University of Sri Jayewardenepura

Faculty of Graduate studies

Solutions to some statistical computing problems

This thesis is submitted by

L.W. Somathilake

As a partial fulfillment of the requirements

for the postgraduate

diploma in statistics

Approval

Name:- L.W. Somathilake.

Degree:-Postgraduate Diploma in Statistics

Examining committee:-

....

Dr.B.M.S.G. Bannaheka(supervisor)
Department of Statistics and Computer Science,
University of Sri Jayewardenepura ,
Nugegoda.

....

Dr. L.A.L.W. Jayasekara(supervisor)
Department of mathematics,
University of Ruhuna,
Matara.

R. A. Dayanand
....

Prof. R.A. Dayananda(Examinor)
Department of Statistics and Computer Science,
University of Sri Jayewardenepura ,
Nugegoda.

....

Mr. P Dias(Examinor)
Department of Statistics and Computer Science,
University of Sri Jayewardenepura ,
Nugegoda.

Date approved :-.....12/11/99..............

Acknowledgment

I wish to express my thanks to my supervisors Dr.B.M.S.G. Bannaheka , Department of Statistics and Computer Science, University of Sri Jayewardenepura and Dr. L.A.L.W. Jayasekara, Department of Mathematics, University of Ruhuna for their useful suggestions and invaluable guidance. I express my special thanks to Dr. L.A.L.W. Jayesekara , Department of Mathematics, University of Ruhuna for giving opportunities to use his own computer for the work of this project.

I would like to thank Mr. M.P.A. Wijayasiri , Head of the Department of Mathematics, University of Ruhuna for allowing me to use computer section of the Department of Mathematics during office hours and after the office hours. I would also like to thank staff members of the department of Mathematics, University of Ruhuna for their help throughout this study.

Contents

Chapter 1

Introduction

The main objective of this dissertation is to write computer programs to solve some statistical problems. We have discussed three problems in this dissertation.

1.1 Problem 1

Suppose we want to obtain shortest confidence intervals for some parameter. Consider the following situations.

1.1.1 Confidence interval for mean μ of a normal distribution

Suppose we want to obtain the 100(1-α)% confidence interval for mean μ of a normal distribution when both μ and σ^2 are unknown. Let s^2 be the sample variance of a sample of size n.

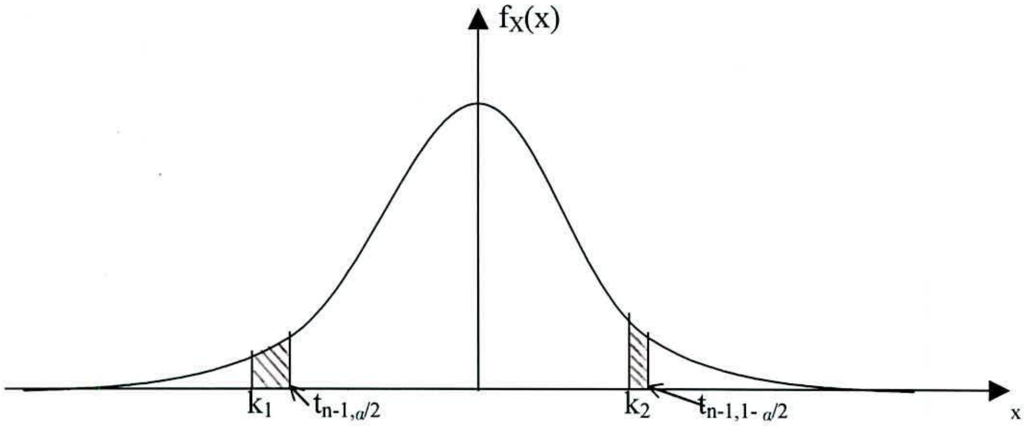
Then we know that $Q = \sqrt{n}\left(\frac{\overline{X}-\mu}{s}\right)$ is a pivotal quantity and Q~t_{n-1}. We can find infinite number of k_1 and k_2 values satisfying $\Pr(k_1 < \sqrt{n}\left(\frac{\overline{X}-\mu}{s}\right) < k_2) = 1-\alpha$

Then 100(1-α)% percent confidence interval for μ is $=\left(\overline{X}-\frac{s}{\sqrt{n}}k_2, \overline{X}-\frac{s}{\sqrt{n}}k_1\right)$

Length of this interval is$=\frac{s}{\sqrt{n}}(k_2-k_1)$

Tail symmetric interval is $=\left(\overline{X}-\frac{s}{\sqrt{n}}t_{n-1,\alpha/2}, \overline{X}+\frac{s}{\sqrt{n}}t_{n-1,\alpha/2}\right)$

Length of this interval is$=\frac{2s}{\sqrt{n}}t_{n-1,\alpha/2}$



The t-distribution is a symmetric one. Therefore as well we can show that tail symmetric confidence interval is the shortest interval.

Suppose $\left(\overline{X}-\frac{s}{\sqrt{n}}k_2, \overline{X}-\frac{s}{\sqrt{n}}k_1\right)$ is an another 100(1-a)% confidence interval. Then we have to consider two cases.

Case-1:- $k_1 < -t_{n-1,\alpha/2}$

Case-2:- $k_1 > -t_{n-1,\alpha/2}$

If $k_1 < -t_{n-1,\alpha/2}$, then $k_2 < t_{n-1,\alpha/2}$ and $\left|-t_{n-1,\alpha/2}-k_1\right| > \left|t_{n-1,\alpha/2}-k_2\right|$ (see figure 1)

These imply $-t_{n-1,\alpha/2}-k_1 > t_{n-1,\alpha/2}-k_2$,

i.e. $k_2 - k_1 > 2t_{n-1,\alpha/2}$

So $\frac{s}{\sqrt{n}}(k_2-k_1) > \frac{2s}{\sqrt{n}}t_{n-1,\alpha/2}$

Therefore, tail symmetric confidence interval for μ is shorter than any other confidence interval for μ.

Similarly we can show that tail symmetric confidence interval for μ is shorter than any other confidence interval for μ in case-2 also.

1.1.2. Confidence interval for variance of normal distribution

Suppose we want to obtain 100(1-a)% confidence interval for variance σ^2 of a normal distribution. Let s^2 be the sample variance of a sample of size n.

Then we know that $Q = \dfrac{(n-1)s^2}{\sigma^2}$ is a pivotal quantity and $Q \sim \chi^2_{n-1}$. Infinite number of k_1 and k_2 values can be obtained satisfying $\Pr\left(k_1 < \dfrac{(n-1)s^2}{\sigma^2} < k_2\right) = 1-\alpha$.

Then 100(1-a)% confidence interval for variance $\sigma^2 = \left(\dfrac{(n-1)s^2}{k_2}, \dfrac{(n-1)s^2}{k_1}\right)$

100(1-a)% tail symmetric confidence interval for σ^2 is $= \left(\dfrac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \dfrac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right)$

Let a=0.1, and n=16 then the 90% tail symmetric interval is $= \left(\dfrac{15s^2}{\chi^2_{15,0.05}}, \dfrac{15s^2}{\chi^2_{15,0.95}}\right)$

Length of tail symmetric interval $= 15s^2\left(\dfrac{1}{\chi^2_{15,0.95}} - \dfrac{1}{\chi^2_{15,0.05}}\right) = 15s^2\left(\dfrac{1}{7.26} - \dfrac{1}{25}\right) = 1.4661s^2$

$\left(\dfrac{15s^2}{\chi^2_{15,0.075}}, \dfrac{15s^2}{\chi^2_{15,0.975}}\right)$ is also a 90% confidence interval for σ^2.

Length of this interval $= 15s^2\left(\dfrac{1}{\chi^2_{15,0.975}} - \dfrac{1}{\chi^2_{15,0.075}}\right) = 15s^2\left(\dfrac{1}{7.97} - \dfrac{1}{27.5}\right) = 1.3366s^2$

Therefore tail symmetric confidence interval is longer than the other confidence interval. That is tail symmetric confidence interval is not the shortest confidence interval for σ^2.

1.1.3 Confidence interval for proportion of two variances of normal distributions.

Let σ_1^2 and σ_2^2 be variances of two normal distributions(say population 1 and population 2) . Let m and s_1^2 be the sample size and sample variance of a sample from population 1. Let n and s_2^2 be the sample size and sample variance of a sample from population 2. Then we know that $\dfrac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \sim F^{m-1}_{n-1}$

Suppose we want to find the 100(1-α)% confidence interval for σ_2^2/σ_1^2. Then we can find infinite number of k_1 and k_2 values satisfying $\Pr(k_1 < \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} < k_2) = 1-\alpha$.

Let α=0.1, m=16 and n =21
Then the 90% tail symmetric confidence interval for σ_2^2/σ_1^2 is given by

$$\left(F_{15,0.95}^{20}\frac{s_2^2}{s_1^2}, F_{15,0.05}^{20}\frac{s_2^2}{s_1^2} \right)$$

Length of this interval is $= \frac{s_2^2}{s_1^2}\left(F_{15,0.05}^{20} - F_{15,0.95}^{20}\right) = \frac{s_2^2}{s_1^2}(2.3275 - 0.4539) = 1.8736\frac{s_2^2}{s_1^2}$

$\left(F_{15,0.925}^{20}\frac{s_2^2}{s_1^2}, F_{15,0.025}^{20}\frac{s_2^2}{s_1^2} \right)$ is also a 90% confidence interval for σ_2^2/σ_1^2

Length of this interval is $= \frac{s_2^2}{s_1^2}\left(F_{15,0.025}^{20} - F_{15,0.925}^{20}\right) = \frac{s_2^2}{s_1^2}(2.089 - 0.3885) = 1.6955\frac{s_2^2}{s_1^2}$

Therefore tail symmetric confidence interval is longer than other confidence interval. That is tail symmetric confidence interval is not the shortest confidence interval for σ_2^2/σ_1^2.

According to the above examples tail symmetric confidence interval is not the shortest confidence interval for some parameters. Therefore we have written computer programs to obtain the shortest confidence intervals for such parameters. Details of the method that I have used is contained in Chapter-2. The corresponding computer program is contained in Appendix-2.

1.2 Problem 2

In one-way and two-way analysis of variance multiple comparison of mean can be done using Tukey's test. Tukey's test is available in Minitab for multiple comparison of one way analysis of variance. But Tukey's test for multiple comparison of two-way analysis of variance is not available in Minitab. Therefore we have written Minitab macros for the multiple comparison of two-way analysis of variance. Details of multiple comparison and Tukey's test is contained in Chapter-3. Corresponding Minitab program is contained in Appendix-1.

1.3 Problem 3

There is a newly developed test (Leslie Jayasekara and Takashi Yanagawa, 1994) to compare distributions of two independent categorical variables. This test is applicable for categorical data. It has been shown that the power of this test is higher than that of the other two-sample tests(Pearsion Chi squared test, Nair's location test, Nairs dispersion test, Cumulative Chi squared test). Calculation of the test statistic of Q_t test manually is tedious. There is no computer program available for this yet. Therefore I have written a computer program for the Q_t test. Details about Q_t test is contained in Chapter-4 and corresponding computer program is contained in Appendix-2.

Chapter 2

Shortest Confidence Interval for some parameters

2.1 Confidence interval

Let $X_1, X_2, \ldots X_n$ be a random sample from the density $f_X(x, \theta)$. Let $T_1=g(X_1, X_2, \ldots X_n)$ and $T_2=h(X_1, X_2, \ldots X_n)$ be two statistics satisfying $T_1< T_2$ for which $\Pr(T_1< \tau(\theta)<T_2)=\gamma$ where γ does not depend on θ. The random interval (T_1, T_2) is called a 100γ percent confidence interval for $\tau(\theta)$. γ is called the confidence coefficient. T_1 and T_2 are called the upper and lower confidence limits respectively

Suppose that $x_1,x_2 \ldots x_n$ is an observed sample from $f_X(x,\theta)$ and let $t_1=g(x_1, x_2, \ldots x_n)$ and $t_2=h(x_1, x_2, \ldots x_n)$. Then the observed interval (t_1, t_2) is also called a 100γ percent confidence interval for $\tau(\theta)$.

2.2 Shortest Confidence Interval

Suppose we are interested in finding the shortest confidence interval for some parameter(say θ). Let $X_1, X_2, \ldots X_n$, be a random sample from the density $f_X(x, \theta)$. Let $Q=q(X_1, X_2, \ldots X_n, \theta)$ be a pivotal quantity. Suppose that the distribution of the pivotal quantity is known (say $f(q)$). If Q is a linear function of θ and if the distribution of Q is symmetric, tail symmetric confidence intervals lead to shortest confidence interval for θ. But if the distribution of Q is skewed, we can not use the tail symmetric confidence limits to obtain shortest interval for θ. In this dissertation, we develop necessary tools to obtain shortest confidence intervals in such cases. Here we derive shortest confidence limits for population variance and ratio of population variances of Normal distributions.

2.2.1 Method of finding the shortest confidence interval

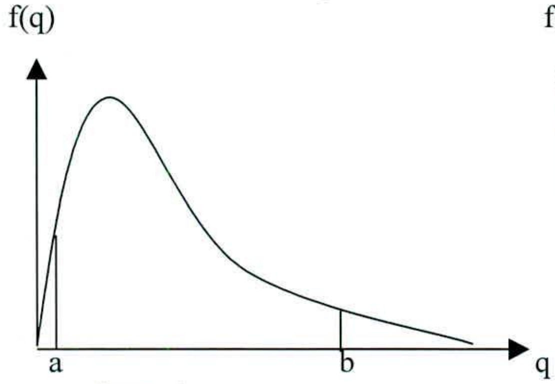


figure 1

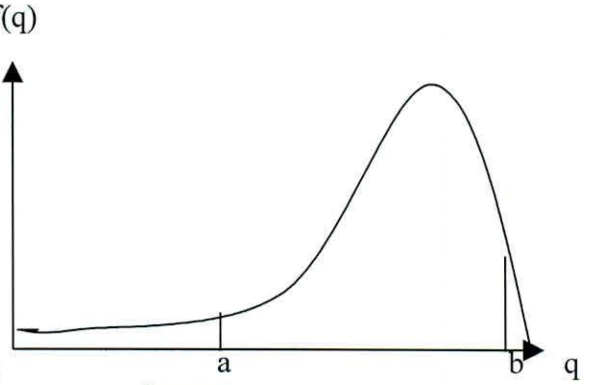


figure 2

Suppose we want to find a confidence interval for a parameter θ, using a pivotal quantity Q which has non-symmetric distribution.

Suppose that

(i) Q is linear in θ

(ii) $\theta \geq 0$

(iii) The distribution of Q is smooth and unimodel (as either in figure 1or 2)

Case I

Suppose the distribution is right skewed as in figure -1.

Let f be the density function of Q and, a and b be the values such that $\Pr(Q<a)=\alpha/2$ & $\Pr(Q>b)=\alpha/2$.

Then $f(a)>f(b)$. $\longrightarrow$ (1)

Let $a_1 = a-h$ (Where h is a small positive value) and select b_1 Such that

$$1-\alpha = \int_{a_1}^{b_1} f(q)dq$$

Then ,

$(f(a_1)+f(a))(a-a_1)/2=(f(b_1)+f(b))(b-b_1)/2 \longrightarrow (2)$ as $h \to 0$

If $f(a_1)>f(b_1)$, then

(1) and (2) imply that $a-a_1<b-b_1$

i.e. $b_1-a_1<b-a$

So, if $f(a_1)>f(b_1)$ then taking $a=a_1$ and $b=b_1$ repeat the above process until $f(a_1)-f(b_1)<0.000001$.

Case 2

Suppose distribution of Q is left skewed as in figure -2.

Let a and b be the values such that $\Pr(Q<a)=\alpha/2$ & $\Pr(Q>b)=\alpha/2$.

Then $f(a)<f(b) \longrightarrow (3)$

Let $a_1 = a+h$ (Where h is a small positive value) and select b_1 Such that

$$1-\alpha = \int_{a_1}^{b_1} f(q)dq$$

Then,

$(f(a_1)+ f(a)) (a_1 -a)/2= (f(b_1)+ f(b)) (b_1 -b)/2, \longrightarrow(4)$ as $h \longrightarrow 0$

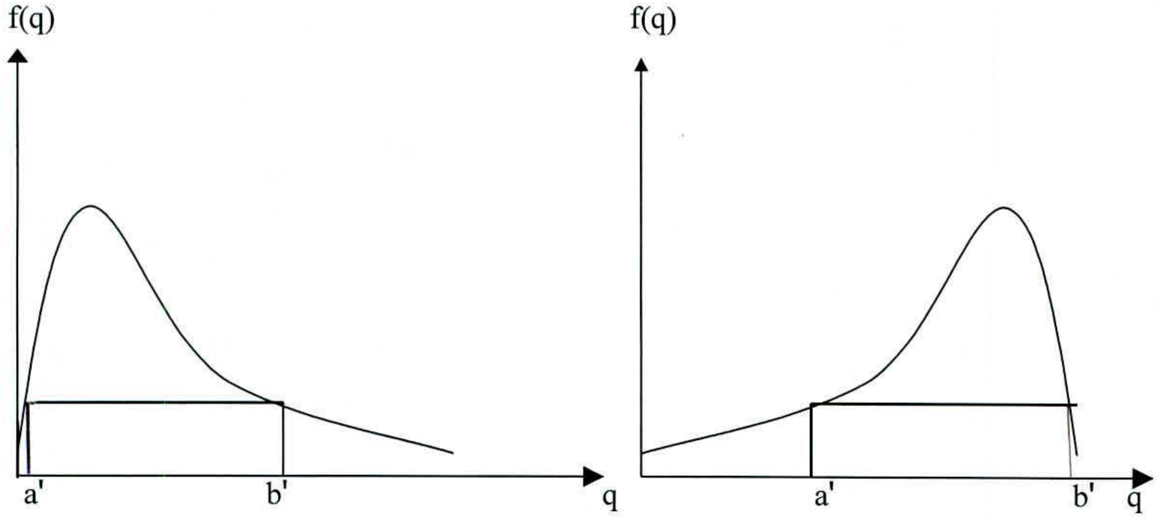
If $f(a_1)< f(b_1)$, Then (3) & (4) imply that $b_1 -b< a_1 -a$

i.e. $b_1 - a_1 < b-a$

So, if $f(a_1)< f(b_1)$,

then taking $a= a_1$ and $b= b_1$ repeat the above process until $f(a_1)- f(b_1)<0.000001$.

According to above Case-1 and Case-2 the new interval (a_1 , b_1) is shorter than tail symmetric interval (a, b).



We have to show that (a', b') is the shortest $100(1-\alpha)$ percent interval.

To show this assume that there is another $100(1-\alpha)$ percent interval (say (a_2 , b_2)) which is shorter than above interval. Then $b_2 - a_2 < b'- a'$

Then $a_2 < a'$ or $a_2 > a'$

So we have to consider four cases.

Case i :- $a_2 < a'$ and left skewed distribution

Case ii :- $a_2 < a'$ and right skewed distribution

Case iii :- $a_2 > a'$ and left skewed distribution

Case iv :- $a_2 > a'$ and right skewed distribution

Case-i

In this case $f(a_2) < f(b_2)$ since $f(b') = f(a')$

$$\text{i.e. } f(a_2) + f(a') < f(b_2) + f(b') \longrightarrow \text{(a)}$$

Also

$$(f(a_2) + f(a'))\,(a' - a_2))/2 = (f(b_2) + f(b'))\,(b' - b_2)/2$$

$$\text{i.e. } (f(a_2) + f(a'))\,(a' - a_2)) = (f(b_2) + f(b'))\,(b' - b_2) \longrightarrow \text{(b)}$$

From (a) and (b) we have $a' - a_2 > b' - b_2$

i.e. $b_2 - a_2 > b' - a'$

This is a contradiction. Similarly considering cases (ii), (iii) , and (iv)we can show that (a' , b') is the shortest 100(1-α)% interval. If Q is linear in θ, then we can obtain the shortest 100(1-α)% confidence interval for θ using (a', b') interval.

2.3. Shortest confidence interval for the variance of normal distribution

Suppose we want to obtain the shortest confidence interval for variance σ^2 of a normal distribution. Let s^2 be the sample variance of a sample of size n.

Let $X = \dfrac{(n-1)s^2}{\sigma^2}$

Then $X \sim \chi^2_{n-1}$ and so, it is a pivotal quantity. But this X is not linear in σ^2. Therefore. The above method cannot be used directly to construct the shortest 100(1-α)% confidence interval for σ^2.

Therefore let us consider the transformation Y=1/X. This transformation is one to one onto transformation ((0,∞) onto (0,∞)).

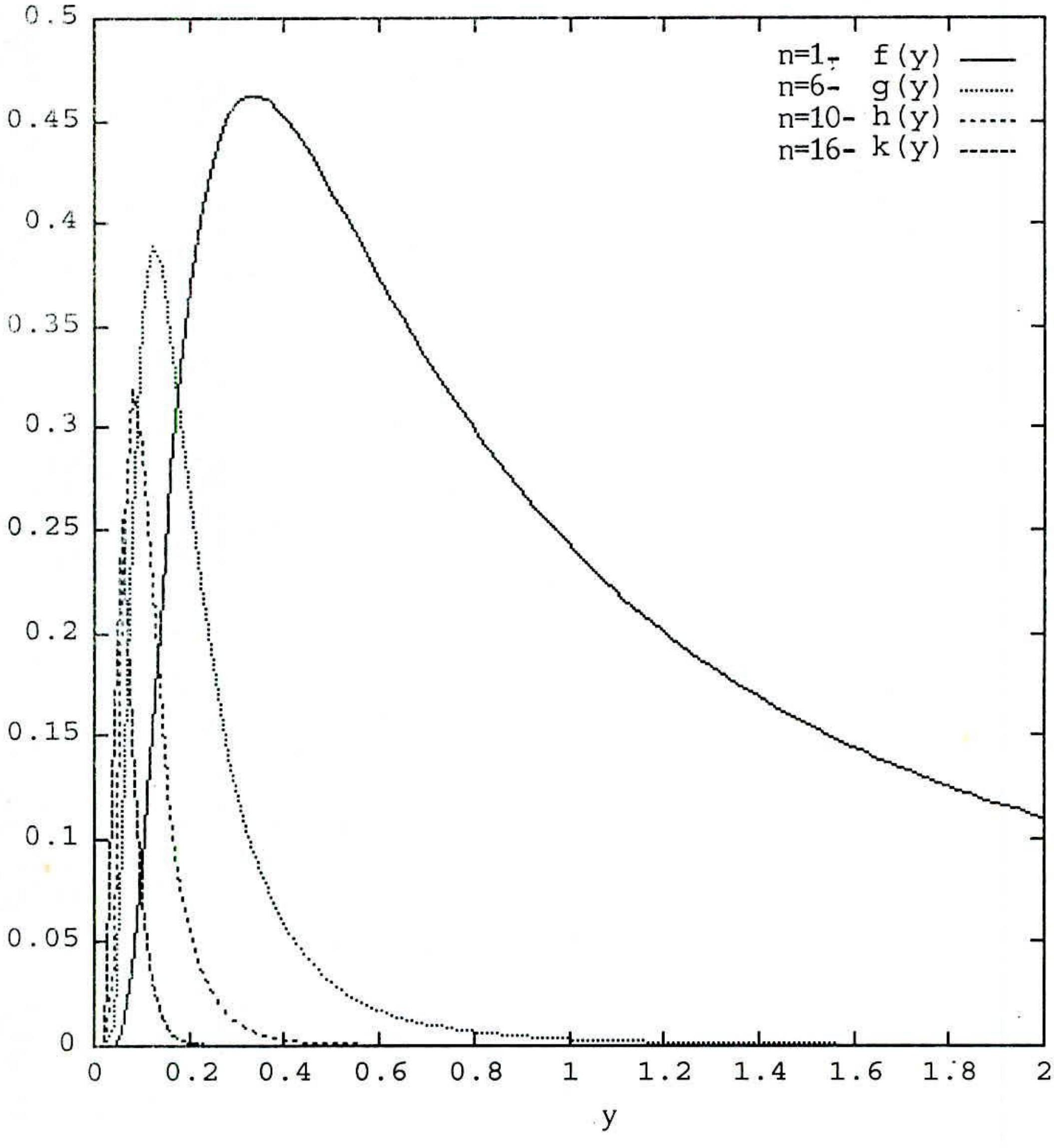
Therefore $f_Y(y) = f_X(x(y)) \left|\dfrac{dx}{dy}\right|$

Therefore $f_Y(y) = f_X(x(y))\left|\frac{dx}{dy}\right|$

Since $f_X(x) = \dfrac{x^{(\frac{n}{2}-1)} e^{-\frac{x}{2}}}{\Gamma(n/2)2^{\frac{n}{2}}}$ and $\left|\dfrac{dx}{dy}\right| = \dfrac{-1}{y^2}$,

$$f_Y(y) = \frac{e^{\frac{-1}{2y}}}{\Gamma(n/2)2^{\frac{n}{2}} y^{(\frac{n}{2}+1)}}$$

Then $f_y(y)$ is a skewed distribution. Graphs of $f_y(y)$ for different n values are as follows.



Since Y is linear in σ^2 we can obtain shortest γ% interval for σ^2 using above method taking Y as our pivotal quantity. Let (k,k_2)be the shortest γ% interval.

Then

$\Pr(k_1<Y<k_2)=\gamma$

$$\Pr(k_1 < \frac{\sigma^2}{(n-1)s^2} < k_2) = \gamma$$
$$\Pr\left((n-1)s^2 k_1 < \sigma^2 < (n-1)s^2 k_2\right) = \gamma$$

So $((n-1)s^2 k_1, (n-1)s^2 k_2)$ gives the shortest γ% confidence interval for σ^2.

Let k_1 and k_2 corresponding to sample size n and confidence level γ be denoted by $l_{n,\gamma}$ and $r_{n,\gamma}$.(l to denote left and r to denote right).

Then, the γ percent shortest confidence limits for σ^2 is $\left((n-1)s^2 l_{n,\gamma}, (n-1)s^2 r_{n,\gamma}\right)$

The values of by $l_{n,\gamma}$ and $r_{n,\gamma}$ are given in table 1 for different values of n and γ.