ISBN: 978-81-947839-7-8

1. Sampling Fundamentals

Swarup Bhattacharjee

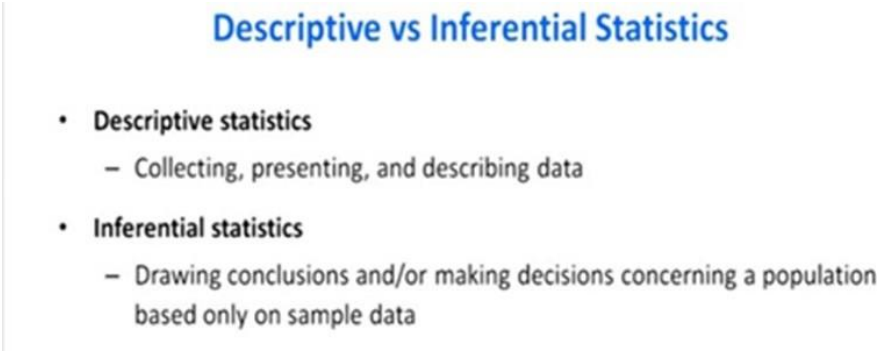
Associate Professor,
ICV College Dept. of Mathematics,
Govt. Of Tripura, Agartala.

1.1 Introduction:

Before providing the idea of sampling we define population first. Population is a group of items, units. It can be finite or infinite or hypothetical.

Examples: Workers in a factory is an example of finite population. All stars in the sky is an example of infinite population.

In other words, if the number of items constituting population is fixed, it is known as finite population. If the population consists of an infinite number of items, it is called infinite population.



1.2 Sample:

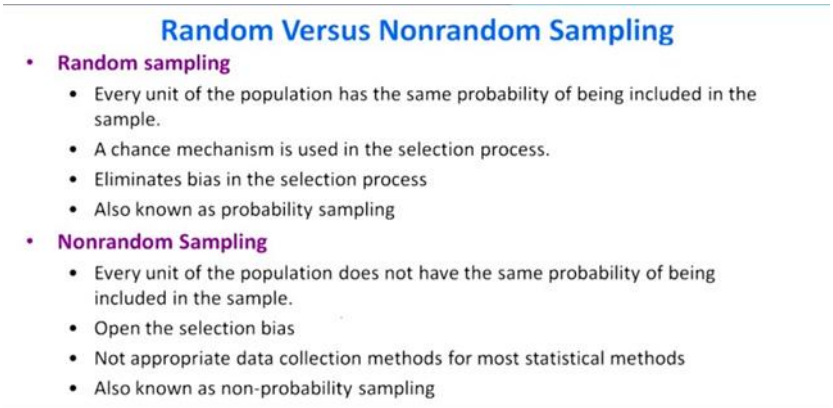
A finite subset of items in population is called sample. Sampling is a part of our day-to-day life. For example, a housewife takes one or two grains of rice from cooking pan and decides whether the rice is cooked or not.

Why Sampling?:

Complete enumeration is much more expensive and time consuming. More errors are happened due to greater volume of work.

Random Sampling:

A random sample is one in which each unit of population has an equal chance of being included in it. If sample size in n and population size in N, there are NCn possible samples.



Population Parameter:

It is a population constant from the population-(i) Population mean(μ), (II)Population variance(σ2) etc.

(Sample) Statistics:

It is a statistical measure computed from sample observations. It is also a random variable but not necessarily parameter.

1. Sample mean($\bar{x}$), 2.sample variance(s^2)

Definitions: Let x1, x2……x_n be a random sample of size n from a population of size N then

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

- If n ≥ 30, the sampling is known as large sample.
- If n < 30, the sampling is known as small sampling.

Difference between a statistic and a parameter?

The difference between a statistic and a parameter is that statistics describe a sample. A parameter describes an entire population.

Sampling Distribution:

A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

Errors

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.
- Sampling Error occurs when the sample is not representative of the population
- Non-sampling Errors
  - Missing Data, Recording, Data Entry, and Analysis Errors
  - Poorly conceived concepts , unclear definitions, and defective questionnaires
  - Response errors occur when people so not know, will not say, or overstate in their answers

Standard Error:

The standard error (SE) is very similar to standard deviation (square root of variance). Both are measures of spread. The higher the number, the more spread out your data is.

To put it simply, the two terms are essentially equal—but there is one important difference. While the standard error uses statistics (sample data) standard deviations use parameters (population data).

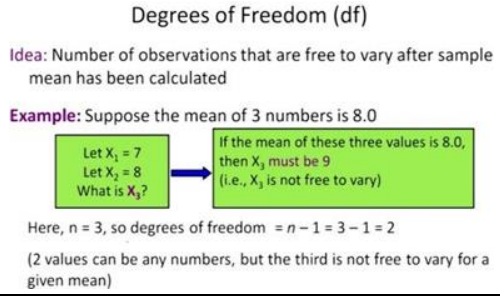
Formula for Standard Error (SE):

$SE(\bar{x}) = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$, s^2, n-: sample size, N: population size

If N is large then 1/N is negligible. Hence $SE(\bar{x}) = s/\sqrt{n} = s/\sqrt{n}$

Degrees of Freedom:

Degrees of Freedom (df) refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample. If n is the number of independent observation of a random sample and k is the number of population parameters which are calculated using sample data, then df=n-k.



Simple Random Sampling with Replacement (Srswr):

If the units are drawn one by one in such a way that a unit drawn at a time is replaced by back to the population before the next draw, it is known as srswr.

In this type of sampling from a population size N, the probability of selection of unit is 1/N in each draw.

Simple Random Sampling without Replacement (Srswor)

In this method the unit selected once is not included in the population at any subsequent draw. The probability of drawing a unit from a population of N units at rth draw is 1/(N-r+1).

In simple random sampling, the probability of selection of any sample of size n from a population of size N is

$1/ {}^{N}C_{n.}$

1.3 Estimator:

An estimator is a rule or a function of variates for estimating the population parameters. An estimator is itself a random variable.

For example, estimator for mean is $\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$ which depends on sample values xi.

- If a random sample of size n is taken from a population having the mean μ and variance σ^2, then x a random variable whose distribution is mean μ.
- The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that min (np, n (1-p))> 5, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

For the random samples we take from the population, we can compute the mean of the sample means:

$$\mu_{\bar{X}} = \mu$$

and the standard deviation of the sample means:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

If the Population is not Normal- Central Limit Theorem

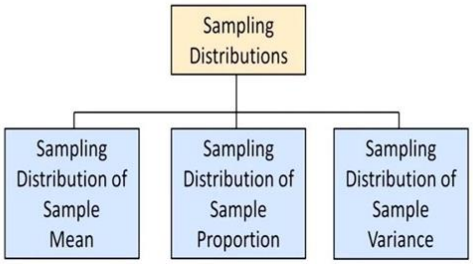
We can apply the Central Limit Theorem:

- Even if the population is not normal,
- sample means from the population will be approximately normal as long as the sample size is large enough.

Properties of the sampling distribution:

$\mu_{\bar{x}} = \mu$ And $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

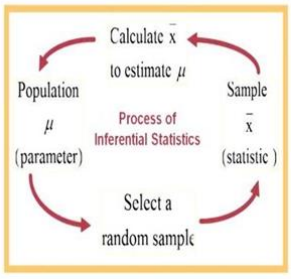
Types of sampling distributions



Distribution of the Sample Mean:

Sampling Distribution of $\bar{X}$

Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.



The statistic used to estimate the mean of a population, μ, is the sample mean. $\bar{X}$

If X has a distribution with mean μ, and standard deviation σ, and is approximately normally distributed or n is large, then $\bar{X}$ is approximately normally distributed with mean μ and **standard Error** $\sigma/\sqrt{n}$.

When σ Is Known:

If the standard deviation, σ, is known, we can transform **known** to an approximately standard normal variable, the test statistic:

$$Z = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Example:

If μ=20, and σ=5. Suppose we draw a sample of size n=16 from this population and want to know how likely we are to see a sample average greater than 22, that is P ($\bar{X}$ > 22)?

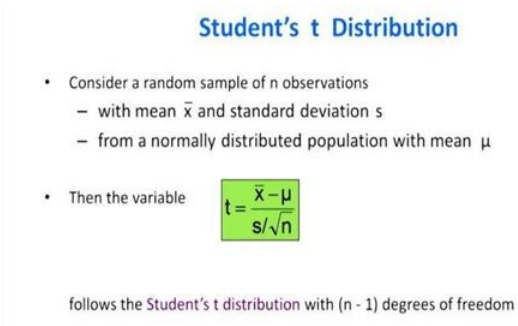
$$z = \frac{22 - 20}{\left(\frac{5}{\sqrt{16}}\right)} = 1.6$$

So the probability that the sample mean will be >22 is the probability that Z is > 1.6 We use the Z table to determine this:

P(> 22) = P(Z > 1.6) = 0.0548.

When σ Is Unknown:

If the standard deviation, σ, is **unknown**, we cannot transform to $\bar{X}$ standard normal. However, we can estimate σ using the sample standard deviation, *s*, and transform $\bar{X}$ to a variable with a similar distribution, the ***t distribution***. There are actually many t distributions, indexed by degrees of freedom (df). As the degrees of freedom increase, the t distribution approaches the standard normal distribution.



If X is approximately normally distributed, then test statistics:

$$t = \frac{\bar{X} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$

has a t distribution with (n-1) degrees of freedom (df)

Example:

In the previous example we drew a sample of n=16 from a population with μ=20 and σ=5. We found that the probability that the sample mean is greater than 22 is P (> 22) = 0.0548. Suppose that is unknown and we need to use s to estimate it. We find that s = 4. Then we calculate t, which follows a t-distribution with df = (n-1) = 24.

$$t = \frac{22 - 20}{\left(\frac{4}{\sqrt{16}}\right)} \quad = 2.0$$

If samples values are not independent

- If the sample size n is not a small fraction of the population size N, then individual sample members are not distributed independently of one another
- Thus, observations are not selected independently
- A correction is made to account for this:

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} \quad \text{or} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

1.3.1 Unbiased Estimate:

A statistics t=t(x1,x2,…….xn) a function of samples values x1,x2,…….xn is an unbised estimate of population parameter θ, if E(t)= θ. In other words if E(statistic)=parameter.

Sampling distribution of sample proportion:

P equal to the proportion of populations having some characteristics, we can call it as P is the population proportion. This sample proportion we are going to call it as a small. It provides an estimate of P.

Z-Value for Proportions

Standardize $\hat{p}$ to a Z value with the formula:

$$Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}} = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Example

- if P = .4 and n = 200, what is $P(.40 \le \hat{p} \le .45)$?

Find: $\sigma_{\hat{p}}$

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{.4(1-.4)}{200}} = .03464$$

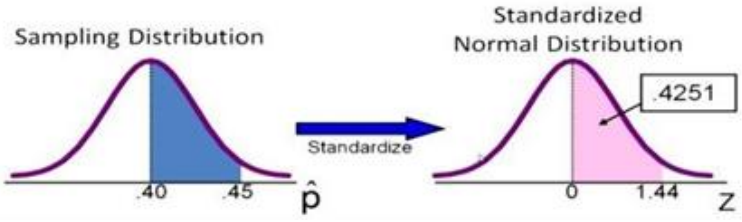
Convert to standard normal:

$$P(.40 \le \hat{p} \le .45) = P\left(\frac{.40-.40}{.03464} \le Z \le \frac{.45-.40}{.03464}\right)$$
$$= P(0 \le Z \le 1.44)$$

Example

if P = .4 and n = 200, what is $P(.40 \le \hat{p} \le .45)$?

Use standard normal table: $P(0 \le Z \le 1.44) = .4251$



Z value is 0 to 1.44 which we got 0.4251. So, now we have seen this one we will go to the sampling distribution of sample variance.

- The sampling distribution of sample variance has the mean population variance. So, what is the meaning in that one is, from the population, you take different sample for that sample you find the sample variance we know of that sample variance is equal to population variance but when you take the from the normal population, if you take some sample, then, you find the sample variance.

Sample Variance

- Let $x_1, x_2, \ldots, x_n$ be a random sample from a population. The sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- the square root of the sample variance is called the sample standard deviation
- the sample variance is different for different random samples from the same population

Sampling Distribution of Sample Variances

- The sampling distribution of s^2 has mean σ^2

$$E(s^2) = \sigma^2$$

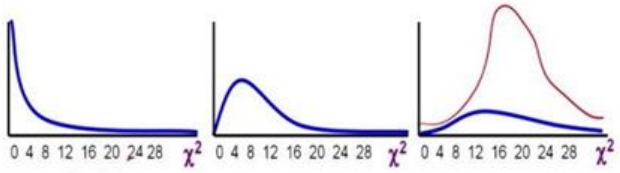
- If the population distribution is normal then

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

has a χ^2 distribution with $n - 1$ degrees of freedom

The Chi-square Distribution

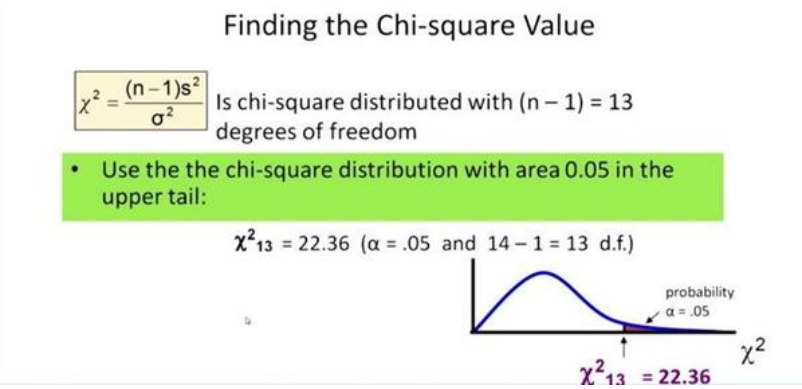
- The chi-square distribution is a family of distributions, depending on degrees of freedom: d.f. = n – 1



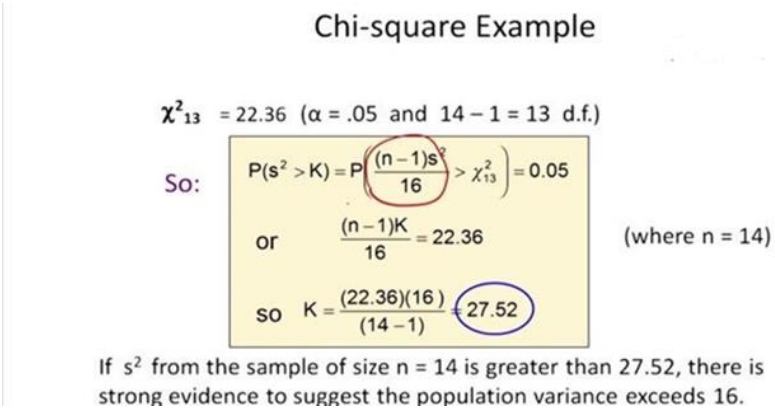
The chi-square example:

A commercial freezer must hold their selected temperature with a little variation specification called for a standard deviation of no more than 4 degrees that is the variance 16 degree square you should not exceed 16, and the standard deviation 4. For a sample of 14 freezers is to be tested what is the upper limit of the sample variance such that the probability of exceeding this limit given that the population standard deviation is 4 is less than 0.05.

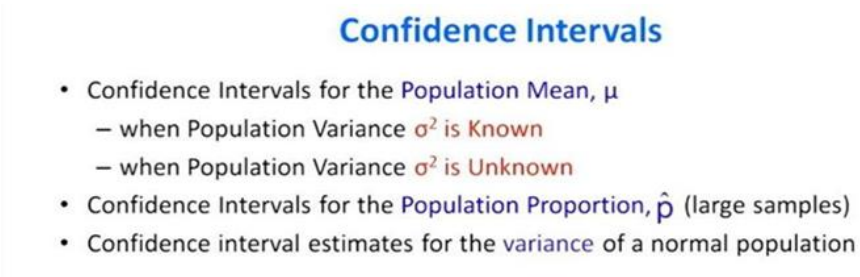
- What is it asking, what is the probability of sample variance that the, the probability of exceeding this limit is less than 0.05?



So, first thing is we have to find out the Chi square value for n minus 1 degrees of freedom. This is a chi-square distribution there are 14 sample is n the degrees of freedom is for 13. 14 minus 1 13, so, the corresponding alpha is equal to 0.05, is 22.36.



1.3.2 Confidence Interval Estimation:



- In the confidence interval, what we are going to see the confidence intervals for the population mean there are two possibilities: When the population variance Sigma square is known other case is when population variance Sigma square is unknown.
- Confidence Interval, How much uncertainty is associated with the point estimate of the population parameter because when we say, the temperature is 35 degree how much uncertainty is associated with that point estimate. That uncertainty is expressed with the help of confidence interval. An estimate provides more information about the population characteristics than does a point estimate.

So, when compared to point estimate, interval estimate is giving more information about the population. Such interval estimates are called confidence intervals.

So, for example, if we say this is the population we are taking different sample say, the population mean may be say 40.

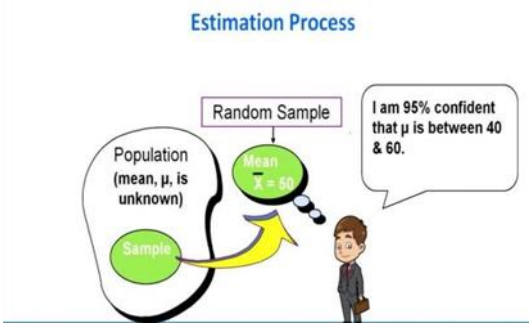
We have taken various sample with help of sample mean, we can predict what will be the lower limit and upper limit of this population mean.

For example, if we say, 35 to 45 this interval is nothing but confidence interval.

Confidence Interval and Confidence Level

- If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 - \alpha)\%$ confidence interval of θ.
- The quantity $(1 - \alpha)$ is called the confidence level of the interval (α between 0 and 1)
  - In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 - \alpha)\%$ of intervals calculated this way.
  - The confidence interval calculated in this manner is written as $a < \theta <$ b with $100(1 - \alpha)\%$ confidence

Estimation Process

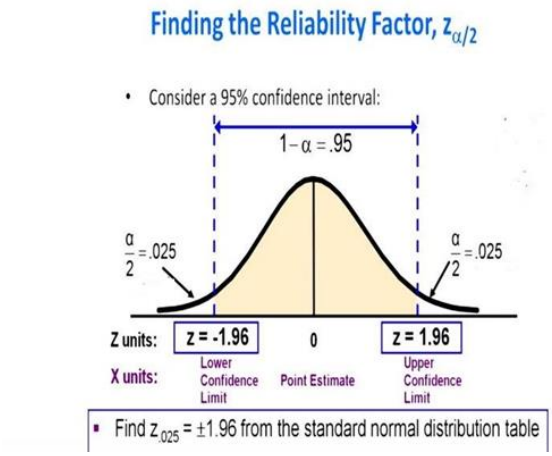


General Formula

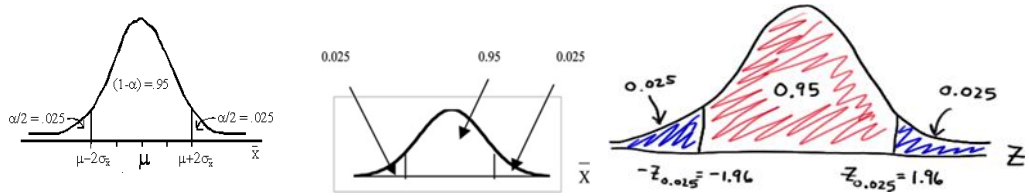
- The general formula for all confidence intervals is:

Point Estimate ± (Reliability Factor)(Standard Error)

- If you use a standard error, $\sigma/\sqrt{n}$, so x-bar + or - Z $(\sigma/\sqrt{n})$ is nothing but the formula for confidence interval. So, when you say + it is upper limit if it is - it is lower limit.
- We look at how to find out the reliability factor that is Z $\alpha/2$. For example, if I suppose, if you want to know something at 95% confidence level, so this is 95% confidence level so the remaining is 5%, when you divide this 5% by 2 see the right hand side you will get is 0.025. The left hand she will get 0.025. When you look at the Z table, when the right hand side is 0.025, the corresponding Z value is 1.96 on right hand side.



The 95% confidence interval for μ



Confidence interval for μ when σ known:

- Assumptions
  - Population variance σ^2 is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

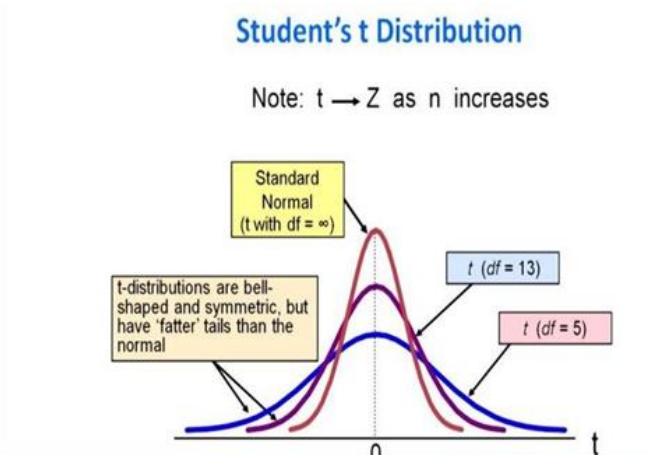
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence interval for μ when σ unknown:

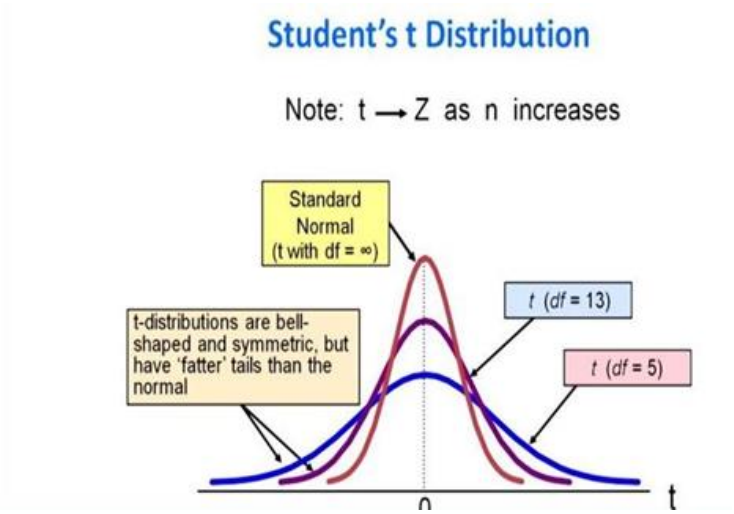
- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

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

where $t_{n-1,\alpha/2}$ is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail



When the sample size increases for the t test so the behavior of Z distribution t distribution is same



Example

A random sample of n = 25 has $\bar{x} = 50$ and s = 8. Form a 95% confidence interval for μ

– d.f. = n – 1 = 24, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$

The confidence interval is

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

$$50 - (2.0639)\frac{8}{\sqrt{25}} < \mu < 50 + (2.0639)\frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

Confidence Intervals for the Population Variance

The (1 - α)% confidence interval for the population variance is

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

- An example you are testing the speed of batch of computer processors you collect the following data, sample sizes 17 sample mean is 3004 samples, standard deviation is 74 assume the population is normal determined 95% confidence interval for σXbar2, here σXbar2 is nothing but lower limit upper limit of the sampling variance.

Example

You are testing the speed of a batch of computer processors. Y collect the following data (in Mhz):

Sample size	17
Sample mean	3004
Sample std dev	74

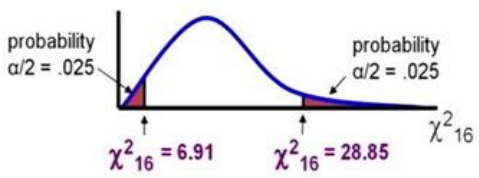
Assume the population is normal. Determine the 95% confidence interval for σ_x^2

Finding the Chi-square Values

- n = 17 so the chi-square distribution has (n – 1) = 16 degrees of freedom
- α = 0.05, so use the the chi-square values with area 0.025 in each tail:

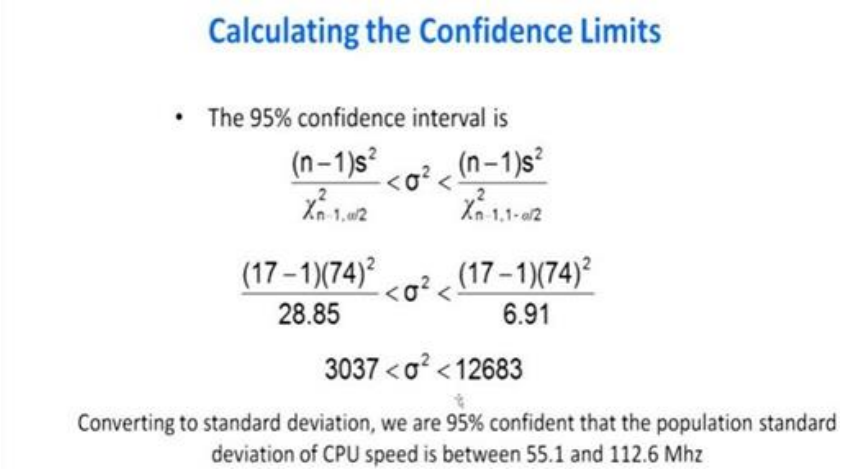
$\chi^2_{n-1,\alpha/2} = \chi^2_{16,0.025} = 28.85$

$\chi^2_{n-1,1-\alpha/2} = \chi^2_{16,0.975} = 6.91$



So, n equal to 17 then chi square distribution has the n – 1, 16 degrees of freedom when alpha equal to 0.05 because it is we are finding upper limit lower limit we got 2 divided by 2 so 0.025.

so, when it is alpha by 2 it is 28.25 so what will happen this is the right side limit when you want to know the left side limit you have to, in the chi square table when area equal to 1 - 0.025 that area you have to find out that probability when the degrees of freedom is 16 so corresponding is 6.91.



1.4 Acknowledgements:

I would like to express thanks and gratitude to the person mentioned hereby for obtaining concept from their contribution to the said topics:

1. Prof. Ramesh Anbanandam
2. Prof. G.S.S Bhishma Rao
3. Prof. Sailes Bhushan Choudhuri
4. S.C. Gupta
5. B. L. Agarwal