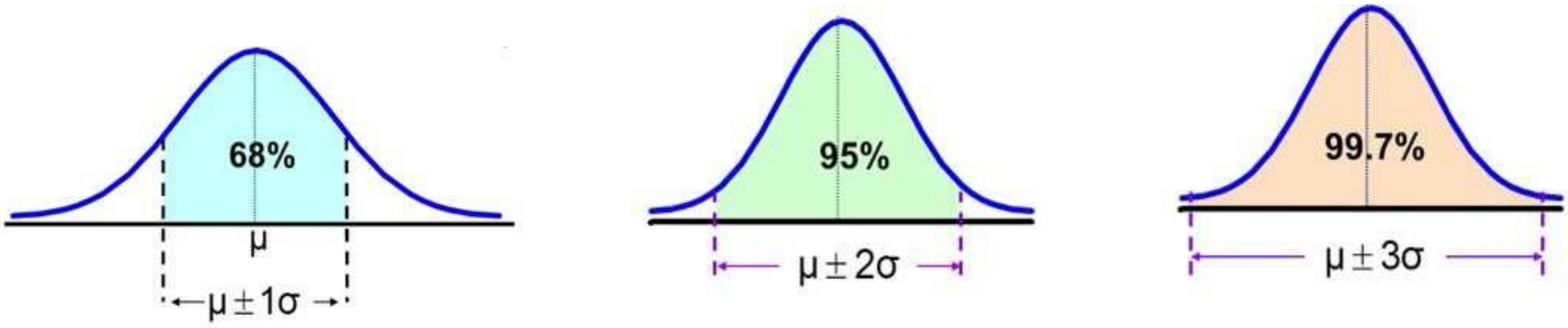
Fundamentals of Business Statistics

Hypothesis Testing

Normal Distribution - Recap

- Following are the approximate numbers



Normal Distribution - Recap

- The maximum daily temperature in City A in the month of February is normally distributed with mean 40 °C and standard deviation of 3 °C
- The maximum daily temperature in City B in the month of February is normally distributed with mean 30 °C and standard deviation of 2 °C

Questions

- What is the probability that in City A, the max. temperature on a day will be less than 43 °C
- What is the probability that in City B, the max. temperature on a day will be less than 32 °C

Normal Distribution - Recap

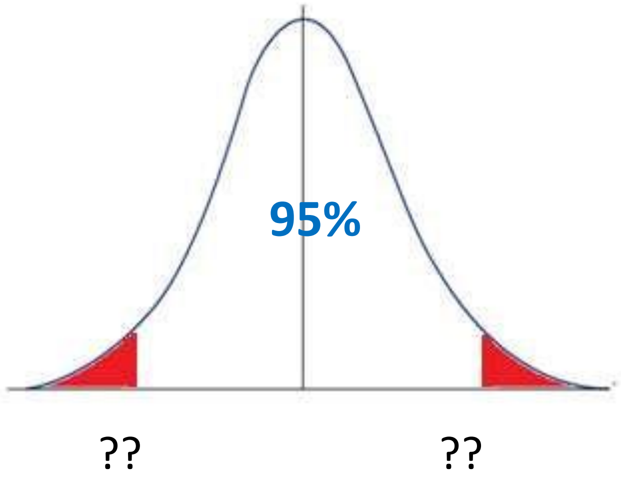
- The Standard Normal Variable is defined as follows:

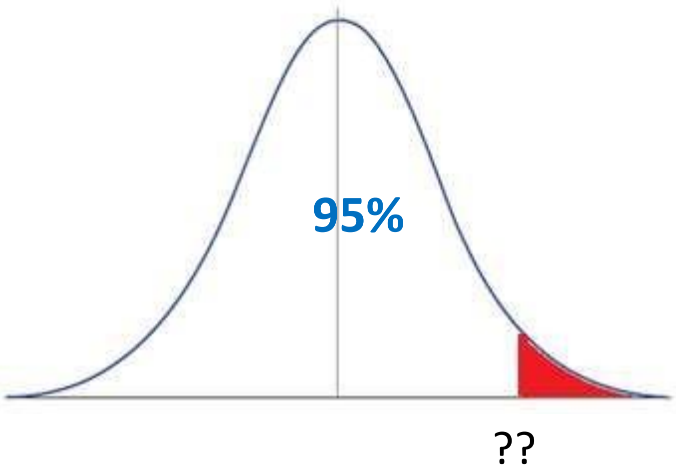
$$z = \frac{x - \mu}{\sigma}$$

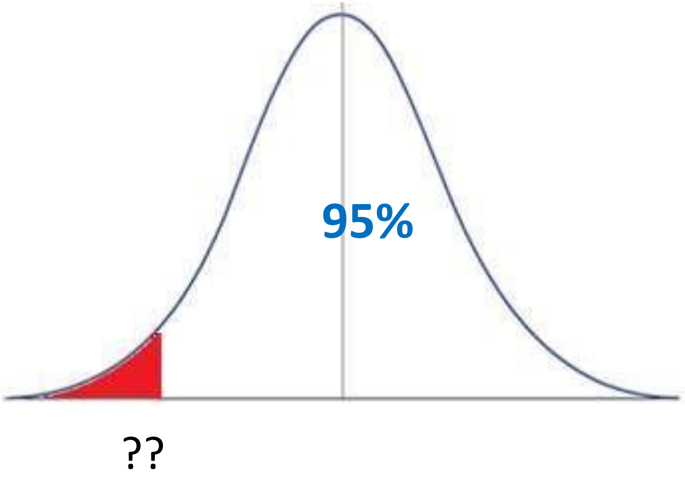
- Please note that Z is a pure number independent of the unit of measurement. The random variable Z follows a normal distribution with mean=0 and standard deviation =1.

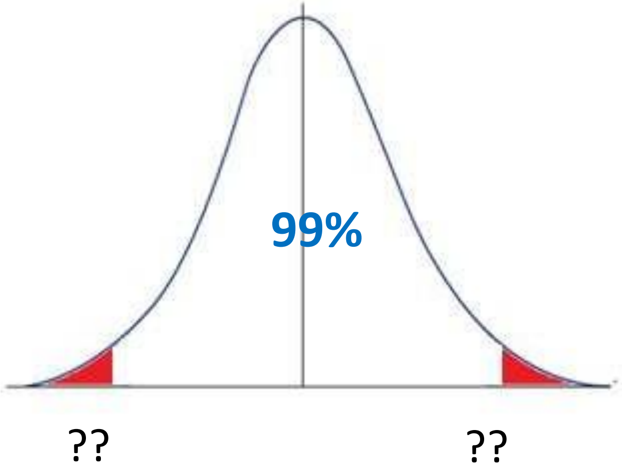
Exercise

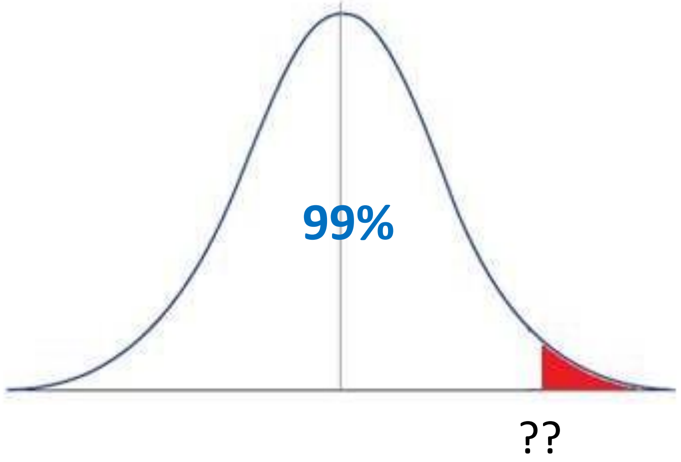
Find Some of the Commonly Used Numbers

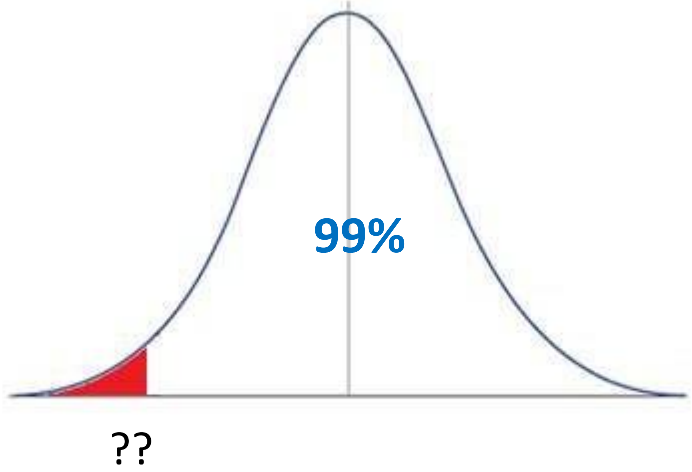












Preliminaries

Sampling Distribution-A Conceptual Framework

- The probability distribution of all the possible values a “sample statistic” can take is called sampling distribution, of the statistic.

- Sample mean and sample proportion based on a random sample are examples of sample statistic(s).

Concept of Standard Error

- The standard deviation of the sample statistic is called the Standard Error of the Statistic.
- The standard deviation of the distribution of the sample means is called the Standard Error of the mean.
- Likewise, the standard deviation of the distribution of the sample proportions is called the Standard Error of the proportion.

Sampling Distribution of Mean-Normal Population

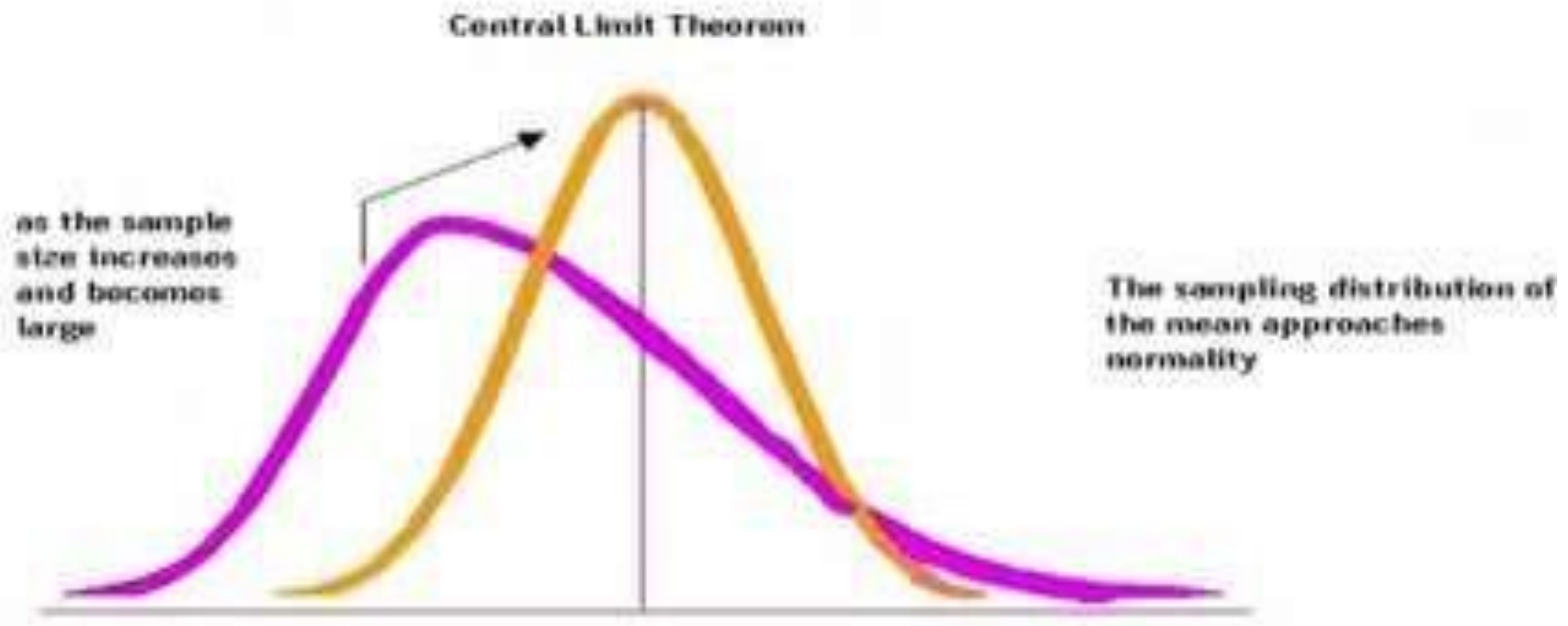
- If X1, X2, X3,…….,Xn are n independent random samples drawn from a Normal Population with Mean = µ and Standard Deviation = σ, then the sampling distribution of X follows a Normal Distribution with Mean = µ , and Standard Deviation = $\frac{\sigma}{\sqrt{n}}$

$\frac{\sigma}{\sqrt{n}}$ is known by the term Standard Error.

Central limit theorem

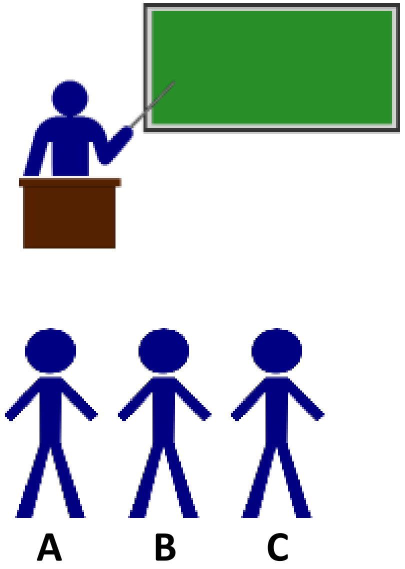
- The distinguishing and unique feature of the central limit theorem that irrespective of the shape of the original distribution, the sampling distribution of mean will approach a normal distribution as the size of the sample increases and becomes large.

Central Limit Theorem-Picture



Hypothesis - Basics

Hypothesis and Probability



- What is the probability that name of A will not be picked (with replacements) in 12 attempts?
- Answer: Probability = $(2/3)^{12} = 0.0077 = 0.77\%$

What is a statistical hypothesis?

- A statistical hypothesis is a statement about a population parameter. It may or may not be True. The analyst has to ascertain the truth of the hypothesis.

Null and Alternative hypothesis

- A Null Hypothesis is status quo. It is so formulated that its rejection leads to the conclusion which is Alternative hypothesis.

- Researchers and Decision makers generally want to prove the Alternative hypothesis.

Null and Alternative hypothesis

Null Hypothesis

- H_0
- Status quo
- Always has equality (permitted signs <= or = or >=)
- Refers to a value of population parameter (and not sample statistic)

Alternate Hypothesis

- H_a or H_1
- **The change claim / what we want to prove**
- Never has equality (permitted signs > or ≠ or <)
- Also refers to value of population parameter (and not sample statistic)

Null and Alternative hypothesis

- Does the data provide ‘sufficient’ confidence to Reject H_0
- Rejecting H_0 means Accept H_1

- Possible results of Hypothesis test:
  - **Reject H_0** i.e. Accept H_1

  OR

  - Fail to Reject H_0 (**Accept H_0**)

- Fail reject H_0 **does not** mean that we have proven H_0 to be true

Hypothesis Formulation Exercises

State the Null and Alternative Hypothesis for the following:

a) Store manager believes that the average waiting time for the customers of Smart Supermarket at the checkouts has become worse and it is more than 15 minutes. Formulate the hypothesis.

b) ATV company suspects that, the proportion of households owning Smart TVs in Chennai is more than 5%?

c) Is the average expenditure per household on eating out significantly higher in Bangalore than in Calcutta?

d) A pharmaceutical company has developed a new improved drug. The company claims that it takes less than 12 minutes for the drug to enter patient's bloodstream. What should be Null and Alternate hypothesis to convince the FDA to approve this claim?

Errors associated with Hypothesis Formulation and Testing:

- Are the decision (e.g. Rejecting the Null Hypothesis or Failing to Reject the Null Hypothesis) made using Hypothesis test, always correct?

Type I error and Type II error

		Truth	
		H_0 True	H_o False (H_1 True)
Statistical Decision	Reject H_0	Type-I Error (α)	No Error
	Fail to Reject H_0 (Accept H_0)	No Error	Type-II Error (β)

		Truth	
		Not Guilty (H_0)	Guilty (H_1)
Decision	Guilty (Reject H_0)	Error	No Error
	Not Guilty (Do not reject H_0)	No Error	Error

Analogy

Type I error and Type II error

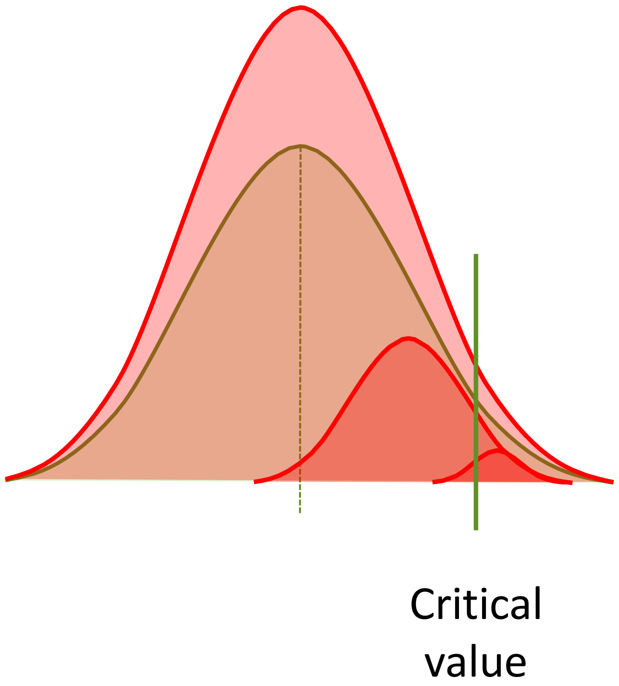
Level of significance (α) is the probability of the occurrence of a Type 1 error

		Truth	
		H_0 True	H_o False (H_1 True)
Statistical Decision	Reject H_0	Type-I Error (α)	No Error
	Do not reject H_0 (Accept H_0)	No Error	Type-II Error (β)

Confidence level = 1-α

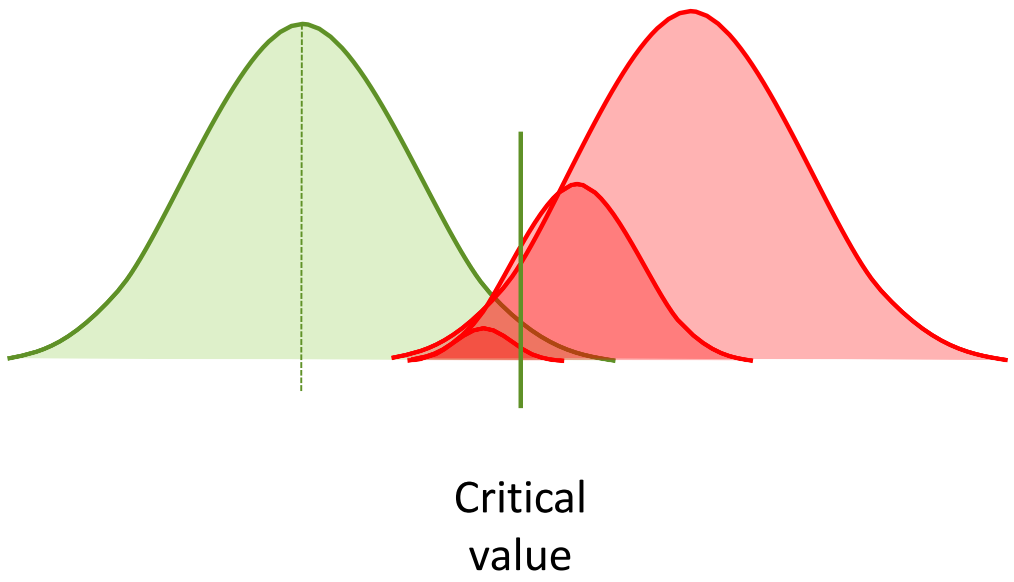
Type-I Error

- Type-I error– Rejecting Null Hypothesis when it is True.

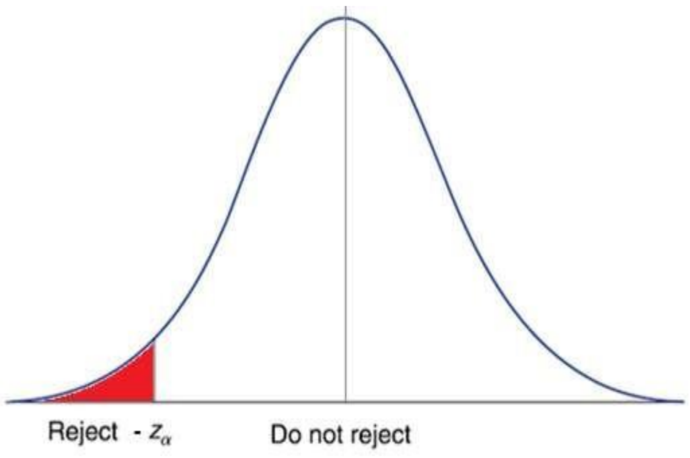


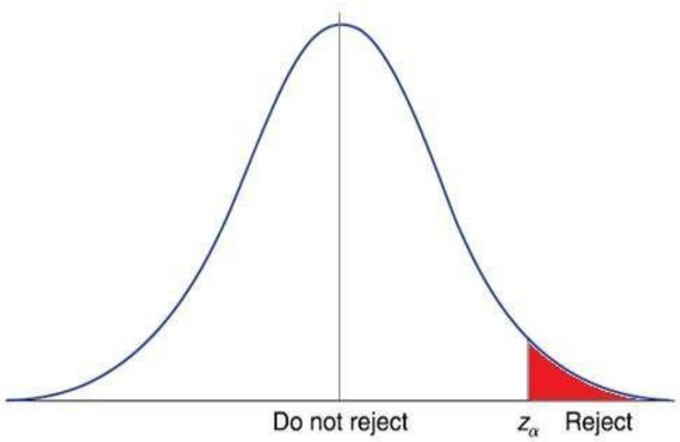
Type-II Error

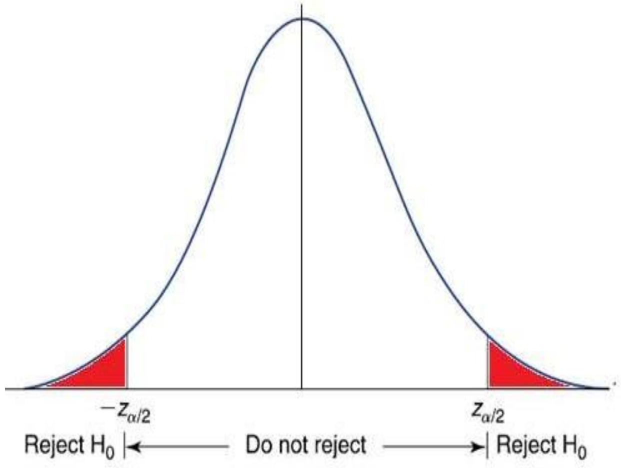
- Type-II error– Failure to Reject Null Hypothesis when it is False. i.e. Accepting Null Hypothesis when it is False.



One-tailed v/s Two-tailed test: Accept and Reject zones







- Lower tail test.
- H_1: $\mu <$ ……

- Upper tail test.
- H_1: $\mu >$ ……

- Two tail test.
- H_1: $\mu \neq$ ……

Z-Test of Mean (sigma is known)

Test Statistic : **Z statistic =** $\frac{(\overline{X}-\mu)}{(\sigma/\sqrt{n})}$

In the subsequent examples of Z-test, we will see how to formulate the hypothesis, set the significance level, calculate the p-value and decide whether to accept or reject the null hypothesis.

Example – Processing time [Z Test]

Tom is working in a credit card processing company as a team leader. His team is responsible to validate certain data for new credit card applications. The time spent by his team on an application is normally distributed with average 300 minutes and standard deviation 40 minutes.

Tom and his team worked on process improvement to reduce the time spent in processing new applications. After implementing the improvements, Tom checked the time spent by his team on randomly selected 25 new card applications. The average time spent is 290 min. Tom is happy that, though it is a small improvement, it is a step in right direction. He shares the good news with his manager Lisa. But Lisa in not convinced about the improvement.

At 95% confidence, is the processes really improved?

Example – Processing time

- Step 1: Given: n = 25, $\overline{x}$ = 290, σ = 40

- Step 2: Formulate Hypothesis

  $H_0 : \mu = 300$

  $H_1 : \mu < 300$

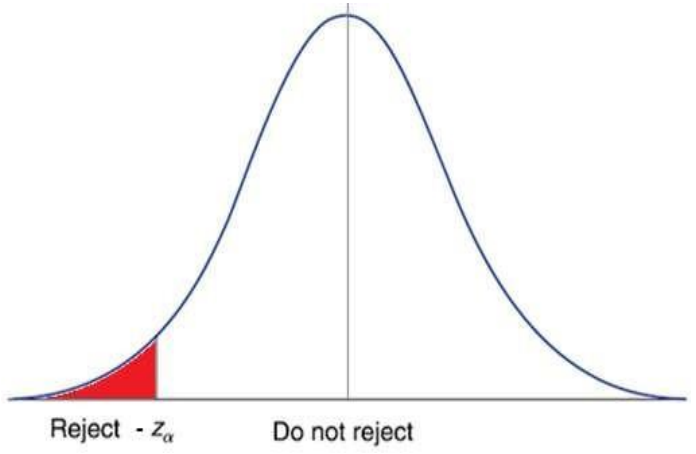
- Step 3: Define test statistic

  $Z_{stat} = (\overline{x} - \mu) \,/\, (\sigma / \sqrt{n})$

  $= (290 - 300) / (40/\sqrt{25}) = -10 / 8 = -1.25$

Example – Processing time

- Step 4: Draw diagram



- Step 5: (critical value method)

  Determine critical values

  α = 5% = 0.05.

  $-Z_\alpha = -1.64$

- Step 6: (critical value method)

  Check whether Zstat value is in accept or reject region and make decision

  Since Zstat is in Accept region, H_0 is <u>not</u> rejected. i.e. H_0 is accepted.

  i.e. <u>At 95% confidence, we cannot claim that the process is improved</u>

Example – Processing time

- Step 5: (p-value method)
  - Find p-value.
  - P-value = 0.1056

- Step 6: (p-value method)
  - Compare p-value with α. If p-value < α then Reject H_0 else Accept H_0

Since p-value <u>not less than</u> α, therefore H_0 is <u>not</u> rejected. i.e. H_0 is accepted.

i.e. <u>At 95% confidence, we cannot claim that the process is improved</u>

Volume of Liquid

- John is a quality control analyst in a plant which required to fill 500 ml of liquid in bottles. Past studies have revealed that the bottles are filled with standard deviation 3 ml.
- John wants to check if the volume of liquid filled in bottles has changed. If the volume has changed then it is required to halt the production and reconfigure the machines. John takes 40 random samples and measures the volume of filled liquid. The sample mean is 501.5 ml.
- At 95% confidence level, is it required to reconfigure the machines?

Example – Liquid volume

- Step 1: Given: σ = 3ml, n = 40, $\bar{x} = 501.5$

- Step 2: Formulate Hypothesis

  H_0 : No change volume. μ = 500ml.

  H_1 : Change volume $\mu \neq$ 500ml.

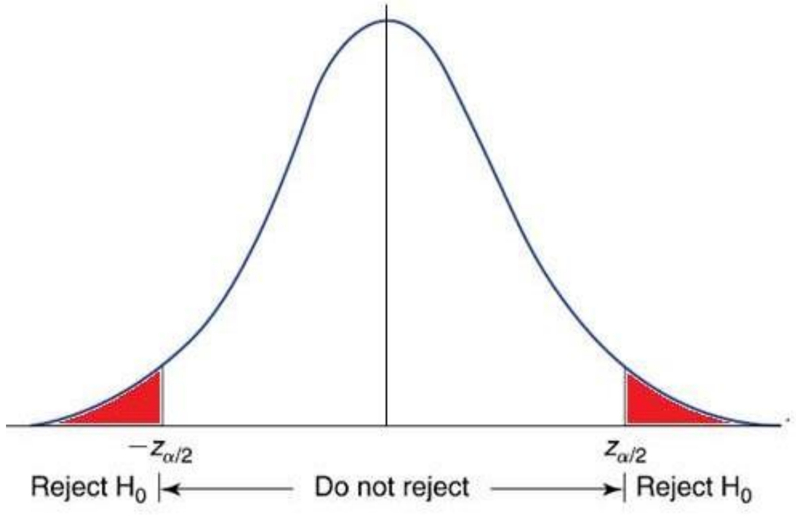
- Step 3: Define test statistic

  $Z_{stat} = (\bar{x} - \mu) / (\sigma / \sqrt{n})$

  $= (501.5\text{-}500) / (3/\sqrt{40}) = 1.5 / (3 / 6.324) = 3.1622$

Example – Liquid volume

- Step 4: Draw diagram



- Step 5 (critical value method):

  Determine critical values

  α = 5% = 0.05. This is two tailed test.

  $Z_{\alpha/2} = 1.96$.

  $-Z_{\alpha/2} = -1.96$.

- Step 6 (critical value method):

  Compare whether Z_{stat} value is in reject region and make decision

  Since $Z_{stat} = 3.1622$ is in Reject region. H_0 is rejected. i.e. H_1 is accepted.

Example – Liquid volume

- Step 5 (p-value method):
  - Find p-value. *If p-value is less than α, then Reject the Null Hypothesis*.
  - p-value = The probability of getting sample statistic equal of or more extreme than the observed sample statistic, if the Null Hypothesis is true
  - P-value = 0.0015654

- Step 6 (p-value method):
  - Since p-value < α, therefore Reject null hypothesis i.e. H_1 is accepted

Launching a Product Line into a New Market Area

Karen, product manager for a line of apparel, to introduce the product line into a new market area.
Survey of a random sample of 400 households in that market showed a mean income per household of $30,000. Standard deviation for the sample of 400 households is $8,000.
Karen strongly believes the product line will be adequately profitable only in markets where the mean household income is greater than $29,000. Should Karen introduce the product line into the new market?

Example – Product Launch

Step 1: Given: $n = 400$, $\bar{x} = 30{,}000$, $s = 8000$

Step 2: Formulate Hypothesis

$H_0 : \mu <= 29{,}000$

$H_1 : \mu > 29{,}000$

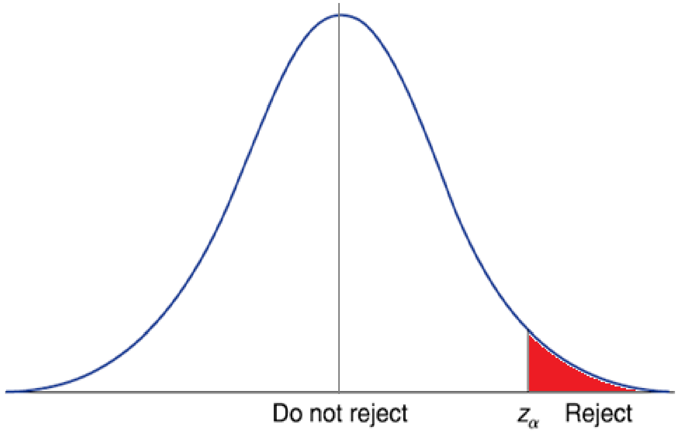
Step 3: Define test statistic

$Z_{stat} = (\bar{x} - \mu) / (\sigma / \sqrt{n})$ But σ is not known

$= (30000 - 29000) / (8000/\sqrt{400}) = 1000 / (8000 / 20) = 2.5$

Example – Product Launch

Step 4: Draw diagram



Step 5: (critical value method)

Determine critical values

α = 5% = 0.05. This is upper tail test.

Z_α = 1.64

Step 6: (critical value method)

Compare whether Zstat value is in reject region and make decision

Since Zstat is in Reject region, H0 is rejected. i.e. H1 is accepted.

Sample size determination for Mean

Confidence Interval Estimate is

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Sampling error is

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

α depends on confidence level. If e is given and σ is known then solve for n to find the sample size required

Confidence Interval

The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)

Where:

- Point Estimate is the sample statistic estimating the population parameter of interest

- Critical Value is a table value based on the sampling distribution of the point estimate and the desired confidence level

- Standard Error is the standard deviation of the point estimate

Sample size determination for Mean

Confidence Interval Estimate is

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Sampling error is

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

α depends on confidence level. If e is given and σ is known then solve for n to find the sample size required

Sample Size-Problem 1

- A marketing manager of a fast food restaurant in a city wishes to estimate the average yearly amount that families spend on fast food restaurants. He wants the estimate to be within Rs 100 with a confidence level of 99%. it is known from an earlier pilot study that the standard deviation of the family expenditure on fast food restaurant is Rs 500. How many families must be chosen for this problem?

Sample Size-Problem 1

Solution

$$e = Z_{\alpha T_2} \frac{\sigma}{\sqrt{n}}$$

$$100 = 2.58 \frac{500}{\sqrt{n}}$$

$$\mathrm{n} = 166$$

t Test

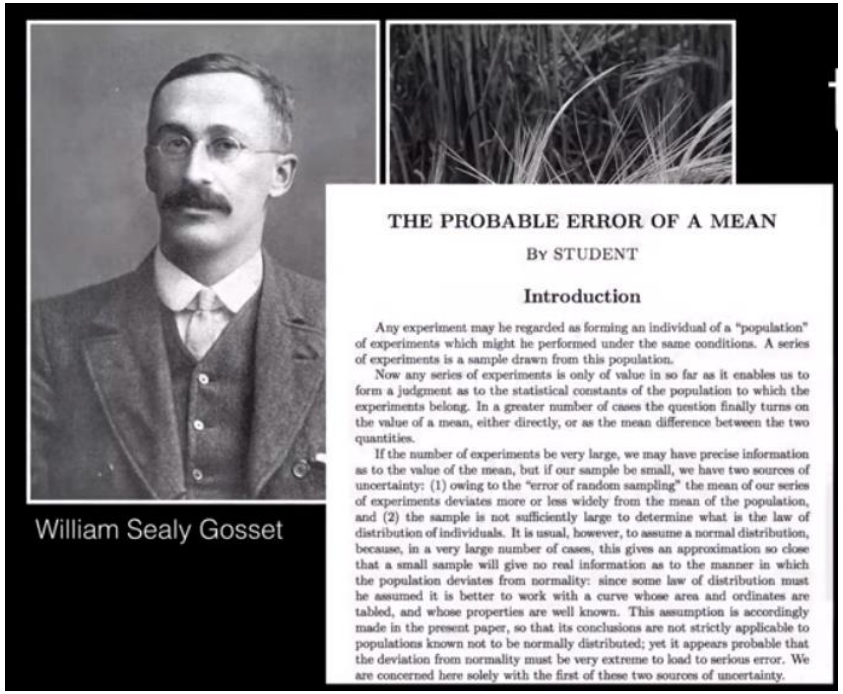
- Developed by William Gosset

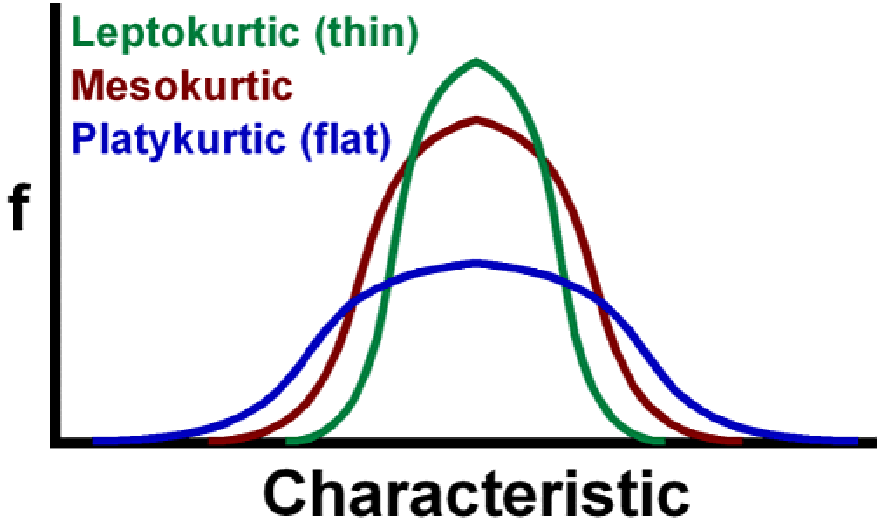
- Useful when Population standard deviation is not known and Population is Normally distributed

T Test

Developed by William Gosset

- If the population standard deviation is unknown and the sample size is small, you instead use the sample standard deviation S.
- Because of this change, you use the t-distribution instead of the Z-distribution to test the null hypothesis about the mean.
- When using the t-distribution you must assume the population you are sampling from follows a normal distribution.
- All other steps, concepts, and conclusions are the same.





t-test requires Degrees of Freedom (df)

Degrees of freedom is the number of values that are free to vary when the value of some statistic, like $\bar{X}$ or $\hat{\sigma}^2$, is known.

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

If the mean of these three values is 8.0,
then X_3 must be 9
(i.e., X_3 is not free to vary)

Here, n = 3, so degrees of freedom = n – 1 = 3 – 1 = 2

(2 values can be any numbers, but the third is not free to vary for a given mean)

t Test -One Sample - Example

Experian Marketing Services reported that the typical American spends a mean of 144 minutes (2.4 hours) per day accessing the Internet via a mobile device. (Source: The 2014 Digital Marketer, available at ex.pn/IkXJjfX.) In order to test the validity of this statement, you select a sample of 30 friends and family. The results for the time spent per day accessing the Internet via mobile device (in minutes) are stored in InternetMobileTime

a. Is there evidence that the population mean time spent per day accessing the Internet via mobile device is different from 144 minutes? Use the p-value approach and a level of significance of 0.05.

b. What assumption about the population distribution is needed in order to conduct the t test in (a)?

Problem 9.35 from the Textbook adapted for Classroom Discussion(Chapter 9-page 314)

Example – Internet Mobile Time

- Step 1: Given: $n = 30$, data in excel

- Step 2: Formulate Hypothesis

  $H_0 : \mu = 144$ $H_1 : \mu \neq 144$

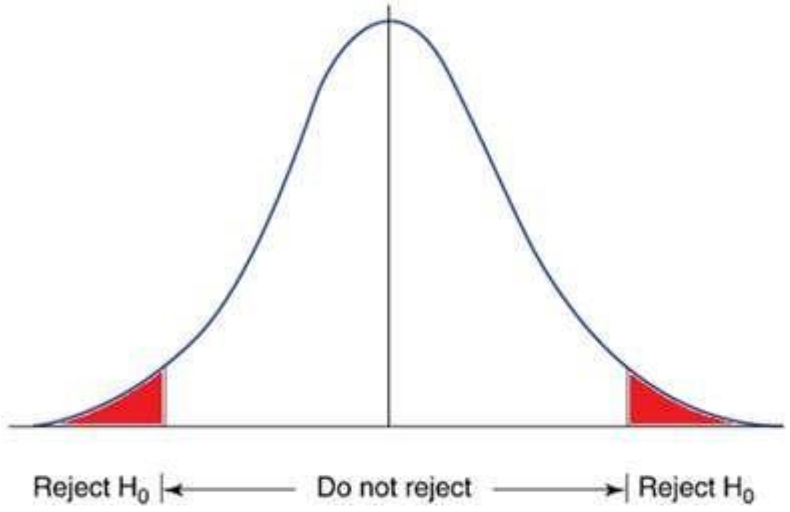
Step 3: Define test statistic

$t_{stat} = (\bar{x} - \mu) / (s / \sqrt{n})$

$= 1.2246$

Example – Internet Mobile Time

- Step 4: Draw diagram



- Step 5: (critical value approach)

  Determine critical values

  α = 5% = 0.05. This is two tailed test.

  $t_{\alpha/2} = 2.045$

  $-t_{\alpha/2} = -2.045$

- Step 6: (critical value approach)

  Compare whether tstat value is in reject region and make decision

  Since t_{stat} is in Accept region, H_0 is not rejected.

Example – Internet Mobile Time

- Step 5 (p-value approach):
  - Find p-value
  - P-value = 0.23055

- Step 6 (p-value approach):

  Since p-value > α, therefore do not reject Null Hypothesis

T Test-Two Independent Sample

- A hotel manager looks to enhance the initial impressions that hotel guests have when they check in. Contributing to initial impressions is the time it takes to deliver a guest's luggage to the room after check-in. A random sample of 20 deliveries on a particular day were selected in Wing A of the hotel, and a random sample of 20 deliveries were selected in Wing B. The results are stored in **Luggage** . Analyze the data and determine whether there is a difference between the mean delivery times in the two wings of the hotel. (Use a = 0.05.)
- Problem 10.83 from the Textbook adapted for Classroom Discussion(Chapter 10-page 387)

Example – Luggage

- Step 1: Given: $n_1 = 20$, $n_2 = 20$, data of observations in excel
- Step 2: Formulate Hypothesis

$H_0 : \mu_1 = \mu_2$ i.e. $\mu_1 - \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$ i.e. $\mu_1 - \mu_2 \neq 0$

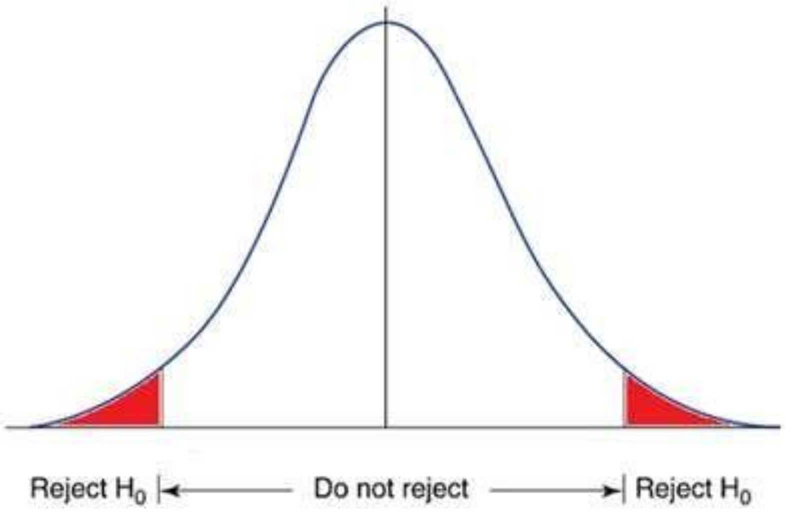
- Step 3: Define test statistic

$$t_{stat} = \frac{(\overline{X_1} - \overline{X_2}) - (\mu1 - \mu2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \text{where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

$= 5.1615$

Example – Luggage

- Step 4: Draw diagram



- Step 5 (p-value approach):
  - P-value = 0.00008 (output of t.test)

- Step 6 (p-value approach):

  Since p-value < α, therefore reject Null Hypothesis

Paired t Test

- Useful for comparing means of two related populations. Repeated measures taken on a same individual or same item. Objective is to compare the mean before and after

	Sample 1	Sample 2	Difference
	X11	X21	D1 = X11 - X21
	X12	X22	D2 = X12 - X22
	X13	X23	…
	…	…	
	X1n	X2n	Dn = X1n - X2n
Mean	$\overline{X1}$	$\overline{X2}$	$\overline{\boldsymbol{D}}$
Standard Deviation			$\mathbf{S_D}$
True mean	μ_1	μ_2	μ_D

$$t_{stat} = (\overline{D} - \mu_D) / (S_D / \sqrt{n})$$