

Data Science Concepts

Lesson01B–Basic of Statistics

Objective

After completing this lesson you will be able to:

- Explain the basic concepts of statistics
- Understand the application of these concepts in statistical modelling



Random Variable

Concept of random variable comes up to model variability around different businesses. Say,

- The return on an investment in a one-year period
- The price of an equity
- The number of customers entering a store on a given day
- The sales volume of a store on a particular day
- The turnover rate at your organization next year
- Cost of treatment at a hospital
- Attrition rate in IT company next year

A random variable is a numerically valued variable which takes on different values with given probabilities.

Type of Random Variable

Continuous Random Variable

May take uncountable number of possible values

- The return on an investment in a one-year period
- Cost of treatment at a hospital
- Attrition rate in IT company next year

Discrete Random Variable

Takes countable number of possible values

- Number of customer count entering a store
- Number of washing machine sold
- The sales volume of a store on a particular day



Randomness of a random variable is defined using **probability distribution**. Probability distribution, specifies the likelihood for a random variable to assume a particular value.

Probability distribution function

If X denotes a random variable with x being a possible value which X can take then:

- The probability distribution in case of a Discrete variable is called **Probability mass function**. It is depicted by $P(X = x)$ for all possible values of x
- The probability distribution in case of a continuous variable is called **Probability density function**.
 - It is depicted by $f(X = x)$ such that probability of observing a value h is given by $f(x) \cdot h \approx P(x < X \leq x + h)$ for small positive h .



The probability mass function specifies the actual probability, while the probability density function specifies the probability rate; both can be viewed as a measure of “likelihood”.

Probability Mass Function basic requirement

- The probability mass function should satisfy the following two requirements:

$$0 \leq p(X = x) \leq 1 \text{ for all } x$$

$$\sum_{\text{for all } x} P(x) = 1$$

- The probability density function should satisfy the following two requirements:

$$f(x) \geq 0 \text{ for all } x$$

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Example of Discrete Random Variable

Empirical data can be used to estimate the probability mass function

#TVs	#Household	x	$P(x)$	CDF	Expected Value	Expected Variance
0	1218	0	0.012	0.012	0.000	0.052
1	32379	1	0.319	0.331	0.319	0.375
2	37961	2	0.374	0.705	0.748	0.003
3	19387	3	0.191	0.896	0.573	0.160
4	7714	4	0.076	0.972	0.304	0.279
5	2842	5	0.028	1.000	0.140	0.238
101501			1		2.08	1.107

- Expected Value interpretation: The average number of TVs in a large number of randomly-selected households will approach the expected value 2.08 (long run average).
- Expected variance interpretation: On an average it is expected that the squared deviation from mean will be 1.107.

Some useful standard distributions

Some of the useful discrete distributions are:

- Bernoulli, Binomial, Poisson

Some of the useful continuous distributions are:

- Uniform, Exponential, Normal, Weibull

The Normal Distribution: Gaussian Distribution

Many years ago [in 1893] I called the Laplace-Gaussian curve the normal curve, which name, while it avoids the international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another abnormal.

- Karl E. Pearson, 1920

The Evolution of Normal Distribution

- Astronomy was the first science troubled by error in measurement
 - January 1, 1801; a celestial event was witnessed by Italian astronomer Giuseppe Piazzi.
 - Thought to be a new planet which disappeared behind the sun in just six weeks time.
 - Not enough observation taken to determine its orbit
- Where will Ceres, the new planet, most likely appear after an year?
 - Gauss proposed that an area of the sky be searched that was quite different from those suggested by the other astronomers. He turned out to be right.
 - Calculated the area based on the probability density of the error curve

The Evolution of Normal Distribution

Gauss devised the error curve based on the following assumption

- Small errors are more likely than large errors.
- For any real number ϵ the likelihood of errors of magnitudes ϵ and $-\epsilon$ are equal.
- In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

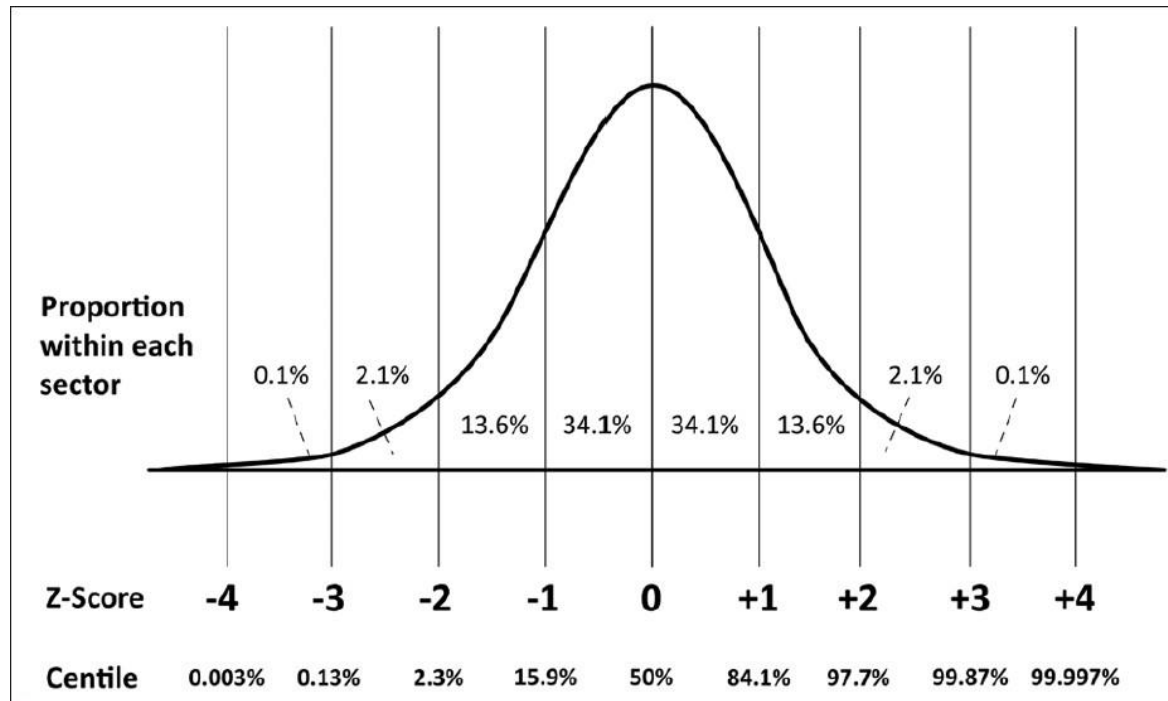
[Link: How the shape of the distribution was identified?](#)

The Evolution of Normal Distribution

With complex derivation, the function which defines the normal distribution looks like

$$f(x) = \frac{1}{\sigma * \sqrt{2\pi}} * e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

This is known as probability density function.



What is Normally distributed?

Any outcome (random variable) which is influenced by a combination of numerous external factors but none of the factors having a dominating effect on the outcome tends to follow a normal distribution.

- Height and Weight of the person
- Marks obtained by students in a class
- Error in measurement
- Price change follows a near normal distribution ($\sim t$ – distribution)
 - Risky proposition: Price change in stock market may not be perfectly normal.
 - The world is tightly interconnected and in the market there are few entities that can exert much influence on the price, under particular circumstances.

Any naturally occurring phenomena may tend to normally distributed.

Properties of Normal distribution

If several *independent* random variables are normally distributed, then their sum will also be normally distributed.

- The mean of the sum will be the sum of all the individual means

$$E(S) = E(X_1) + E(X_2) + \dots; X_1, X_2 \dots \text{are independent random variable}$$

- by virtue of the independence, the variance of the sum will be the sum of all the individual variances

$$V(S) = V(X_1) + V(X_2) + \dots; X_1, X_2 \dots \text{are independent random variable}$$

Standard Normal Variable

- Any normal random variable $X \sim N(\mu, \sigma^2)$ can be transformed into a standard normal variable.
- A standard normal variable is represented by Z where $Z \sim N(0, 1^2)$

$$Z = (X - \mu) / \sigma$$

Why Transformation?

For easier computation. Say, random variable $X \sim N(50, 10^2)$. Suppose we want the probability that X is greater than 60. Find $P(X > 60)$?

Transform to get the Z value and then lookup the Z table to find the probability.

Application–Example

Suppose that the time it takes the electronic device in the car to respond to the signal from the toll plaza is normally distributed with mean 160 microseconds and standard deviation 30 microseconds. What is the probability that the device in the car will respond to a given signal within 100 to 180 microseconds?

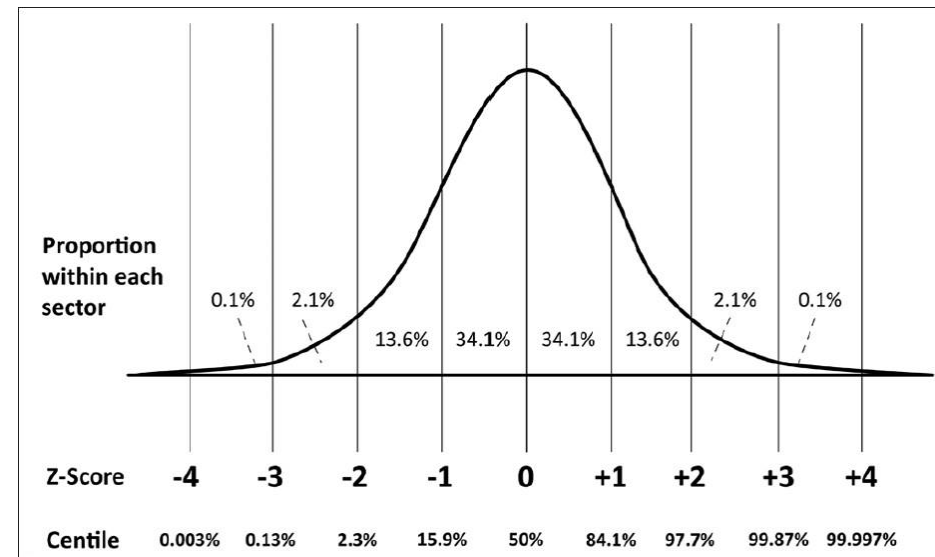
$$P(100 < X < 180) = P\left(\frac{100-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{180-\mu}{\sigma}\right)$$

$$=P\left(\frac{100-160}{30} < Z < \frac{180-160}{30}\right) = P(-2 < Z < 0.666)$$

$$= \{ \text{NORM.S.DIST}(0.666, \text{TRUE}) \} - \{ \text{NORM.S.DIST}(-2, \text{TRUE}) \}$$

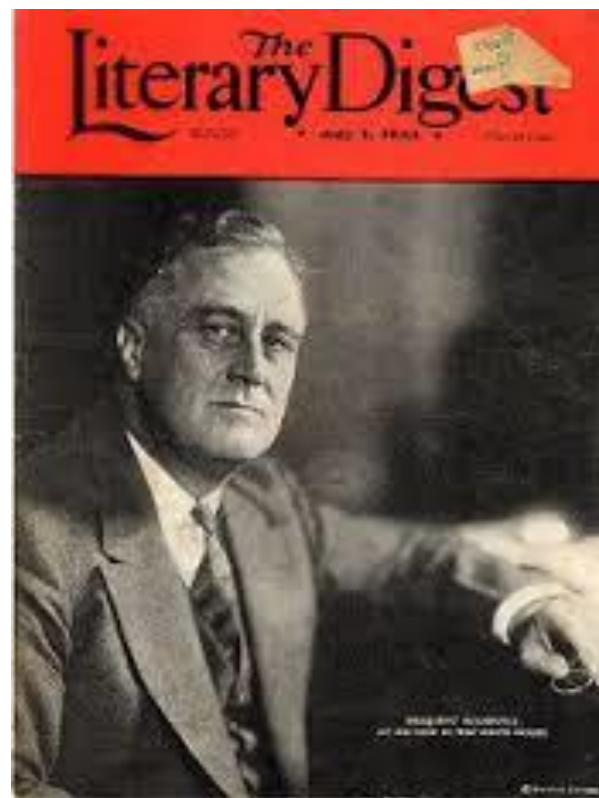
$$= 0.7474 - 0.0227$$

$$= 0.7247$$



Sampling


Sampling –What Went Wrong!!!



- 1936 Presidential election: [Landon vs. Roosevelt](#)
- Survey mail: 10 million
- Response received: 2.4 million
- Prediction:
 - Roosevelt: 43% vote
- Actual:
 - Roosevelt: 62% vote

The Infamous *Literary Digest* Poll, and the Election of 1936

■ At the same time, a young man named George Gallup sampled only 50 000 people and predicted that Roosevelt would win. Gallup's prediction was ridiculed as naive. After all, the *Digest* had predicted the winner in every election since 1916, and had based its predictions on the largest response to any poll in history. But Roosevelt won with 62% of the vote. The size of the *Digest*'s error is staggering. How could they have been so far off?

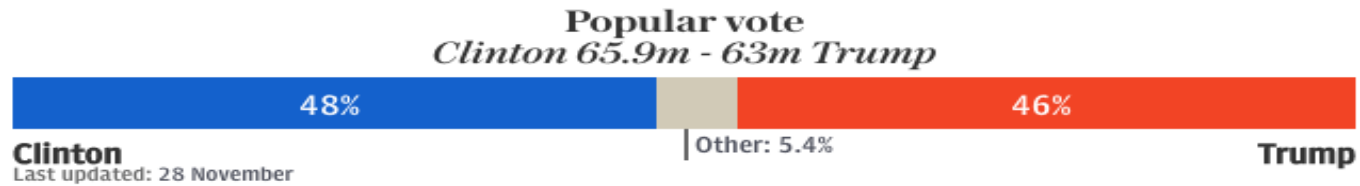


George Gallup (1901–1984)

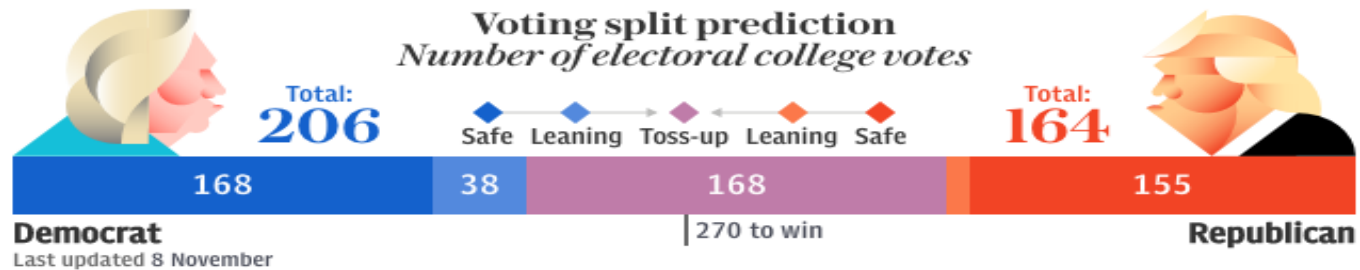
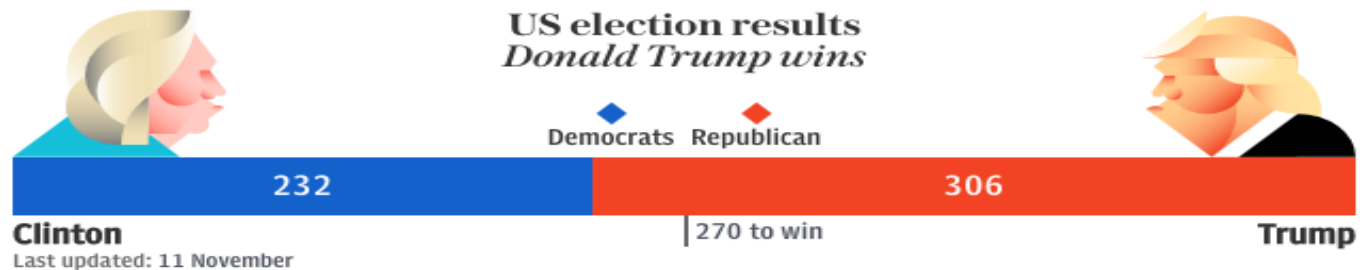
Class, Data Lab, and Statistics

11

Sampling –What Went Wrong!!!

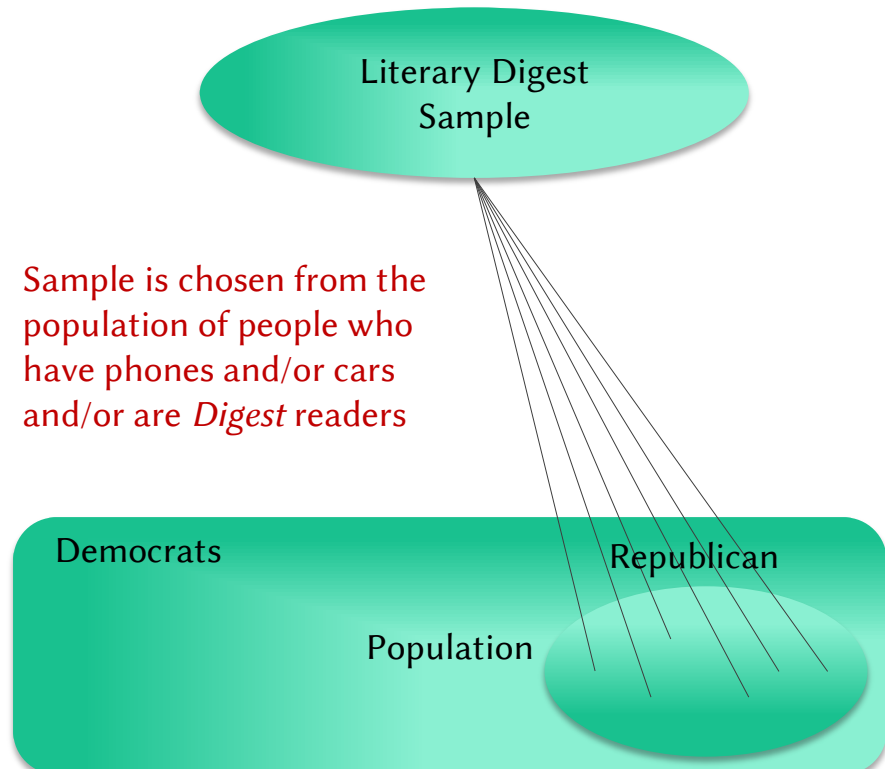
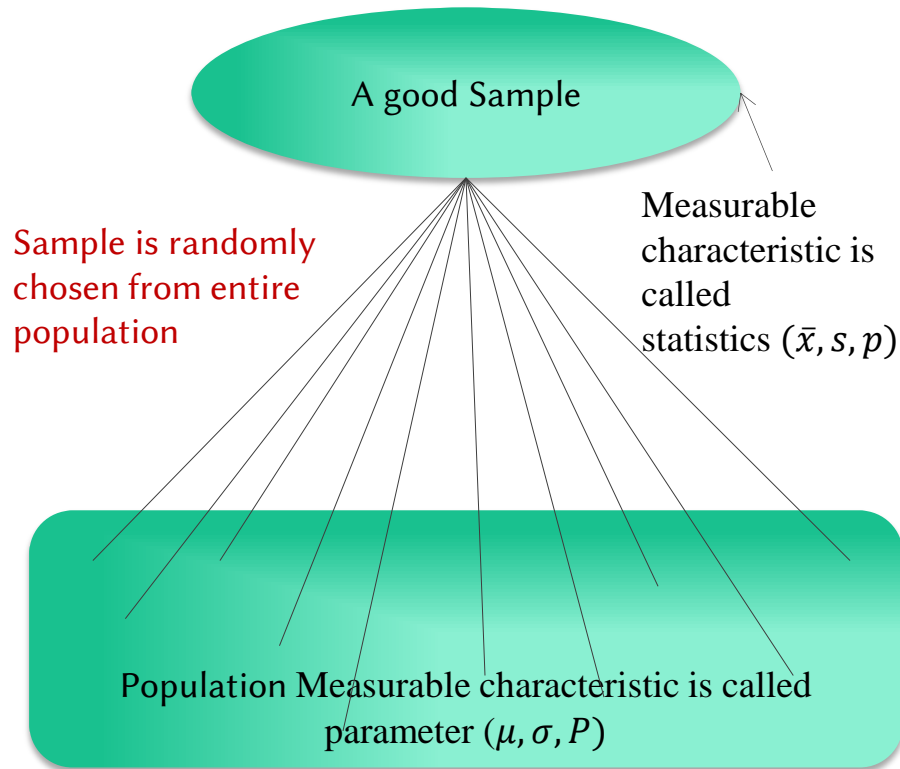


Where did the error come?



- 10 out of 93 polls predicted Trump victory. Six states which swung the fortune for Trump.
- LA Times/ USC economist Kaptean: Online system response was better than phone call survey in swing states (Non college educated white)

A good sampling procedure



The sample should be a good estimator of the population parameter. Thus the sample must be random in order to use statistics to learn things about the population.

Sampling Steps

- Identification of Target population
 - Study to understand the attrition behavior of IT professional.
- Deciding the sampling frame
 - Whom to go for data.
 - **Selection bias**
- Determine the sample size
 - How many instances to pick
 - **Response bias**
- Sampling method
 - Random sampling
 - Stratified sampling
 - Cluster sampling

Central Limit Theorem

-Laplace, 1810

Central Limit Theorem

Generally, the sample mean (\bar{X}) derived in repeated sampling from a **normally distributed population** with mean μ and standard deviation σ will follow a normal distribution with mean $\bar{X} = \mu$ and standard deviation $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ for any sample size n .

Central limit theorem:

The sample mean (\bar{X}) derived in repeated sampling from a **population** with mean μ and standard deviation σ will follow a normal distribution with mean $\bar{X} = \mu$ and standard deviation $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ for **large sample size** $n > 30$.

Confidence Interval concept

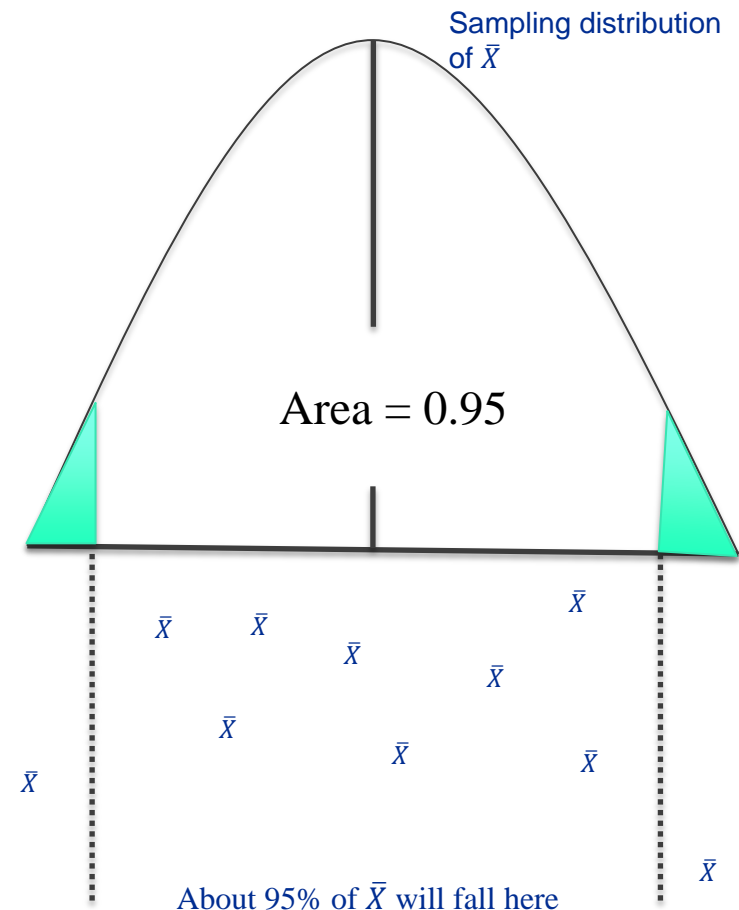
- If sample is drawn from a [normal population](#) or a [large sample](#) is used, then by the rules of normal distribution, before the sampling, there is .95 probability that sample mean \bar{X} will fall within the interval

$$\mu \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

- After sampling, about 95% of the values of \bar{X} obtained in large number of repeated sampling will fall in the interval defined by equation above.
- \bar{X} falls within the interval defined above if and only if μ happens to be within

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

- 95% confident that population mean lies within the above range. CLT in action.



Degree of Freedom

Objective is to estimate the population parameters (mean, std deviation etc.) from the sample. The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

Population mean (height in ft)= 6

- Sampling of one person: height is 8 ft
 - Variance = $(8-6)^2 = 4$

This estimate is based on one piece of information, so $df = 1$

- Sampling of one person: height is 5 ft
 - Variance = $(5-6)^2 = 1$

This estimate is based on one piece of information, so $df = 1$

- Population variance = 2.5 with $df = 2$

Population mean not known

- Sampling lead to 8ft and 5ft as two data points
 - Mean = 6.5
 - Variance estimate 1 = $(8-6.5)^2 = 2.25$
 - Variance estimate 2 = $(5-6.5)^2 = 2.25$
- The two estimates are not independent so $df \neq 2$ but $df = 1$



The degrees of freedom for an estimate is equal to the [number of values in the sample](#) minus the [number of parameters estimated in route](#) to the estimate in question.

Hypothesis Testing

Blackout Babies

On November 9, 1965 a massive blackout darkened homes and businesses from Michigan to New York to Canada, affecting 50 million people.

On Wednesday August 10 , 1966 New York Times published an article, which claimed that the “Births Increased 9 months after the Blackout”.



If you drink Complan, you grow taller

Horlicks MAKES KIDS
TALLER, STRONGER*, SHARPER**



TALLER STRONGER* SHARPER**

© 2015 Horlicks Pvt. Ltd. All rights reserved. Horlicks is a registered trademark of Nestlé India Ltd. *Based on a study conducted by Nestlé India Ltd. **Based on a study conducted by Nestlé India Ltd.

Horlicks is a registered trademark of Nestlé India Ltd.



Dettol kills 99.9% of the germs

NEW



clini-care¹⁰
with active naturol shield



Lifebuoy clini-care 10 with its revolutionary **Activ Naturol Shield** technology offers **10 times better germ protection** and **10 times more skin care** than leading hygiene soap.*

The Mother of all Doubts!



Is the "AXE Effect" Really Real ?

Interesting Hypothesis

- Good looking couples are more likely to have girl child(ren)!
- Married people are more happier than singles!!!
- Vegetarians miss fewer flights.
- Black cars have more chance of involving in an accident than white cars in moon light.
- Women use camera phone more than men.
- Left handed men earn more money!
- Smokers are better sales people.
- Those who whistle at workplace are more efficient.

What is Hypothesis Testing

A **hypothesis test** is a statistical **test** that is used to check whether there is enough evidence in a sample of data to infer that a certain statement/claim/condition is true for the entire population.

Hypothesis is a claim. The process of verifying the claim is Hypothesis Testing.

Basis of Hypothesis testing–CLT

- Sample is coming from a population which is I.I.D (Independent and Identically distributed) and is an unbiased estimator of the population

And

- Regardless of the population, Sampling distribution of mean will follow a normal distribution with mean μ and standard deviation σ/\sqrt{n}

Government of West Bengal approaches Die Another Day (DAD) Hospital and requests to provide cardiac related treatments for 150,000 for all the patients supported by a government scheme.

Should DAD Hospital accept the offer?

What do we know?

The average cost of treatment based on the sample estimate = 198723 (approx. 198000)

Sample size = 248

Assume standard deviation of the population is known and equals 10000

Null Hypothesis

- Hypothesis testing has two hypotheses, **null** and **alternative**.
- **Null hypothesis** is a statement about a population parameter, such as population mean or proportion.
- In most cases, we believe that the null hypothesis is true. Decisions are made about the null hypothesis.

In context of DAD Hospital

$$H_0: \mu \leq 150000$$

Alternate Hypothesis

- **Alternative hypothesis** is complement of null hypothesis.
- Alternative hypothesis is also known as research hypothesis.

In context of DAD Hospital

$$H_a: \mu > 150000$$

Objective of Testing

1. There are two hypotheses, the null and the alternative hypotheses.
2. The procedure begins with the assumption that the null hypothesis is true.
3. The goal is to determine whether there is enough evidence to infer that the alternative hypothesis is true, or the null is not likely to be true.
 - Evidence is collected from the sample data by calculating something as test statistics
 - Probability value is calculated (p-value) using the test statistics
 - Significance level is defined (α) which forms the basis for decision
4. There are two possible decisions:
 - Conclude that there is enough evidence to support the alternative hypothesis. Reject the null.
 - Conclude that there is *not* enough evidence to support the alternative hypothesis. Fail to reject the null.

Test Statistics

The **test statistic** is a mathematical expression that allows researchers to determine the likelihood of obtaining sample outcomes if the null hypothesis were true.

- The value of the test statistic is used to make a decision regarding the null hypothesis. In case of DAD hospital:

$$\text{Test statistics} = (198000 - 150000) / 10000 = 4.8$$

In excel:

- $\text{Test statistics} = 1 - \text{NORM.DIST}(198000, 150000, 10000, \text{TRUE}) = 7.9 \times 10^{-7}$
- $\text{Test statistics} = 1 - \text{NORMDIST}(198000, 150000, 10000, \text{TRUE}) = 7.9 \times 10^{-7}$
- $\text{Test statistics} = 1 - \text{NORM.S.DIST}(4.8, \text{TRUE}) = 7.9 \times 10^{-7}$
- $\text{Test statistics} = 1 - \text{NORMSDIST}(4.8) = 7.9 \times 10^{-7}$

P-value

A **p value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. The p value for obtaining a sample outcome is compared to the level of significance.

In case of DAD Hospital, the p -value will mean:

If the mean is 150000 and SD is 10000, what is the probability of observing 198000 as the cost of treatment in the sample?

The answer is 7.9×10^{-7}

Significance (Rejection Criteria)

- Significance level (α) is a decision criteria used for rejection of null hypothesis.

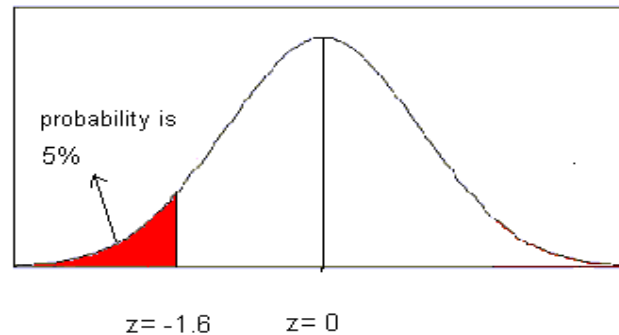
$\alpha = \text{probability (rejecting the null hypothesis | the null hypothesis is true)}$

- In most cases, we set the null hypothesis to be 5%
 - Means we are taking a risk of 5%. 5 out of 100 cases, we may go wrong in rejecting the null hypothesis when it is actually true.

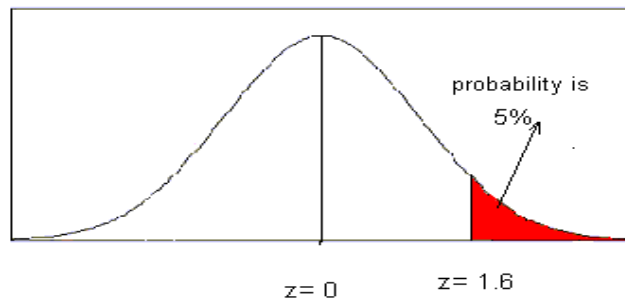
When p-value is less than significance (α), we reject the null hypothesis

Critical Value

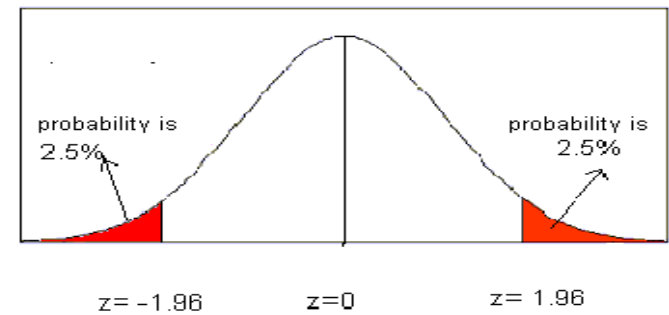
A **critical value** is a cutoff value that defines the boundaries beyond which less than 5% of sample means can be obtained if the null hypothesis is true. Sample means obtained beyond a critical value will result in a decision to reject the null hypothesis.



Lower one tailed $H_1: \mu_o < \mu_h$



Upper one tailed $H_1: \mu_o > \mu_h$



Two tailed test $H_1: \mu_o \neq \mu_h$

Type I and Type II Error

- A Type I error occurs when **we reject** a true null hypothesis. As an analogy, a Type I error occurs when the jury convicts an innocent person.
- **$P(\text{Type I error}) = \alpha$** [usually 0.05 or 0.01]
- A Type II error occurs when we don't reject a false null hypothesis [**accept the null hypothesis**]. That occurs when a guilty defendant is acquitted.
- **$P(\text{Type II error}) = \beta$**

Type I and Type II Error

Truth in the Population	Decision	
	Retain the Null	Reject the Null
	Correct (1 - α)	Type I Error α
Null is True		
Null is False	Type II Error β	Correct (1- β) Power of Test

Terms in Hypothesis Test

Term	Explanation
α	Conditional probability of incorrectly rejecting H_0 , when it is actually true (Type I Error)
β	Conditional probability of failing to reject H_0 , when it is false (Type II Error)
$1 - \beta$ (Power)	Conditional probability of correctly rejecting H_0 , when H_1 is true
p-value	Evidence against the null hypothesis in favor of alternative hypothesis. Smaller p-value indicates stronger evidence against null hypothesis

Z - Test

One independent sample Z-test

The **one-independent sample z test** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance.

When z –test is used:

Population is normal distribution & population standard deviation is known

Test Statistic – Z test

- The **z statistic** is an inferential statistic used to determine the number of standard deviations in a standard normal distribution that a sample mean deviates from the population mean stated in the null hypothesis.

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

$$Z = \frac{\text{Estimated value of mean} - \text{Hypothesis value of mean}}{\text{Standard deviation of sampling distribution}}$$

Time Spent on Texting...

- Magazine women's health published an article in which they claimed that female students spend 94.6 minutes everyday “**texting**” with a standard deviation of 26.8 minutes.

<http://www.womenshealthmag.com/life/hours-you-spend-on-your-phone>

Checking the hypothesis

- Data was collected from 57 female students from City college of Business and the average amount spent by them on texting was 89.2 minutes.
- Check the hypothesis that the time spent by female students is less than 94.8 minutes.

Null and Alternative Hypothesis

- H_0 : Time spend by female students on texting is less than or equal to 94.8 minutes.
- H_A : Time spend by female students on texting is greater than 94.8 minutes

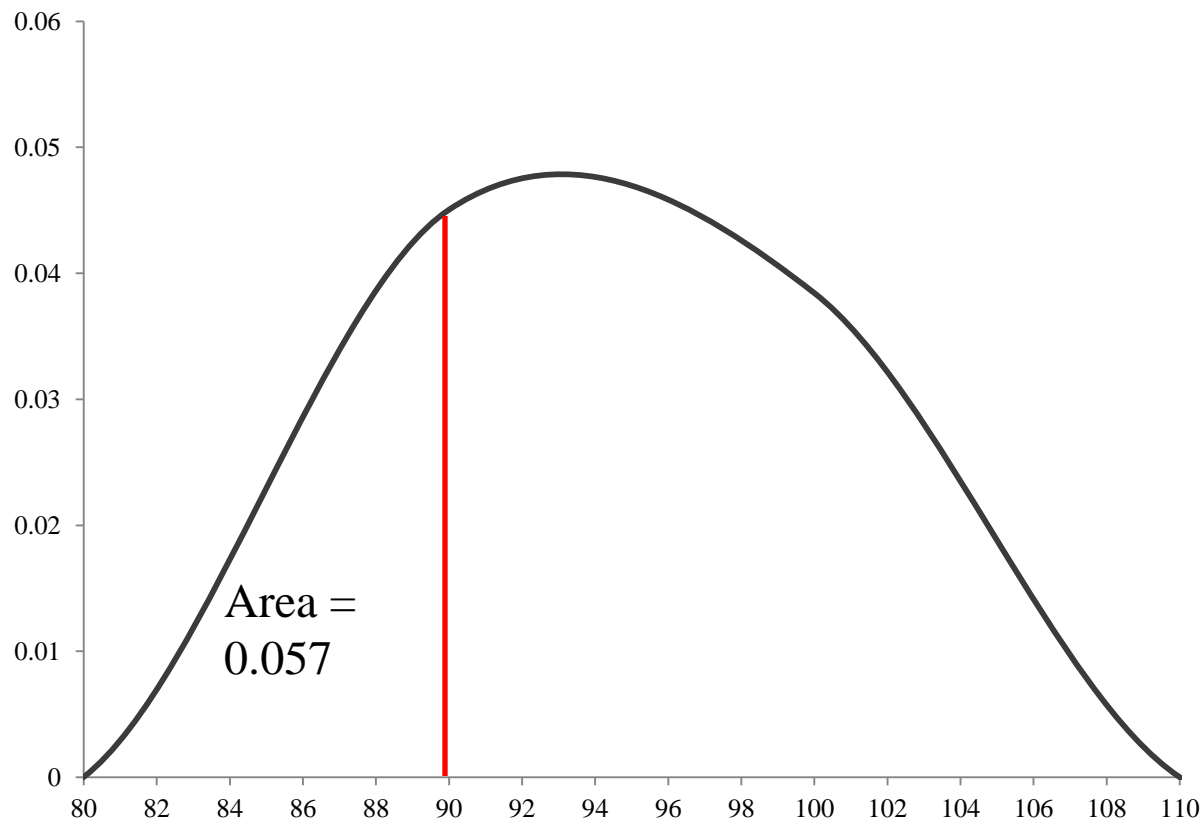
$$H_0: \mu \leq 94.8$$

$$H_A: \mu > 94.8$$

Z - Statistic

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{89.2 - 94.8}{26.8 / \sqrt{57}} = -1.57$$

$$p\text{-value} = 0.057$$



What it means...

If mean value time spent on texting is 94.8 minutes with a standard deviation of 26.8 minutes (as being claimed by the magazine), then probability of observing 89.2 minutes in the sample is 5.7%.

Since p-value (0.057) is greater than 0.05 (significance value), we retain the null hypothesis (or fail to reject the null hypothesis).

Implies, women spend less than 94.8 minutes texting.

Two-Sample Z test

- Two sample z-test is used for comparing two populations.
- For example, we would like to check whether the time spent on using mobile phone is different for men and women.

Two Sample Z test with Equal variance

- Null Hypothesis (H_0): $\mu_1 = \mu_2$
- Alternative hypothesis (H_1): $\mu_1 \neq \mu_2$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Exercise—Mobile Usage between Men and Women

- A sample of 78 men revealed that they spent 4.6 hours on average in a day on using their mobile phones.
- A sample 56 women revealed that they spent 6.2 hours on average in a day on using their mobile phones.
- Assume that the population standard deviation is 1.1 hours.
- Check the hypothesis that there is no difference between men and women on usage of mobile phones.

t-Test

T-Distribution

- T-distribution is a continuous probability distribution that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown

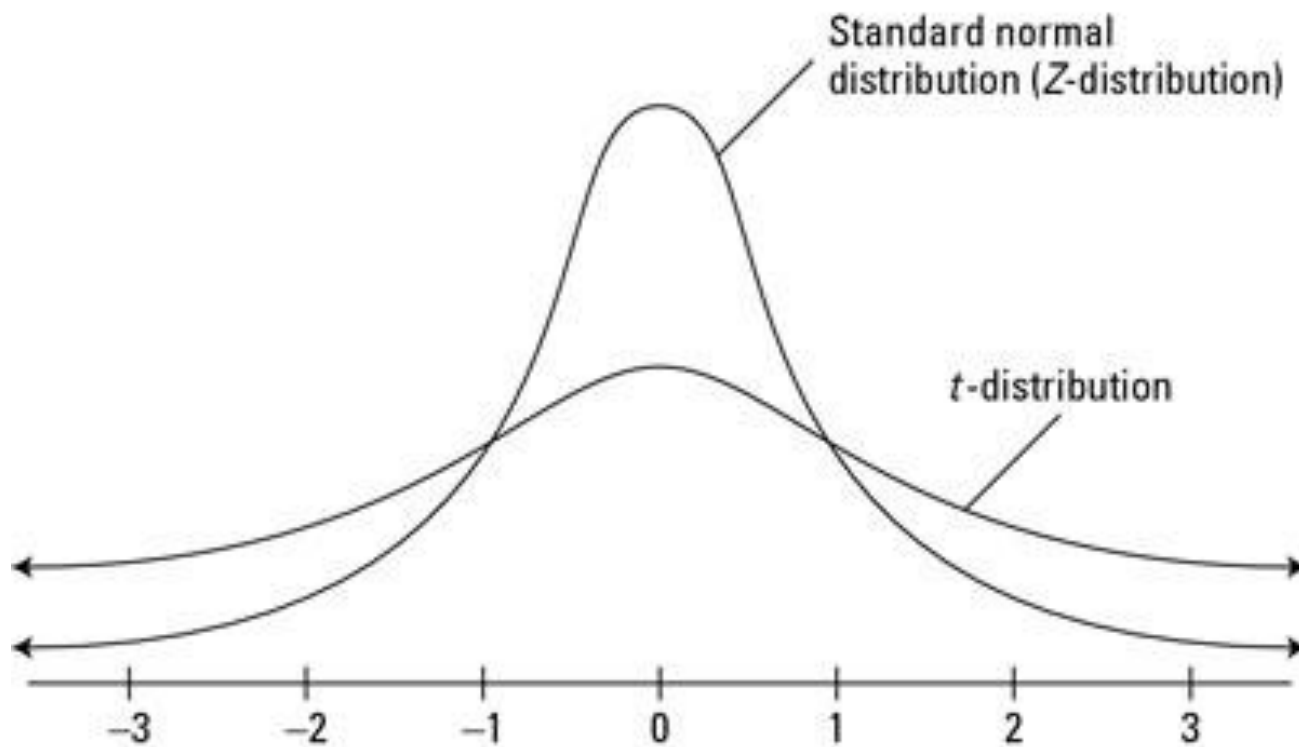
t-distribution

- t-distribution is a family of continuous probability distribution that arises when estimating the mean of sample from a population of normal distribution in which sample size is small and population standard deviation is unknown.

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

Z test Vs T-test

- Z test is used when the population is normal (or large sample) and σ is known.
- T-test is used when population is normal and σ is estimated from sample.



Treatment cost of patients – t distribution

$$H_0 : \mu \leq 150000$$

$$H_1 : \mu > 150000$$

150000 is the cost offered by the government

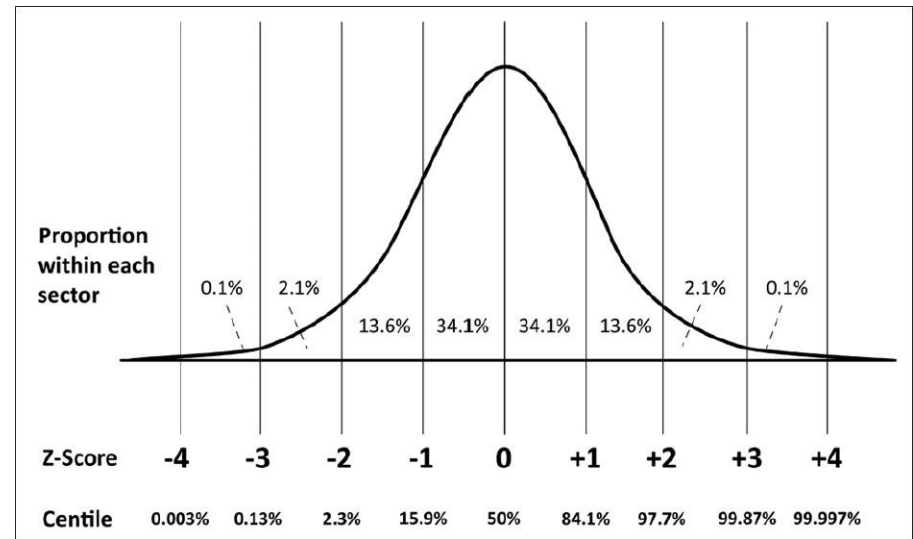
$$\bar{X} = 198723$$

$$\mu = 150000$$

$$S = 122587$$

$$n = 248$$

$$Z = \frac{(198723 - 150000)}{122587 / \sqrt{248}} = 6.25$$



More Hypothesis Testing–Binary Decision

- Remember: The value being tested is the value in the null hypothesis.
- Remember: The value being tested is a parameter, a population value
- The decision maker either rejects the null hypothesis or fails to reject the null hypothesis.

Two-tailed and One tailed tests

- If the alternative hypothesis uses an equal sign, this indicates a two tailed test(non directional).
 - In this case, the region of rejection is located in both tails.
- If the alternative hypothesis uses a greater or less than sign ($<>$), this is a directional test.
 - In this case, the region of rejection is located is one tail of the sampling distribution

Central Limit Theorem for Proportions

- Let p be the probability of success and q is the probability of failure.

Sampling distribution of proportion will follow a normal distribution with mean p and standard deviation $\sqrt{(p * q)/n}$

Z-test for Proportion

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

Example on test of proportion

The CEO of the hospital believes that the probability that the treatment cost exceeds 150000 in at least 25% of the cases.

Validate the claim

Solution:

$$H_0: P \leq 0.25$$

$$H_1: P > 0.25$$

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{0.52 - 0.25}{\sqrt{0.25 \times 0.75 / 248}} = 9.8$$

Exercise—z test for Proportions

- An E-commerce company believes that 10% of all their customers return the products (jewelry) after using them.
- In a sample of 220 customers, 45 customers are estimated to have returned the product after using it.

Check the hypothesis that the proportion of customers who return the product after using it is other than 10%.

ANNOVA

ANNOVA

ANNOVA (Analysis of variance) is a statistical test to determine whether means of two or more groups differ (used when X is categorical and Y continuous)

One way ANNOVA:

Hospital want to determine whether the cost of treatment differs based on Gender Type

Two way ANNOVA:

Hospital want to determine whether the cost of treatment differs based on Gender and Past medical history

ANNOVA–Hypothesis

Null hypothesis H_0 : All means are equal

Alternative hypothesis H_1 : At least one mean is different

$P\text{-value} \leq \alpha$: The differences between some of the means are statistically significant. Reject the Null Hypothesis

$P\text{-value} > \alpha$: The differences between the means are not statistically significant. Retain the Null Hypothesis

Co-variance and Correlation

Covariance and Correlation Coefficient

- Covariance** is a statistical measure of the degree to which the two variables move together. The sample covariance is calculated as :

$$\text{COV}_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

- Correlation** coefficient is a measure of the strength of the linear relationship between two variables. The correlation coefficient is given by:

$$r_{xy} = \frac{\text{COV}_{xy}}{\sigma_x \sigma_y}$$

- Population correlation is denoted by ρ (rho). Sample correlation is denoted by r . Features of ρ and r
 - Unit free and ranges between -1 and 1
 - The closer to -1, the stronger the negative linear relationship
 - The closer to 1, the stronger the positive linear relationship
 - The closer to 0, the weaker the linear relationship

Y (Exp)	X (Inc)	Y' = Y-Y(Avg)	X' = X-X(Avg)	X'*Y''
700	800	-410	-900	369000
650	1000	-460	-700	322000
900	1200	-210	-500	105000
950	1400	-160	-300	48000
1100	1600	-10	-100	1000
1150	1800	40	100	4000
1200	2000	90	300	27000
1400	2200	290	500	145000
1550	2400	440	700	308000
1500	2600	390	900	351000
1110	1700			2E+06

Covariance	186666.6667
Correlation	0.980847369

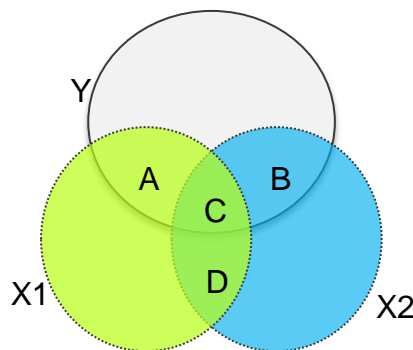
Partial and Semi-Partial Correlation

- Partial correlation** coefficient measures the relationship between two variables (say Y and X1) when the influence of all other variables (say X2, X3, ..., Xn) connected with these two variables (Y and X1) are removed.

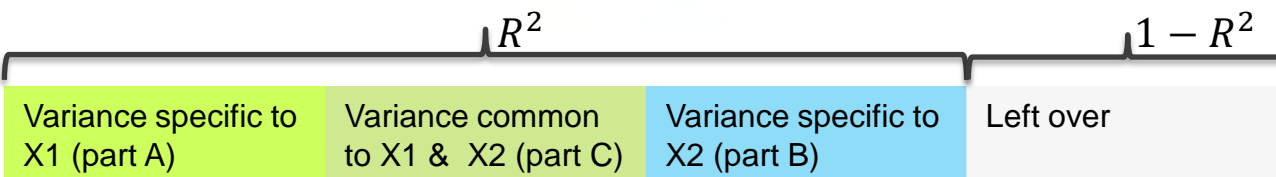
$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Correlation between y1 and x2, when the influence of x3 is removed from both y1 and x2.

- Part correlation** (or semi partial) coefficient measures the relationship between two variables (say Y and X1) when the influence of all other variables (say X2, X3, ..., Xn) connected with these two variables (Y and X1) are removed from one of the variables (X1).

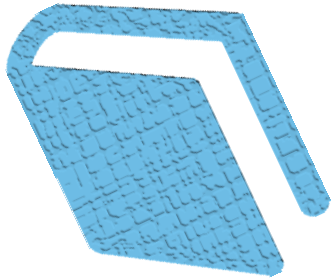


$$sr_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{23}^2}}$$



Summary

Summary of the topics covered in this lesson:



- We always deal with sample dataset and the objective is to give a meaningful estimate of the population parameters through sample statistics.
- Central limit theorem is the building block for interpreting the outcome of many advanced statistical techniques.

QUIZ TIME

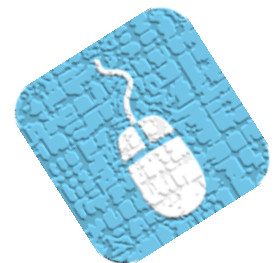


Quiz Question 1

Quiz 1

The population mean is known and 11 people from the population are selected at random to estimate the standard deviation. DF of standard deviation will be:

- a. 11
- b. 9
- c. 10
- d. None of the above



Quiz Question 1

Quiz 1

The population mean is known and 11 people from the population are selected at random to estimate the standard deviation. DF of standard deviation will be:

- a. 11
- b. 9
- c. 10
- d. None of the above

Correct answer is:

a

Since population mean is known, there is no intermediate estimate to arrive at the estimate of standard deviation. Hence degree of freedom will be 11.

End of Lesson01B–Basic of Statistics

