



# Statistical Inference

By

Amritansh



# Statistical Inference

The process of making guess/inference about the truth from a sample






# Central Limit Theorem

- In probability theory, the **central limit theorem (CLT)** establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.
- Suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic mean of the observed values is computed.

If this procedure is performed many times, the central limit theorem says that the distribution of the average will be closely approximated by a normal distribution. A simple example of this is that if one flips a coin many times the probability of getting a given number of heads in a series of flips will approach a normal curve, with mean equal to half the total number of flips in each series. (In the limit of an infinite number of flips, it will equal a normal curve.)

# Useful Interpretation

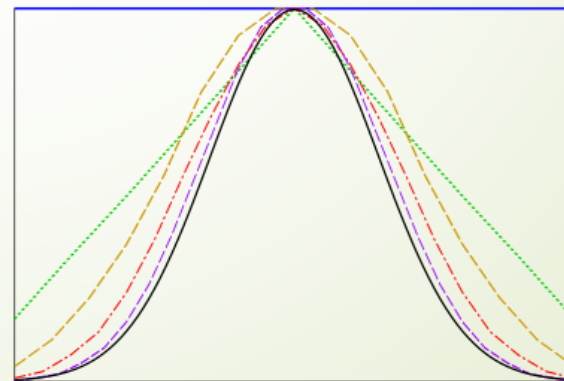
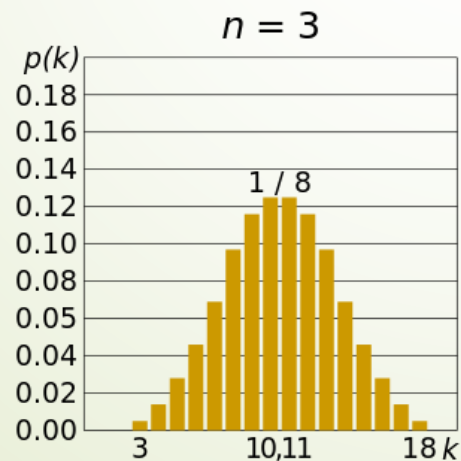
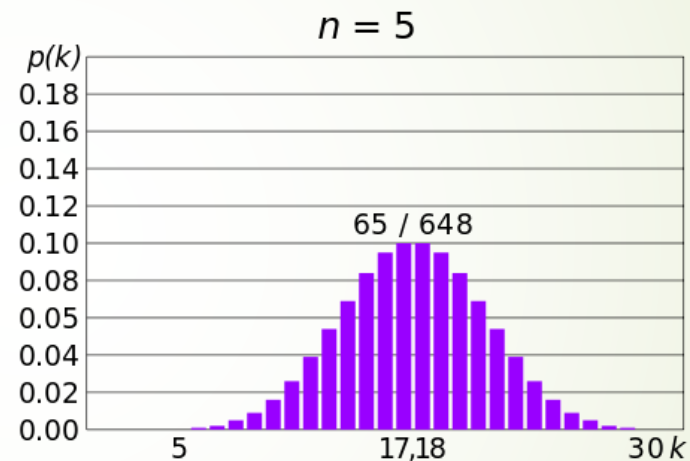
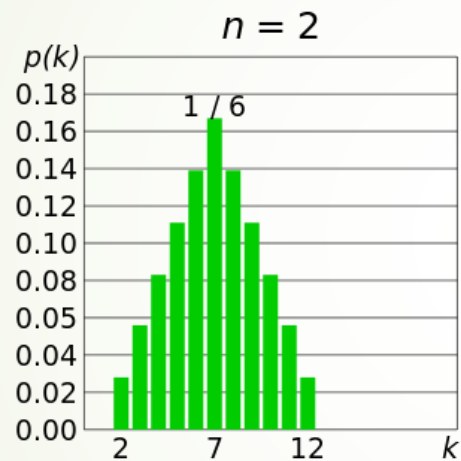
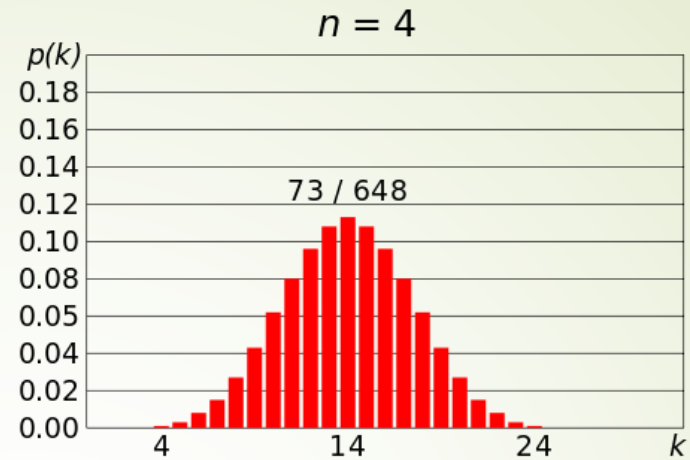
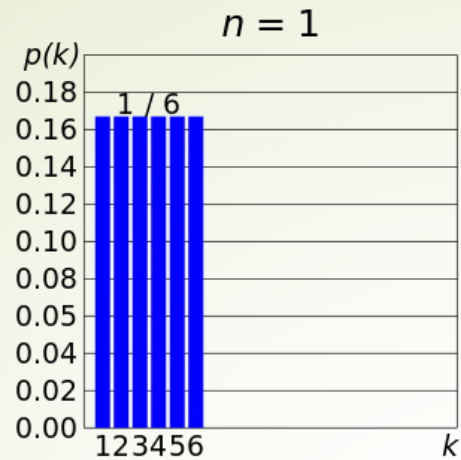
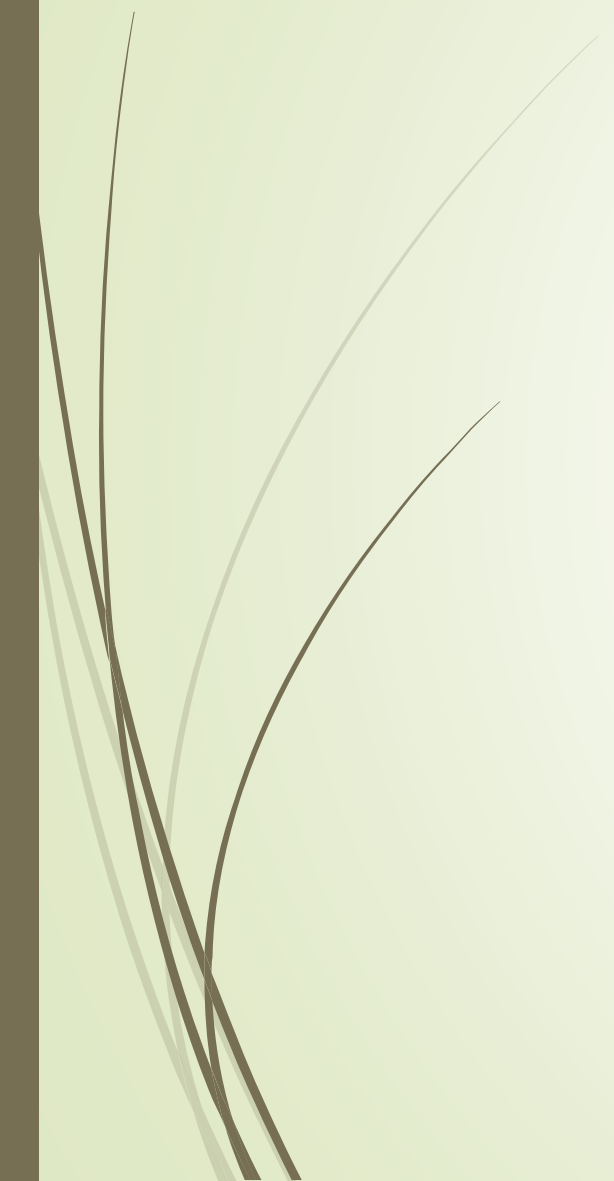

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

*Where,*

$\sigma_x$  = Std Deviation of Sample x

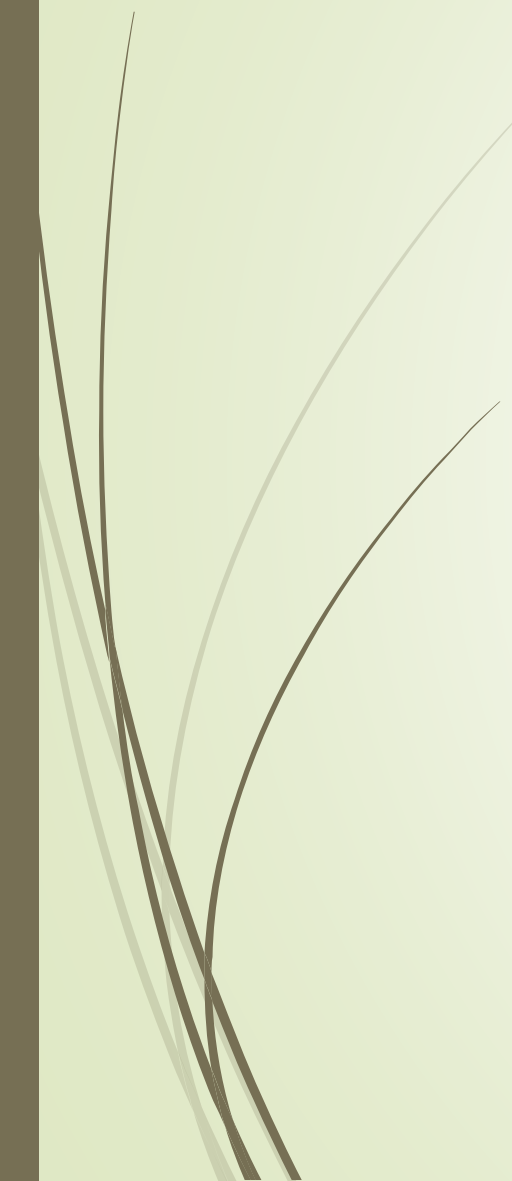
$\sigma$  = Std Deviation of population

$n$  = sample size





# Interval Estimation

- Population Mean:  $\mu$  Known
  - Population Mean:  $\mu$  Unknown
  - Population Std Dev:  $\sigma$  Unknown
  - Determining the Sample Size
  - Population Proportion
- 





# Margin of Error and Interval Estimate

- ▶ A point estimator cannot be expected to provide the exact value of the population parameter
- ▶ An interval estimate can be computed by adding and subtracting a margin of error to the point estimate

## **Point Estimate $\pm$ Margin of Error**

- ▶ The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

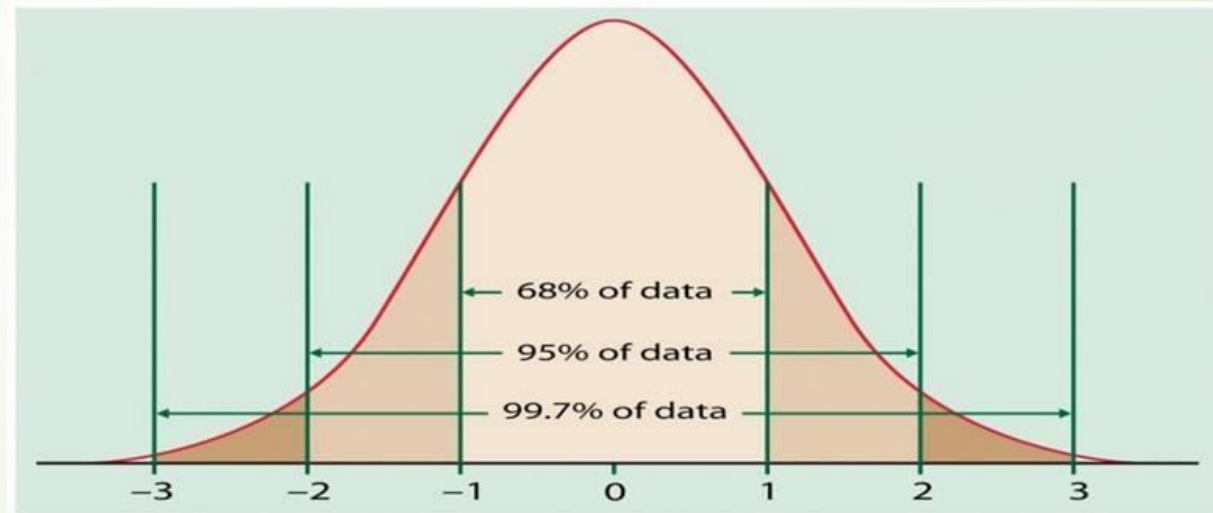


- 
- 
- In order to develop an interval estimate of a population mean, the margin of error must be computed using either: the population standard deviation  $\sigma$  , or the sample standard deviation  $s$
  - $\sigma$  is rarely known exactly, but often a good estimate can be obtained based on historical data or other information.
  - We refer to such cases as the  $\sigma$  known case.

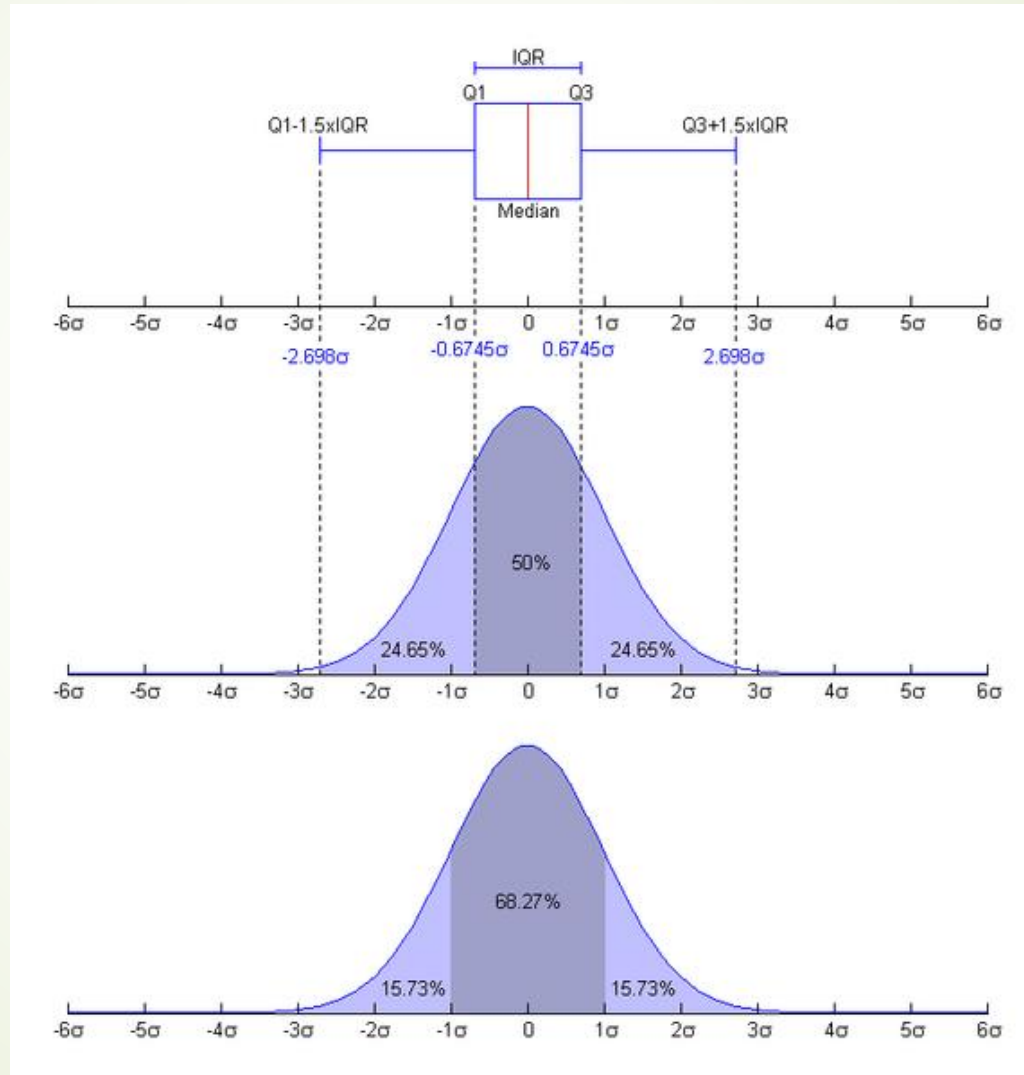


# Z-distribution

- In statistics, the **Z-distribution** is used to help find probabilities and percentiles for regular normal **distributions** (X). It serves as the standard by which all other normal **distributions** are measured. The **Z-distribution** is a normal **distribution** with mean zero and standard deviation 1



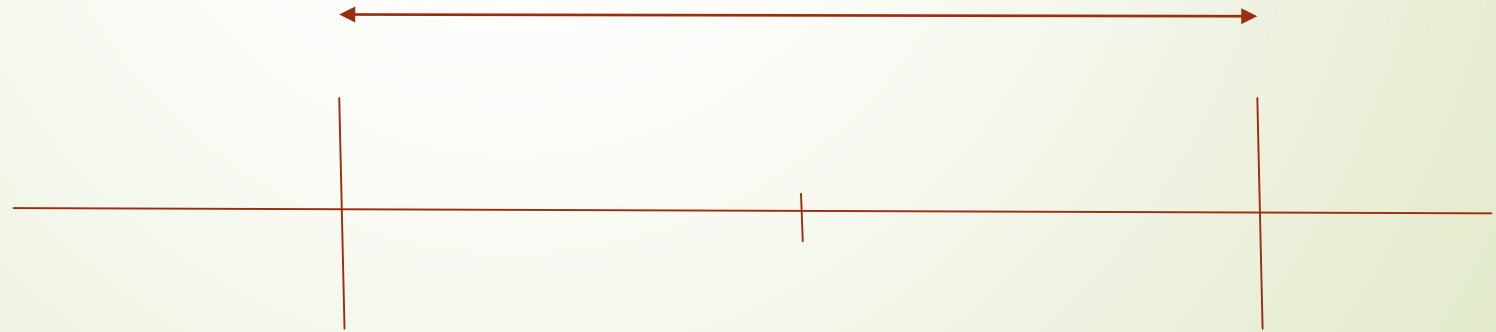
# A new picture



# Interval Estimate of a Population Mean


- After we found a point estimate of the population mean, we would need a way to quantify its accuracy. Here, we discuss the case where the population variance  $\sigma^2$  / std deviation  $\sigma$  is assumed known

Sampling distribution of  $\bar{X}_{\text{mean}}$





# How to Calculate Z


$$Z = \frac{x - \mu}{\sigma}$$



## Error Value

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

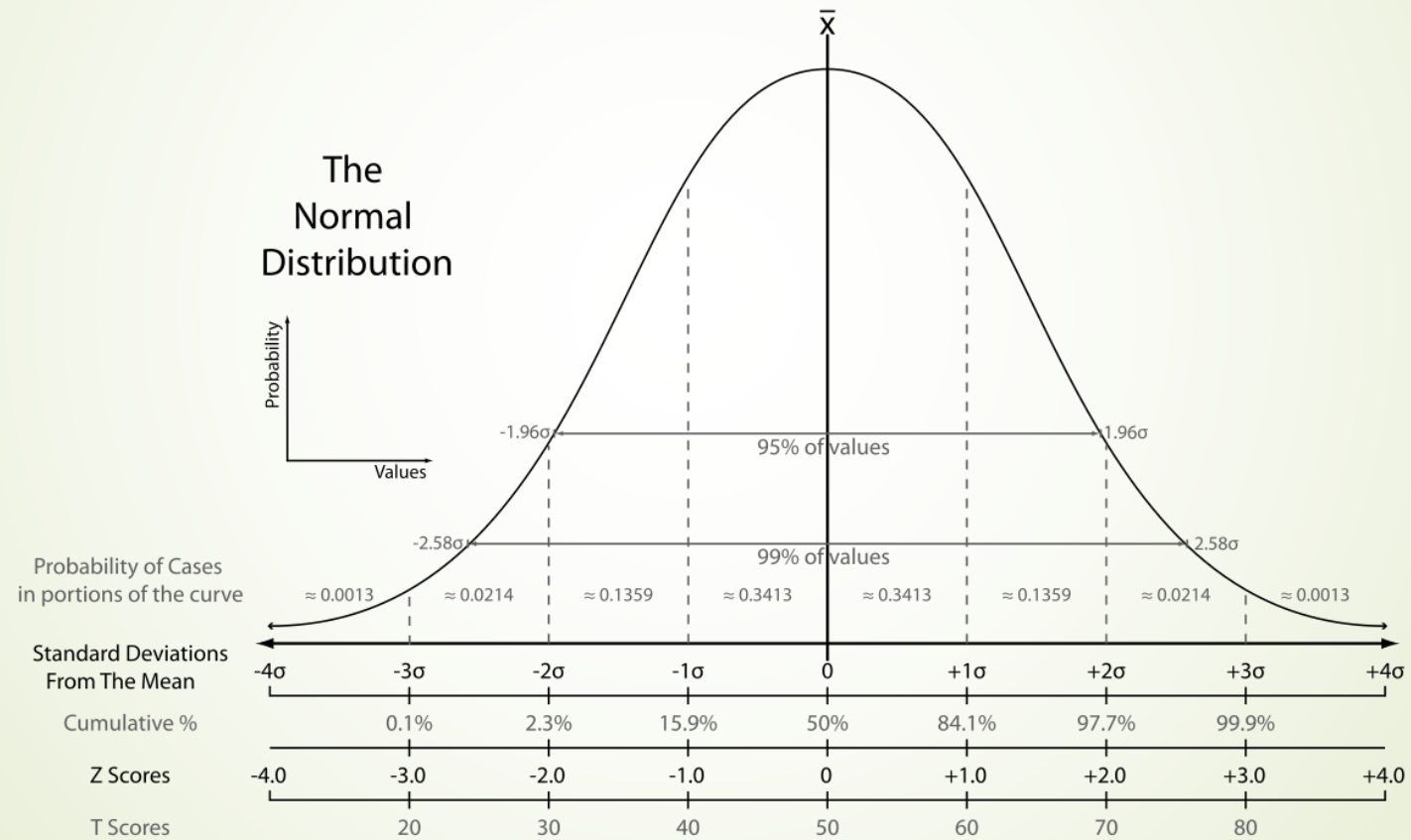
1 -  $\alpha$  is the confidence coefficient

$z_{\alpha/2}$  is the z value providing an area of  $\alpha/2$  in the upper tail of the standard normal probability distribution

$\sigma$  is the population standard deviation

$n$  is the sample size

# Confidence levels





# Z-score table

	50 %	80%	90 %	95 %	99 %
Z- Score	<b>0.674</b>	<b>1.282</b>	<b>1.645</b>	<b>1.960</b>	<b>2.576</b>



# Meaning of Confidence

► Because 90% of all the intervals constructed using  $\bar{X}_{\text{mean}} \pm 1.645 \sigma$  will contain the population mean, we say we are 90% confident that the interval contains the population mean  $\mu$ .

We say that this interval has been established at the 90% confidence level.

The value .90 is referred to as the confidence coefficient.

Similarly 95% Confidence Interval (C.I.) means  $\bar{X}_{\text{mean}} \pm 1.96 \sigma$  will contain the population mean  $\mu$



# Question to solve

Discount Sounds (company) has 260 retail outlets throughout the United States. The firm is evaluating a potential location for a new outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

A sample of size  $n = 36$  was taken;

the sample mean income is \$41,100. The population is not believed to be highly skewed.

The population standard deviation is estimated to be \$4,500

Find the Confidence Interval when Confidence Coefficient is 0.9 and 0.95



# Interval Estimate of a Population Mean:


$\sigma$  Unknown


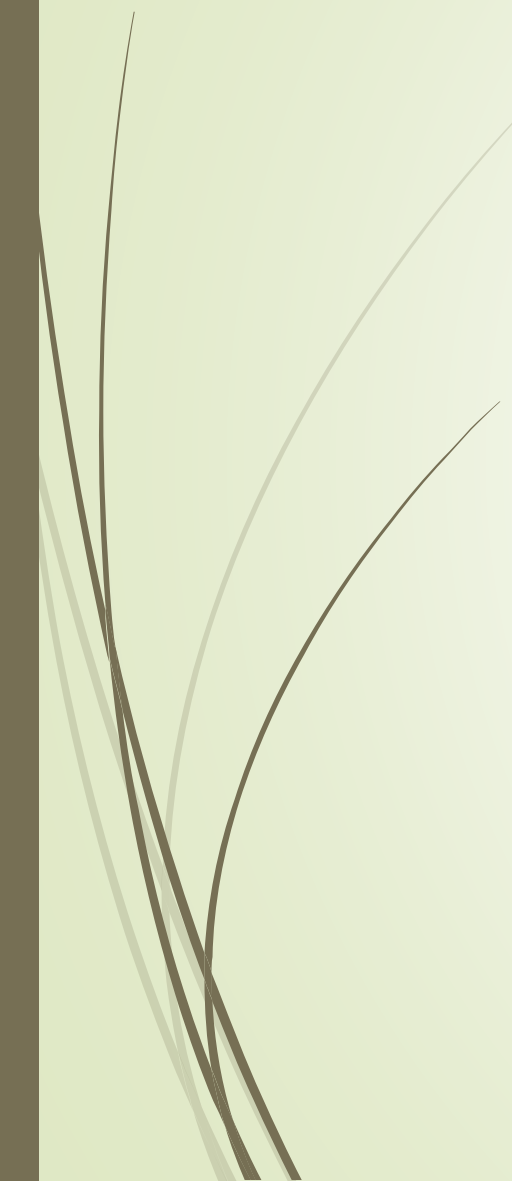
- ▶ If an estimate of the population standard deviation  $\sigma$  cannot be developed prior to sampling, we use the sample standard deviation  $s$  to estimate  $\sigma$ .
- ▶ This is the  $\sigma$  unknown case.
- ▶ In this case, the interval estimate for  $\mu$  is based on the  $t$  distribution.


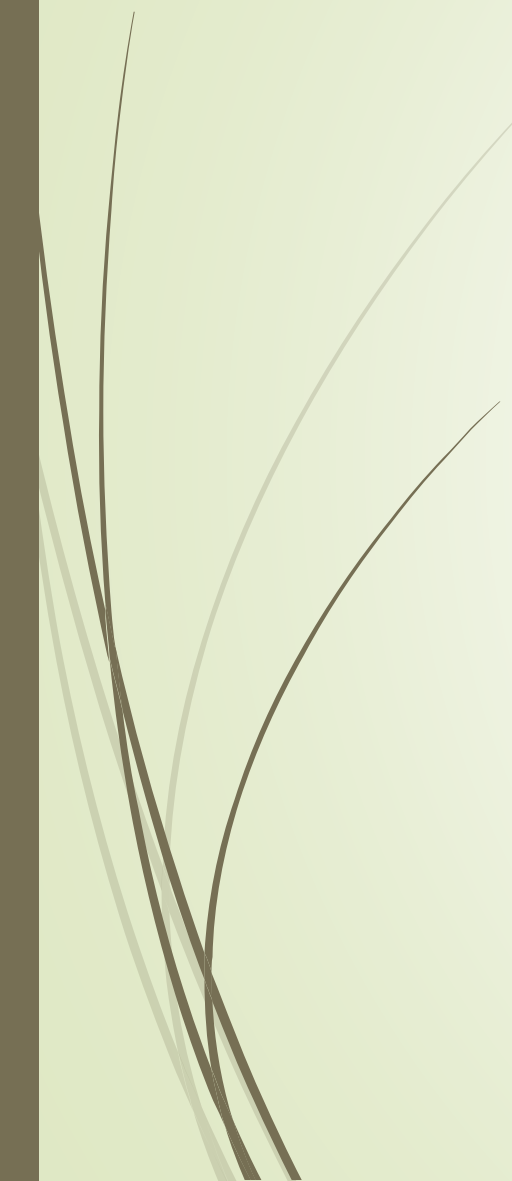
(We'll assume for now that the population is normally distributed.)



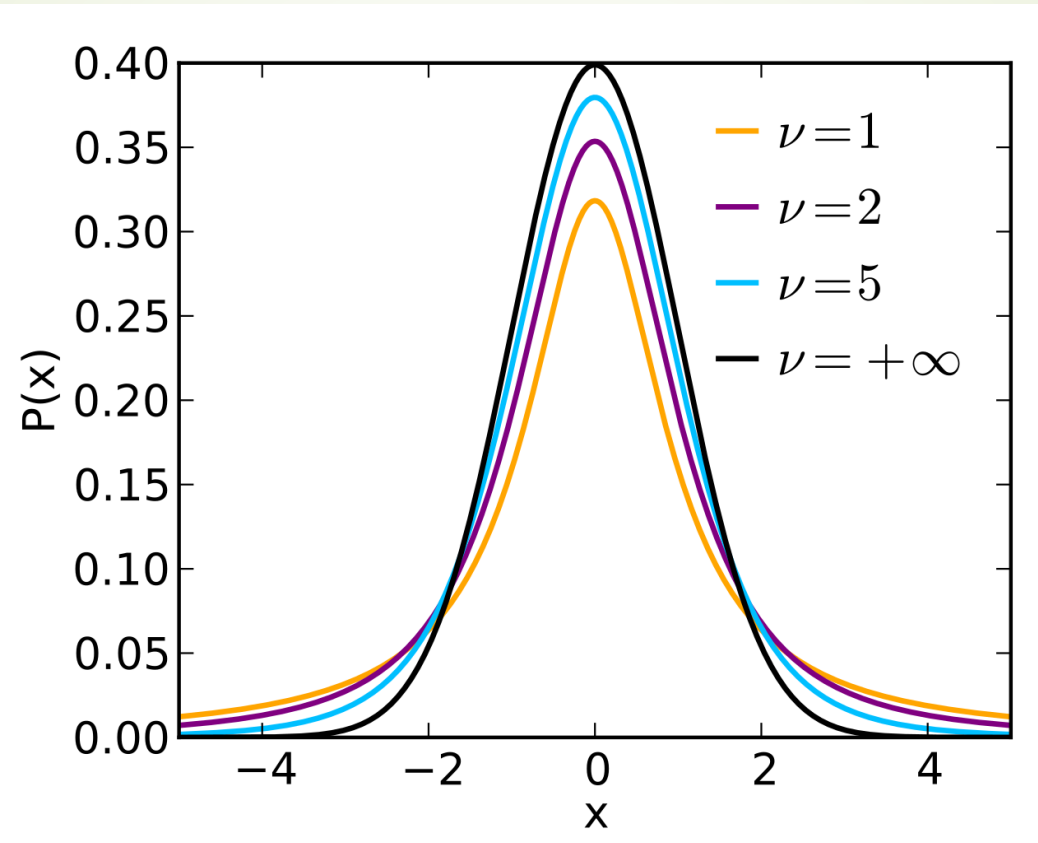
# t-distribution

- William Gosset, writing under the name “Student”, is the founder of the t distribution.
  - He developed the t distribution while working on small-scale materials and temperature experiments.
- 

- 
- 
- The  $t$  distribution is a family of similar probability distributions.
  - A specific  $t$  distribution depends on a parameter known as the degrees of freedom
  - Degrees of freedom refer to the number of independent pieces of information that go into the computation of  $s$
  - A  $t$  distribution with more degrees of freedom has less dispersion
  - As the degrees of freedom increases, the difference between the  $t$  distribution and the standard normal probability distribution becomes smaller and smaller.

- 
- 
- ▶ The  $t$  test (also called Student's T Test) compares two averages (means) and tells you if they are different from each other. The  $t$  test also tells you how significant the differences are; In other words it lets you know if those differences could have happened by chance
  - ▶ **A very simple example:** Let's say you have a cold and you try a naturalistic remedy. Your cold lasts a couple of days. The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week. You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you *really* want to know is, are these results repeatable? A  $t$  test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.





$$\nu = \text{dof} = n - 1$$



# Interval Estimate of a population mean

$\sigma$  Unknown

$$\Rightarrow \bar{x} \pm t_{\alpha/2} s / \sqrt{n}$$

Where,

$t$  = the  $t$  value providing an area of  $\alpha/2$  in the upper tail of a  $t$  distribution with  $n - 1$  degrees of freedom

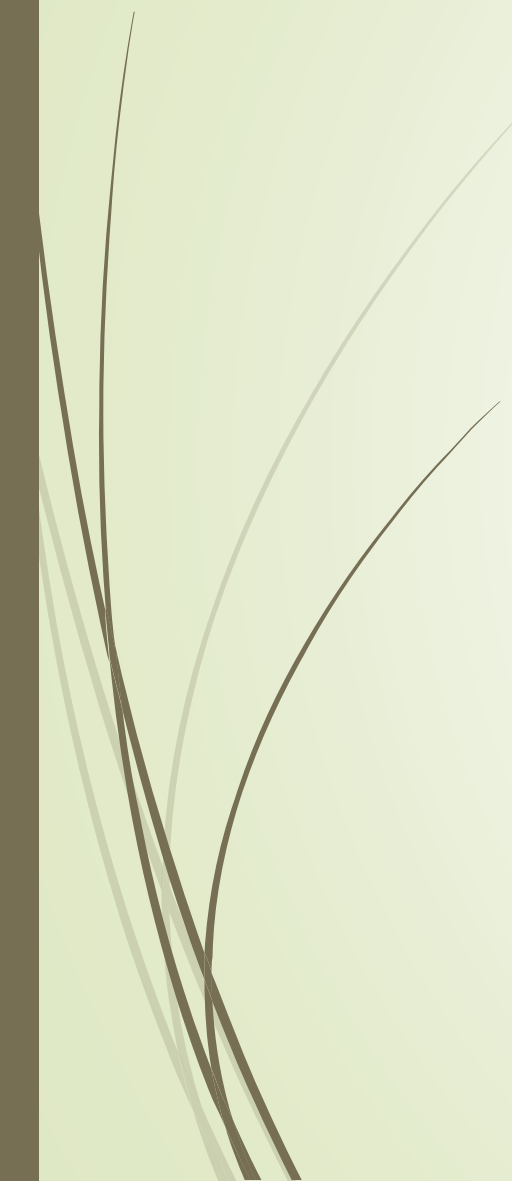
$s$  = the sample standard deviation



## Example: Apartment Rents

A reporter for a student newspaper is writing an article on the cost of off-campus housing. A sample of 16 one-bedroom apartments within a half-mile of campus resulted in a sample mean of \$750 per month and a sample standard deviation of \$55.

Let us provide a 95% confidence interval estimate of the mean rent per month for the population of one- bedroom efficiency apartments within a half-mile of campus. We will assume this population to be normally distributed





# Sample Size for an Interval Estimate of Population Mean

- ▶ Let  $E$  = the desired margin of error
- ▶  $E$  is the amount added to and subtracted from the point estimate to obtain an interval estimate
- ▶ If a desired margin of error is selected prior to sampling, the sample size necessary to satisfy the margin of error can be determined



## Some Formulae

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

Where n = necessary sample size



# Strategy

- The Necessary Sample Size equation requires a value for the population standard deviation  $\sigma$
- If  $\sigma$  is unknown, a preliminary or planning value for  $\sigma$  can be used in the equation.
  - Use the estimate of the population standard deviation computed in a previous study
  - Use a pilot study to select a preliminary study and use the sample standard deviation from the study
  - Use judgment or a “best guess” for the value of  $\sigma$



# Question to solve

- ▶ Recall that Discount Sounds is evaluating a potential location for a new retail outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location. Sample Size for an Interval Estimate of a Population Mean

Suppose that Discount Sounds' management team wants an estimate of the population mean such that there is a .95 probability that the sampling error is \$500 or less. How large a sample size is needed to meet the required precision?