efron suggested a bias-corrected version, which is given in theorem 3.

Theorem 3:: Defining $z_{bi} = (N_{bi} - 1)(\hat{p_b} - \tilde{p})$, the (upward) bias in $\hat{V}_{\mathbf{IJ}}$ can be estimated as

$$\hat{\delta}_o = \frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}(z_{bi} - \bar{z}_i)^2$$

where $\bar{z}_i = \sum_{b=1}^{B} z_{bi}/B$. Further assuming approximate independence of N_{bi} and $\hat{p_b}$, a computationally more efficient bias estimate is given by

$$\hat{\delta} = \frac{n-1}{B^2}\sum_{b=1}^{B}(\hat{p_b} - \tilde{p})^2 = \frac{n-1}{B^2}\mathbf{P}_B^{\perp}\hat{\mathbf{p}}$$

Thus the bias-corrected (unbiased) **IJ** estimate of variance is

$$\hat{V}_u = \hat{V}_{\mathbf{IJ}} - \hat{\delta} = \left\{\mathbf{j}_n^T\mathbf{N}^T - (n-1)\hat{\mathbf{p}}^T\right\}\mathbf{P}_{\mathbf{B}}^{\perp}\ \hat{\mathbf{p}}/\mathbf{B^2}$$

According to theorem 3, another set of $(1-\alpha)\times 100\%$ bias-corrected **IJ** confidence interval for the expected p-value p can be given by:

$$\tilde{p} \pm z_{1-\frac{\alpha}{2}} SE_u$$

with

$$SE_u = \sqrt{\hat{V}_u} = \frac{1}{B}\sqrt{\{\mathbf{j}_n^T\mathbf{N}^T - (n-1)\hat{\mathbf{p}}^T\}\,\mathbf{P}\frac{1}{B}\hat{\mathbf{p}}}$$

It is important to emphasize that all the four sets of confidence intervals are based on one single bootstrapping procedure. With the exclusion of the quantile Confidence Interval, the lower/upper bounds for the SE-based CIs may occasionally go out of the range of [0, 1]. In this case, the simplest approach is to replace the lower bound with 0 or the upper bound with 1. Optionally, one may apply the entire procedure to the log transformed p, say, p'= logit(p). After obtaining the CI for p', say, (L, U), we transform it back to a CI for, (expit(L), expit(U)). This is again because the log function is monotonically increasing and so is its inverse function.

Chapter 4

Simulations

This section presents simulation studies designed to explore or investigate the performance of the proposed methods. Random data with effect size of 0.5 was generated for the two sample t-test.Two sample sizes n $\in \{50, 200\}$ and two choices of B $\in \{200, 2000\}$ were considered in this study. For each test, a total of 500 simulation runs were taken. We then carry out two-sample t-test bootstrap with the assumption that variances are equal to compute and transform p-values. The results including the bagging estimate $\tilde{p_b}$ and four sets of confidence intervals are extracted.

4.1 Using an effect size of 0.5

With an effect size of 0.5, i.e. standardized difference between two treatment groups is 0.5, we carry out simulation studies considering cases where B is small as well as large. Upon critical analysis, it was observed that the p-values associated with an effect size of 0.5 are very small and this accounted for SD of the bagging estimator being greater than SD of the transformed p-values. As a result, the notion that the bagging estimator has much reduced variation is not realized here. It can also be seen that with small B, the averaged bootstrap SE is slightly larger than the empirical SD of the transformed p-values whereas with large B, the same averaged value is slightly smaller than the empirical SD of transformed p-values.

Table 4.1: B=200, Delta=0.5, n=200

nrun	pvalue_0	teststat_0	SE	pvaluebagging	teststatbagging	SE_0	SEc
1	1.212405e-06	5.916352	2.176528	1.300942e-04	6.131698	3.614900	1.914130
2	1.604339e-06	5.794704	2.052723	2.538906e-04	6.094200	3.372956	1.735681
3	1.795738e-04	3.745757	1.694679	5.145572e-03	3.939970	2.823593	1.507253
4	9.661547e-07	6.014953	2.123370	2.598863e-04	6.168198	3.621473	2.040880
5	3.264914e-11	10.486128	2.801735	1.466809e-06	10.688847	4.658856	2.474479
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
495	5.850541e-09	8.232804	2.236522	1.904467e-06	8.650800	3.765839	2.062133
496	1.330172e-08	7.876092	2.874167	2.199498e-05	8.333473	4.985115	2.907473
497	1.074617e-09	8.968746	2.464003	2.349440e-06	9.270868	4.219796	2.398954
498	4.600288e-13	12.337215	3.075573	2.967633e-08	12.429617	5.566965	3.494918
499	9.804094e-05	4.008593	1.847031	4.055639e-03	4.237294	3.219443	1.895481
500	7.514117e-07	6.124122	2.089157	8.290408e-04	6.210913	3.597139	2.067756

SD(teststat) = 2.122556

Mean(SE)= 2.178735

SD(teststatbagging) = 2.140764

Mean(SE_{IJ}) = 3.761448

Mean(SE_{IJ_c})= 2.168658

Table 4.2: B=2000, Delta=0.5, n=200

nrun	pvalue_0	teststat_0	SE	pvaluebagging	teststatbagging	SE_0	SEc
1	1.212405e-06	5.916352	2.176528	3.243560e-04	6.080448	2.245018	2.038755
2	2.946168e-07	6.530743	2.278224	1.652524e-04	6.789316	2.530584	2.317091
3	1.633279e-04	3.786940	1.684228	7.173212e-03	3.961780	1.831980	1.670488
4	1.256532e-05	4.900826	1.921359	2.336115e-03	5.069231	2.091270	1.907170
5	1.951016e-08	7.709739	2.487570	6.914860e-05	7.998425	2.708540	2.470283
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
495	8.340682e-05	4.0787984	1.7278241	3.792772e-03	4.2909674	1.8229579	1.6516328
496	2.167827e-06	5.6639753	2.1337504	9.679282e-04	5.8349182	2.3138632	2.1085814
497	6.191895e-06	5.2081764	2.0747653	1.460650e-03	5.4473598	2.2346723	2.0335711
498	4.337231e-05	4.3627875	1.8529739	2.844441e-03	4.6116784	1.9752643	1.7936070
499	2.604017e-10	9.5843563	2.7943880	4.537485e-06	9.8960926	3.1300144	2.8705318
500	3.902814e-07	6.4086221	2.2147387	2.877734e-04	6.5536352	2.4308501	2.2205767

SD(teststat) = 2.267368

Mean(SE)= 2.156006

SD(teststatbagging) = 2.271519

Mean(SE_{IJ}) = 2.337379

Mean(SE_{IJ_c})= 2.129831

4.2 Using an effect size of 0.1

Here, we also carry out simulation studies considering cases where B is small and where B is large as well. It is observed that the p-values associated with an effect size of 0.1 are relatively large as compared to an effect size of 0.5. With p-values being large, the reduced variation property of the bagging estimator is fully achieved because the SD of the bagging estimator is always less than SD of the transformed p-values irrespective of size B.

Again, it can clearly be observed that with small B, the averaged bootstrap SE is slightly greater than the empirical SD of the transformed p-values whilst the same averaged values with large B is slightly less than the empirical SD of transformed p-values.

Table 4.3: B=200, Delta=0.1, n=200

nrun	pvalue_0	teststat_0	SE	pvaluebagging	teststatbagging	SE_0	SEc
1	0.2760693364	0.5589818286	0.7915194	0.37041305	0.7614293	1.2972690	0.66278417
2	0.4636175009	0.3338401784	0.5735534	0.42156738	0.5932261	0.9046065	0.40658602
3	0.8551321612	0.0679667596	0.3772579	0.50285400	0.4125509	0.5685579	0.20184745
4	0.3920420042	0.4066673993	0.6136251	0.40991187	0.6362375	0.9946451	0.49181916
5	0.0078078627	2.1074678327	1.3100886	0.05956879	2.2968279	2.1620756	1.12588119
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
495	0.0635156367	1.197119344	0.9234243	0.139202762	1.4389055	1.5624362	0.86518873
496	0.0742557402	1.129269968	1.1926090	0.200430860	1.4464987	2.0541279	1.18157025
497	0.0329956732	1.481543006	1.1223203	0.103806121	1.7030913	1.9403064	1.12447578
498	0.0006782702	3.168597233	1.6672743	0.014952113	3.3142842	3.0232282	1.90313659
499	0.7765339271	0.109839565	0.4624899	0.499384113	0.4583337	0.6597999	0.10365876
500	0.2206692106	0.656258259	0.6863111	0.310129657	0.8420537	1.1205969	0.56634368

SD(teststat) = 0.6557654

Mean(SE)= 0.7631976

SD(teststatbagging) = 0.6195870

Mean(SE_{IJ}) = 1.2573140

Mean(SE_{IJ_c})= 0.6642033

Table 4.4: B=2000, Delta=0.1, n=200

nrun	pvalue_0	teststat_0	SE	pvaluebagging	teststatbagging	SE_0	SEc
1	0.2760693364	0.558981829	0.7067256	0.354868733	0.7574722	0.6819535	0.60453938
2	0.1587453524	0.799298981	0.8563251	0.265435486	1.0269691	0.8973633	0.81181406
3	0.9019940012	0.044796351	0.4461160	0.497971557	0.4410512	0.2255227	0.10571636
4	0.6232843039	0.205313810	0.5070222	0.470255549	0.4981545	0.3549827	0.27340919
5	0.0953315736	1.020763238	0.9462779	0.201433405	1.2454042	0.9832160	0.88778509
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
495	0.799384770	0.097244130	0.4155352	0.49217862	0.4364366	0.2280583	0.13259047
496	0.350769345	0.454978369	0.6592396	0.38612387	0.6828648	0.6002247	0.52307836
497	0.456541594	0.340519649	0.5873609	0.41268154	0.6093836	0.5097018	0.43704039
498	0.950259297	0.022157873	0.4277428	0.51162075	0.4201226	0.2011396	0.06304030
499	0.012290539	1.910429063	1.3204283	0.07112228	2.1568125	1.4499709	1.32467175
500	0.390442803	0.408442578	0.6332206	0.40098311	0.6392324	0.5643493	0.48840241

SD(teststat) = 0.7345061

Mean(SE)= 0.7550737

SD(teststatbagging) = 0.6877221

Mean(SE_{IJ}) = 0.7012176

Mean(SE_{IJ_c})= 0.6191575

With regards to the results obtained, it is rational to resort to an effect size of 0.1 instead of 0.5 because the p-values associated with an effect size of 0.1 are relatively large which helps in achieving the reduced variation property of the bagging estimator. However, the averaged bootstrap SE is slightly larger than the empirical SD of the transformed p-values when B is small, in which case bias correction really helps.

Chapter 5

Real Data Exploration

In this chapter, we experiment our methodology with real data in R software package known as PBC. This data set is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival as stated by López-Ratón et al. (2014)
Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. We made use of age by categorizing individuals who are 65 years old and below as a group and those whose ages are above 65 as another group. We then compared the cholesterol level between the two groups of individuals to see if there is significant difference.

The following figures show the various types of transformation employed in this study with $B = 2000$ bootstrap samples. It is observed that the logistic transformation yields more precise results as compared to the $-\log_{10}$ (logworth) and Φ^{-1} (inverse-phi) transformations. The confidence intervals produced with the use of logistic transformation are narrower than the ones produced by the other two transformation methods.

5.1 Confidence Intervals Plot Using Logworth Transformation

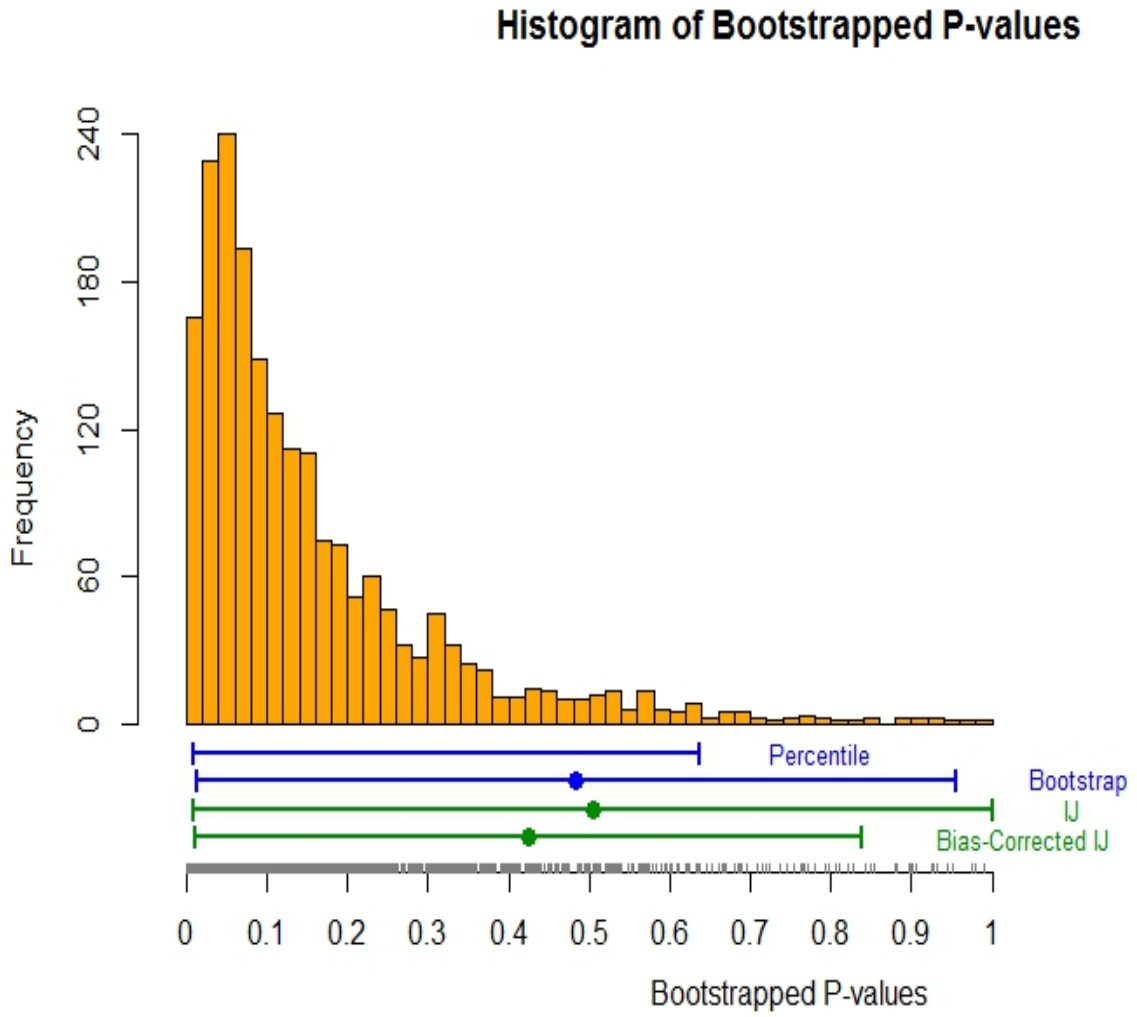

Figure 5.1: Logworth Transformation Plot

5.2 Confidence Intervals Plot Using Inverse-phi transformation

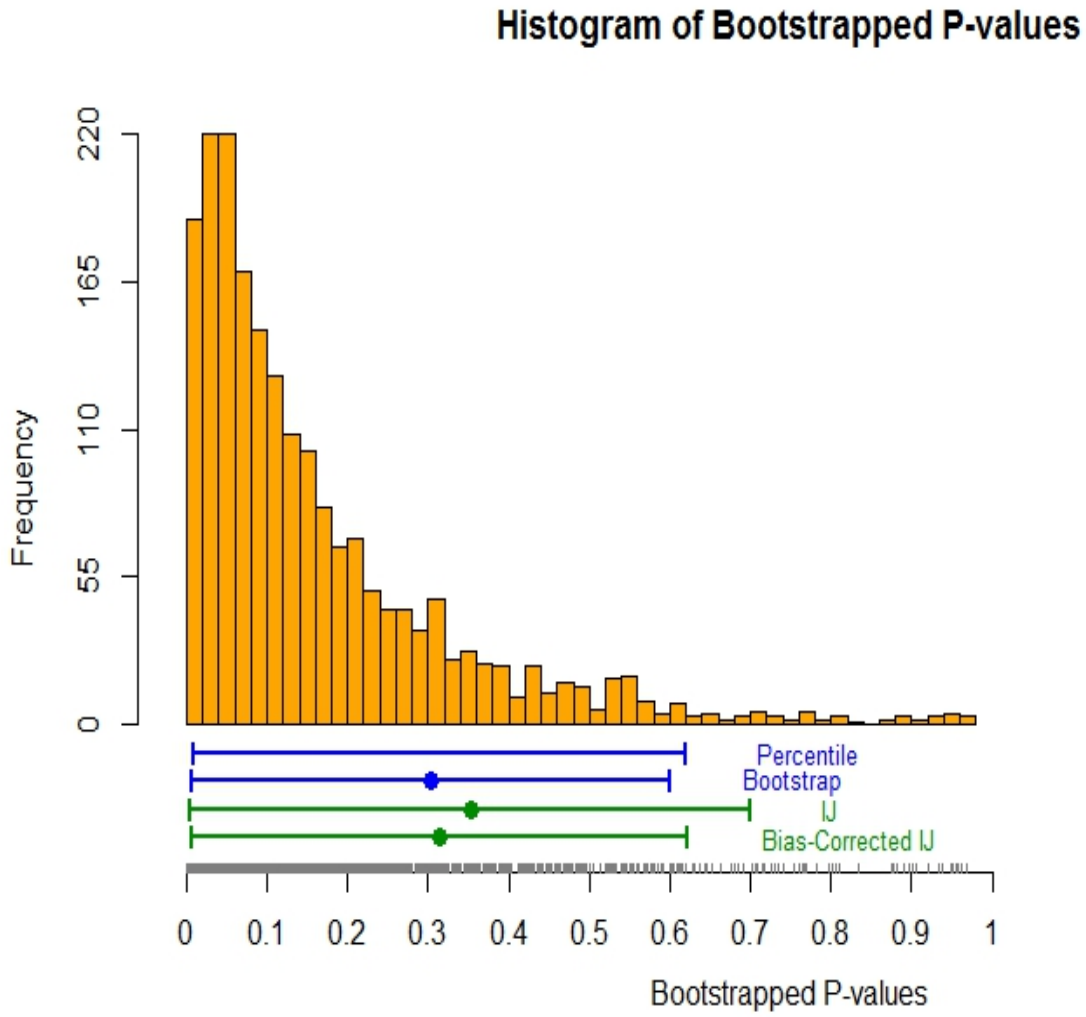

Figure 5.2: Inverse-phi Transformation Plot

5.3 Confidence Intervals Plot Using Logit Transformation

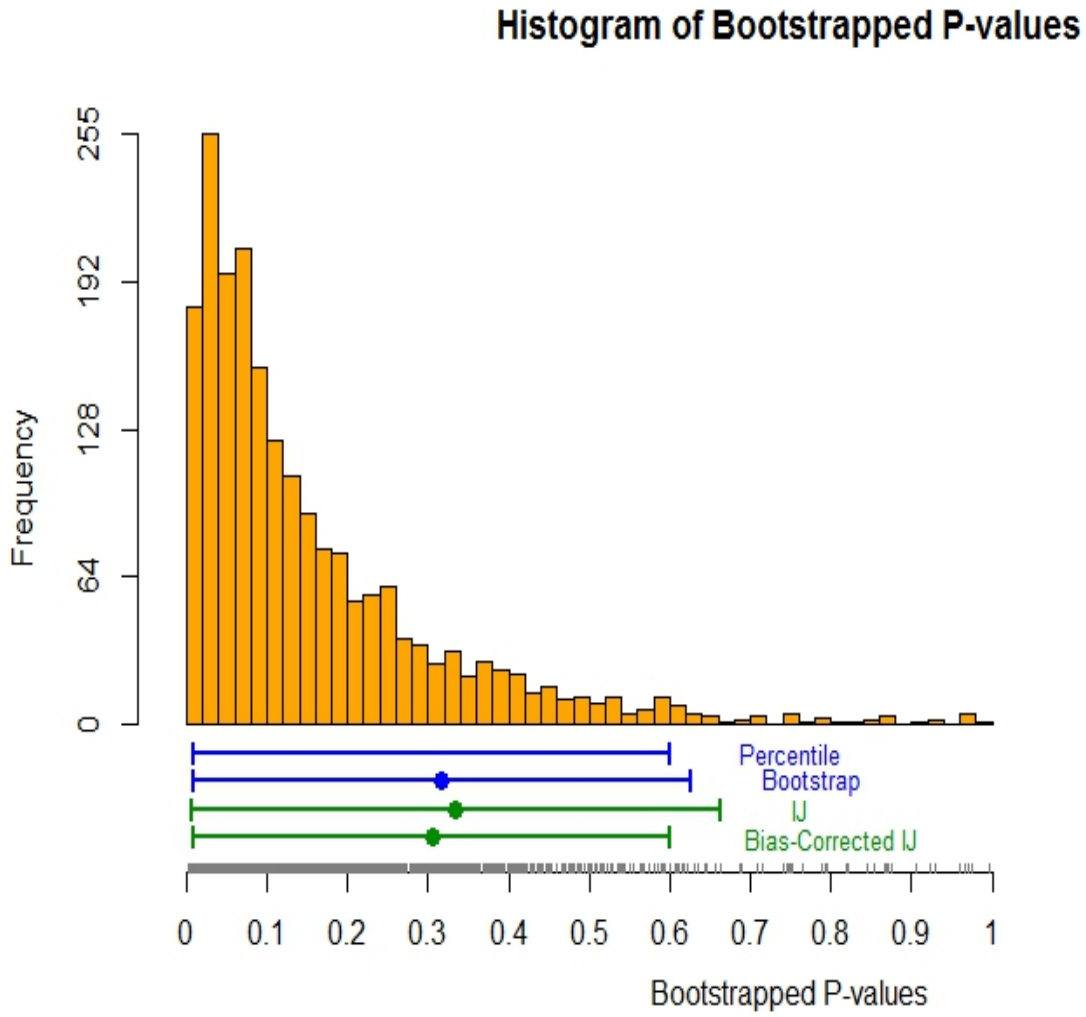

Figure 5.3: Logit Transformation Plot

5.4 Algorithm

Algorithm 1: Psudo-Code for Infinitesimal Jackknife Intervals of Expected P-Value

Input: data $\mathcal{D}$

Output: IJ confidence intervals for the expected p-value

Set: B – the number of bootstrap samples and $h(\cdot)$ – an increasing transformation function

begin:

Perform the significance test with original data $\mathcal{D}$

Obtain the p-value $\hat{p}_0$ and transform $\hat{p}_0 := h(\hat{p}_0)$

Initialize $\tilde{p} := 0$ and incidence matrix $\mathbf{N} = \{0\} \in \mathbb{R}^{B \times n}$.

for $b \leftarrow$ *1 to* B **do**

- Generate the b-th bootstrap sample $\mathcal{D}_b$
- Obtain the b-th row of $\mathbf{N}$ such that N_{bi} equals how many times the i-th observation appears in $\mathcal{D}_b$
- Perform the significance test with $\mathcal{D}_b$
- Obtain the p-value $\hat{p}_b$ and transform $\hat{p}_b := h(\hat{p}_b)$
- Update $\tilde{p} := \tilde{p} + \hat{p}_b$

Compute the bagging estimator $\tilde{p} := \tilde{p}/B$

Compute Infinitesimal Jackknife SE: $SE_{IJ}(\tilde{p}) = \frac{1}{B}\sqrt{\sum_{i=1}^{n}\left[\sum_{b=1}^{B} N_{bi}(\hat{p}_b - \tilde{p})\right]^2}$

Compute bias-corrected Infinitesimal Jackknife SE

$SE_{IJc}(\tilde{p}) = \frac{1}{B}\sqrt{\sum_{i=1}^{n}\left[\sum_{b=1}^{B} N_{bi}(\hat{p}_b - \tilde{p})\right]^2 - (n-1)\sum_{b=1}^{B}(\hat{p}_b - \tilde{p})^2}$

Construct the IJ-based $(1-\alpha)\times 100\%$ CI: $(L, U) = \tilde{p} \pm z_{1-\frac{\alpha}{2}} SE_{IJ}$ or its Biased corrected version $(L, U) = \tilde{p} \pm z_{1-\frac{\alpha}{2}} SE_{IJ_c}$.

Transform the bounds of CI: $L := h^{-1}(L)$ and $U := h^{-1}(U)$.

Chapter 6

Concluding Remarks

6.1 Significance of the Result

It is clear that directly studying the sampling distribution of $\hat{p}$ is generally difficult due to the complexity of the procedures in obtaining $\hat{p}$. However, we have been able to apply a general method of computing standard errors for bagging estimators based on the infinitesimal jackknife (IJ) approach which is useful in constructing the confidence intervals.

The numerical results from our simulation study revealed that the IJ-based SE overestimates the SD of the bagging estimator, in which case the bias correction becomes very useful.

It can also be observed that the bootstrap SE provides good approximation to the SD of the transformed p-values and that bias corrected IJ SE approximates the SD of the bagged transformed p-values.

Therefore, the p-value remains a very useful tool, as long as we are interpreting and communicating its significance appropriately. Interpretation of the expected p-value helps with the usual over-interpretation of p-value. One common problem with this misinterpretation is that the p-value is interpreted alone. However, the expected p-value is the average p-value that one would get when repeating the same study (including the same sample size) for infinite times.

6.2 Recommendations

The main issue with this project is how to deal with small p-values. As a result, several transformation methods were employed to stabilize the p-value but the results obtained were not all that precise. We therefore recommend that, in future project, the entire bootstrap procedure should be applied to the test statistic of the two-sample t-test itself after which it is transformed to obtain the associated p-values for the confidence intervals, instead of initially transforming the p-values. This project is generally applicable to all significance testing; parametric and non-parametric. Hence, we also would like to explore the use of the proposed method in multiple comparison analysis in future research.

Appendix

R codes

A

```
# ###################################################################
# FUNCTIONS FOR CONSTRUCTING CI OF P-VALUE FOR TWO-SAMPLE t TEST
# ###################################################################


# ================================================================
# FUNCTION rdat.twosamplet() GENERATE DATA FOR TWO-SAMPLE t TEST
# ================================================================
rdat.twosamplet <- function(n0=100, n1=100, delta=0.5, sigma=1){
n <- n0 + n1
mu0 <- 0
mu1 <- mu0 + delta
x <- rep(c(0, 1), c(n0, n1))
y <- rep(c(mu0, mu1), c(n0, n1)) + rnorm(n, mean = 0, sd = sigma)
dat <- data.frame(y=y, x=x)
return(dat)
}

# ========================================
# FUNCTION modify.CI() PLACES CI INTO [0,1]
# ========================================
modify.CI <- function(CI){
# if (length(CI)!=2) stop("CI should contain just LB and UB.")
LB <- CI[1]; UB <- CI[2]
LB <- ifelse(LB<0, 0, LB)
UB <- ifelse(UB>1, 1, UB)
CI <- c(LB, UB)
return(CI)
}

# ================================================================
# FUNCTION trans.pvalue AND inverse.trans.pvalue TRANSFORMS P-VALUES
```

```
# ====================================================================
trans.pvalue <- function(pvalue, transform="inversePhi",
alternative="two.sided")
{
if (transform=="logworth") teststat <- -log10(pvalue)
else if (transform=="logit") teststat <- log(pvalue/(1-pvalue))
else if (transform=="inversePhi") {
if (alternative=="two.sided") teststat <- qnorm(1-pvalue/2)
else if (alterantive=="less") teststat <- qnorm(pvalue)
else if (alterantive=="greater") teststat <- qnorm(1-pvalue)
else stop("Sth is wrong with alterantive= arguments")
} else teststat <- pvalue
return(teststat)
}

inverse.trans.pvalue <- function(teststat, transform="inversePhi",
alternative="two.sided")
{
if (transform=="logworth") pvalue <- 10^(-teststat)
else if (transform=="logit") pvalue <- exp(teststat)/(1+exp(teststat))
else if (transform=="inversePhi") {
if (alternative=="two.sided") pvalue <- 2*(1-pnorm(teststat))
else if (alterantive=="less") pvalue <- pnorm(teststat)
else if (alterantive=="greater") pvalue <- 1-pnorm(teststat)
else stop("Sth is wrong with alterantive= arguments")
} else pvalue <- teststat
return(pvalue)
}


# ============================================================================
# FUNCTION Bootstrap.2samplettest COMPUTES AND TRANSFORMS BOOTSTRAPPED P-VALUES
# ============================================================================

Bootstrap.2samplettest <- function(formula, dat, B=2000, transform="inversePhi",
var.equal=TRUE, alternative="two.sided", mu.d=0)
{
fit0 <- t.test(formula, data=dat, alternative=alternative,
mu=mu.d, paired=FALSE, var.equal=var.equal)
p0 <- fit0$"p.value"    # P-VALUE
```

```
teststat0 <- trans.pvalue(p0, transform=transform, alternative=alternative)

# BOOTSTRAP PROCEDURE
n <- NROW(dat)
PVALUE <- TestStat <- rep(0, B); N.Bn <- matrix(0, nrow=B, ncol=n)
for (b in 1:B){
# print(b)
id.b <- sample(1:n, size=n, replace=TRUE)
N.Bn[b, ] <- as.vector(table(factor(id.b, levels=1:n)))
dat.b <- dat[id.b, ]
fit.b <- t.test(formula, data=dat.b, alternative=alternative,
mu=mu.d, paired=FALSE, var.equal=var.equal)
pvalue.b <- fit.b$"p.value"
PVALUE[b] <- pvalue.b
teststat.b <- trans.pvalue(pvalue.b, transform=transform, alternative=alternative)
TestStat[b] <- teststat.b
}
pvalue.bagging <- mean(PVALUE)
# RETURN THE OUTPUT
return(list(pvalue0=p0, teststat0=teststat0, TestStat=TestStat, PVALUE=PVALUE,
N.Bn=N.Bn, pvalue.bagging=pvalue.bagging, transform=transform,
alternative=alternative))
}



# ========================================================================
# FUNCTION PvalueCI() CONSTRUCTS CI FOR P-VALUE WITH TWO-SAMPLE t TEST
# ========================================================================

PvalueCI <- function(bootstrap.output, confidence.level=0.95)
{
TestStat <- bootstrap.output$TestStat
PVALUE <- bootstrap.output$PVALUE
N.Bn <- bootstrap.output$N.Bn
teststat0 <- bootstrap.output$teststat0
pvalue0 <- bootstrap.output$pvalue0
pvalue.bagging <- bootstrap.output$pvalue.bagging
transform <- bootstrap.output$transform
alternative <- bootstrap.output$alternative
B <- length(TestStat); n <- NCOL(N.Bn)
```

```
# BOOTSTRAP SE AND INFINITESIMAL JACKKNIFE SE
teststat.bagging <- mean(TestStat)
V0 <- sum((apply(N.Bn, 2, FUN=cov, y=TestStat, method="pearson"))^2)
bias <- var(TestStat)*(B-1)*(n-1)/(B^2)
Vc <- V0 - bias
SE0 <- sqrt(V0);    # UNCORRECTED SE FOR ENSEMBLE LOGWORTH
SEc <- sqrt(Vc)    # BIAS-CORRECTED SE FOR ENSEMBLE LOGWORTH
SE <- sd(TestStat) # SE FOR LOGWORTH

# FOUR TYPES OF CONFIDENCE INTERVALS
z0 <- qnorm(p=1-(1-confidence.level)/2)
# METHOD I: BOOTSTRAP CI FOR P-VALUE
lb1 <- teststat0 - z0*SE; ub1 <- teststat0 + z0*SE;
LB1 <- inverse.trans.pvalue(lb1, transform=transform, alternative=alternative)
UB1 <- inverse.trans.pvalue(ub1, transform=transform, alternative=alternative)
CI.bootstrap <- modify.CI(sort(c(LB1, UB1)))
print(CI.bootstrap)

# METHOD II: PERCENTILE CI
CI.percentile <- quantile(PVALUE, probs =c((1-confidence.level)/2, 1-(1-confidence.level)/2));

# INFINITESIMAL JACKNIFE CI FOR ENSEMBLE P-VALUE
# UNCORRECTED
lb.u <- teststat.bagging - z0*SE0; ub.u <- teststat.bagging + z0*SE0
LBu <- inverse.trans.pvalue(lb.u, transform=transform, alternative=alternative)
UBu <- inverse.trans.pvalue(ub.u, transform=transform, alternative=alternative)
CI.IJ <- modify.CI(sort(c(LBu, UBu)))
# BIAS-CORRECTED
lb.c <- teststat.bagging - z0*SEc; ub.c <- teststat.bagging + z0*SEc
LBc <- inverse.trans.pvalue(lb.c, transform=transform, alternative=alternative)
UBc <- inverse.trans.pvalue(ub.c, transform=transform, alternative=alternative)
CI.IJc <- modify.CI(sort(c(LBc, UBc)))


# OUTPUT
return(list(pvalue0=pvalue0, teststat0=teststat0, SE=SE,
CI.bootstrap=CI.bootstrap, CI.percentile=CI.percentile,
pvalue.bagging=pvalue.bagging, teststat.bagging=teststat.bagging,
SE0=SE0, SEc=SEc, CI.IJ=CI.IJ, CI.IJc=CI.IJc,
```

```
PVALUE=PVALUE, confidence.level=confidence.level))
}


# ============================================================
# FUNCTION print.PvalueCI PRINTS THE RESUTLS FROM PvalueCI
# ============================================================
print.PvalueCI <- function(PvalueCI.output){
# PRINT OUT THE RESULTS
cat("-----------------------------------------------------\n")
cat("The P-value for this two-sample t test is", PvalueCI.output$pvalue0, ";\n")
cat("The ", PvalueCI.output$confidence.level, " Bootstrap CI for P-Value is
 (", PvalueCI.output$CI.bootstrap[1], ", ", PvalueCI.output$CI.bootstrap[2], ");\n", sep="")
cat("The", PvalueCI.output$confidence.level, "Bootstrap Quantile CI for P-Value is ",
 PvalueCI.output$CI.percentile, ".\n")
cat("-----------------------------------------------------\n")
cat("The Bagging P-value is", PvalueCI.output$pvalue.bagging, "\n")
cat("The ", PvalueCI.output$confidence.level,
 "Uncorrected IJ CI for Bagging P-Value is ", PvalueCI.output$CI.IJ, ";\n")
cat("The ", PvalueCI.output$confidence.level, "Bias-Corrected IJ CI for Bagging P-Value is ",
 PvalueCI.output$CI.IJc, ".\n\n")
}
```

B

```
library(survival)
library(OptimalCutpointsCI)
help(package="OptimalCutpointsCI")
?plot.BootstrapCI

source("FunctionsN-PValueCI.R")

data(pbc)
dat <- pbc
names(pbc)
dat$agegrp <- ifelse(pbc$age <= 65, 0, 1)

#set.seed(1226)
B <- 2000;
transform <- "inversePhi"
```

```
bootstrap.output <- Bootstrap.2samplettest(formula=chol~agegrp, dat=dat, B=B,
transform=transform, alternative="two.sided")
PvalueCI.output <- PvalueCI(bootstrap.output, confidence.level=0.95)
names(PvalueCI.output)




# EXTRACTING RESULTS
a0 <- 5; b0 <- a0 - 1 + 0.2; gap <- 0.2
cex.text <- 0.8
pvalue.bootstrap <- PvalueCI.output$PVALUE
CI.percentile <- PvalueCI.output$CI.percentile
CI.bootstrap <- PvalueCI.output$CI.bootstrap
pvalue0 <- PvalueCI.output$pvalue0
CI.IJ <- PvalueCI.output$CI.IJ
CI.IJc <- PvalueCI.output$CI.IJc

# PLOT THE RESULTANT CIs
hist.out <- hist(pvalue.bootstrap, nclass = 50, plot = FALSE)
max.p <- max(hist.out$counts)
min.p <- min(hist.out$counts)
range.p <- max.p - min.p
plot(hist.out, ylim = c(-range.p/5, max.p), xlim = c(0,
1.5), col = "orange", main ="Histogram of Bootstrapped P-values",
xlab = "", ylab = "Frequency", xaxt = "n", yaxt = "n")
mtext(text = "Bootstrapped P-values", side = 1, line = 2.5)
axis(2, at = seq(0, max.p, length.out = 5), labels = ceiling(seq(0,
max.p, length.out = 5)))
axis(1, at = 0:10/10, labels = 0:10/10)
rug(jitter(pvalue.bootstrap), ticksize = 0.01, side = 1, col = "gray50")
x.percentile <- (-range.p/a0)/b0
# ---------------
# PERCENTILE CI
# ---------------
LB.percentile <- as.numeric(CI.percentile[1])
UB.percentile <- as.numeric(CI.percentile[2])
arrows(x0 = LB.percentile, y0 = x.percentile, x1 = UB.percentile,
y1 = x.percentile, length = 0.05, angle = 90, lwd = 2,
```

```
code = 3, col = "blue")
text(x = UB.percentile + gap - 0.05, y = x.percentile, labels = "Percentile",
cex = cex.text, col = "blue")
# ---------------
# BOOTSTRAP CI
# ---------------
x.bootstrap <- (-range.p/a0) * 2/b0
LB.bootstrap <- as.numeric(CI.bootstrap[1])
UB.bootstrap <- as.numeric(CI.bootstrap[2])
arrows(x0 = LB.bootstrap, y0 = x.bootstrap, x1 = UB.bootstrap,
y1 = x.bootstrap, length = 0.05, angle = 90, lwd = 2,
code = 3, col = "blue")
points(x=(LB.bootstrap + UB.bootstrap)/2, y = x.bootstrap, pch = 19,
cex = 1.2, col = "blue")
text(x = UB.bootstrap + gap - 0.05, y = x.bootstrap, labels = "Bootstrap",
cex = cex.text, col = "blue")
# ---------------
# IJ CI
# ---------------
x.IJ <- (-range.p/a0) * 3/b0
LB.IJ <- as.numeric(CI.IJ[1])
UB.IJ <- as.numeric(CI.IJ[2])
arrows(x0 = LB.IJ, y0 = x.IJ, x1 = UB.IJ, y1 = x.IJ, length = 0.05,
angle = 90, lwd = 2, code = 3, col = "green4")
points(x = (LB.IJ+UB.IJ)/2, y = x.IJ, pch = 19,
cex = 1.2, col = "green4")
text(x = UB.IJ + gap - 0.1, y = x.IJ, labels = "IJ", cex = cex.text,
col = "green4")
# ---------------
# IJc CI
# ---------------
x.IJ.unbiased <- (-range.p/a0) * 4/b0
LB.IJ.unbiased <- as.numeric(CI.IJc[1])
UB.IJ.unbiased <- as.numeric(CI.IJc[2])
arrows(x0 = LB.IJ.unbiased, y0 = x.IJ.unbiased, x1 = UB.IJ.unbiased,
y1 = x.IJ.unbiased, length = 0.05, angle = 90, lwd = 2,
code = 3, col = "green4")
points(x = (LB.IJ.unbiased+UB.IJ.unbiased)/2, y = x.IJ.unbiased,
pch = 19, cex = 1.2, col = "green4")
text(x = UB.IJ.unbiased + +gap, y = x.IJ.unbiased, labels = "Bias-Corrected IJ",
```

```
cex = cex.text, col = "green4")
```

References

Baker, M. (2016). Statisticians issue warning over misuse of p values. *Nature News*, 531(7593):151.

Boos, D. D. and Stefanski, L. A. (2011). P-value precision and reproducibility. *The American Statistician*, 65(4):213–221.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

López-Ratón, M., Rodríguez-Álvarez, M. X., Cadarso-Suárez, C., Gude-Sampedro, F., et al. (2014). Optimalcutpoints: an r package for selecting optimal cutpoints in diagnostic tests. *J Stat Softw*, 61(8):1–36.

Nuzzo, R. (2015). How scientists fool themselves–and how they can stop. *Nature News*, 526(7572):182.

Patel, R. and Terasaki, P. I. (1969). Significance of the positive crossmatch test in kidney transplantation. *New England Journal of Medicine*, 280(14):735–739.

Sackrowitz, H. and Samuel-Cahn, E. (1999). P values as random variables—expected p values. *The American Statistician*, 53(4):326–331.

Su, X., Fan, J., Levine, R. A., Nunn, M. E., and Tsai, C.-L. (2016). Sparse estimation of generalized linear models (glm) via approximated information criteria. *arXiv preprint arXiv:1607.05169*.

Wasserstein, R. L., Lazar, N. A., et al. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.

Woolston, C. (2015). Psychology journal bans p values. *Nature News*, 519(7541):9.

Zhang, Z., Shi, X., Xiang, X., Wang, C., Xiao, S., and Su, X. (2019). Bootstrap confidence intervals for the optimal cutoff point to bisect estimated probabilities from logistic regression. *Statistical methods in medical research*, page 0962280219864998.

Curriculum Vitae

Emmanuel Kofi Abrefa was born on March 13, 1989, the first son of James Abrefa and Veroica Mensah. He graduated from Saint John's School, Ghana in 2009. He holds a Bachelor of Science (BS) degree in Actuarial Science from the University of Cape Coast (UCC). Emmanuel was awarded the maiden eSyllabus Scholarship for Africa in 2014 and he followed that up by claiming an equally-prestigious Prudential Support Award for making the top five best graduating students a year later. He was appointed Teaching Assistant at the Mathematics and Statistics Department at the University of Cape Coast during his national service. He was actively engaged in programs like Actuarial sensitization to High Schools, Conferences, Workshops and Seminars, as a means to increase students interests as well as to broaden their knowledge-base in Actuarial Science and its benefits to the nation as a whole. These workshops helped him improved his interpersonal skills and challenged his creativity and analytical skills. In the Spring of 2018, he entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Statistics he worked as a Teaching Assistant at the Mathematical Sciences Department. During his second semester in graduate school, Emmanuel started his thesis work titled "Confidence Intervals for the Expected P-value" which was supervised by his mentor, Prof. Xiaogang Su. After obtaining his Master's degree, Emmanuel will pursue his doctoral degree in Statistics at North Dakota State University in Fargo.

Email: ekabrefa@miners.utep.edu