In the Name of God

Proceeding of

the First Seminar on

Evidential Inference

Department of Statistics

Ferdowsi University of Mashhad, Iran,

12 June, 2019

Preface

Phrases such as “evidential inference” and “statistical evidence”, which one encounters in resent statistics literature refer to a new statistical paradigm, in which statistical inference is based only one probabilistic model and data and is not affected by subjective components such as loss functions and prior distributions. The forerunner of this select is (retired) professor Richard M. Royall, whose book titled “Statistical Evidence” subtitled “A Likelihood Paradigm” contains the first principles of this new methodology of statistical inference. Although the principle axis of this approach is the likelihood function, it not only depends on “Likelihood Principle”, it is also shaped by another similar (but different) principle, called “The Law of Likelihood”. Tests and confidence intervals (these names are appropriate in this new methodology) are defined and interpreted quite differently as compared with these ideas in classical statistics. Although this paradigm has his own critics (some of whom express their dissatisfaction with Royall’s methods in the discussion par of Royall (2000), is seems quite possible that this new school of statistical inference is going to find place among other school such as Bayes and Neyman-Pearson.)

Mahdi Emadi (Chair)
June 2019

Topics

The aim of the seminar is to provide a forum for presentation and discussion of scientific works covering theories and methods in the field of evidential inference in a wide range of areas:

1. Information criteria for measuring evidence
2. Aspects of differences and similarities between Bayesian and evidential school
3. Criteria for supporting statistical hypotheses by data
4. Troubles with p-value
5. The likelihood principle and the law of likelihood
6. Common mistakes in statistical inference
7. Measuring the strength of statistical evidence
8. The laws of evidential inference
9. Paradigms for statistics

Scientific Committee

1. Arashi, M., Shahrood University of Technology
2. Ashoori, M., Research Institute of Wisdom and Philosophy
3. Chinipardaz, R., Shahid Chamran University of Ahvaz
4. Dastbaravarde, A., Yazd University
5. Doostparast, M., Ferdowsi University of Mashhad
6. Emadi, M., Ferdowsi University of Mashhad (Chair)
7. Habibirad, A., Ferdowsi University of Mashhad
8. Jamalizadeh, A., Shahid Bahonar University of Kerman
9. Meshkani, M. R., Shahid Beheshti University
10. Rasooli, A., University of Zanjan
11. Taheri, S. M., University of Tehran
12. Yousefzadeh., F., University of Birjand
13. Zamanzadeh, E., University of Isfahan

Organizing Committee

1. Ahmadi, J., Ferdowsi University of Mashhad
2. Amini, M., Ferdowsi University of Mashhad
3. Azarnoosh, H. A., Ferdowsi University of Mashhad
4. Bahrami, A., Ferdowsi University of Mashhad
5. Baratpoor, S., Ferdowsi University of Mashhad
6. Bozorgniya, A., Ferdowsi University of Mashhad
7. Doostparast, M., Ferdowsi University of Mashhad
8. Emadi, M., Ferdowsi University of Mashhad (Chair)
9. Fakoor, V., Ferdowsi University of Mashhad
10. Fashandi, M., Ferdowsi University of Mashhad
11. Habibirad, A., Ferdowsi University of Mashhad
12. Jabari Nooghabi, H., Ferdowsi University of Mashhad
13. Jabari Nooghabi, M., Ferdowsi University of Mashhad
14. Khashyar Manesh, K., Ferdowsi University of Mashhad
15. Mohtashami Borzadaran, G. R., Ferdowsi University of Mashhad
16. Niroomand, H. A., Ferdowsi University of Mashhad
17. Razmkhah, M., Ferdowsi University of Mashhad
18. Rezaeiroknabadi, A. B., Ferdowsi University of Mashhad
19. Sadegi, H., Ferdowsi University of Mashhad
20. Sadeghpour Gildeh, B., Ferdowsi University of Mashhad
21. Sal Moslehian, M., Ferdowsi University of Mashhad
22. Soheili, A., Ferdowsi University of Mashhad
23. Tabatabaei, M. M., Ferdowsi University of Mashhad

Student Organizing Committee

1. Ahmadi Nadi, A., PhD student in Statistics, Ferdowsi University of Mashhad

2. Alami, T., PhD student in Statistics, Ferdowsi University of Mashhad
3. Baratnia, M., PhD student in Statistics, Ferdowsi University of Mashhad
4. Esfahani, M., PhD student in Statistics, Ferdowsi University of Mashhad
5. Eskandani, A. H., BSc student in mathematics, Ferdowsi University of Mashhad
6. Hooti, F., PhD student in Statistics, Ferdowsi University of Mashhad
7. Kazempoor, J., PhD student in Statistics, Ferdowsi University of Mashhad
8. Khobi, M., BSc in Photography, University of Tehran
9. Mohammadi, M., PhD student in Statistics, Ferdowsi University of Mashhad
10. Mohammadian, Z., PhD student in Statistics, Ferdowsi University of Mashhad
11. Rasouli, S. H., BSc student in mathematics, Ferdowsi University of Mashhad
12. Sabokkhiz, M. J., MSc student in Statistics, Ferdowsi University of Mashhad

Contents

The First Seminar on
Evidential Inference
12 June 2019

Goodness of Fit Test based on Statistical Evidence for Scale Family with Censored Data

Habibirad, A. [1]

[1]Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad

Abstract

The life distributions coming from scale family are of great significance in the area of life testing studies, which has attracted interest from many researchers. Moreover, goodness of fit procedures have been developed in the literature when the available samples are censored. In this paper, a new test statistic based on statistical evidence is proposed to goodness of fit test the distributions from the scale family with censored data. The simulation results in order to compare the test powers are presented and finally, the use of the proposed test is shown in some illustrative examples.

Keywords: Censored data, Goodness of fit test, Statistical evidence.

1 Introduction

The goodness of fit, of a statistical model describes how well it fits into a set of observations. The goodness of fit indices summarize the discrepancy between the observed values and the values expected under a statistical model. It is used to test if sample data fits a distribution

[1] ahabibi@um.ac.ir

from a certain population. In other words, it tells us if our sample data represents the data we would expect to find in the actual population.

The statistical analysis of what are variously referred to as lifetime, survival time, or failure time data is an important topic in many areas, including the biomedical, engineering, and social sciences. Applications of lifetime distribution methodology range from investigations of the durability of manufactured items to studies of human diseases and their treatment. Some methods of dealing with lifetime data are quite old, but starting about 1970 the field expanded rapidly with respect to methodology, theory, and fields of application. Software packages for lifetime data analysis have been widely available since about 1980, with the frequent appearance of new features and packages.

Suppose, in a life-testing experiment, n items are placed on the test. The failure times observed from such a life-test, $X_{(1)} \leq ... \leq X_{(n)}$, are the order statistics from a random sample of size n from a parametric distribution with probability density function (pdf) $f(x;\theta)$ and cumulative distribution function (cdf) $F(x;\theta)$, where $\theta \in \mathbb{R}$. However, one may not continue the experiment until the last failure since the waiting time for the final failure is unbounded (Muenz and Green, 1977). For this reason, in some cases, the life-testing experiment is usually terminated when the rth failure, $X_{(r)}$, is observed, which is referred to as a type II censoring scheme. This censoring model saves time and cost, but some information about the underlying parameters is lost in the censored data (Zheng and Park, 2004). So, the inference based on type II censored data will naturally be less efficient than that based on the complete data of n observations. More than the above specified scheme, there exist some other different sorts of censoring schemes such as random censoring, hybrid censoring (Epstein, 1954) and progressively type II censoring (Balakrishnan and Aggarwala, 2000).

Let $X_{(1)}, ..., X_{(n)}$ denote the ordered values of the random sample $X_1, ..., X_n$ (failure times). In type II plan, observations terminate after the rth failure occurs. So we only observe the r smallest observations in a random sample of n items. The likelihood function based on $X_{(1)}, ..., X_{(r)}$ is given by (Arnold et al. 1992)

$$L_{typeII} = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(X_i)[1 - F(x_r)]^{(n-r)}.$$

An important role of the statistical analysis in science is interpreting observed data as evidence, that is assessing What do the data say?

Although standard statistical methods (hypothesis testing, estimation, confidence intervals) are routinely used for this purpose, the theory behind those methods contains no defined concept of evidence and no answer to the basic question when is it correct to say that a given body of data represent evidence supporting one statistical hypothesis against another? (Royall, 1997, 2000). Emadi and Arghami (2003) and Emadi et al. (2007) have studied some measures of support of statistical hypotheses. Doostparast and Emadi (2006), Arashi and Emadi (2008) have studied some measures of support of statistical hypotheses based on independent and identically distribution (iid) observations and record statistics. Habibirad et al. (2006) generalized the concept of expected true statistical evidence based on the law of likelihood. So, when the objective of the study is to produce statistical evidence for one hypothesis against another, it is desirable to have a measure of performance of the experiments E1 and E2.

If the experimenters object is obtaining statistical evidence about some competing hypotheses, then she/he would like to know the potential true evidence in the available data.

The outline of this paper is as follows. In Section 2, we specify the scale family as model and the likelihood functions based on type II censored data. The criteria of statistical evidence introduce in Section 3, and we show the exponentiality test and Rayleigh tes results under weibull alternative distribution by a simulation study in Section 4. The performance of the considered tests for a real data is evaluated in section 5.

2 Introducing the Model

We consider the distribution from the scale family with probability density function (pdf) $f(x;\theta)$ and the cumulative distribution function (cdf) $F(x;\theta)$ given by

$$f(x;\theta) = \frac{1}{\theta} g(\frac{x}{\theta}); x > 0; \quad \theta > 0,$$

and

$$F(x;\theta) = G(\frac{x}{\theta}); x > 0; \quad \theta > 0,$$

respectively, where θ is the scale parameter, and the $g(x)$ and $G(x)$ are the standard pdf and cdf for the scale family respectively. So the likelihood function based on $X_{(1)}, ..., X_{(r)}$ is

$$L_{typeII} = \frac{n!}{(n-r)!} (\frac{1}{\theta})^r \prod_{i=1}^{r} g(\frac{x_i}{\theta}) [1 - G(\frac{x_r}{\theta})]^{(n-r)}$$

where r is the number of failures.

Consider the following hypotheses

$$H_0 : f(x) = f_0(x,\theta) \quad v.s. \quad H_1 : f(x) \neq f_0(x,\theta),$$

where $f_0(x,\theta)$ is the distribution from the scale family and the unknown θ, will be estimated by the maximum likelihood estimator.

2.1 Exponentiality Test

A large number of recent results pertaining to lifetime tests are obtained based on the assumption that the lifetime of a system is described by an exponential distribution. Testing methods to detect exponential distribution patterns still attract much attention and are the topic of a large amount of recent researches. Many authors provide test statistics for detecting departures from the hypothesis of exponentiality against specific or general alternatives. For example, see Hanis (1976), Henze and Meintanis (2002b), Ebrahimi et al. (1992), Ebrahimi (1998) and Baratpour and Habibirad (2012).

A random variable X follows the exponential distribution if and only if it has pdf and cdf, respectively

$$f_0(x,\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, \qquad F_0(x,\theta) = 1 - e^{-\frac{x}{\theta}},$$

where, $x > 0, \quad \theta > 0$, and θ is unknown.

Then, the likelihood function associated with type II censored data is as

$$L_{TypeII}(\theta) = \frac{n!}{(n-r)!}(\frac{1}{\theta})^r \exp\left(-\frac{1}{\theta}\left[\sum_{i=1}^{r} x_{(i)} + (n-r)x_{(r)}\right]\right), \tag{2.1}$$

where r is the number of failures and the maximum likelihood estimator of θ is

$$\hat{\theta} = \sum_{i=1}^{r} x_{(i)}/r.$$

2.2 Rayleigh Test

The Rayleigh distribution is a special case of the Weibull distribution with a scale parameter of 2 and a suitable model in various areas including reliability, life testing, and survival analysis. The square of a Rayleigh random variable with a shape parameter 1 is equal to a chi square random variable with 2 degrees of freedom. Also, the square root of an exponential random variable has the Rayleigh distribution. Also, the Rayleigh distribution is widely used in the physical sciences to model wind speed, wave heights and sound/light radiation and has been used in medical imaging science, to model noise variance in magnetic resonance imaging. For more information about the applications and properties of the Rayleigh distribution, we refer the interested readers to Siddiqui (1962) and Johnson et al. (1994).

A random variable X follows the Rayleigh distribution if and only if it has pdf and cdf, respectively

$$f_0(x;\theta) = \frac{x}{\theta^2}\exp\left(-\frac{x^2}{2\theta^2}\right), \qquad F_0(x) = 1 - \exp\left(-\frac{x^2}{2\theta^2}\right),$$

where, $x > 0, \quad \theta > 0$, and θ is unknown.
Then, the likelihood function associated with Type II censored data is as

$$L_{typeII}(\theta) = \frac{n!}{(n-r)!}(\frac{1}{\theta})^{2r}\Pi x_i \exp\left(- \frac{1}{2\theta^2}\left[\sum_{i=1}^{r} x_{(i)}^2 + (n-r)x_{(r)}^2 \right]\right), \tag{2.2}$$

where r is the number of failures. It is easy to show that the maximum likelihood estimator of θ is

$$\hat{\theta} = \sqrt{\sum_{i=1}^{r} x_{(i)}^2/2r}.$$

3 Creteria Statistical Evidence

Statistical evidence is represented and interpreted by the law of like-lihood and its strength is measured by the likelihood ratio. The law of likelihood explains that the strength of statistical evidence for one hypothesis over another is measured by their likelihood ratio, (Blume, 2002).
Let $p(\underline{x};\theta)$ be the joint pdf of n iid observations from a distribution with pdf $f(x;\theta)$, then the likelihood ratio

$$\lambda = \frac{p(\underline{x};\theta_0)}{p(\underline{x};\theta_1)}$$

measures the strength of evidence favorable to the simple hypothesis $H_0 : \theta = \theta_0$ against the simple hypothesis $H_1 : \theta = \theta_1$, (Royall, 1997, 2000).
Another measure of expected true statistical evidence, we use $abc(\eta)$, defined by Emadi and Arghami (2003), as

$$abc(\eta) = E_0(\eta) - E_1(\eta), \tag{3.1}$$

where $\eta = \lambda/(1+\lambda)$ and $E_i(\eta)$ is the expected value of η under $H_i, i = 0, 1$.
Suppose E_1 and E_2 are two experiments (or sampling schemes) with (approximately) the same cost, having outcomes $\underline{x}$ and $\underline{y}$, which are the realizations of random vectors $\underline{X}$ and $\underline{Y}$, with densities $p(\underline{x};\theta)$ and $q(\underline{x};\theta)$, respectively, where θ is an unknown parameter.
When the objective of the study is to produce statistical evidence for one hypothesis against another (in the above sense), it is desirable to have a measure of performance of the experiments E_1 and E_2. This can be defined as, (Habibirad et al., 2006)

$$S_\varphi(E) = E_{\theta_0}\varphi(\lambda) + E_{\theta_1}\varphi(1/\lambda),$$

where $\varphi(.)$ is a non decreasing function.
(i) If

$$\varphi(t) = \begin{cases} 1, & t \geq K \\ 0, & t < K \end{cases} \tag{3.2}$$

$S_{\varphi}(E)$ is the sum of the probabilities of observing strong true evidence under H_0 and H_1, as

$$S_{1\varphi}(E) = P_{\theta_0}(\lambda \geq K) + P_{\theta_1}(\lambda < 1/K),$$

where K is arbitrary and is usually between 8 and 32 (Royall, 1997).
(ii) If $\varphi(t) = t/(1+t)$, then

$$S_{2\varphi}(E) = abc(E)$$

the area between the cumulative distribution function (cdf) curves (under H_0 and H_1) of $\eta = \lambda/(1+\lambda)$ (Emadi and Arghami, 2003).
(iii) If $\varphi(t) = \log(t)$, then

$$\begin{aligned} S_{3\varphi}(E) &= E_{\theta_1}\left[\log \frac{p(\underline{X};\theta_1)}{p(\underline{X};\theta_0)}\right] + E_{\theta_0}\left[\log \frac{p(\underline{X};\theta_0)}{p(\underline{X};\theta_1)}\right] \\ &= D(p_{\theta_1}, p_{\theta_0}) + D(p_{\theta_0}, p_{\theta_1}) \\ &= J(p_{\theta_1}, p_{\theta_0}), \end{aligned} \tag{3.3}$$

where $D(p_{\theta_1}, p_{\theta_0})$ and $J(p_{\theta_1}, p_{\theta_0})$ are, respectively, asymmetric and symmetric KullbackLeibler divergance (information) of p_{θ_1} and p_{θ_0}. In this article, it is the last of the above three criteria that we shall use by $S_{\varphi}(E)$.

4 Simulation

In this part, we conducted a simulation study to compute λ, $S_1\varphi$, $S_2\varphi$ and $S_3\varphi$ as four criteria in statistical evidences. We considered $n = 10, 15$ and 20 for some different r, $(r \leqslant n)$. We used 50,000 Monte Carlo simulations to compute mentioned creteria. The results for exponential and Rayleigh tests with Weibull alternative are presented in Tables 1 and 2, respectively.
The results in Tables 1 and 2 show, statistical evidence supports the H_0 more than H_1 and obviously, when the value of n and r increase, the statistical evidence for supporting the H_0 is increasing, too.

Table 1: Evidencial criteria for exponentiality test when alternative is Weibull.

		W(0.5)				W(1.5)			
n	r	λ	$S_1\varphi$	$S_2\varphi$	$S_3\varphi$	λ	$S_1\varphi$	$S_2\varphi$	$S_3\varphi$
10	5	3.168	0.602	0.366	3.273	20.273	1.101	0.179	1.112
	7	9.479	0.638	0.524	4.564	37.949	1.130	0.261	1.545
	9	33.518	0.891	0.651	6.023	52.387	1.120	0.338	2.002
15	8	5.289	0.339	0.451	6.283	147.612	1.313	0.245	2.122
	12	58.673	0.717	0.727	8.572	468.712	1.428	0.409	2.850
	14	225.414	0.973	0.807	10.063	664.298	1.139	0.480	3.319
20	10	55.18	0.484	0.406	8.823	128.454	1.530	0.235	2.976
	15	104.257	0.593	0.757	11.368	160.760	1.378	0.450	3.773
	18	756.351	0.926	0.865	13.122	532.720	1.182	0.558	4.364

Table 2: Evidencial criteria for Rayleigh Ttest when alternative is Weibull.

		W(0.5)			
n	r	λ	$S_1\varphi$	$S_2\varphi$	$S_3\varphi$
10	5	7.935	0.213	0.261	8.645
	7	42.836	0.236	0.358	13.247
	9	322.458	0.289	0.462	18.889
15	8	96.47	0.081	0.187	17.278
	12	227.015	0.173	0.345	25.980
	14	480.672	0.232	0.414	32.363
20	10	12.348	0.031	0.104	24.568
	15	178.779	0.113	0.245	34.443
	18	472.420	0.194	0.350	42.173

5 Real example

Lawless (2011) analyzed an example with data presented in Wilk et al. (1962). These data consisted of lifetimes of transistors obtained from an accelerated life test. The lifetimes are singly type II censored and come from a sample of size n = 34, with three censored observations. The lifetimes (in weeks) are given in Table 3 (three of them are censored and denoted by asterisks). As can be seen, the data are heavily rounded off.

We computed the introduced criteria in Section 3 for the real data set in Table 3 and presented the results in Table 4.

The results of Table 4 indicate evidencial criteria for real data set, λ, $S_1\varphi$, $S_2\varphi$ and $S_3\varphi$, support that H_0 hypothesis versus H_1. Thus it is concluded that the data is from an exponential distribution with type II censored data.

Table 3: Wilk data

3	4	5	6	6	7	8	8	9	9	9	10	10	11	11	11	13
13	13	13	13	17	17	19	19	25	29	33	42	42	52	52*	52*	52*

Table 4: Evidencial criteria for the real data set

	W(0.5)	W(0.7)	W(1.5)
λ	13842.5	106.009	5.848
$S_1\varphi$	1.049	1.50	0.558
$S_2\varphi$	0.984	0.835	0.711
$S_3\varphi$	27.595	8.3565	4.761

References

[1] Ahmad, I. A., & Alwasel, I. A. (1999). A Goodnessoffit Test for Exponentiality Based on the Memoryless Property. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), 681-689.

[2] Alwasel, I. (2001). On goodness of fit testing of exponenttality using the memoryless property. Journal of Nonparametric Statistics, 13(4), 569-581.

[3] Arashi, M. & Emadi, M. (2008). Evidential inference based on record data and inter-record times. Statist. Papers, 49, 291-301.

[4] Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). A first course in order statistics (Vol. 54). Siam.

[5] Balakrishnan, N., Balakrishnan, N., & Aggarwala, R. (2000). Progressive censoring: theory, methods, and applications. Springer Science & Business Media.

[6] Balakrishnan, N., Rad, A. H., & Arghami, N. R. (2007). Testing exponentiality based on Kullback-Leibler information with progressively Type-II censored data. IEEE Transactions on Reliability, 56(2), 301-307.

[7] Baratpour, S., & Rad, A. H. (2012). Testing goodness-of-fit for exponential distribution based on cumulative residual entropy. Communications in Statistics-Theory and Methods, 41(8), 1387-1396.

[8] Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. Statistics in Medicine, 21, 2563-2599.

[9] Doostparast, M. & Emadi, M. (2006). Statistical evidence methodology for model acceptance based on record values. J. Korean Statist. Soc., 35, 167-177.

[10] Ebrahimi, N., Habibullah, M., & Soofi, E. S. (1992). Testing exponentiality based on Kullback-Leibler information. Journal of the Royal Statistical Society. Series B (Methodological), 739-748.

[11] Epstein, B. (1954). Truncated life tests in the exponential case. The Annals of Mathematical Statistics, 555-564.

[12] Emadi, M., Ahmadi, J. & Arghami, N. R., (2007). Comparing of record data and random observation based on statistical evidence. Statist. Papers, 48, 1-21.

[13] Emadi, M. & Arghami, N. R. (2003), Some measures of support for statistical hypotheses. J. Stat. Theory Appl., 2, 165-176.

[14] Habibirad, A. H., Yousefzadeh, F., & Balakrishnan, N. (2011). Goodness-of-fit test based on Kullback-Leibler information for progressively Type-II censored data. IEEE Transactions on Reliability, 60(3), 570-579.

[15] Habibirad, A., Arghami, N. R. and Ahmadi, J. (2006). Statistical evidence in experiments and in record values. Comm. Statist.-Theory Meth., 35, 1971-1983.

[16] Lawless, J. F. (2011). Statistical models and methods for lifetime data (Vol. 362). John Wiley & Sons.

[17] Muenz, L. R., & Green, S. B. (1977). Time savings in censored life testing. Journal of the Royal Statistical Society. Series B (Methodological), 269-275.

[18] Noughabi, H. A., & Balakrishnan, N. (2015). Goodness of fit using a new estimate of Kullback-Leibler information based on Type II censored data. IEEE Transactions on Reliability, 64(2), 627-635.

[19] Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. London: Chapman & Hall, New York.

[20] Royall, R. (2000). On the probability of observing misleading statistical evidence. J. Amer. Statist. Assoc., 95, 760-780.

[21] Torabi, H., Montazeri, N. H., & Gran, A. (2018). A wide review on exponentiality tests and two competitive proposals with application on reliability. Journal of Statistical Computation and Simulation, 88(1), 108-139.

[22] Wilk, M. B., Gnanadesikan, R., & Huyett, M. J. (1962). Estimation of parameters of the gamma distribution using order statistics. Biometrika, 49(3/4), 525-545.

[23] Yousefzadeh, F., & Arghami, N. R. (2008). Testing exponentiality based on type II censored data and a new cdf estimator. Communications in Statistics-Simulation and Computation, 37(8), 1479-1499.

[24] Zheng, G., & Park, S. (2004). On the Fisher information in multiply censored and progressively censored data. Communications in Statistics-Theory and Methods, 33(8), 1821-1835.

The First Seminar on
Evidential Inference
12 June 2019

Evidence Tests and Optimal Sample Size Determination for Risk Measures

Yousefzadeh, F. [1]

[1]Department of Statistics, Faculty of Mathematics and Statistics, University of Birjand

Abstract

In actuarial applications we often work with risk measures for insurance products. In this paper, some criteria based on statistical evidence are proposed to test a risk measures. Optimal sample size in which the substantial evidence reaches a desired level is determined. Also, a simulation is presented to illustrate the results.

Keywords: Evidence test, Empirical Likelihood, Jackknife, Risk measure.

1 Introduction

In life insurance and finance, quantifying risks is an essential task to price an insurance product or manage a financial portfolio. In general, a risk measure is erected to be a mapping from a set of risks to the set of real numbers. Some well known risk measures include coherent risk measures (Yaari (1987); Artzner (1999)), distortion risk measures, Wangs premium principle, and proportional hazards transform risk measures; see Wang et al (1997); Wang

[1] fyousefzadeh@birjand.ac.ir

(1998); Wirch and Hardy (1999), and Necir and Meraghni (2009) for references. Jones and Zitikis (2003) defined a large class of risk measures associated with a risk variable X with distribution function F as,

$$R(F) = \int_0^1 F^{-1}(t)\psi(t)dt, \tag{1.1}$$

where the generalized inverse function of F is denoted by F^{-1} , and ψ is a non negative function chosen for showing the objective opinion about the risk loading. Different choices of ψ result in different risk measures. For example, tail value-at-risk has $\psi(t) = I(t > \alpha)/(1-\alpha)$ with $0 < \alpha < 1$, the proportional hazards transform risk measure has $\psi(t) = r(1-t)^{r-1}$, and Wangs premium principle has $\psi(t) = g'(1-t)$, where g is an increasing convex function with derivatives over $[0, 1]$; see Jones and Zitikis (2003) for details. Other choices of the function ψ can be found in Jones and Zitikis (2007). Jones and Zitikis (2003) also introduced a related quantity to illustrate the right tail, left tail, and two sided deviations, which is defined as

$$r(F) = \frac{R(F)}{E(X)}. \tag{1.2}$$

Note that the general definition of distortion measures as mentioned in Wang and Young (1998) and Wirch and Hardy (1999) includes the two widely used risk measures: value-at-risk (VaR) and tail value-at-risk (T-VaR). However the class defined by (1.1) excludes the VaR.

In this paper, we focus on the evidence tests of the risk measure and its related quantity defined in (1.1) and (1.2), respectively.

Statistical inference for $R(F)$ and $r(F)$ plays an important role in the applications of risk measures. nonparametric estimation by replacing F^{-1} and $E(X)$ by the sample quantile function and sample mean, respectively, are proposed by Jones and Zitikis (2003) and he derived the asymptotic normality. Therefore, confidence intervals for $R(F)$ and $r(F)$ can be constructed via estimating the asymptotic variance. For comparing the two risk measures. Jones and Zitikis (2007) investigated the nonparametric estimation of the parameter associated with distortion-based risk measures. In this paper we investigate the possibility of applying an empirical likelihood method for constructing nonparametric tests for $R(F)$ and $r(F)$. The empirical likelihood method is a nonparametric likelihood approach for statistical inference, which has been shown to be powerful in interval estimation and hypothesis testing. We refer to Owen (2001) for an overview of the method. However, it is known that the empirical likelihood method is not effective in dealing with nonlinear functionals. Recently, Jing et al (2009) proposed a so called jackknife empirical likelihood method to deal with nonlinear functionals. **?** formulate a jackknife sample based on estimating the nonlinear functional and then apply the empirical likelihood method for a mean to the jackknife sample. Since the risk measure $R(F)$ and its related quantity $r(F)$ are nonlinear functionals, we propose employing the jackknife empirical likelihood method to use in evidence tests, for these two quantities. Note that the profile empirical likelihood method can be employed for some special risk measures such as VaR and T-VaR because one can simply linearize them; for a study of VaR and T-VaR see Baysal and Staum (2008). Interpreting observed data as evidence has a key role of the statistical analysis in science, that is assessing What do the data say? Although standard statistical methods (hypothesis testing, estimation, confidence intervals) are routinely used for this purpose, the theory behind those methods contains no defined concept of evidence and no answer to the basic question when is it correct to say that a given body of data represent evidence supporting one statistical hypothesis against another? (Royal (2000)).

Let $p(\mathbf{x},\theta)$ be the joint probability density function (pdf) of n iid observations from a distribution with pdf $f(x,\theta)$ then the likelihood ratio

$$\lambda = \frac{p(\mathbf{x},\theta_0)}{p(\mathbf{x},\theta_1)},$$

measures the strength of evidence favorable to the simple hypothesis $H_0 : \theta = \theta_0$ against the simple hypothesis $H_1 : \theta = \theta_1$, (Royal (2000)). Some measures of support of statistical hypotheses have been studied by Emadi and Arghami (2003) and Emadi et al (2007). Doostparast and Emadi (2006), Arashi and Emadi (2008) have studied some measures of support of statistical hypotheses based on independent and identically distribution (iid) observations and record statistics. Habibirad et al (2006) generalized the concept of expected true statistical evidence based on the law of likelihood. Hashempour (2017) studied the statistical evidences in lifetimes of dynamic r-out-of-n systems, which are modeled by sequential order statistics (SOS).

To produce statistical evidence for one hypothesis against another one can find a measure of performance of the experiments for the objective of the study . This can be defined as

$$S_\varphi(E) = E_{\theta_0}(\varphi(\lambda)) + E_{\theta_1}(\varphi(1/\lambda)),$$

where $\varphi(.)$ is a non decreasing function.

(i) If

$$\varphi(t) = \begin{cases} 1 & if\ t \geq k, \\ 0 & if\ t \leq k, \end{cases}$$

$S_\varphi(E)$ is the sum of the probabilities of observing strong true evidence under H_0 and H_1, where K is arbitrary and is usually between 8 and 32 (Royall, 1997). (ii) If $\varphi(t) = \frac{t}{1+t}$, then $S_\varphi(E) = abc(E)$ the area between the cdf curves (under H_0 and H_1). (Emadi and Arghami (2003). (iii) If $\varphi(t) = log(t)$, then

$$\begin{aligned} S_\varphi(E) &= E_{\theta_0}\left(\log\frac{p(\mathbf{x},\theta_0)}{p(\mathbf{x},\theta_1)}\right) + E_{\theta_1}\left(\log\frac{p(\mathbf{x},\theta_1)}{p(\mathbf{x},\theta_0)}\right) \\ &= D(p_{\theta_0},p_{\theta_1}) + D(p_{\theta_1},p_{\theta_0}) \\ &= J(p_{\theta_0},p_{\theta_1}) \end{aligned}$$

where $D(.)$, $J(.)$ are, respectively, asymmetric and symmetric Kullback-Leibler (KL) distance (information).

The rest of the paper is organized as follows. Section 2, proposes the jackknife empirical likelihood function based on data for statistical evidence tests. Optimal sample size that guarantees the evidence reaches a desired level is obtained in Section 3. Then we compute the average of the proposed criteria under different sample sizes in Section 4.

2 Evidence tests for risk measures

Consider the following hypotheses

$$H_0 : R = R_0 \text{ V.S. } H_1 : R = R_1$$

and

$$H_0 : r = r_0 \text{ V.S. } H_1 : r = r_1$$

Throughout, we assume that $X_1, ..., X_n$ are independent non negative random variables with continuous distribution function $F(x)$. Put $\Psi(t) = \int_0^t \psi(s)ds$. When $R(F) < \infty$, we have $t\Psi(1) - \Psi(F(t)) \to 0$ as $t \to \infty$. Thus we can write the risk measure defined in 1.1 as

$$R = R(F) = \int_0^\infty \Psi(1) - \Psi(F(t))dt.$$

The empirical distribution function as $F_n(x) = \frac{1}{n}I(X_j \leq x)$ is defined. Then Jones and Zitikis (2003) proposed estimating $R(F)$ and $r(F)$ by $\hat{R}_n = \int_0^\infty(\Psi(1) - \Psi(F_n(t)))dt$, and $\hat{r}_n = \frac{n\int_0^\infty(\Psi(1)-\Psi(F_n(t)))dt}{\sum_{i=1}^n X_j}$, respectively, and showed that $\sqrt{n}(\hat{R}_n - R) \to^d N(0, \sigma_1^2)$ and $\sqrt{n}(\hat{r}_n - r(F)) \to^d N(0, \sigma_2^2)$ under some regularity conditions, where $\sigma_1^2 = Q_F(\Psi, \Psi)$, $\sigma_2^2 = \frac{1}{\mu^2}(Q_F(\Psi, \Psi) - 2r(F)Q_F(\Psi, 1) + (r(F))^2 Q_F(1, 1))$ and $Q_F(a, b) = \int_0^\infty \int_0^\infty (F(x \wedge y) - F(x)F(y))a(F(x))b(F(y))dxdy$, where $a(.)$, $b(.)$ are two functions on $[0, 1]$.

Here, we apply the jackknife empirical likelihood method developed by Jing et al (2009). This procedure is easy to implement and is described as follows.

Define $F_{n,i} = \frac{1}{n-1}\sum_{j=1,j\neq i}^n I(X_j \leq x)$ and $\hat{R}_{n,i} = \int_0^\infty(\Psi(1) - \Psi(F_{n,i}(t)))dt$ for $i = 1, ..., n$. Then the jackknife sample is defined as

$$Y_i = n\hat{R}_n - (n-1)\hat{R}_{n,i}, i = 1, ..., n.$$

Now, we apply the empirical likelihood method to the above jackknife sample. That is, we define the jackknife empirical likelihood function for $\theta = R(F)$ as

$$L_1(\theta) = \sup\{\prod_{i=1}^n (np_i) : p_i \geq 0, \text{ for } i = 1, ..., n; \sum_{i=1}^n p_i = 1; \sum_{i=1}^n p_i Y_i = \theta\}.$$

By the Lagrange multiplier technique, we have $p_i = (\lambda_1 + \lambda_2 Y_i)^{-1}$, where $\lambda_1 = \lambda_1(\theta), \lambda_2 = \lambda_2(\theta)$ satisfies$\sum_{i=1}^{n} \frac{Y_i}{\lambda_1+\lambda_2 Y_i} = \theta, \sum_{i=1}^{n} \frac{1}{\lambda_1+\lambda_2 Y_i} = 1$.
Next, we consider the related quantity $r(F) = R(F)/\mu$, where $\mu = E(X_1)$. Alternatively, we consider the quantity $R - \theta\mu$ with $\theta = r(F)$. Then, this quantity can be estimated by $\hat{R}_n - \theta n^{-1} \sum_{i=1}^{n} X_i = \hat{R}_n - \theta \int_0^\infty x dF_n(x)$.
As before, we define the jackknife sample as

$$n(\hat{R}_n - \theta \int_0^\infty x dF_n(x)) - (n-1)(\hat{R}_{n,i} - \theta \int_0^\infty x dF_{n,i}(x)) = Y_i - \theta X_i.$$

for $i = 1, ..., n$, where the Y_i are defined as above. So the jackknife empirical likelihood function for $\theta = r(F)$ is defined as

$$L_2(\theta) = \sup\{\prod_{i=1}^{n}(np_i) : p_i \geq 0, \quad \text{for} \quad i = 1, ..., n; \quad \sum_{i=1}^{n} p_i = 1; \quad \sum_{i=1}^{n} p_i(Y_i - \theta X_i) = 0\}.$$

Statistical evidence is represented and interpreted by the law of likelihood and its strength is measured by the likelihood ratio. The law of likelihood explains that the strength of statistical evidence for one hypothesis over another is measured by their likelihood ratio, (**?**). By using (ii), as a measure of expected true statistical evidence, we use $S_\varphi(E)$, defined by Emadi and Arghami (2003).Utilizing (iii) we have $S_\varphi(E)$ measure as symmetric Kullback-Leibler (KL) distance (information) of p_1 and p_0 .

3 Optimal sample size

Suppose a random sample $X_1, ..., X_n$ are independent non negative random variables with continuous distribution function $F(x)$. Here, we try to get an optimal value for n by minimizing $P_D = \min\{D_1, D_2\}$ where D_1 and D_2 are decisive and correct evidences defined by

$$D_1 = P(\lambda \geq k | H_0 \text{is true})$$

and

$$D_2 = P(\lambda \leq 1/k | H_1 \text{is true})$$

As mentioned by De Santis (2004), a sample size that guarantees P_D reaches a desired level ξ , is often enough to also bound the probabilities of weak and misleading evidences. Hence, for chosen $\xi \in (0, 1)$ and k, we then need to solve the following optimization problem:

$$n = \min\{n : P_D \geq \xi\}.$$

As an illustration, Table 1 presents optimal sample size for some selected values of k, ξ. There are various choices for k. Following De Santis (2004), we considered $k = 3, 7, 8$ in Table 1. The optimal sample size is increasing in k and ξ .

Table 1: Optimal Sample size for $R(F), r(F)$ based on some different k, ξ, a

			Lognormal		Gamma		Weibull	
ξ	k	a	R	r	R	r	R	r
0.7	3	0.55	30	20	10	50	25	20
		0.85	30	20	10	35	35	55
	7	0.55	40	30	10	50	25	25
		0.85	35	30	10	35	45	55
	8	0.55	40	35	10	50	25	25
		0.85	45	35	10	35	40	60
0.8	3	0.55	45	35	15	150	30	40
		0.85	55	40	15	55	50	95
	7	0.55	55	40	15	140	40	40
		0.85	60	45	15	60	55	95
	8	0.55	60	45	15	155	35	40
		0.85	60	45	15	60	60	100
0.9	3	0.55	80	50	25	220	50	50
		0.85	80	55	20	110	70	130
	7	0.55	90	60	55	230	50	60
		0.85	95	60	20	115	80	135
	8	0.55	110	70	55	235	55	65
		0.85	110	70	20	115	95	155

4 Simulation study

In this section, we examine the finite-sample behavior of the proposed jackknife empirical likelihood method in terms of the values of the evident measures, and compare them. We focus on the proportional hazards transform risk measure with $\psi(s) = a(1-s)^{(a-1)}$ and choose $a = 0.55, 0.85$ for simulation. Since the Lognormal distribution, Weibull distribution, and Gamma distribution are widely used in fitting the losses data in insurance (see Klugman et al (2008), data is simulated from these distributions. We draw 10000 random samples of sizes $n = 10, 20, 30$. The results are presented in Tables 2 and 3. The results based on three criteria statistical evidences (i), (ii) and (iii) in Tables 2 and 3 show that simulated data support that H_0 hypothesis versus H_1. Evidences based on lognormal distribution for $R(F)$ is more than $r(F)$, However that is decreased based on Gamma and weibull distributions.

Table 2: Computed criteria for $R(F)$ based on some different n, a

		Lognormal		Gamma		Weibull	
n	a	abc	$J(p_{\theta_0}, p_{\theta_1})$	abc	$J(p_{\theta_0}, p_{\theta_1})$	abc	$J(p_{\theta_0}, p_{\theta_1})$
10.00	0.55	0.91	11.74	0.38	9.44	0.25	10.87
10.00	0.85	0.45	16.39	0.14	5.11	0.05	7.14
20.00	0.55	0.97	11.91	0.43	10.54	0.31	11.47
20.00	0.85	0.49	16.59	0.16	6.06	0.10	8.11
30.00	0.55	0.98	12.79	0.51	11.83	0.38	12.60
30.00	0.85	0.53	17.29	0.20	7.93	0.12	9.30

Table 3: Computed criteria for $r(F)$ based on some different n, a

		Lognormal		Gamma		Weibull	
n	a	abc	$J(p_{\theta_0}, p_{\theta_1})$	abc	$J(p_{\theta_0}, p_{\theta_1})$	abc	$J(p_{\theta_0}, p_{\theta_1})$
10.00	0.55	0.56	12.80	0.69	5.12	0.63	4.11
10.00	0.85	0.27	14.18	0.31	4.32	0.04	0.82
20.00	0.55	0.65	13.38	0.71	6.13	0.66	4.94
20.00	0.85	0.32	14.27	0.35	5.12	0.07	0.89
30.00	0.55	0.75	13.45	0.73	7.42	0.67	5.36
30.00	0.85	0.35	14.72	0.38	6.12	0.20	0.90

5 Real Data Analysis

Grzegorz et al (2005) analyzed auto-insurance bodily injury liability data that are given in Table 4. The results of Table 5 indicate that based on two criteria statistical evidence (i), (ii) and real data set support that H_0 hypothesis versus H_1.

Table 4: Auto-insurance bodily injury liability data

0.045	0.047	0.070	0.075	0.077	0.092	0.117	0.117	0.140	0.145
0.149	0.165	0.167	0.169	0.180	0.180	0.199	0.202	0.212	0.225
0.230	0.242	0.264	0.275	0.285	0.290	0.363	0.384	0.400	0.400
0.413	0.414	0.416	0.425	0.425	0.430	0.430	0.431	0.450	0.460
0.486	0.514	0.531	0.540	0.556	0.564	0.600	0.605	0.605	0.650

Conclusion

In this paper, we have proposed the jackknife empirical likelihood function based on data for statistical evidence tests. Also Optimal sample size that guarantees the evidence reaches

Table 5: Computed criteria for $R(F), r(F)$ based on some different a

	$R(F)$		$r(F)$	
a	abc	$J(p_{\theta_0}, p_{\theta_1})$	abc	$J(p_{\theta_0}, p_{\theta_1})$
0.55	1.00	13.98	0.49	16.13
0.85	0.94	12.23	0.34	10.87

a desired level is obtained. Then we computed the average of the proposed criteria under different sample sizes.

References

Arashi, M. and Emadi, M. (2008). Evidential inference based on record data and interrecord times. Statist. Papers, 49, 291-301.

Artzner, P., 1999. Application of coherent risk measures to capital requirements in insurance. North American Actuarial Journal 3, 1129.

Baysal, R.E., Staum, J., 2008. Empirical likelihood for value-at-risk and expected shortfall. Journal of Risk 11 (1), 332.

Castillo E, Hadi AS, Balakrishnan N, Sarabia JM (2005) Extreme value and related models with applications in engineering and science. Wiley, Hoboken

Chen, J., Peng, L., Zhao, Y., 2009. Empirical likelihood based confidence intervals for copulas. Journal of Multivariate Analysis 100, 137151.

Claeskens, G., Jing, B., Peng, L., Zhou, W., 2003. Empirical likelihood confidence regions for comparison distributions and ROC curves. The Canadian Journal of Statistics 31 (2), 173190.

Csorgo, M., Horvath, L., 1993. Weighted Approximations in Probability and Statistics. Wiley, Chichester.

De Santis F (2004) Statistical evidence and sample size determination for Bayesian hypothesis testing. J Stat Plan Inference 124:121144

Doostparast, M. Emadi, M. (2006). Statistical evidence methodology for model acceptance based on record values. J. Korean Statist. Soc., 35, 167-177.

Emadi, M., Ahmadi, J. Arghami, N. R., (2007). Comparing of record data and random observation based on statistical evidence. Statist. Papers, 48, 1-21.

Emadi, M. Arghami, N. R. (2003), Some measures of support for statistical hypotheses. J. Stat. Theory Appl., 2, 165-176.

Grzegorz A. Rempala PhD and Richard A. Derrig PhD (2005) Modeling hidden exposures in claim severity via the Em Algorithm, North American Actuarial Journal, 9:2, 108-128,

Habibirad, A., Arghami, N. R. and Ahmadi, J. (2006). Statistical evidence in experiments and in record values. Comm. Statist.-Theory Meth., 35, 1971-1983.

Hashempour, M. Evidences in lifetimes of sequential r-out-of n systems and optimal sample size determination for Burr XII populations, Statistics Opt. Inform. Comput., vol. 5, pp. 147-157, 2017.

Jing, B., Yuan, J., Zhou, W., 2009. Jackknife empirical likelihood. Journal of American Statistical Association 104, 12241232.

Jones, B.L., Zitikis, R., 2003. Empirical estimation of risk measures and related quantities. North American Actuarial Journal 7, 4454.

Jones, B.L., Zitikis, R., 2005. Testing for the order of risk measures: an application of L-statistics in an actuarial science. Metron LXIII, 193211.

Jones, B.L., Zitikis, R., 2007. Risk measures, distortion parameters, and their empirical estimation. Insurance: Mathematics and Economics 41, 279297.

Kaiser, T., Brazauskas, V., 2007. Interval estimation of actuarial risk measures. North American Actuarial Journal 10 (4), 249268.

Klugman, S.A., Panjer, H.H., Willmot, G.E., 2008. Loss Distributions: From Data to Decisions. Wiley.

Klugman, S.A., Panjer, H.H., Willmot, G.E., 2008. Loss Distributions: From Data to Decisions. Wiley.

Necir, A., Meraghni, D., 2009. Empirical estimation of the proportional hazard premium for heavy tailed claim amounts. Insurance: Mathematics and Economics 45, 4958.

Owen, A., 2001. Empirical Likelihood. Chapman and Hall, CRC.

Peng, L., Qi, Y., Wang, R., Yang, J., 2012. Jackknife empirical likelihood method for some risk measures and related quantities. Insurance: Mathematics and Economics 51 (1), 142 150

Royall, R. (2000). On the probability of observing misleading statistical evidence. J. Amer. Statist. Assoc., 95, 760-780.

Shorack, G.R., Wellner, J.A., 1986. Empirical Processes with Applications to Statistics. Wiley, New York.

Wang, S., 1998. An actuarial index of the right-tail risk. North American Actuarial Journal 2, 88101.

. Wang, S., Young, V.R., 1998. Ordering risks: expected utility theory versus Yaaris dual theory of risk. Insurance: Mathematics and Economics 22, 145161.

Wang, S., Young, V.R., Panjer, H.H., 1997. Axiomatic characterization of insurance prices. Insurance: Mathematics and Economics 21, 173183.

Wirch, J.L., Hardy, M.R., 1999. A synthesis of risk measures for capital adequacy. Insurance: Mathematics and Economics 25, 337347.

Yaari, M.E., 1987. The dual theory of choice under risk. Econometrica 55, 95115.

Comparison Between Bayesian Lower Bound and P-value

Abtahi, A. [1] Chinipardaz, R. [2]

[1]Department of mathematics and Statistics, Shiraz Branch, Islamic Azad university, Shiraz
[2]Department of Statistics, University of Shahid Chamran, Ahvaz,Iran

Abstract

For deciding a hypothesis, in significant test p-value and in Bayesian test Bayes factor or posterior probability of null hypothesis is used. It is well known that, in univariate case the posterior probability of null hypothesis is typically much larger than the p-value. It has been shown that the difference is depend on sample size, type of hypothesis, and prior distribution that was chosen. In multivariate case, the difference is also depend on dimension of the parameter, which is to be tested. Dependence of significance testing and Bayesian testing on sample size and specified prior distribution is criticized by some Bayesian researches. They suggested to compare p-value with lower bound of posterior probability of null hypothesis on a class of prior distribution. This paper reviews the problem of testing a point null hypothesis. Of interest is the relationship between p-value and Baysisan measures of evidence against the null hypothesis. The lower bound of the posterior of null hypothesis over three classes of prior distribution are derived and compared with the p-value.

Keywords: Bayes factor, Lower bound of posterior probability, P-value.

[1] Asieh-abtahi@yahoo.com
[2]

1 Introduction

p-values indicate the probability that departure of some magnitude from the null hypothesis occurs if the null is actually true. Bayes factors provide a measure of the strength of the evidence in favor/against the null. For testing any hypothesis about distributional parameters, Bayes factors are computed from the exact same input that is used to compute a p-value under null hypothesis significance testing plus, occasionally, the researchers choice of a prior distribution.

To conduct a Bayesian test, a prior distribution must be chosen. A reasonable prior, in the point null hypothesis is given by

$$g(\theta) = \pi_0 I_{\{\theta=\theta_0\}}(\theta) + (1-\pi_0) g_1(\theta) I_{\{\theta=\theta_0\}}(\theta) \tag{1.1}$$

where π_0 is prior probability of H_0 to be true, $I(.)$ is indicator function and $g_1(\theta)$ is a multivariate density function which describes how the prior mass is spread out over the alternative hypothesis. Using this representation, the weighted likelihood ratio or Bayes factor in support of H_0 against H_1 is given by

$$B(x) = \frac{L(\theta=\theta_0|x)}{\int_{\Theta_1} L(\theta|x) g_1(\theta) d\theta} = \frac{m(x|H_0)}{m(x|H_1)}, \tag{1.2}$$

where $m(x|H_0)$ and $m(x|H_1)$ are the marginal likelihood densities of X for H_0 and H_1, respectively. Thus the posterior probability of H_0 is obtained by

$$P(H_0|x) = [1 + \frac{1-\pi_0}{\pi_0} . \frac{1}{B(x)}]^{-1}. \tag{1.3}$$

A reasonable choice for π_0, would be 1/2, assigning equal prior probability to the two hypotheses. However there seems to be no agreeable objective density for g_1 which use by all Bayesian, even though the posterior probability is due to choice of g_1 is almost arbitrary. This criticized by classical statistician and is discussed (Berger and Perichi, (2001, 2004)).

The lower bound of the posterior probability of null hypothesis is given by

$$\underline{P}(H_0|x) = \inf_{g_1 \in G} P(H_0|x) = [1 + \frac{1-\pi_0}{\pi_0} . \frac{1}{\inf\limits_{g_1 \in G} B(x,G)}]^{-1}. \tag{1.4}$$

where G is the class of distribution that g_1 is belonging to it. This class of distribution should be larger enough to include all plausible densities, but not so large that includes unreasonable densities and also should be impartial towards H_0 and H_1.

There is a substantial literature on the controversy between Bayesian and Classical procedures for point null hypothesis testing.The most famous of these difficulties is the paradox discovered by Jeffreys (1939), Lindley (1957) and Bartlett (1957). This paradox arises when parameters from the prior distribution appear in the posterior odds ratio or Bayes factor, so that reasonable variations in the prior distribution (especially increasing or decreasing the prior variance) lead to substantial changes in the test results. This leads to difficulties in specifying scientifically objective prior distributions that could be widely accepted as appropriate for precise hypothesis testing. A large literature exists attempting to develop informative priors for precise hypothesis testing that mitigate the practical adverse effects of the mentioned paradox (see Berger and Perichi, (2001, 2004), Perichi, (2005)). That these effects can be large in practice, and no satisfactory solution has been found, is a serious problem for Bayesian hypothesis testing (Villa and Walker(2017)).

This paper, acording to the work of Chinipardaz and Abtahi(2008), compare p-value and Bayesian evidence in multivariate normal distribution. In section 2 the lower bound of posterior probability is considered when the prior distribution is belonged to one of three important classes. Section 3, draws conclusions. In section 4 the calibrated p-value is adapted to multivariate normal distribution and compared with lower bound.

2 Lower bound of the posterior probability in three classes

In this section, the lower bound of the posterior probability of null hypothesis over three classes of prior distribution are derived and compared with p-value.

2.1 All Distribution (G_A))

Theorem 1. *Let $\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}$ be a random sample from $N_d(\theta, \Sigma)$ (Σ is known). To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ when g_1 in G_A, then*

$$\inf_{\pi \in G_A} B(x, G_A) = \underline{B}(x, G_A) = exp\{-\frac{T(x)}{2}\}, \tag{2.1}$$

and

$$\underline{P}(H_0|x, G_A) = [1 + \frac{1-\pi_0}{\pi_0}.\frac{1}{\underline{B}(x, G_A)}]^{-1} = .[1 + \frac{1-\pi_0}{\pi_0}.exp\{-\frac{T(x)}{2}\}]^{-1}, \tag{2.2}$$

where G_A is the class of all distributions and $T(x)$ is the observed value of test statistic in significance test.

Proof. To proof, note that

$$\sup_{\pi \in G_A} \int f(x; \theta, \Sigma) g_1(\theta) d\theta = (\frac{2\pi}{n})^{-\frac{d}{2}} |\sum|^{-\frac{1}{2}},$$

Because $\bar{X}$ is the maximum likelihood estimator of θ. Then

$$\inf_{\pi \in G_A} B(x, G_A) = \underline{B}(x, G_A) = \frac{m(x|H_0)}{m(x|H_1} = exp\{-\frac{T(x)}{2}\}.$$

□

Table 1 gives some p-value and their corresponding bounds obtained from (6). The table show that for $d \geq 2$ the lower bound is smaller than p-value. It is completely different result with $d = 1$ which is given in Berger and Selke (1987). As d tends to be large the lower bound tends to be zero and yet with fix p-value. Indeed, this result is true for any regular sampling model and realistic classes of proir densities.

Table 1: Comparison between $\underline{P}(H_0|x, G_A)$ and p-value and callibration of p-value ($\pi_0 = 1/2$).

p-value	0.001	0.010	0.050	0.100
$[1 + (-eplogp)^{-1}]^{-1}$	0.0184	0.1113	0.2894	0.3849
$d = 1$	0.0044	0.0350	0.1278	0.2054
$d = 2$	0.0010	0.0099	0.0476	0.0909
$d = 5$	0.0000 5	0.0005 6	0.0039	0.0098
$d = 10$	0.0000	0.0000	0.0001	0.0003
$d = 25$	0.0000	0.0000	0.0000	0.0000
$d = 40$	0.0000	0.0000	0.0000	0.0000

Now let $T(x)$ be a observed valued test for which p-value is p, with some algebra the following lemma be obtained:

Lemma 1. *Under the condition of Theorem 2.1, lower bound is larger than p-value if and only if*

$$\chi^2_d(p) \leq ln(\frac{1-p}{p})^2, \tag{2.3}$$

where $\chi^2_d(p)$ is the upper p probability point for the chi-square distribution with d degrees of freedom.

Note that only the left-hand side of equation (7) is clearly depended on degree of freedom. Therefore, validity of (7) is depend on the dimension of x, In the case of $d = 1$, the equation is valid for

$$|z_p| \leq \sqrt{2ln(\frac{1-p}{p})}, p \leq \frac{1}{2}$$

where z_p is the upper p probability point for the normal distribution. The different inequality is given in theorem 2 of Berger and Selke (1987).

2.2 Symmetric and Unimodal Distribution

Although using G_A is simple and contains all density function, but it may also concludes unreasonable priors. A reasonable class of g_1 would be the class of G_{US}, unimodal, symmetric distributions about θ_0.

Theorem 2. *Suppose X is d-variate normal mean $\theta = (\theta_1, \theta_2, ..., \theta_d)'$ and identity covariance matrix. To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ we have*

$$\inf_{g_1 \in G_{US}} B(x, G_{US}) = \frac{exp\{-\frac{1}{2}(x-\theta_0)^2\}}{\sup\limits_{k} \frac{1}{v(k)} \int_{|\theta-\theta_0| \leq k} exp\{-\frac{1}{2}(x-\theta)^2 d\theta\}},$$

and

$$\underline{P}(H_0|x, G_{US}) = [1 + \frac{1-\pi_0}{\pi_0} . \frac{1}{\underline{B}(x, G_{US})}]^{-1}$$

where $v(k)$ is the volume of a ball of radius k.

Proof. See Delampady (1989). □

Table 2 gives some selected values of $\underline{P}(H_0|x, G_{US})$, for various d and for t corresponding to certain p-value.

Table 2: Comparison between $\underline{P}(H_0|x, G_A)$ and p-value and callibration of p-value ($\pi_0 = 1/2$).

p-value	0.001	0.010	0.050	0.100
$[1 + (-eplogp)^{-1}]^{-1}$	0.0184	0.1113	0.2894	0.3849
$d = 1$	0.0179	0.1093	0.2904	0.3916
$d = 2$	0.0141	0.0891	0.2582	0.3630
$d = 5$	0098	0.0761	0.2350	0.3391
$d = 10$	0.0093	0.0721	0.2264	0.3292
$d = 25$	0.0092	0.0692	0.2214	0.3235
$d = 40$	0.0091	0.0680	0.2183	0.3200

Note, the discrepency between p-value and the lower bound of posterior null hypothesis $\underline{P}(H_0|x, G_A)$ appears to be always larger than the corresponding p-value. Indeed, $\underline{P}(H_0|x, G_A)$ decreases but remains almost constant as the distribution d increase.

2.3 Normal Distribution

Theorem 3. *Suppose that in theorem 2.1, g_1 belongs to the class of distributions with density $N_d(0, \frac{1}{k}\sum)$. Then*

$$\underline{B}(x, G_{NOR}) = (\frac{T(x)}{d})^{\frac{d}{2}} exp\{\frac{d}{2} - T(x)\}$$

and

$$\underline{P}(H_0|x, G_{NOR}) = [1 + \frac{1-\pi_0}{\pi_0}(\frac{d}{T(x)})^{\frac{d}{2}} exp\{\frac{T(x)}{2} - \frac{d}{2}\}]^{-1}.$$

Proof. To proof, note that

$$\underline{B}(x, G_{NOR}) = (\frac{n+k}{n})^{\frac{d}{2}} exp\{-\frac{n^2}{2(n+k)}(x-\mu_0)' \sum^{-1}(x-\mu_0)\} = (\frac{n+k}{n})^{\frac{d}{2}} exp\{-\frac{n^2}{2(n+k)} T(x)\}$$

with respect to x, using some easily calculations we have $B(x, G_{NOR})$. Now, to minimize

$$\underline{B}(x, G_{NOR}) = \inf_k B(x, G_{NOR}) = (\frac{T(x)}{d})^{\frac{d}{2}} exp\{\frac{d}{2} - \frac{T(x)}{2}\}$$

and

$$\underline{P}(H_0|x, G_{NOR}) = [1 + \frac{1-\pi_0}{\pi_0}(\frac{d}{T(x)})^{\frac{d}{2}} exp\{\frac{T(x)}{2} - \frac{d}{2}\}]^{-1}$$

□

$\underline{P}(H_0|x, G_{NOR})$ computed for various d with $\pi_0 = 1/2$. The result is shown in Table 3. One can observe that for every d, $\underline{P}(H_0|x, G_{NOR})$ is larger than the corresponding p-value, similar to other classes, and it decreases as d increases and remain almost constant. In this class the difference between lower bound as posterior probability and corresponding p-value is much larger than the other classes. This is because G_{NOR} is more restricted than two other classes.

3 Comparison of the lower bounds

In this paper we computed lower bounds of posterior probability of null hypothesis over three classes of prior (G_A, G_{US}, G_{NOR}). Note that

Table 3: Comparison between $\underline{P}(H_0|x, G_{NOR})$ and p-value and calibration of p-value ($\pi_0 = 1/2$).

p-value	0.001	0.010	0.050	0.100
$[1 + (-eplogp)^{-1}]^{-1}$	0.0184	0.1113	0.2894	0.3849
$d = 1$	0.0235	0.1333	0.3213	0.4112
$d = 2$	0.0184	0.1113	0.2894	0.3850
$d = 5$	0.0144	0.0925	0.2596	0.3580
$d = 10$	0.0125	0.0835	0.2440	0.3434
$d = 25$	0.0112	0.0772	0.2331	0.3329
$d = 40$	0.0104	0.0730	0.2250	0.3250

$$\underline{P}(H_0|x, G_A) \leq \underline{P}(H_0|x, G_{US}) \leq \underline{P}(H_0|x, G_{NOR}).$$

The computed lower bounds and corresponding p-value is given in Figure1. In symmetric classes of priors (G_{US}, G_{NOR}), the lower bound of posterior probability of null hypothesis, like univariate case is larger than p-value and almost constant as d increases. However, if all priors are considered, the results would be contrary, i.e., p-value is larger than lower bound. In this case the lower bound tends to zero as d increases. The class G_A may be too large, leading to excessively small lower bounds. Notice that, this class includes many functional forms which probably leads to an excessive bias in favor of alternative hypothesis.

4 Creating Agreement between Bayesian and Significant test

P-values are commonly though to imply considerably great evidence against H_0 than is actually warranted. Sellke et al.(2001), using some examples claimed that, this result occurs because the alternative hypothesis is not considered in significance test. The most important conclusion achieved is the p-value should not be used directly and has to be calibrated. In this regard, Sellke er al. (2001) offered a calibration for p-value,p, given by

$$\alpha(p) = [1 + [-eplog(p)]^{-1}]^{-1}, p < 1/e. \tag{4.1}$$

In fact, they suggest $(-eplog(p))$ as an approximation value for lower bound of Bayes factor for H_0 to H_1. When $\pi_0 = 1/2$, $\alpha(p)$ given in (8) can be considered as prior distribution on the alternative hypothesis.

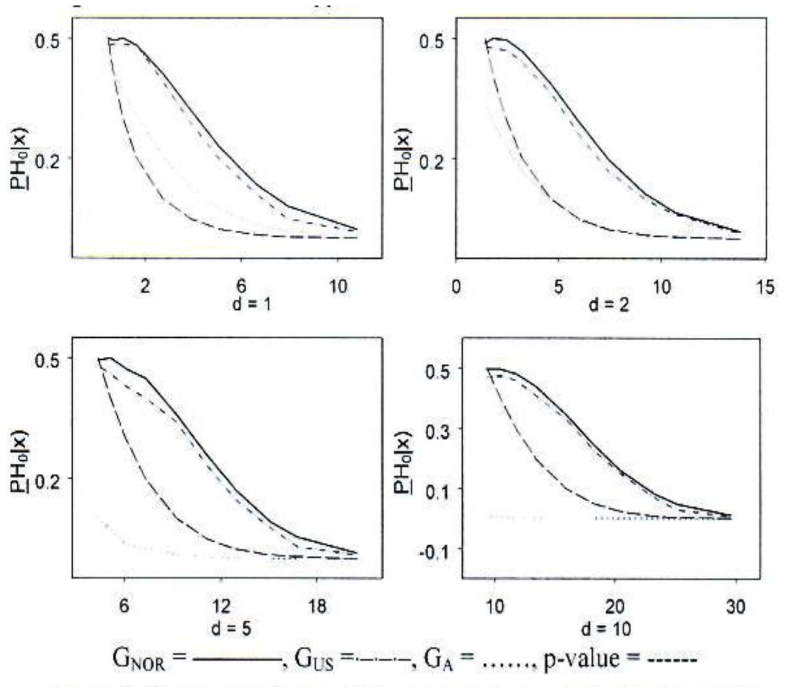

Figure 1: Comparison between the lower bound of posterior probability and corresponding p-value.

From the result given in section 3, the proposed calibration can be used for higher dimension (Table 1-3). From these tables one can see that the calibration seems to maintain very close to lower bound on G_{US}.

The difference between $\alpha(p)$ and $\underline{P}(H_0|x, G_{US})$ shows that, using calibrated p-value in significance testing and using the reasonable classes of priors for Bayesian testing may leads to an agreements between two approaches in multivariate normal distribution.

References

Bartlett, M. S. (1957), A Comment on D. V. Lindleys Statistical Paradox, *Biometrika* 44, 533-534.

Berger, J. O. and Perichi, L.R. (2004), Training Samples in objective Bayesian model selection, *In: Lahiri, P. (ed.), IMS Lecture Notes - Monograph Series*, 32:3, 841-869.

Berger, J. O. and Perichi, L.R. (2001), Objective Bayesian methods for model selection: Introduction and comparison, *Annals of Statistics*, 38, 135-207.

Berger, J. O. and Sellke, T. (1987), Objective Bayesian methods for model selection:

Introduction and comparison, *Journal of the American Statistical Association*, 82(397), 112-122.

Delampady, M. (1989), Lower bound on Bayes factors for invariant testing situation, *Journal of Multivariate Analysis*, 28,227-246.

Chinipardaz, R. and Abtahi, A. (2008), TTesting a point null hypothesis: the irreconcilability of p values and evidence, *Pakestanian journal statistics* 24(2) 123-133.

Jeffreys, H. (1939), *Theory of Probability,* Oxford University Press, Ox- ford. (1st ed. 1939, 2nd ed. 1961).

Lindley, M. S. (1957), A Statistical Paradox. Biometrika, *Biometrika* 44, 187-192.

Perichi, L.R. (2005), Model selection and hypothesis testing based on objective probabilities and Bayes factors, *Handbook of Statistics* 25, 115-149.

Sellke, T. and Bayarri,M.j. and Berger, J.O. (2001). Calibration of p-values for precise null hypotheses, *Amer. Statistician*,55, 62-71.

Villa, C. and Walker (2017), On the mathematics of the Jeffreys-Lindley paradox, *Communications in Statistics: Theory and Methods* 46:24, 12290- 12298.