Statistical Inference (Estimation)

Where we have been...

In the preceding class we discussed fundamental ideas and techniques of sampling and their distributions. Sampling distributions allow us to make probability statements about statistics.

Here we begin with the study of some problems of mathematical statistics. Statisticians deal with the problems of planning and designing experiments, of collecting information, and of deciding how best the collected information should be used. Suppose we seek information about some numerical characteristics of a population. For reasons of time or cost we may not wish or be able to study each element of the population. Our objective is to draw conclusions about the unknown population characteristics...

Where we have been...

On the basis of information on some sample characteristics of a suitably selected sample. In a typical statistical problem, we have a random variable (r.v.) X that describes the population under investigation. But, its distribution is not known. There will be two possibilities:

1. X has a cdf F_θ with a known functional form (except for the parameter θ) where θ may be a vector.
2. The distribution of X is completely unknown.

In the first case, let Θ be the set of all possible values of the unknown parameter θ . Then the job of statistician is to decide on the basis of a sample which member(s) of the family $\{F_\theta : \theta \in \Theta\}$ can represent the distribution of X.

Where we are going...

This type of problems are called *parametric statistical inference*. The case in which nothing is known about the functional form of the distribution function F of X is much more difficult and falls into the domain of *nonparametric statistics.*

For now, we consider the 1^{st} one when one or more parameters associated with F will be unknown. Some examples are the following:

✓ X has an exponential distribution $\text{Exp}(\theta)$ where θ is unknown.

✓ X has an binomial distribution $B(n, p)$ where n is known but p is unknown.

Where we are going...

✓ X has a normal distribution $N(\mu,\sigma^2)$ where both the parameters are unknown.

✓ X has a Poisson distribution $P(\lambda)$ where λ is unknown.

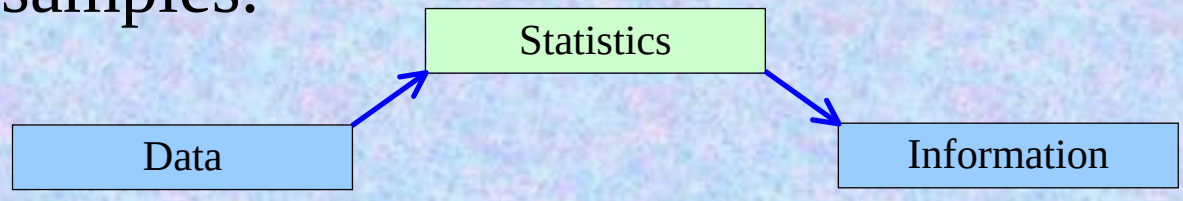
In almost all realistic situations parameters are unknown.

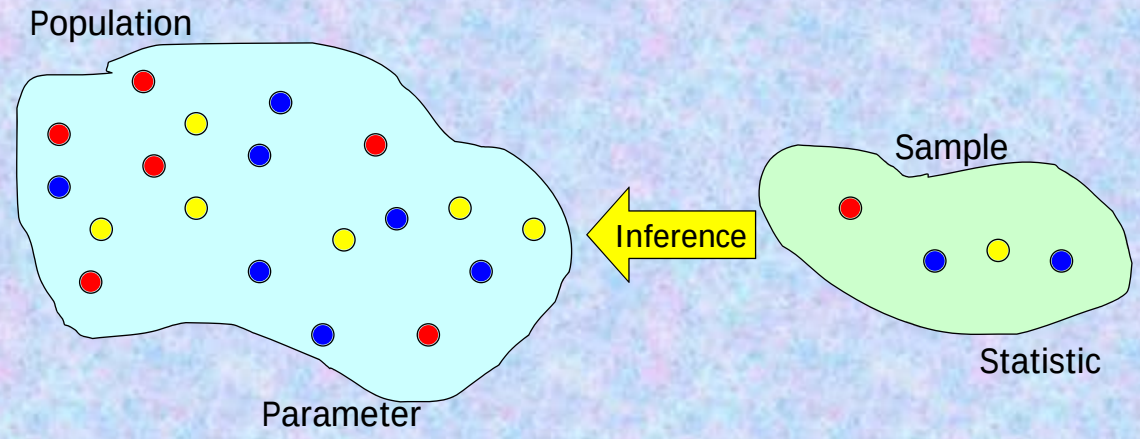
We will use the sampling distribution to draw inferences about the unknown population parameters.

Sampling distributions allow us to make probability statements about statistics.

Statistical Inference...

Statistical inference is the process by which we acquire information and draw conclusions about populations from samples.





In order to do inference, we require the skills and knowledge of descriptive statistics, probability distributions, and sampling distributions.

Estimation...

There are two types of inference: estimation and hypothesis testing; ***estimation*** is introduced first.

The objective of estimation is to determine the ***approximate value*** of a population parameter on the basis of a sample statistic.

E.g., the sample mean ($\overline{X}$) is employed to ***estimate*** the population mean (μ).

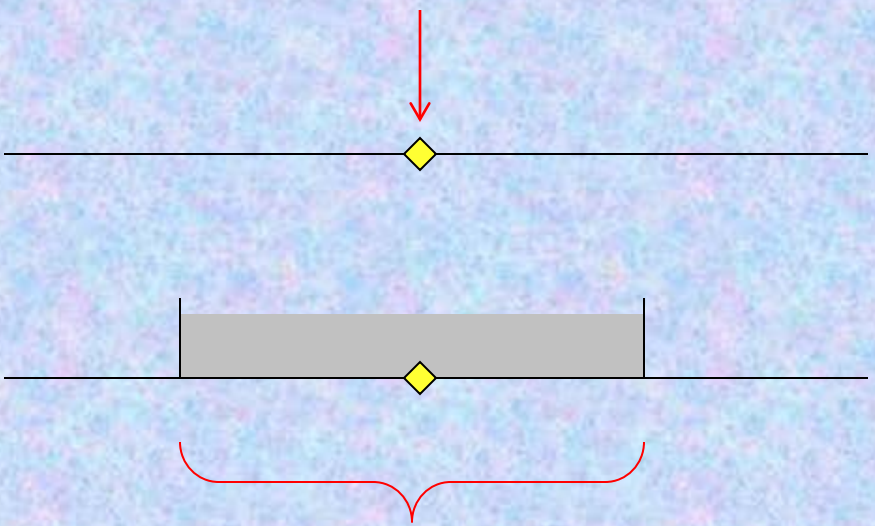
Estimation...

The objective of estimation is to determine the ***approximate value*** of a population parameter on the basis of a sample statistic.

There are two types of estimators:

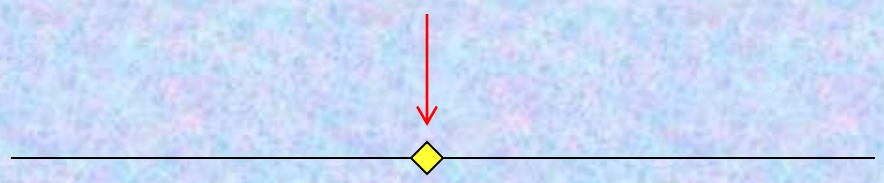
Point Estimator

Interval Estimator



Point Estimator…

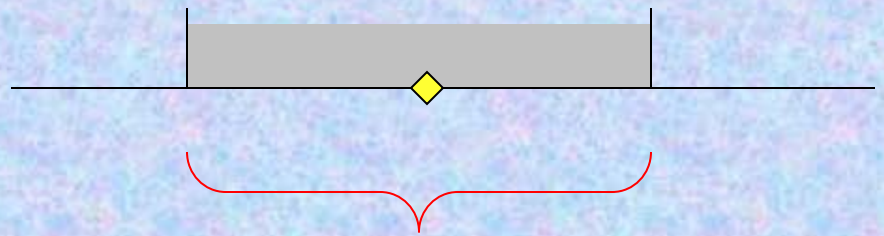
A ***point estimator*** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



We saw earlier that point probabilities in continuous distributions were virtually zero. Likewise, we'd expect that the point estimator gets closer to the parameter value with an increased sample size, but point estimators don't reflect the effects of larger sample sizes. Hence we will employ the ***interval estimator*** to estimate population parameters…

Interval Estimator...

An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.



That is we say (with some ___% certainty) that the population parameter of interest is between some lower and upper bounds.

Point & Interval Estimation...

For example, suppose we want to estimate the mean summer income of a class of business students. For n = 25 students, $\bar{x}$ is calculated to be 400 $/week.

point estimate

interval estimate

An alternative statement is:

The mean income is ***between*** 380 and 420 $/week.

Point Estimator

Maximum Likelihood Estimation (MLE):

Definition: Let $X_1, X_2, ..., X_n$ be an iid random sample of size n drawn from a population with pdf (pmf) $f(x;\theta_1,\theta_2,...,\theta_k)$ Then the function $L(x_1,x_2,...,x_n;\theta_1,\theta_2,...,\theta_k)=\prod_{i=1}^{n} f(x_i;\Theta)$ is called the *likelihood function*.

Thus, if the function L has a unique maximum for

$$\theta_1=\hat{\theta}_1(x_1,x_2,...,x_n),\ \theta_2=\hat{\theta}_2(x_1,x_2,...,x_n),\cdots,\theta_k=\hat{\theta}_k(x_1,x_2,...,x_n)$$

Then the statistics $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_k$ are called the *maximum likelihood estimates* of $\theta_1,\theta_2,\cdots,\theta_k$.

Since $L>0$, maximizing L amounts to maximizing $\log L$, the equations (called *likelihood eqns*) for which are

$$\frac{\partial \log L}{\partial \theta_1}=0=\frac{\partial \log L}{\partial \theta_2}=\cdots=\frac{\partial \log L}{\partial \theta_k}$$

Method of MLE

Find the Point Estimator of the following population parameters by MLE:

Binomial: p

Poisson: μ

Normal: μ and σ

Exponential: λ or μ

Uniform $U[0,\theta]$ or $U\left[\theta-\frac{1}{2},\theta+\frac{1}{2}\right]$

X follows the distribution with pdf

$$f(x,\theta)=\begin{cases} y_0 e^{-\beta(x-\alpha)}; \alpha \le x < \infty, \beta > 0, \ y_0 \ const \\ 0, \ elsewhere \end{cases}$$

Qualities of Estimators...

Qualities desirable in estimators include unbiasedness, consistency, and relative efficiency:

An ***unbiased estimator*** of a population parameter is an estimator whose expected value is equal to that parameter.

An unbiased estimator is said to be ***consistent*** if the difference between the estimator and the parameter grows smaller as the sample size grows larger.

If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to be ***relatively efficient***.

Unbiased Estimators...

An ***unbiased estimator*** of a population parameter is an estimator whose expected value is equal to that parameter.

E.g. the sample mean $\overline{X}$ is an ***unbiased*** estimator of the population mean μ , since:

$$E(\overline{X}) = \mu$$

Similarly, the sample median is an ***unbiased*** estimator of the population mean μ since:

$$E(\text{Sample median}) = \mu$$

Consistency...

An unbiased estimator is said to be ***consistent*** if the difference between the estimator and the parameter grows smaller as the sample size grows larger.

E.g. $\overline{X}$ is a ***consistent*** estimator of μ because:

$V(\overline{X})$ is σ^2/n

That is, as n grows larger, the variance of $\overline{X}$ grows smaller.

Similarly, Sample median is a ***consistent*** estimator of μ because:

V(Sample median) is $1.57\sigma^2/n$

Interval Estimation...

Applications to Normal (μ, σ) Population

Let us produced the following general probability statement about X

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

And we know that the sampling distribution of $\overline{X}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$ (for infinite population)

Thus $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is (approximately) standard normally distributed.

Estimating μ when σ is known...

Thus, substituting Z we produce

$$P(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

We expressed the following

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

With a little bit of different algebra we have

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Estimating μ when σ is known...

This

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

is still a probability statement about $\bar{X}$. However, the statement is also a confidence interval estimator of μ.

The probability $1 - \alpha$ is the confidence level, which is a measure of how frequently the interval will actually include μ.

Estimating μ when σ is known...

The interval can also be expressed as

Lower confidence limit = $\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

Upper confidence limit = $\left(\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$

For Finite Population:

The confidence interval takes the form

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} z_{\alpha/2}, \bar{x} + \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} z_{\alpha/2} \right)$$

Example

The Doll Computer Company makes its own computers and delivers them directly to customers who order them via the Internet.

To achieve its objective of speed, Doll makes each of its five most popular computers and transports them to warehouses from which it generally takes 1 day to deliver a computer to the customer.

This strategy requires high levels of inventory that add considerably to the cost.

Example

To lower these costs the operations manager wants to use an inventory model. He notes demand during lead time is normally distributed and he needs to know the mean to compute the optimum inventory level.

He observes 25 lead time periods and records the demand during each period.

The manager would like a 95% confidence interval estimate of the mean demand during lead time. Assume that the manager knows that the standard deviation is 75 computers.

Example

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

Example

"We want to estimate the ***mean*** demand over lead time with 95% confidence in order to set inventory levels…"

IDENTIFY

Thus, the parameter to be estimated is the population mean:μ

And so our confidence interval estimator will be:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Example

COMPUTE

In order to use our confidence interval estimator, we need the following pieces of data:

$\bar{x}$	370.16
$z_{\alpha/2}$	1.96
σ	75
n	25

Calculated from the data…

$1-\alpha = .95,\ \therefore \alpha/2 = .025$

$so\ \ z_{\alpha/2} = z_{.025} = 1.96$

Given

therefore: $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm z_{.025}\frac{75}{\sqrt{25}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 29.40$

The **lower** and **upper** confidence limits are 340.76 and 399.56.

Example

INTERPRET

The estimation for the mean demand during lead time lies between 340.76 and 399.56 — we can use this as input in developing an inventory policy.

That is, we estimated that the mean demand during lead time falls between 340.76 and 399.56, and this type of estimator is correct 95% of the time. That also means that 5% of the time the estimator will be incorrect.

Incidentally, the media often refer to the 95% figure as “19 times out of 20,” which emphasizes the ***long-run*** aspect of the confidence level.

Interval Width...

A wide interval provides little information.

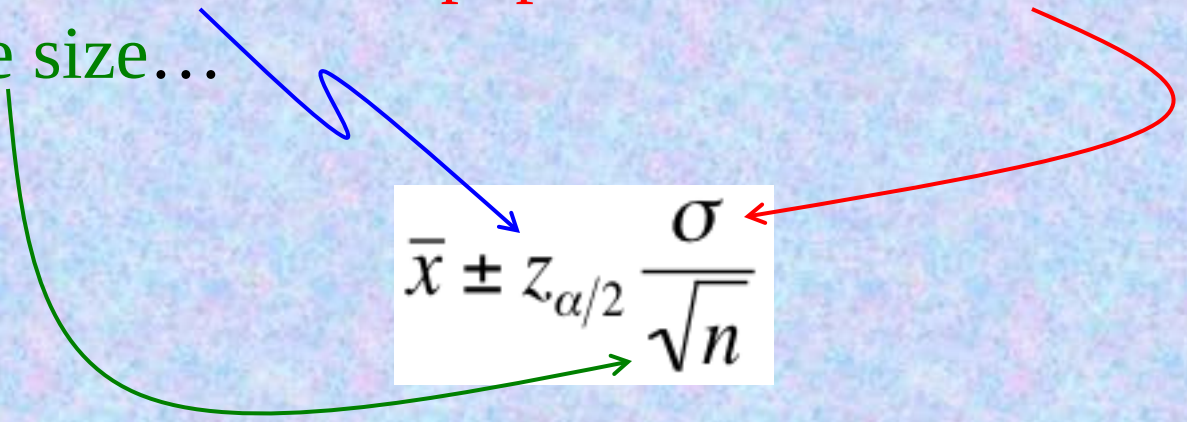
For example, suppose we estimate with 95% confidence that an accountant's average starting salary is between $15,000 and $100,000.

Contrast this with: a 95% confidence interval estimate of starting salaries between $42,000 and $45,000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

Interval Width...

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

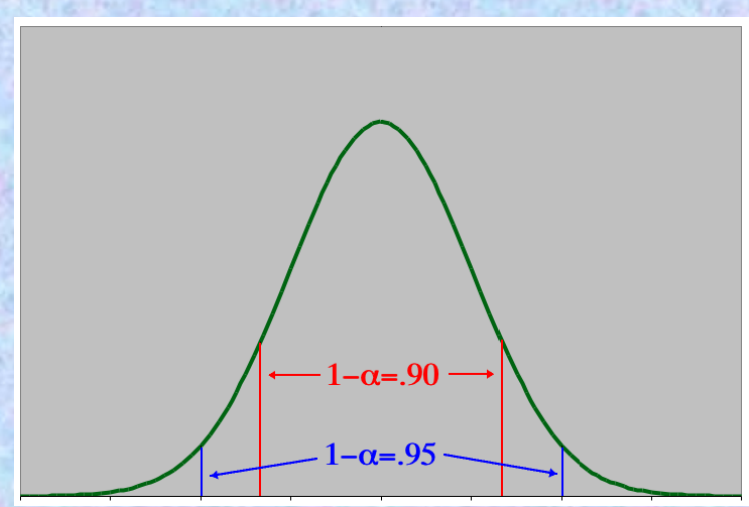


Interval Width...

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

A larger confidence level produces a **w i d e r** confidence interval:

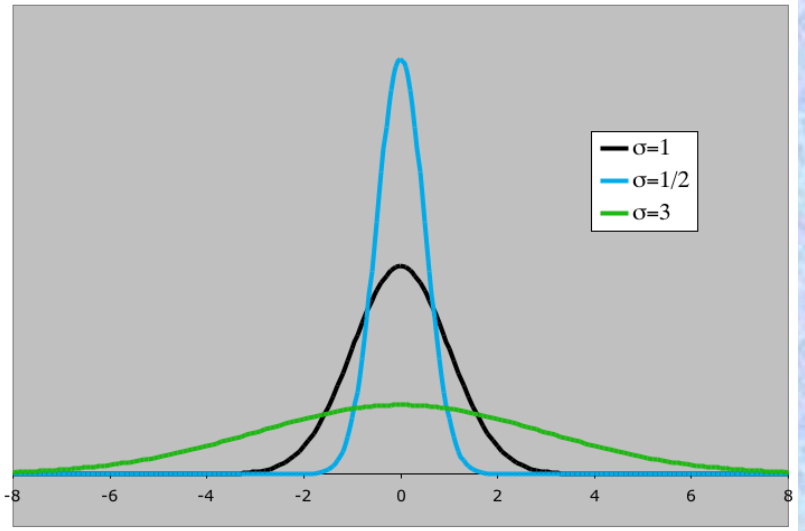


Interval Width...

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

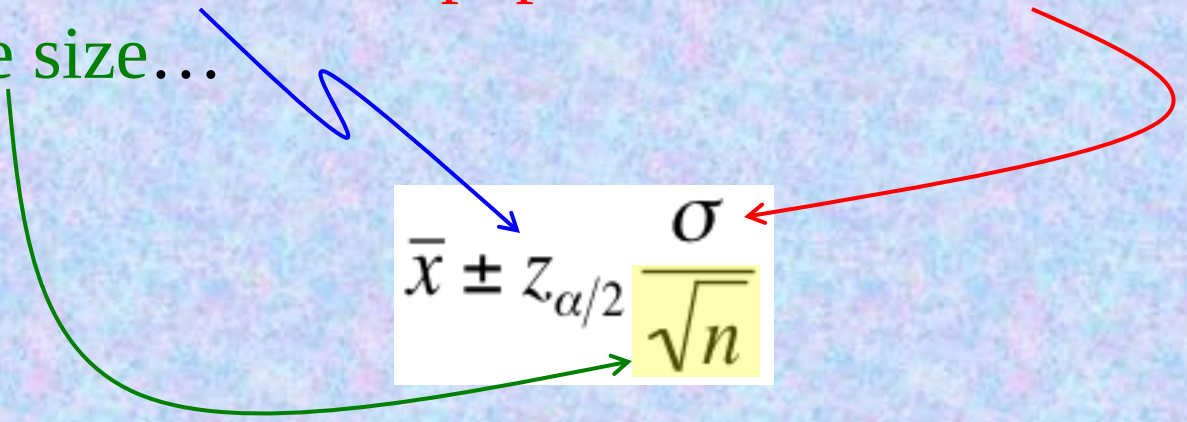
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Larger values of σ produce **w i d e r** confidence intervals



Interval Width...

The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…



Increasing the sample size decreases the width of the confidence interval while the confidence level can remain unchanged.

Note: this also increases the ***cost*** of obtaining additional data

Home Work

In an effort to estimate the mean amount spent per customer for dinner at a major Atlanta restaurant, data were collected for a sample of 49 customers. Assume a population S.D. of $5.

i) At 95% confidence, what is the margin of error?

ii) If the sample mean is $24.80, what is the 95% C.I. for population mean?

Example on Finite Population

A researcher wants to measure the income level of employees working in igl. The total employee strength of igl is 1200. A random sample of 50 employees reveals that the average income of sampled employees is Rs. 15,000. Historical data reveals that the S.D. of the income of the employees is approx. Rs. 1500. Construct a 99% C.I. for obtaining the average income of all the employees working here.

Estimating μ when σ is unknown...

In this case the confidence interval of mean is given by

For infinite population:

$$\left(\bar{x}-\frac{S}{\sqrt{n}}t_{\alpha/2},\bar{x}+\frac{S}{\sqrt{n}}t_{\alpha/2}\right)$$

For finite population:

$$\left(\bar{x}-\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}[t_{\alpha/2}\ or\ z_{\alpha/2}],\bar{x}+\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}[t_{\alpha/2}\ or\ z_{\alpha/2}]\right)$$

Here $t_{\alpha/2}$ is obtained from student's t-distribution table for degrees of freedom $\nu = n-1$ and on using

$$P(t > t_{\alpha/2}) = \frac{\alpha}{2}$$

Example...

- A stock market analyst wants to estimate the average return on a certain stock. A random sample of 15 days yields an average (annualized) return of $\bar{x} = 10.37\%$ & a S.D. of S=3.5%, Assuming a normal population of returns, give a 95% C.I. for the average return on this stock.
- The Westview High School nurse is interested in knowing the average height of seniors at this school, but she doesn't have enough time to examine the records of all 430 seniors. She randomly selects 48 students. She finds the sample mean to be 64.5 inches and the S.D. to be 2.3 inches. Construct a 90% C.I. for the average height of the senior students.

Confidence Interval for σ

For estimating σ, we choose χ^2-distribution with $\nu = n-1$ d.f. and obtain the confidence interval as

$$\left(S\sqrt{\frac{n-1}{\chi^2_{\varepsilon_2}}}, S\sqrt{\frac{n-1}{\chi^2_{\varepsilon_1}}} \right)$$

where

$$P(0 < \chi^2 < \chi^2_{\varepsilon_1}) = \frac{\varepsilon}{2} \quad \text{and} \quad P(\chi^2 > \chi^2_{\varepsilon_2}) = \frac{\varepsilon}{2}$$

Example...

In an automated process, a machine fills cans of coffee. If the average amount filled is different from what it should be, the machine may be adjusted to correct the mean. If the variance of the filling process is too high, however the machine is out of control & needs to be repaired. Therefore, from time to time regular checks of the variance of the filling process are made. This is done by randomly sampling filled cans, measuring their amounts, and computing the sample variance. A random sample of 30 cans gives an estimate of the sample variance as 18,540 Give a 95% C.I. for population variance.

Home Work

Seven laboratory determinations of the value of g , the acceleration due to gravity, at Kolkata gave a mean 977.51cm/sec^2 and a S. D. 4.42 cm/sec^2. Now it is known that the population of the measured values of any physical quantity subject to experimental errors has a normal distribution whose mean is the true value of the quantity. Assuming this fact, find 95% confidence interval for the true value of g. Also find 95% C.I. for the population S.D.