TEST OF STUDENT'S T DISTRIBUTION USING KERNEL DENSITY ESTIMATION UNDER RSS

Sameer Ahmad Hassan Al-Subh
Department of Mathematics and Statistics, Mutah University, Karak, Jordan
Email: salsubh@mutah.edu.jo, salsubh@yahoo.com

ABSTRACT: In this paper, the null hypothesis is *a Student's t distribution is tested. A goodness of fit (gof) test statistics involving Kullback-Leibler information (KLI) which is found based on kernel density estimation is used. The performance of the test under ranked set sampling (RSS) agianst simple random sampling (SRS) is investigated. Several alternative distributions are considered under the alternative hypothesis. Based on a simulation, it is found that the test is more efficient under RSS than SRS for the distributions considered.*

Keywords: Goodness of fit test; Kullback-Leibler Information; Kernel density function; Student's t distribution; Ranked set sampling; Simple random sampling

.INTRODUCTION

In probability and statistics, Student's t distribution is any member of a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. It was developed by William Sealy Gosset under the pseudonym Student. The Student's t distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails and a bit shorter and fatter. This makes it useful for understanding the statistical behavior of certain types of ratios of random quantities, in which variation in the denominator is amplified and may produce outlying values when the denominator of the ratio falls close to zero.

McIntyre [12] introduced a sampling scheme called Ranked Set Sampling (RSS). RSS produces a sample which is more informative about the population of interest than simple random sampling (SRS). This technique can be described as follows. Select m random samples each of size m from the population of interest. From the i^{th} sample detect, using a visual inspection, determine the i^{th} order statistic and choose it for actual quantifications, say, Y_i, where $i = 1,...,m$. Assuming the ranking is perfect, RSS is the set of the order statistics $Y_1,...,Y_m$. The technique could be repeated r times to get more observations. The resulting measurements form an RSS of size rm. A comprehensive survey about developments in RSS can be found in [2, 3]).

Many works have been done for identifying certain distribution based on various gof test. A comprehensive survey for gof tests based on SRS can be found in [6]. Although many works have been carried out on gof test under RSS, the gof tests based on data collected via RSS technique and its modifications have not been given much attention in the literature. [9] proposed a method to improve the power of the chi-square test for gof based on RSS. They used the KLI measure to compare the data collected by both SRS and RSS. Also, they conducted a simulation study for the power of chi-square test of the method. [4] studied the empirical distribution function EDF GOF tests of Laplace distribution under Extreme Ranked Set Sample (ERSS).

This paper introduces a method for gof test which involves the use of KLI as obtained based on kernel density estimator [8 , 1]. Others [7], have proposed a method of finding the optimal bandwidth using the exact mean squared error (MSE) and mean integrated squared error (MISE) for estimation of normal densities.[10] has applied the kernel method when conducting gof test. Although kernel density estimator is often used to approximate the data distribution, its used for finding the KLI measure has not been explored.

This paper is organized as follows. In Section 2, we define the kernel density estimator and the selection of the optimal value of h and we define the gof test statistics involving KLI. Then, we apply the test on Student's t distribution using two algorithms to calculate the percentage points and the power function of the test at an alternative distribution. In Section 3, a simulation study is conducted to study the power and efficiency of this test statistics under RSS relative to SRS counterpart. We state our conclusions in Section 4.

1. MATERIAL AND METHODS

2.1 Kernel density estimation and bandwidth selection

Let $X_1, X_2,...,X_n$ be a random sample of size n from the distribution function $F(x)$ with unknown pdf $f(x)$. Then, the kernel density estimator of $f(x), x \in R$ is defined by [14] as

$$f(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \qquad (2.1)$$

where $K(.)$ is called the kernel function and h is called the bandwidth that controls the degree of smoothing applied to the data. We need to determine K and h to find the Kernel estimator. The kernel function K is usually assumed to be a symmetric function, such as in the case of student's t distribution. The following conditions are satisfied:

a. $\int_{-\infty}^{\infty} K(x)dx = 1,$ indicating that the kernel has a unit mass.

b. $\int_{-\infty}^{\infty} tK(t)dt = 0,$ indicating that the kernel has zero first moment.

(2.2)

c. $\int_{-\infty}^{\infty} t^2K(t)dt = k_2 \neq 0,$ and $k_2 < \infty,$

indicating that the kernel has a finite non-degenerate second moment.

The kernel method is widely used in nonparametric density estimation particularly for determining a kernel estimator for the unknown pdf $f(x)$ [13] pointed out that the choice of the bandwidth parameter h is crucial for an effective performance of the kernel estimator. Since the kernel estimator of pdf, $\hat{f}(x)$, depend on the choice of bandwidth, many methods have been suggested to determine the bandwidth. In our case, we define the value of h which minimizing the mean integrated square error $(MISE)$ given by [15]

$$MISE\left(\hat{f}(x)\right)=E\left\{\int_{-\infty}^{\infty}\left(\hat{f}(x)-f(x)\right)^2 dx\right\}=\int_{-\infty}^{\infty}\left(Bias\left(\hat{f}(x)\right)\right)^2 dx+\int_{-\infty}^{\infty}Var\left(\hat{f}(x)\right)dx,$$

where Bias

$$\hat{f}(x)=E\left(\hat{f}(x)\right)-f(x) \text{ and } Var\left(\hat{f}(x)\right)=E\left(\hat{f}^2(x)\right)-\left[E\left(\hat{f}(x)\right)\right]^2.$$

Substituting the value of the integrated square bias and the value of the integrated variance, then the asymptotic $MISE$ given by

$$AMISE=\frac{h^4}{4}k_2^2\int_{-\infty}^{\infty}f''^2(x)dx+\frac{1}{nh}\int_{-\infty}^{\infty}K^2(t)d(t).$$

We can obtain the optimal value of h, h_{opt}, (see [14]), by minimizing the $AMISE$ with respect to h to have

$$h_{opt}=k_2^{-2/5}\left\{\int_{-\infty}^{\infty}K^2(t)dt\right\}^{1/5}\left\{\int_{-\infty}^{\infty}f''^2(t)dt\right\}^{-1/5}n^{-1/5}, \tag{2.3}$$

where $k_2=\int_{-\infty}^{\infty}t^2K(t)dt, \quad 0<k_2<\infty.$

Note that $h_{opt}\to 0 \text{ as } n\to\infty$.

Since h_{opt} depends on the unknown pdf $f(x)$, $f(x)$ has to be estimated. The quantity $\int_{-\infty}^{\infty}f''^2(t)dt$ can be estimated by $\int_{-\infty}^{\infty}\hat{f}''^2(t)dt$.

2.2 Kullback-Leibler information (KLI)

We use the KLI number (see [11]) to test $H_o:F(x)=F_o(x)$ for all x against $H_1:F(x)\neq F_o(x)$ *for some* x. The information theory defines the KLI as follows. Let $f_0(x)$ and $f_1(x)$ be two density functions induced by two hypotheses, say H_0 and H_1 respectively. The KLI number of the two densities $f_0(x)$ and $f_1(x)$, denoted by $I(f_0,f_1)$, is given by

$$I(f_0,f_1)=\int_{-\infty}^{\infty}f_0(x)\text{Log}\frac{f_0(x)}{f_1(x)}dx. \tag{2.4}$$

The quantity $I(f_0,f_1)$ describes the amount of 'Information' lost for approximating $f_0(x)$ using $f_1(x)$. The larger value of $I(f_0,f_1)$ indicates the greater disparity between $f_0(x)$ and $f_1(x)$. It known that $I(f_o,f_1)=0$ if and only if $f_o(x)\equiv f_1(x)$ for all $x>0$. Hence a test for H_o *vs* H_1 can be designed as follows. Reject H_o *vs* H_1 if $I(f_o,\hat{f}_1)$ is large, where $\hat{f}_1(x)$ is the kernel density estimator of $f_1(x)$.

2.3 Testing for Student's t distribution

To test the hypothesis $H_o:F(x)=F_o(x)\ \forall x$ vs $H_1:F(x)\neq F_o(x)$ for some x where $F_o(x)$ is a Student's t distribution function. We consider two cases:

a) SRS Case:

Let

$$K(x)=f_o(x)=\frac{\Gamma\left((v+1)/2\right)}{(\pi v)^{0.5}\Gamma(v/2)\left[1+\left(x^2/v\right)\right]^{((v+1)/2)}}, -\infty<x<\infty, v>1.$$

$f(x)$ = a p.d.f under H_1,

$k_2=\int_{-\infty}^{\infty}x^2K(x).dx,$

then the bandwidth h can be found by

$$h=k_2^{-2/5}\left\{\int_{-\infty}^{\infty}K^2(x)dx\right\}^{1/5}\left\{\int_{-\infty}^{\infty}f''^2(x)dx\right\}^{-1/5}n^{-1/5}, \tag{2.5}$$

and the kernel density estimator can be obtained by

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}f_o((x-X_i)/h). \tag{2.6}$$

Then we defined test statistics T by incorporating the kernel density estimator in the KLI measure to have

$$\text{T}=\int_{-\infty}^{\infty}\hat{f}(x)Log\left(\frac{\hat{f}(x)}{f_o(x)}\right)dx, \tag{2.7}$$

We can reject H_o if $T>d_\alpha$,

where d_α is the $(1-\alpha)100$ th percentage point of the distribution of T under H_o.

b) RSS Case:

Let $Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_r^{(i)}$ be r iid i^{th} order statistics, $i=1, \ldots, m$. Thus, the pdf of $Y_j^{(i)}$ can be given by (see [5])

$$g_i(y)=\frac{m!}{(i-1)!(m-i)!}F^{i-1}(y)\left(1-F(y)\right)^{m-i}f(y).$$

An estimator for $g_i(y)$ can be obtained using the kernel estimator,

$$\hat{g}_i(y;h_i)=\frac{1}{rh_i}\sum_{j=1}^{r}K((y-Y_j^{(i)})/h_i). \tag{2.8}$$

Thus, the pdf under RSS can be estimated based on the kernel estimator as given by

$$\hat{f}_{RSS}(y)=\frac{1}{m}\sum_{i=1}^{m}\hat{g}_i(y;h_i). \tag{2.9}$$

The kernel function $K(y)$ can be chosen as follows. Let

$K(y) = f_o(y) =$ pdf of Student's t or

$$K(y) = \frac{m!}{(i-1)!(m-i)!} F_o^{i-1}(y)\left(1 - F_o(y)\right)^{m-i} f(y). \quad (2.10)$$

Then, the optimal value of h_i can be found by

$$h_i = k_2^{-2/5} \left\{ \int_{-\infty}^{\infty} K^2(t)dt \right\}^{1/5} \left\{ \int_{-\infty}^{\infty} \hat{f}_i^{''2}(t)dt \right\}^{-1/5} n^{-1/5}. \quad (2.11)$$

Hence we reject H_o if

$$T^* = \int_{-\infty}^{\infty} \hat{f}_{RSS}(y) Log\left(\frac{\hat{f}_{RSS}(y)}{f_o(y)} \right) dy, \quad (2.12)$$

is reasonably large.

A simulation is conducted to show that test statistics T^* is more powerful than the test statistics T when comparing samples of the same size under student's t distribution. The power of the T^* test statistics can be calculated according to the equation

Power of $T^*(H) = P_H(T^* > d_\alpha)$, (2.13)

where H is a cdf under the alternative hypothesis H_1 and d_α is the $(1-\alpha)100$ th percentage point of the distribution of T^* under H_o. We will calculate the efficiency of the test statistics as a ratio of powers given by

$$eff\,(T^*,\, T) = \frac{\text{power of } T^*}{\text{power of } T}. \quad (2.14)$$

Hence T^* is more powerful than T if $eff\,(T^*,\, T) > 1$.

2.4 Algorithm for Power Comparison

Let $v = 5$. To compare the powers of T^* and T; the following algorithm is designed to calculate the percentage points: Calculate h in formula (2.5).

1. Let $Y_1, \ldots, Y_r$ be a random sample from $F_o(y)$.
2. Calculate the formula (2.6).
3. Calculate the value of T as in (2.7).
4. Repeat the steps (1-4) 10, 000 times to get $T_1, \ldots, T_{10,000}$.
5. Determine the percentage point d_α of T which is given by the $(1-\alpha)100$ th quantile of $T_1, \ldots, T_{10,000}$.

Secondly, to calculate the power of T at H, we need to use simulation. So, we design the following algorithm:

1. Calculate h in formula (2.5).
2. Let $Y_1, \ldots, Y_r$ be a random sample from H, a distribution under H_1.
3. Calculate the formula (2.6).
4. Calculate the value of T as in (2.7).
5. Repeat the steps (1-4) 10, 000 times to get $T_1, \ldots, T_{10,000}$.
6. Calculate Power of $T(H) \approx \frac{1}{10,000} \sum_{t=1}^{10,000} I(T_t > d_\alpha)$, where $I(.)$ stands for indicator function.

2. RESULTS AND DISCUSSION

Based on a Monte Carlo simulation of 10,000 iterations, the power of each test is approximated according to the algorithm of Section 5. In the case of student's t distribution under RSS, we can't find the optimal bandwidth values. So, we used the same values of bandwidths as found in SRS case. We compared the efficiency of the tests for different samples sizes: $r = 5,\ 10,\ 15,\ 20,\ 25,\ 30$, set size: $m = 3$ and different alternative distributions: $Normal = N(0,1), Cauchy(0,1), Logistic = Lo(0,1),\ Student\ T = S(10),\ Extreme\ Value = Ext(0,1),\ Lognormal = Log(0,1), Chi - Square - chi(5), Beta = Be(1,3),$ $Gamma = G(1,2),$ $Weibul = W(1,2)\ and\ Exponential = E(5).$

The comparisons are made for the cases when the data are quantified via minimum, maximum and median. For Lognormal, Chi-Square , Beta, Gamma, Weibul and Exponential distributions, computations show that the efficiency of all tests equal one. The Simulation results are presented in the Tables (1)-(5).

Table 1. The values of h under SRS for $n = 5,\ 10,\ 15,\ 20,\ 25,\ 30$

H	SRS					
	n					
	5	10	15	20	25	30
$N(0,1)$	.615	.535	.493	.466	.446	.430
$C(0,1)$	.600	.522	.482	.455	.435	.419
$Lo(0,1)$	.952	.828	.764	.721	.690	.665
$ST(10)$	.619	.539	.497	.469	.449	.433
$Ext(0,1)$	.595	.518	.477	.451	.431	.416
log(0, 1)	.137	.120	.111	.104	.100	.096
$Chi(5)$	.099	.087	.080	.075	.072	.070
$Be(1,3)$	.220	.192	.177	.167	.159	.154
$G(1,2)$	1.035	.901	.831	.785	.750	.723
$W(2,2)$	.576	.501	.462	.436	.417	.402
$E(5)$	.104	.090	.083	.078	.075	.072

Table 2. The values of h under RSS for $r = 5,\ 10,\ 15, 20, 25, 30$

H	RSS						
	r						
		5	10	15	20	25	30
$N\,(0,1)$	Min	.423	.368	.339	.320	.306	.295
	Med	.676	.589	.543	.513	.490	.473
	Max	.423	.368	.339	.320	.306	.295
$C\,(0,1)$	Min	.520	.453	.417	.394	.377	.363
	Med	.746	.649	.599	.565	.541	.521
	Max	.520	.453	.417	.394	.377	.363
$Lo\,(0,1)$	Min	.693	.603	.556	.525	.502	.484
	Med	1065	.928	.855	.807	.772	.745
	Max	.693	.603	.556	.525	.502	.484
$ST\,(10)$	Min	.619	.539	.497	.469	.449	.433
	Med	.619	.539	.497	.469	.449	.433
	Max	.619	.539	.497	.469	.449	.433
$Ext\,(0,1)$	Min	.360	.313	.289	.272	.261	.251
	Med	.718	.625	.576	.544	.520	.501
	Max	.557	.485	.447	.422	.404	.390
$\log(0,1)$	Min	.082	.072	.066	.062	.060	.057
	Med	.345	.300	.277	.262	.250	.241
	Max	.411	.358	.330	.311	.298	.287
$Chi\,(5)$	Min	.060	.052	.048	.046	.044	.042
	Med	1.665	1.449	1.337	1.262	1.207	1.164
	Max	1.435	1.249	1.152	1.087	1.040	1.003
$Be\,(1,3)$	Min	.059	.051	.047	.044	.043	.041
	Med	.115	.100	.092	.087	.083	.081
	Max	.243	.212	.195	.184	.176	.170
$G\,(1,2)$	Min	.324	.282	.260	.245	.234	.226
	Med	.580	.505	.465	.439	.420	.405
	Max	.665	.579	.534	.504	.482	.465
$W\,(2,2)$	Min	.312	.271	.250	.236	.226	.218
	Med	.622	.541	.499	.471	.451	.435
	Max	.449	.390	.360	.340	.325	.313
$E\,(5)$	Min	.032	.028	.026	.025	.023	.023
	Med	.058	.050	.047	.044	.042	.041
	Max	.067	.058	.053	.050	.048	.047

Table 3. 5% Percentage points for SRS and RSS for $r = 5,\ 10,\ 15, 20, 25, 30,\ m = 3$ and $\alpha = 0.05.$

H	SRS						RSS					
	r											
	5	10	15	20	25	30	5	10	15	20	25	30
$N\,(0,1)$	.657	.389	.291	.240	.203	.180	.479	.304	.234	.198	.170	.156
$C\,(0,1)$	.671	.392	.302	.245	.209	.180	.441	.275	.217	.180	.159	.140
$Lo\,(0,1)$	.575	.375	.273	.225	.195	.171	.482	.316	.238	.197	.156	.155
$ST\,(10)$	.646	.385	.297	.245	.200	.182	.411	.261	.198	.162	.141	.126
$Ext\,(0,1)$	.664	.409	.303	.245	.207	.181	.474	.300	.234	.194	.170	.150
$\log(0,1)$	1.076	.705	.530	.446	.388	.351	.362	.215	.164	.137	.119	.103
$Chi\,(5)$	1.209	.817	.627	.528	.460	.404	.474	.341	.279	.228	.211	.195
$Be\,(1,3)$	.406	.271	.221	.190	.167	.148	.282	.145	.106	.081	.073	.056
$G\,(1,2)$	.507	.360	.272	.230	.206	.184	.259	.155	.118	.097	.083	.074
$W\,(2,2)$	.582	.371	.294	.245	.214	.194	.292	.178	.136	.113	.097	.088
$E\,(5)$	1.218	.808	.618	.514	.443	.405	1.801	1.72	.863	.735	.605	.522

Table 4. The values of Power of test under RSS and SRS for $n = r = 5,\ 10,\ 15, 20, 25, 30,\ m = 3$ **and** $\alpha = 0.05.$

H	SRS, $\alpha = 0.05.$						RSS					
	n						r					
	5	10	15	20	25	30	5	10	15	20	25	30
$N\,(0,1)$	.006	.003	.003	.004	.002	.003	.002	.002	.005	.013	.039	.069
$C\,(0,1)$	.361	.489	.574	.624	.668	.716	.673	.866	.945	.980	.990	1
$Lo\,(0,1)$	.266	.380	.495	586	.669	.737	.549	.791	.911	.965	.980	.995
$ST\,(10)$	.021	.012	.010	.008	.009	.008	.009	.005	.007	.007	.004	.006
$Ext\,(0,1)$	.109	.147	.201	.264	.332	.404	.232	.564	.825	.966	.993	1
$\log(0,1)$	1	1	1	1	1	1	1	1	1	1	1	1
$Chi\,(5)$	1	1	1	1	1	1	1	1	1	1	1	1
$Be\,(1,3)$	1	1	1	1	1	1	1	1	1	1	1	1
$G\,(1,2)$	1	1	1	1	1	1	1	1	1	1	1	1
$W\,(2,2)$	1	1	1	1	1	1	1	1	1	1	1	1

$E(5)$	1	1	1	1	1	1	1	1	1	1	1	1

Table 5. The efficiency of test using RSS relative to SRS for $r = 5,\ 10,\ 15, 20, 25, 30,$ $m = 3$ **and** $\alpha = 0.05.$

H	$\alpha = 0.05.$					
	r					
	5	10	15	20	25	30
$N(0,1)$	0.333	0.667	1.667	3.25	19.50	23
$C(0,1)$	1.864	1.771	1.646	1.571	1.482	1.397
$Lo(0,1)$	2.064	2.082	1.84	0.002	1.465	1.350
$ST(10)$	0.429	0.417	0.700	0.875	0.444	0.750
$Ext(0,1)$	2.128	3.837	4.104	3.659	2.991	2.475
$\log(0,1)$	1	1	1	1	1	1
$Chi(5)$	1	1	1	1	1	1
$Be(1,3)$	1	1	1	1	1	1
$G(1,2)$	1	1	1	1	1	1
$W(2,2)$	1	1	1	1	1	1
$E(5)$	1	1	1	1	1	1

From the above tables, we make the following remarks:

1. The bandwidths are decreasing as the sample size r increases for SRS and RSS methods.
2. The efficiencies in Table 5 are all greater than 1 except for Student's distribution (10), which means that the test statistics under RSS is more powerful than their counterparts in SRS.
3. The efficiency is decreasing as the sample size r increases.

3. CONCLUSION

We have introduced a test for gof when the data is collected via selective order statistics. This test statistics involves KLI measure which is found based on kernel density estimation. We found that the test introduced is more efficient under RSS than SRS for the distributions considered, i.e. the mean information per observation under RSS is larger than the mean information per observation under SRS.

REFERENCES

1. Akaike, H. (1954). An approximation to the density function. Ann. Inst. Statist. Math. (6): 127-132.
2. Alodat, M. T. & Al-Sagheer, O. A. (2007). Estimating the Location and the Scale Parameters using Ranked Set Sampling. J. App. Statis. Sci. Vol. 15(3), 245-252.
3. Alodat, M. T., Al-Rawwash, M. Y. & Nawajah, I. M. (2009). Analysis of Simple Linear Regression via Median Ranked Set Sampling. Metron International Journal of Statistics, LXVII, n.1, 1-18.
4. Al-Subh, S. A. (2018). Goodness of Fit Tests of Laplace Distribution Using Selective Order Statistics. International Journal of Applied Engineering Research,13(7):5508-5514.
5. Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (1992). A First Course in Order Statistics. John Wiley and Sons, New York.
6. D'Agostino, R.B & Stephens, M.A. (1986). Goodness of fit Techniques. Marcel Dekker Inc., New York.
7. Feyer, M. J. (1976). Some errors associated with the nonparametric estimation of density functions. J. Inst. Math. Appl., (18): 371-380.
8. Fix, E. & Hodges, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. Report No. 4, project no.21-29-004, USAF school of Aviation Medicine, Randolph Field, Texas.
9. Ibrahim, K. Alodat, M.T. Jemain, A.A. & Al-Subh, S.A. (2011). Chi-square test for goodness of fit for logistic distribution using ranked set sampling and simple random sampling. Statistica & Applicazioni, IX(2), 111-128.

10. Kim, C., Hong, C., Jeong, M. & Yang, M. (1997). Goodness-of-fit test for density estimation. Comm. Statist. Theory & Method., (26): 2725-2741.
11. Kullback, S. (1959). Information Theory and Statistics. New York, Wiley.
12. McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. Australian Journal of Agricultural Research, (3): 385-390.
13. Scott, D. W. (1992). Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York.
14. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
15. Wand, M. P and Jones, M. C. (1995). Kernel smoothing. Chapman and Hall, London.