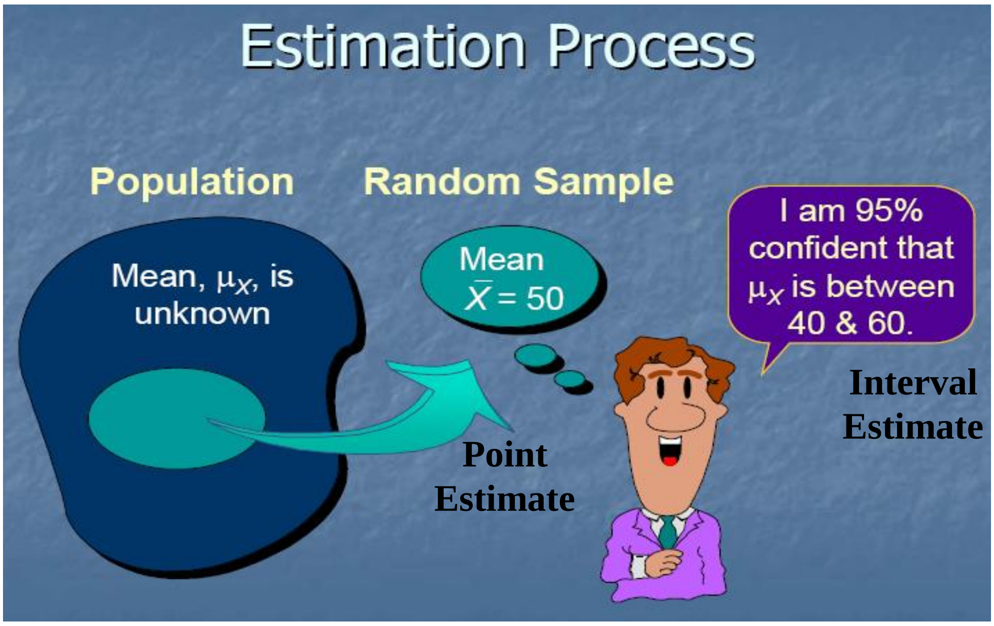
Lecture 12:Estimation

Contents:

12.0 Introduction



Estimation Process

Population

Random Sample

Mean, μ_X, is unknown

Mean $\overline{X} = 50$

I am 95% confident that μ_X is between 40 & 60.

Point Estimate

Interval Estimate

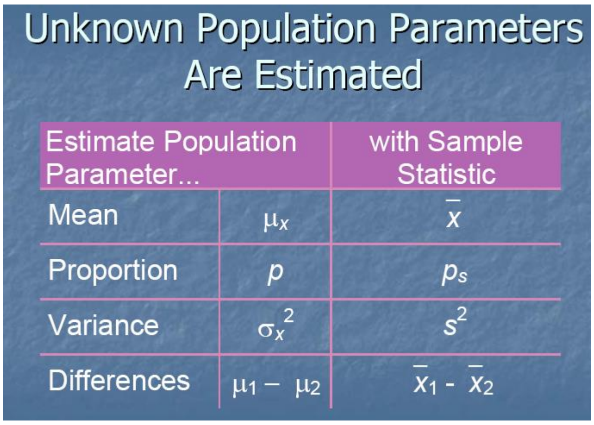
When we use the value of a statistic to estimate a population parameter, we call this point estimation, and the value of the statistic is referred as a point estimate of the parameter.

If one can infer a population parameter with stating the confidence of estimation, namely $(\overline{X} \pm \varepsilon)$ then it is called interval estimation.

12.1. Basics of Point Estimation

A point estimator consists of a single sample statistic that estimates an unknown population parameter.

For instance, the sample mean $\overline{x}$ is a ***point estimator*** of population mean μ, the sample variance s^2 is a ***point estimator*** of population variance σ^2, $\hat{p}$ is a ***point estimator*** of p.



Unknown Population Parameters Are Estimated

Estimate Population Parameter...		with Sample Statistic
Mean	μ_X	$\overline{x}$
Proportion	p	p_s
Variance	σ_X^2	s^2
Differences	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$

Example 1.

Consider the following observations on data packet loss (in percent) for a TCP/IP traffic.
24.46, 25.61, 26.25, 26.42, 26.66, 27.15, 27.31, 27.54, 27.74, 27.94, 27.98, 28.04, 28.28, 28.49, 28.50, 28.87, 29.11, 29.13, 29.50, 30.88

Assume that these samples are taken from a normal distribution.

We can have the following "point estimates" for the population mean μ:

(i) Estimator =sample mean $\bar{X}$,

estimate $\bar{x} = \sum x_i / n = 555.86/20 = 27.793.$

(ii) Estimator = sample median $\tilde{X}$,
estimate = $\tilde{x}$ = (27.94+27.98)/ 2 = 27.960

(iii) Estimator = average of two extreme values
=[min(X_i)+max(X_i)]/2 = (24.46+30.88)/2 = 27.670

(iv) Estimator = 10% trimmed mean = (discard the smallest and largest 10% of the sample and then average):

$$\bar{x}_{tr(10\%)} = \frac{555.86-(24.46+25.61+29.50+30.88)}{16} = 27.838$$

Each one of the estimators (i)-(iv) uses different measure of the center of the sample to estimate μ. **The QUESTION is which of the estimates is closest to the true value?**

Example 2

Consider the following sample of observations on round-trip time for a TCP/IP data traffic (in milliseconds):
44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1
We want to estimate the population variance σ^2.

A natural estimator is the sample variance:

$$\hat{\sigma}^2 = s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{\sum X_i^2 - (\sum X_i)^2 / n}{n-1}$$

The corresponding estimate is

$$\hat{\sigma}^2 = s^2 = \frac{\sum x_i^2 - (\sum x_i)^2 / 8}{7} = 0.251$$

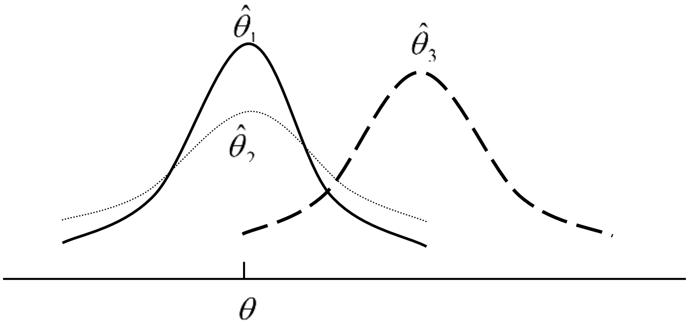
The estimator of σ would then be $\hat{\sigma} = s = \sqrt{0.251} = 0.501$.

Unbiased estimator and accuracy of estimation

If one writes $\hat{\theta} = \theta +$ error of estimation, then an **accurate estimator** would be the one

resulting in small estimation errors, so that estimated values will be near the true value.
A statistic $\hat{\theta}$ is an ***unbiased estimator*** of the parameter θ if and only if $E(\hat{\theta}) = \theta$ for every possible value of θ. If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.

If we consider all possible unbiased estimator of some parameter θ, the one with the ***smallest variance*** is called the ***most efficient estimator* of θ.**



Sampling distribution of different estimators of θ

It is clear that only $\hat{\theta}_1$ and $\hat{\theta}_2$ are **unbiased** estimators, since their distribution are centered at θ.

The estimator $\hat{\theta}_1$ has a **smaller variance** than $\hat{\theta}_2$ and is therefore **more efficient.**

Hence, our choice for an estimator of θ, among the three considered, would be $\hat{\theta}_1$.

Example 3:

If X has the binomial distribution with the parameter n and p, show that the sample proportion $\hat{p} = \frac{X}{n}$ is an unbiased estimator of p.

Solution :

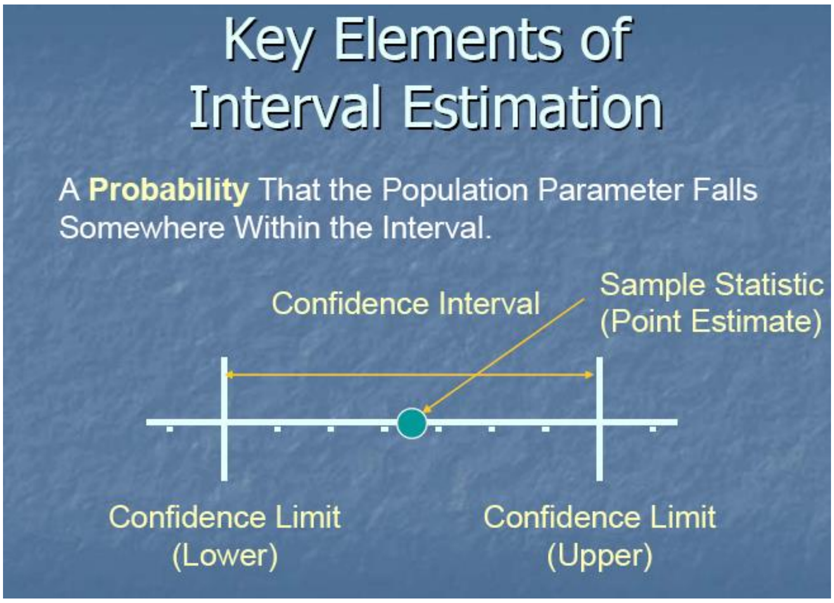
Recall that $X \sim bin(n, \ p) \quad \Rightarrow \quad E(X) = np$

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}.np = p$$

$\therefore \quad \hat{p}$ is an unbiased estimator of p

12.2 Interval Estimation

An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values –an ***interval estimate or confidence interval*** **(CI)**.



Key Elements of Interval Estimation

A **Probability** That the Population Parameter Falls Somewhere Within the Interval.

Confidence Interval

Sample Statistic (Point Estimate)

Confidence Limit (Lower)

Confidence Limit (Upper)

Let θ be a population parameter and $\hat{\theta}$ be the corresponding statistic. From the sampling distribution of $\hat{\theta}$, we will be able to find 2 values such that

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha\,, \quad 0 < \alpha < 1$$

- For $0 < \alpha < 1$, we have ***probability*** of $(1-\alpha)$ of selecting a random sample that will produce an interval containing θ.
- The interval $\theta_L < \theta < \theta_U$ is called $(1-\alpha)100\%$ ***confidence interval of*** θ.
- $(1-\alpha)$ is called the ***confidence coefficient*** or the degree of confidence.
- The endpoints $\hat{\theta}_L$ and $\hat{\theta}_U$ are called the ***lower and upper confident limits***

12.2.1: Single Sample Mean Estimation

We know that , if $\overline{X}$ is the mean of a random sample of size n and $s = \sigma / \sqrt{n}$ is the sample standard deviation from a normal population with the mean μ and the variance σ^2.

In other words, its sampling distribution is a normal distribution with the mean μ and the variance σ^2/n .

Thus we can write $P(|Z| < z_{\alpha/2}) = 1 - \alpha$, where

$$Z = \frac{\overline{X} - \mu}{s}.$$

Here we shall consider two cases where one need to estimate of the mean μ with $(1-\alpha)100\%$ confidence:

- σ is **known** or σ is **unknown but** $n \geq 30$
- σ is **unknown** and $n \leq 30$

Case 1 : *σ is known or σ is unknown but n≥30*

If $\overline{x}$ is the value of the mean of a random sample of size n from a normal population with known variance σ^2, then

$$\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for μ.

If σ is unknown but n≥30, one can use the sample variance s^2 instead of σ^2.

Case 2 : *σ is unknown and n ≤30*

If $\overline{x}$ and s are the mean and standard deviation of a random sample of size n from a normal population with unknown variance σ^2, then

$$\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,\,n-1}\frac{s}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for μ.

Remark of error of estimation

Recall that

$$\mu = \bar{x} \pm error = \bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If $\bar{x}$ is used as an estimate of μ, we can then be $(1-\alpha)100\%$ confident that the error, $\boldsymbol{\varepsilon}$ will not exceed

$$\varepsilon = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

On the other hand, $(1-\alpha)100\%$ confident that the error will not exceed a specified amount $\boldsymbol{\varepsilon}$ when the sample size ***n*** is

$$n = \left(\frac{Z_{\alpha/2}\sigma}{\varepsilon}\right)^2.$$

Example 4:

The 95% confidence interval for the mean μ of a very large normally distributed population is presented as 1.25 cm ± (1.96 × 0.078 cm). The sample size is reported as n = 25, and it is said that the population standard deviation σ, not presented, was used in this calculation. How large a sample would have been required to achieve 99% confidence that the margin of error of the estimate would be e =0.1 cm?

Solution:

Given N=25, the 95% C.I is (1.25 ± 0.078), so $\bar{x} = 1.25$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 0.078$,

thus σ =0.078×5=0.39.

Given that the error, ε = 0.1. The critical value for 99% confidence level, $Z_{\alpha/2} = Z_{0.005} = 2.575$. The minimal number of sample required is

$$n = \left(z_{\alpha/2}\frac{\sigma}{\varepsilon}\right)^2 = ???$$

Example 5:

You work for an airline and are studying the time it takes for passengers to get their baggage after leaving a plane. You know that this population of times are not normally distributed, but rather is somewhat positively skewed, and that σ = 6.4 min. You take a random sample of 45 measurements and get a time of $\overline{x}$ = 26.7 min. What is the approximate 99% confidence interval for the population mean time μ?

Example 6:

A random sample of size n = 16 is taken from a normally distributed population with unknown μ and σ. If the sample has a mean $\overline{x}$ = 27.9 and a standard deviation s = 3.23 , then what is the 99% confidence interval for the data ?

12.2.2: Two samples - Estimating the Difference Between Two Means

Case 1 : σ_1^2 and σ_2^2 *are known or* σ_1^2 and σ_2^2 *are unknown but* $n_1 \geq 30$ and $n_2 \geq 30$ [(*)].

Let $\bar{x}_1$ and $\bar{x}_2$ are the mean of random sample of size n_1 and n_2 from population with **known variance** σ_1^2 **and** σ_2^2 respectively, then $(1 - \alpha)100\%$ confidence interval estimate of $\mu_1 - \mu_2$.

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

***For the case of* σ_1^2 and σ_2^2 *are* unknown *but* $n_1 \geq 30$ and $n_2 \geq 30$ *then* replace σ_1^2 and σ_2^2 *with* s_1^2 and s_2^2.**

Case 2 : σ_1^2 *and* σ_2^2 *are unknown* and $n_1 \leq 30$ and $n_2 \leq 30$

If $\bar{x}_1$, $\bar{x}_2$ are the mean of random sample of size n_1 and n_2 respectively, from approximate normal population with **unknown variance** σ_1^2 and σ_2^2, then $(1-\alpha)100\%$ confidence interval estimate for $\mu_1 - \mu_2$ is given by

- $\sigma_1^2 = \sigma_2^2$ ***(equal variances)***

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2, n_1+n_2-2} . s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2, n_1+n_2-2} . s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

with ***pooled variances***, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

- $\sigma_1^2 \neq \sigma_2^2$ ***(unequal variances)***

$$(\overline{x_1} - \overline{x_2}) - t_{\frac{\alpha}{2}, \nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x_1} - \overline{x_2}) + t_{\frac{\alpha}{2}, \nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

$$w_1 = \frac{s_1^2}{n_1}, \qquad w_2 = \frac{s_2^2}{n_2}, \qquad \nu = \frac{(w_1 + w_2)^2}{\dfrac{w_1^2}{n_1 - 1} + \dfrac{w_2^2}{n_2 - 1}}$$

Example 7:

An experiment was conducted in which two types of engines, A and B were compared. 75 experiments were conducted using engine type A and 50 experiments were done for engine type B. The average gas mileage for engine A was 42 miles per gallon and for engine B was 36 miles per gallon. The sample standard deviations are 8 and 6 for engine A and B respectively. Assuming that the two populations sampled are normal and have ***unequal variance***, find a 90% confidence interval for the difference between the average gas mileage of the two types of engine.

Example 8:

Twelve randomly selected mature citrus trees of one variety have a mean height of 13.8 feet with a standard deviation of 1.2 feet, and fifteen randomly selected mature citrus trees of another variety have a mean height of 12.9 feet with a standard deviation of 1.5 feet. Assuming that the random samples were selected from normal populations with equal variance. Construct a 95% confidence interval for the difference between the true average heights of the two kinds of citrus trees.

12.2.3: Paired observation / Paired t- test

Consider two or more **dependent** samples. For example, to measure the effectiveness of a diet plan, we would select *n* people at random and weigh them both before and after the diet.

The observation would be independent ***between*** pairs, but the observations ***within*** a pair would be dependent because they are taken on the same individual.

If $\bar{d}$ and s_d are the mean and standard deviation of the normally distributed difference of *n* random pairs of measurement, then $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2,\,n-1}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2,\,n-1}\frac{s_d}{\sqrt{n}}$$

where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i\ , \qquad \text{and} \qquad s_d^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right]$$

Example 9:

In a study of the effectiveness of physical exercise in weight reduction, a group of 16 persons engaged in a prescribed program of physical exercise for a month showed the following result:

Weight before, X_1	209	178	169	212	180	192	158	180	170	153
Weight after, X_2	196	171	170	207	177	190	159	180	164	152

Weight before, X_1	183	165	201	179	243	144
Weight after, X_2	179	162	199	173	231	140

Construct 99% confidence interval for the effectiveness of the prescribed program of exercise.

12.2.4: Single sample – Estimating proportion

A point estimator of the population *p* in a binomial experiment is given by the statistic $\hat{p} = \frac{X}{n}$ where *X* represents the number of successes in *n* trials.

If $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$, by central limit theorem, for *n* sufficiently large, $\hat{p}$ is approximately normally distributed with

mean; $\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$

and

variance; $\sigma^2_{\hat{p}} = \mathrm{var}\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{npq}{n^2} = \frac{pq}{n}$.

Therefore,

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

If $\hat{p}$ is the proportion of successes in a random sample of size *n*, then an approximate (1–α)100% confidence interval for the binomial parameter *p* is given by

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

under assumption that $\hat{p} >> \frac{z^2_{\alpha/2}}{2n}$.

Example 10:

A DNA test on *n* = 48 trials samples at a laboratory reports 16 with trace element in the suspect's DNA. Let $\hat{p}$ = 16/48=0.333 be the point estimate of the success *p*. A confidence interval for *p* with the confidence level of approximately 95% is

$$0.333 \pm 1.96\sqrt{\frac{(0.333)(0.667)}{48}} = 0.333 \pm 0.133$$

Note: For *n* samples of *N*-finite population, the confidence intervals are $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$.

If $\hat{p}$ is used as an estimate of *p*, and *q* = 1–*p* we can then be (1–α)100% confident that the error will not exceed the width of the CI, Δ when the sample size is approximately $n \approx \left(\frac{Z_{\alpha/2}\sqrt{\hat{p}\hat{q}}}{\Delta/2}\right)^2$

12.2.5: Two Samples - Estimating the Difference Between Two Proportions

If $\hat{p}_1$ and $\hat{p}_2$ are the proportion of successes in a random sample of size n_1 and n_2, then an approximate $(1-\alpha)100\%$ confidence interval for the binomial parameter p_1-p_2 is given by

$$(\hat{p}_1-\hat{p}_2)-Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}} \quad < \quad p_1-p_2 \quad < \quad (\hat{p}_1-\hat{p}_2)+Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Example 11:

In a random sample of 400 adults and 600 teenagers who watched a certain TV program, 100 adults and 300 teenagers indicated that they liked it. Construct (a) 95%, (b) 99% confidence limits for the difference in proportions of all adults and all teenagers who watched the program and liked it.

Recall the CI for the difference in proportion of the two groups are given by:

$$p1-p2\pm z_c\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$$

Where subscript 1 and 2 refer to teenagers and adults, respectively, and $q_1=1-p_1$, $q_2=1-p_2$. Here $p_1=300/600=0.5$, $p_2=100/400=0.25$ are proportion of teenagers and adults who liked the program.

(a) 95% confidence limits (z_c =1.96):

$$0.50-0.25\pm1.96\sqrt{\frac{(0.5)(0.5)}{600}+\frac{(0.25)(0.75)}{400}}=0.25\pm0.06$$

Therefore, we can be 95% confident that the true proportion lies between 0.19 and 0.31.

(b) 99% confidence limits (z_c =2.58):

$$0.50-0.25\pm2.58\sqrt{\frac{(0.5)(0.5)}{600}+\frac{(0.25)(0.75)}{400}}=0.25\pm0.08$$

Therefore, we can be 99% confident that the true proportion lies between 0.17 and 0.33.

12.2.6: Single sample - Estimating the Variance

An interval estimate of σ^2 can be established by using the statistic

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

If s^2 is the variance of a random sample of size n from a normal population, a $(1-\alpha)100\%$ confidence interval for σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

Example 12:

An experiment produces the following data:
1470, 1510, 1690, 1740, 1900, 2000, 2030, 2100, 2190, 2200, 2290, 2380, 2390, 2480, 2500, 2580, 2700

9.3.7: Two samples - Estimating the Ratio of Two Variance

If σ_1^2 and σ_2^2 are two variance of normal populations, we can establish an interval estimate of $\frac{\sigma_1^2}{\sigma_2^2}$ by using the statistic

$$F = \left(\frac{s_1^2}{\sigma_1^2}\right)\left(\frac{\sigma_2^2}{s_2^2}\right) \sim f_{(\nu_1=n_1-1,\ \nu_2=n_2-1)}$$

and the probability

$$P\left(f_{1-\alpha/2,\,\nu_1,\ \nu_2} < F < f_{\alpha/2,\nu_1,\ \nu_2}\right) = 1-\alpha$$

$$P\left(\ f_{1-\alpha/2,\ \nu_1,\ \nu_2} < \frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} < f_{\alpha/2,\ \nu_1,\ \nu_2}\ \right) = 1-\alpha$$

If s_1 and s_2 are the variance of independent random sample of size n_1 and n_2, respectively, from normal populations, then a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is given by

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{\alpha/2,\, \nu_1,\, \nu_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{1-\alpha/2,\, \nu_1,\, \nu_2}}$$

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{f_{\alpha/2,\, \nu_1,\, \nu_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\cdot f_{\alpha/2,\, \nu_2,\, \nu_1}$$

Example 13:

Given,

$n_1 = 8,\ \nu_1 = 7$; $\bar{x}_1 = 546$ $s_1 = 31$

$n_2 = 4,\ \nu_2 = 3$; $\bar{x}_2 = 492$ $s_2 = 26$

and CL 98% or $\alpha = 0.02$, find a 98% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

Solution: