K2 Analytics
Building Skills, Building Individuals

Statistical Learning

- Rajesh Jakhotia

Earning is in Learning
- Rajesh Jakhotia

About K2 Analytics

At K2 Analytics, we believe that skill development is very important for the growth of an individual, which in turn leads to the growth of Society & Industry and ultimately the Nation as a whole. For this it is important that access to knowledge and skill development trainings should be made available easily and economically to every individual.

Our Vision: *"To be the preferred partner for training and skill development"*

Our Mission: *"To provide training and skill development training to individuals, make them skilled & industry ready and create a pool of skilled resources readily available for the industry"*

*We have chosen Business Intelligence and Analytics as our focus area. With this endeavour we make this presentation on "**Statistical Learning**" accessible to all those who wish to learn Analytics. We hope it is of help to you. For any feedback / suggestion or if you are looking for job in analytics then feel free to write back to us at ar.jakhotia@k2analytics.co.in*

Learning Objectives

1. Why Statistics?
2. Measures of Central Tendency
3. Measures of Dispersion
4. Descriptive Statistics
5. Probability
6. Distribution – Probability, Binomial, Poisson and Normal
7. Central Limit Theorem
8. Hypothesis Testing (Z, t, F, ANOVA)

Why Statistics?

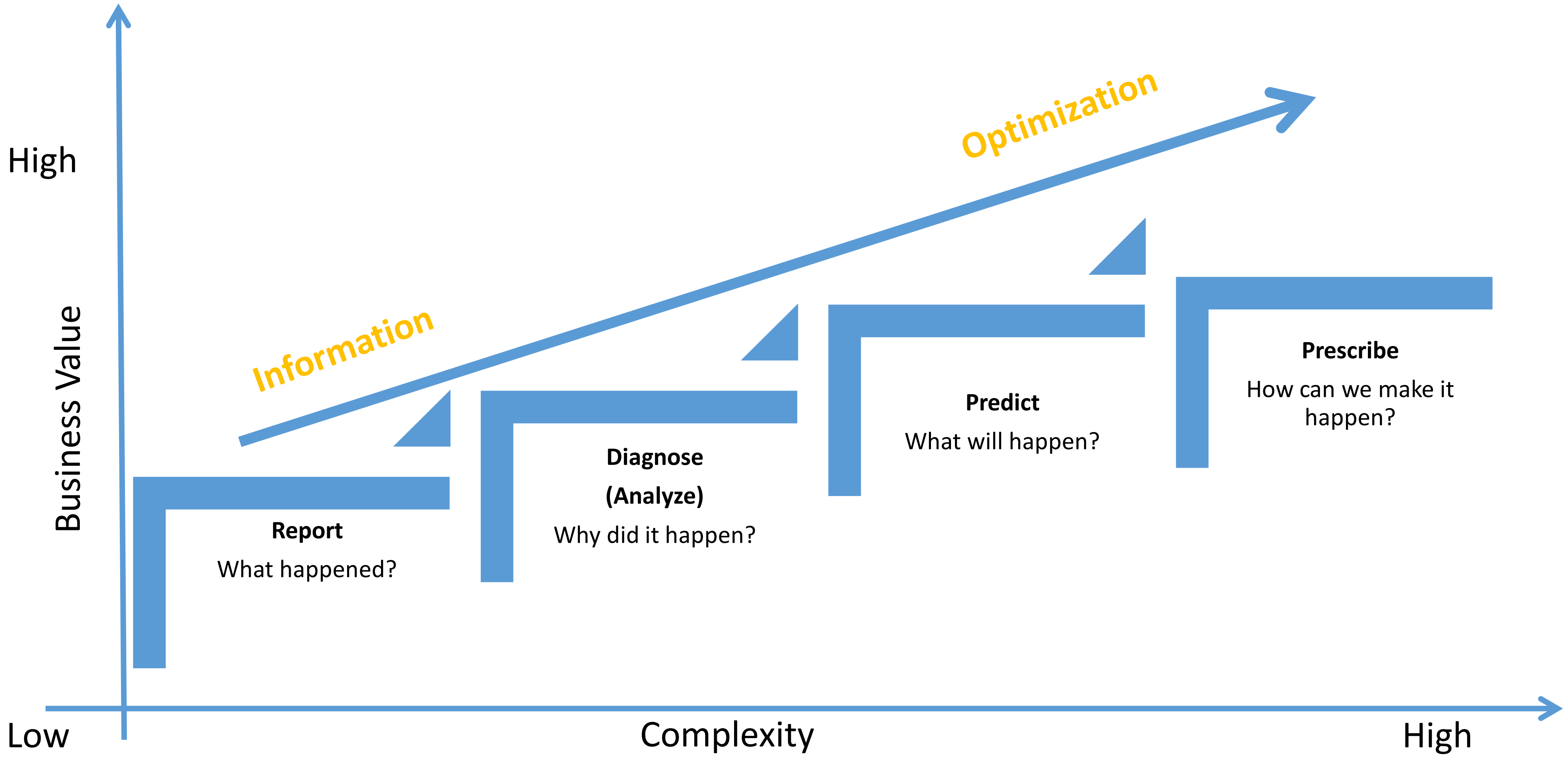
Analytics in our day to day life

Knowingly / Unknowingly we all tend to use analytics in our day-to-day life for decision making

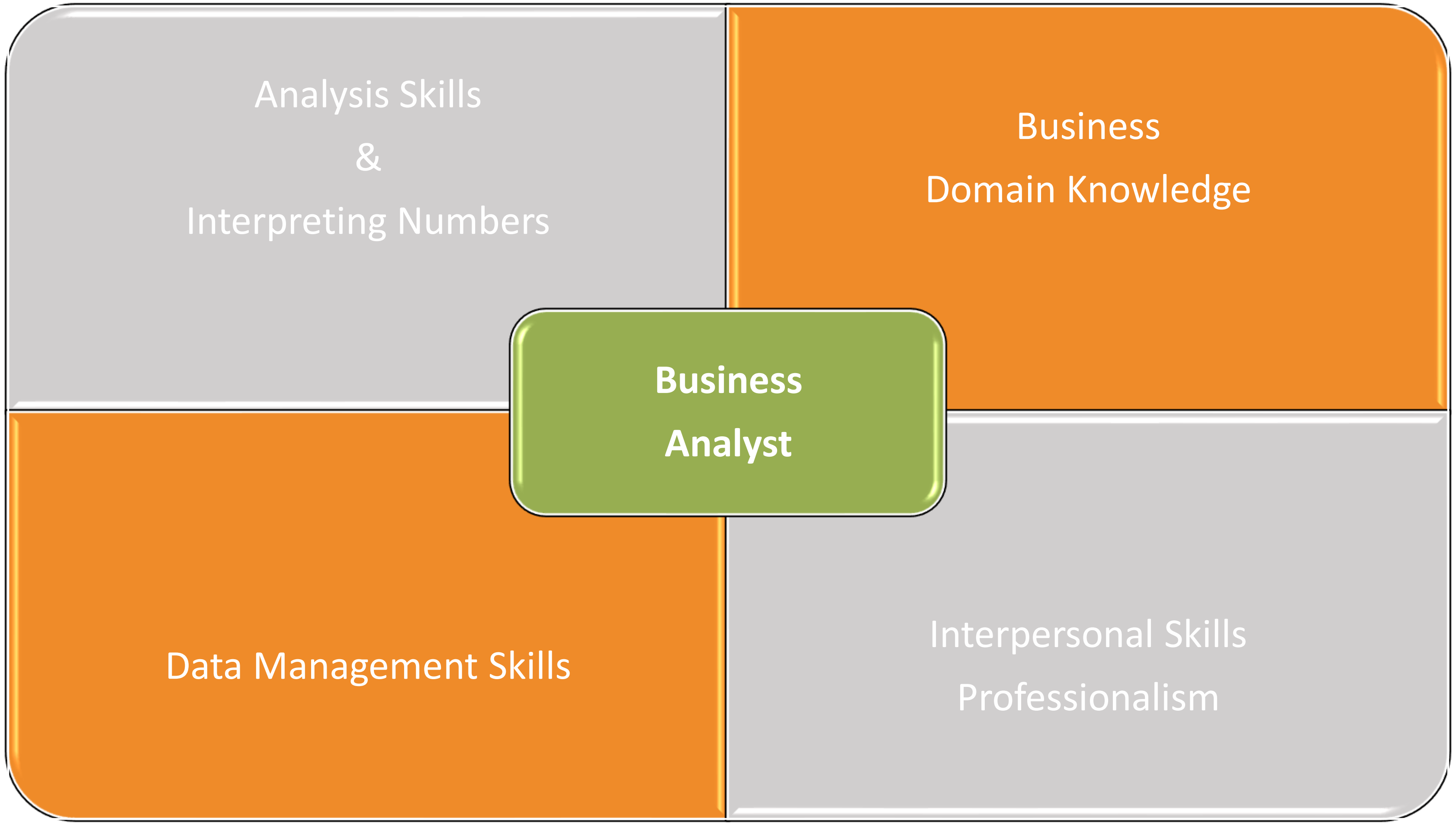
- Buying something in market
- Planning a vacation
- Analyzing a cricket match
- Preparing for exam
- Playing Chess
- Weather Forecast

Analyzing things helps you do better planning and increases the likelihood of you getting the desired results

Analytics in Business: Descriptive – Predictive – Prescriptive



Skills required to be successful in Business Analytics



Understanding Numbers is the Key to Analytics

If you don't know the business,
data can teach you.

……………

If you don't understand the numbers,
data wont help you

Statistics & Data Mining - Definitions

Statistics

- the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Data Mining

- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

http://www.cs.csi.cuny.edu/~imberman/DataMining/Statistics%20vs.pdf

What number skills should I have?

Basic Number Skills

- Types of Numeric Variables
- Measures of Central Tendency: Mean, Median, Mode
- Measures of Dispersion: Std. Deviation, Variance
- Correlation and Covariance
- Probability Concepts
- Normal Distribution
- Hypothesis Testing
- Additive Variables, Count and Ratio

Basic Statistics

Types of Variables

Measures of Central Tendency

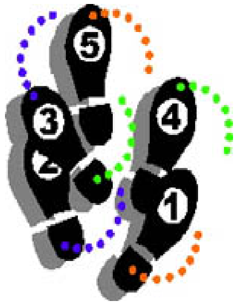
Measures of Dispersion

Ratio, Interval, Cardinal, Ordinal, & Nominative scales

A **cardinal number** tells **"how many."** Cardinal numbers are also known as "counting numbers," because they **show quantity.**

Here are some examples using cardinal numbers:

- 8 puppies
- 14 friends



Ordinal numbers tell the **order of things in a set**—first, second, third, etc. Ordinal numbers do not show quantity. They only **show rank or position.**

Here are some examples using ordinal numbers:

- 3rd fastest
- 6th in line

A **nominal number names something**—a telephone number, a player on a team. Nominal numbers do not show quantity or rank. They are used only to **identify something.**

Here are some examples using nominal numbers:

- jersey number 4
- zip code 02116

http://www.factmonster.com/ipka/A0875618.html

- **Ratio Variable** is a quantitative variable measured on a scale such that **ratios of its values are meaningful** and there in an inherently defined zero value
- **Interval Variable** is a quantitative variable where *ratios* of its values *are not meaningful* and there in *not* an inherently *defined zero value.* e.g Temperature, we cannot say 60^{o} C is 2 times hot than 30^{o} C

Measures of Central Tendency

Mean

Median

Model

Measure of Central Tendency

- Mean
  - Sum of Values divided by Number of Values
  - Also called Arithmetic Average
  - Impacted by outliers; Cannot be used for Categorical variables
- Median
  - Middle value in a distribution when the values are arranged ascending or descending
- Mode
  - Most commonly occurring value in a distribution
  - Can be applied to both Categorical and Numerical Variables

Population Mean	Sample Mean
$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

$$md = x_{\frac{(n-1)}{2}} \quad \text{for n is odd}$$

$$md = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \quad \text{for n is even}$$

Mean, Median, Mode e.g.

- 15 students are there in a Tiny Tots class.
- The age in months of the students is given below:

24	37	38	38	36	39	40	37	38	41	40	36	37	37	39

$$\boldsymbol{Mean\ Age} = \frac{(24+37+38+38+36+39+40+37+38+41+40+36+37+37+39\,)}{15}$$

$$= \mathbf{37.13}$$

Median Age : Sort the values in ascending order. The Middle Value is Median : **38**

24	36	36	37	37	37	37	**38**	38	38	39	39	40	40	41

Mode: Highest Repeating Age **- 37**

2nd Example - Household Expense Data

```python
import pandas as pd
import os

os.getcwd()
os.chdir("C:\chandan\PPT\STATISTICS\Convert to Python")

inc_exp = pd.read_csv("Inc_Exp_Data.csv")
inc_exp.head(10)
```

Index	Mthly_HH_Income	Mthly_HH_Expense	No_of_Fly_Members	Emi_or_Rent_Amt	Annual_HH_Income	Highest_Qualified_Member	No_of_Earning_Members
0	5000	8000	3	2000	64200	Under-Graduate	1
1	6000	7000	2	3000	79920	Illiterate	1
2	10000	4500	2	0	112800	Under-Graduate	1
3	10000	2000	1	0	97200	Illiterate	1
4	12500	12000	2	3000	147000	Graduate	1
5	14000	8000	2	0	196560	Graduate	1
6	15000	16000	3	35000	167400	Post-Graduate	1
7	18000	20000	5	8000	216000	Graduate	1
8	19000	9000	2	0	218880	Under-Graduate	1
9	20000	9000	4	0	220800	Under-Graduate	2
10	20000	18000	4	8000	278400	Under-Graduate	2

Measures of Central Tendency

- What is the Mean Expense of a Household?

```
In [4]: inc_exp.Mthly_HH_Expense.mean()
   ...:
Out[4]: 18818.0
```

- What is the Median Household Expense?

```
In [5]: inc_exp.Mthly_HH_Expense.median()
   ...:
Out[5]: 15500.0
```

- What is the Monthly Expense for most of the Households?

```
In [6]: mth_exp_tmp = pd.crosstab(index=inc_exp["Mthly_HH_Expense"], columns="count")
   ...: mth_exp_tmp.reset_index(inplace=True)
   ...: mth_exp_tmp[mth_exp_tmp['count'] == inc_exp.Mthly_HH_Expense.value_counts().max()]
   ...:
Out[6]:
col_0  Mthly_HH_Expense  count
18                25000      8
```

Measures of Dispersion

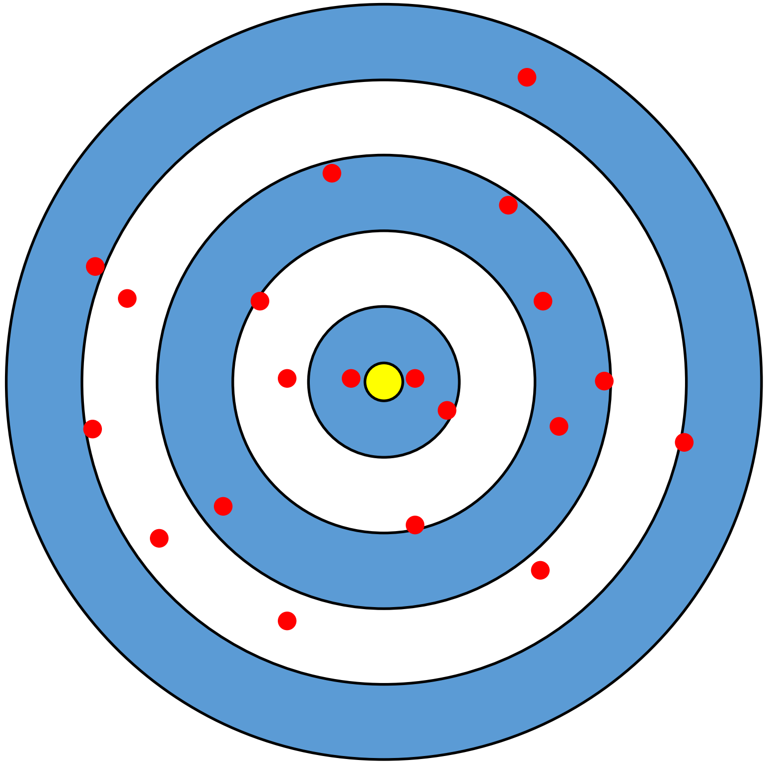
Standard Deviation

Variance & Coefficient of Variation

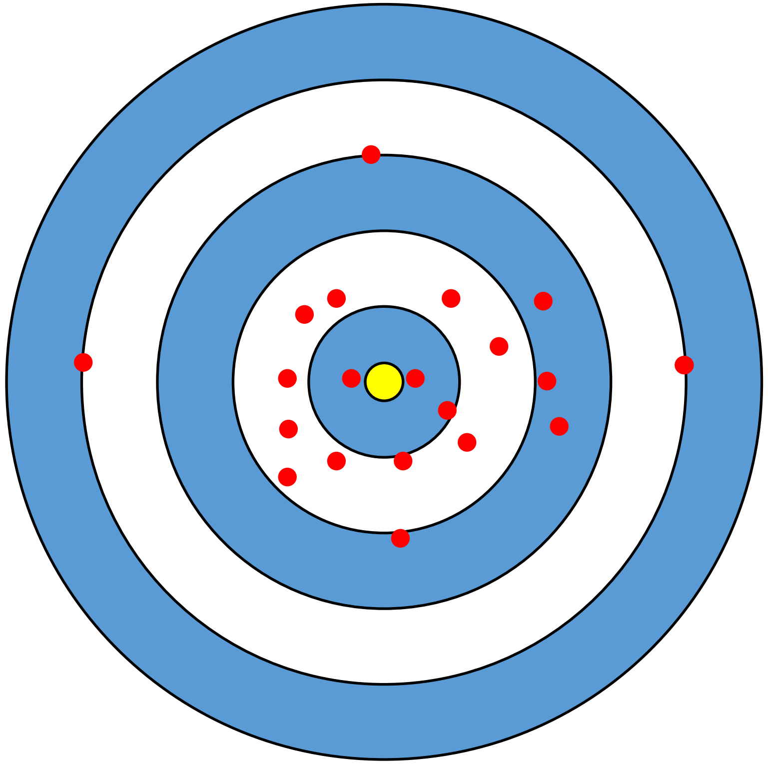
Range & Inter-Quartile Range

Which of the two shooters has high dispersion? Why?

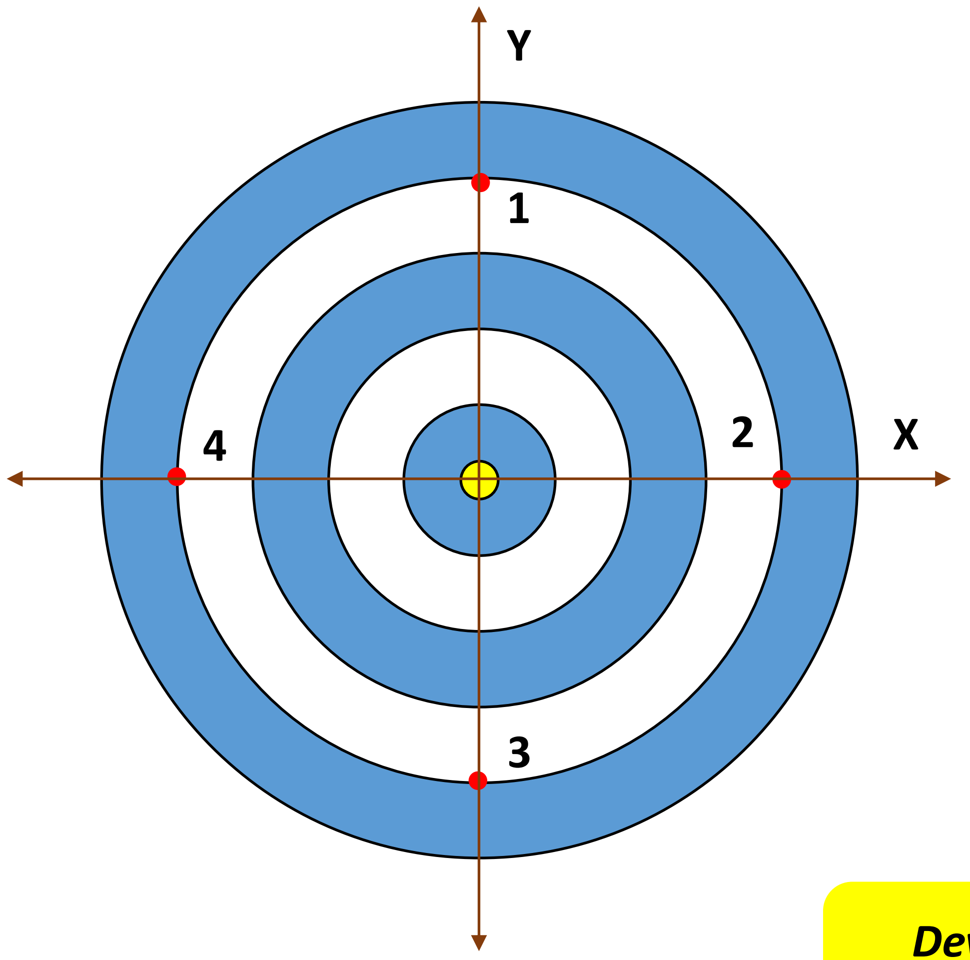
Shooter 1



Shooter 2



Why Deviation measure is required?



Shot No	X Coord	Y Coord
1	0	4
2	4	0
3	0	-4
4	-4	0
Mean	0	0

If we simply take the Mean Statistics, then we may conclude the Shooter has hit the Target

It is clear that only the Mean Statistics may sometime lead to wrong interpretation.

Hence we need the second important parameter estimate, which is called Deviation

Deviation is a measure of difference between the observed value of a variable and some other value, often that is variable's mean

Measure of Dispersion

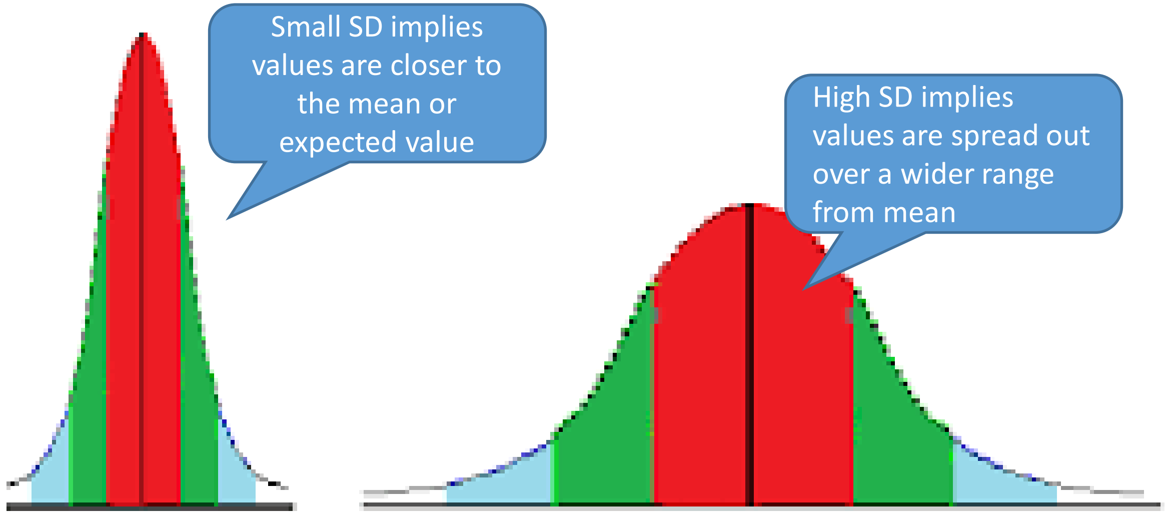
- In **statistics, dispersion** (also called **variability, scatter, or spread**) is the extent to which a distribution is **stretched or squeezed**. Common examples of measures of statistical dispersion are:

  - **Standard Deviation**
  - **Variance**
  - **Inter-Quartile Range**

https://en.wikipedia.org/wiki/Statistical_dispersion

Standard Deviation & Variance

- **Standard Deviation** is a measure used to quantify the amount of variation or dispersion of a set of data values
- Often represented as **SD**, Greek letter **σ** (sigma) or the Latin letter s



https://en.wikipedia.org/wiki/Standard_deviation

Sample SD Formula

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

Population SD Formula

$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$$

Variance & Coefficient of Variation

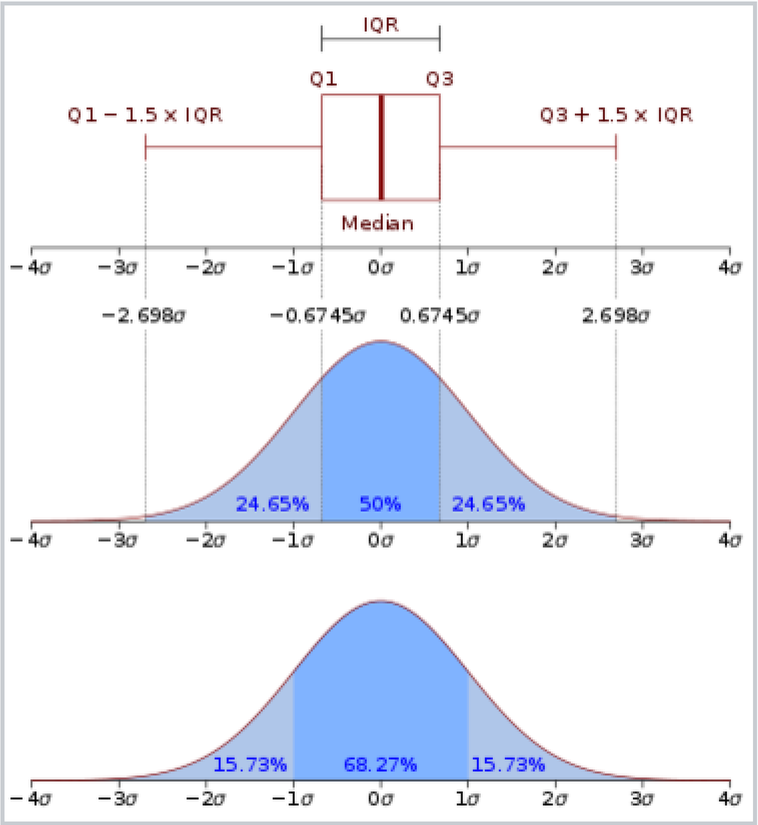
- **Variance** is Square of Standard Deviation
- often represented as:

σ^2, s^2, or $\mathrm{Var}(X)$

- **Coefficient of Variation (CV)** is a measure of relative variability
- It is measured as the ratio of the standard deviation to the mean
- Useful for comparison of variability between two variables or two test
- Can be used for comparison only for ratio-scale variables

Range & Inter-Quartile Range

- Range = Largest Value – Smallest Value
- **Interquartile range** (**IQR**), also called the **midspread** or **middle 50%**
- A measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles
- IQR = $Q_3 - Q_1$



Boxplot (with an interquartile range) and a probability density function (pdf) of a Normal $N(0,\sigma^2)$ Population

https://en.wikipedia.org/wiki/Interquartile_range

Let’s compute SD, Variance and Inter-Quartile Range

```
In [7]: pd.DataFrame(inc_exp.iloc[:,0:5].std().to_frame()).T
   ...:
Out[7]:
   Mthly_HH_Income  Mthly_HH_Expense  No_of_Fly_Members  Emi_or_Rent_Amt  Annual_HH_Income
0     26097.908979      12090.216824           1.517382      6241.434948     320135.792123
```

```
In [8]: pd.DataFrame(inc_exp.iloc[:,0:5].var().to_frame()).T
   ...:
Out[8]:
   Mthly_HH_Income  Mthly_HH_Expense  No_of_Fly_Members  Emi_or_Rent_Amt  Annual_HH_Income
0     6.811009e+08      1.461733e+08           2.302449     3.895551e+07      1.024869e+11
```

Let's compute SD, Variance and Inter-Quartile Range

```
In [9]: summary = inc_exp.describe(include='all')
   ...:
```

Index	Mthly_HH_Income	Mthly_HH_Expense	No_of_Fly_Members	Emi_or_Rent_Amt	Annual_HH_Income	Highest_Qualified_Member	No_of_Earning_Members
count	50	50	50	50	50	50	50
unique	nan	nan	nan	nan	nan	5	nan
top	nan	nan	nan	nan	nan	Graduate	nan
freq	nan	nan	nan	nan	nan	19	nan
mean	4.16e+04	1.88e+04	4.06	3.06e+03	4.9e+05	nan	1.46
std	2.61e+04	1.21e+04	1.52	6.24e+03	3.2e+05	nan	0.734
min	5e+03	2e+03	1	0	6.42e+04	nan	1
25%	2.36e+04	1e+04	3	0	2.59e+05	nan	1
50%	3.5e+04	1.55e+04	4	0	4.47e+05	nan	1
75%	5.04e+04	2.5e+04	5	3.5e+03	5.95e+05	nan	2
max	1e+05	5e+04	7	3.5e+04	1.4e+06	nan	4

EXERCISE : Compute SD, Variance and Inter-Quartile Range in Excel

CoV example

- Suppose you have option to invest in Stock A or Stock B. The stocks have different expected returns and standard deviations. The expected return of Stock A is 15% and Stock B is 10%. Standard Deviation of the returns of these stocks is 10% and 5% respectively.

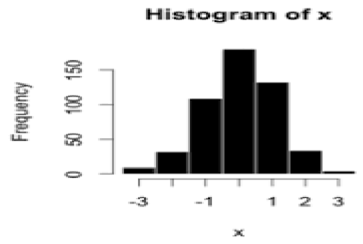
- Which is better investment?

- Stock B would be better investment as its CoV (5% / 10% = 0.5) is less than the CoV of Stock A (10% / 15% = 0.67)

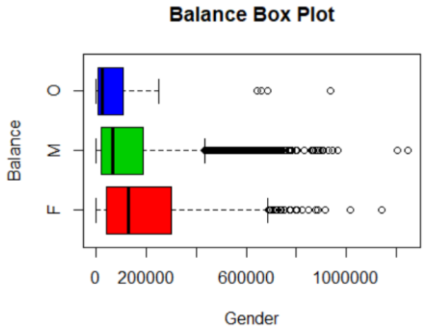
Descriptive Statistics

Descriptive Statistics

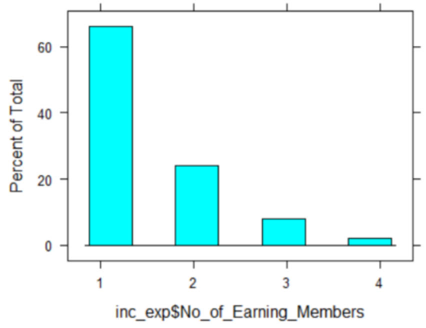
- Categorical – Variable that can take limited & fixed number of values
  - Frequency Distribution
  - Proportions
  - Cross-tabs



- Continuous – Variable whose value is obtained by measuring
  - Measures of Central Tendency
  - Measures of Dispersion



- Discrete – Variable whose value is obtained by counting
  - Frequency Distribution
  - Proportions
  - Mean, Mode

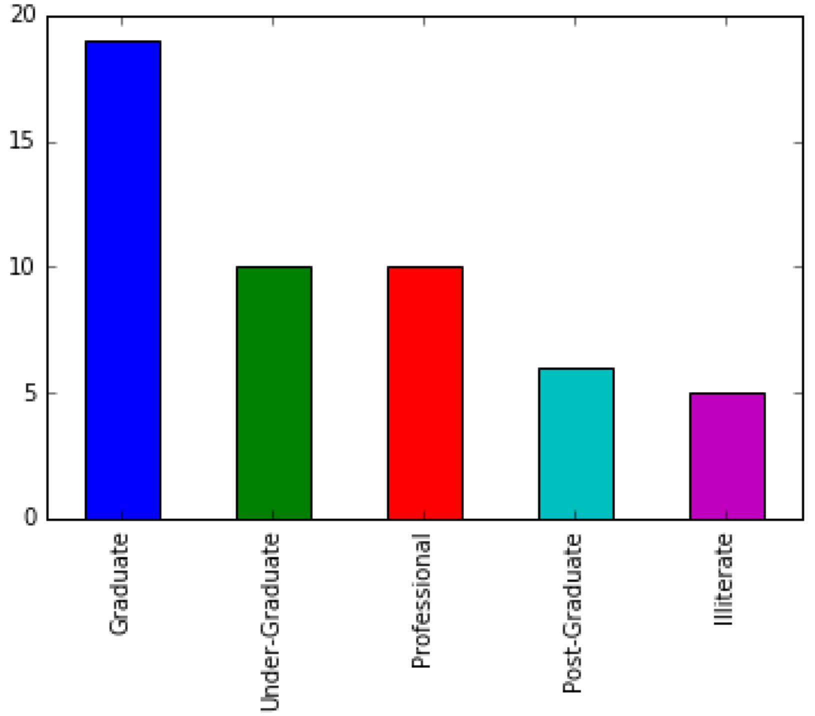


Frequency Distribution & Proportions

```
In [11]: pd.DataFrame(inc_exp['Highest_Qualified_Member'].value_counts().to_frame()).T
Out[11]:
                          Graduate  Professional  Under-Graduate  Post-Graduate  Illiterate
Highest_Qualified_Member        19            10              10              6           5

In [12]: freq = pd.DataFrame(inc_exp['Highest_Qualified_Member'].value_counts())
    ...: freq.reset_index(inplace=True)
    ...: freq.columns = [freq.columns[1], 'count']
    ...: freq['prop'] = freq['count'] / sum(freq['count'])
    ...: freq
    ...:
Out[12]:
  Highest_Qualified_Member  count  prop
0                 Graduate     19  0.38
1             Professional     10  0.20
2           Under-Graduate     10  0.20
3            Post-Graduate      6  0.12
4               Illiterate      5  0.10


In [13]: inc_exp['Highest_Qualified_Member'].value_counts().plot(kind='bar')
    ...:
```



Cross-table

```
In [14]: pd.crosstab(inc_exp.Highest_Qualified_Member,
    ...:             inc_exp.No_of_Earning_Members,margins =True)
    ...:
Out[14]:
No_of_Earning_Members      1   2  3  4  All
Highest_Qualified_Member
Graduate                  14   3  2  0   19
Illiterate                 4   1  0  0    5
Post-Graduate              5   0  1  0    6
Professional               4   5  0  1   10
Under-Graduate             6   3  1  0   10
All                       33  12  4  1   50

In [15]: def percConvert(ser):
    ...:     return round(ser / float(ser[-1]),2)
    ...:
    ...:
    ...: cr_tb_per = pd.crosstab(inc_exp.Highest_Qualified_Member,
    ...:                         inc_exp.No_of_Earning_Members,
    ...:                         margins =True).apply(percConvert, axis=1)
    ...: cr_tb_per.iloc[0:len(cr_tb_per)-1,0:len(cr_tb_per)-2]
    ...:
Out[15]:
No_of_Earning_Members        1     2     3    4
Highest_Qualified_Member
Graduate                  0.74  0.16  0.11  0.0
Illiterate                0.80  0.20  0.00  0.0
Post-Graduate             0.83  0.00  0.17  0.0
Professional              0.40  0.50  0.00  0.1
Under-Graduate            0.60  0.30  0.10  0.0
```

<u>Inference</u>

Absolute number of records are less… still if we have to draw an inference then I may say that, Professional Family have relatively more number of earning members

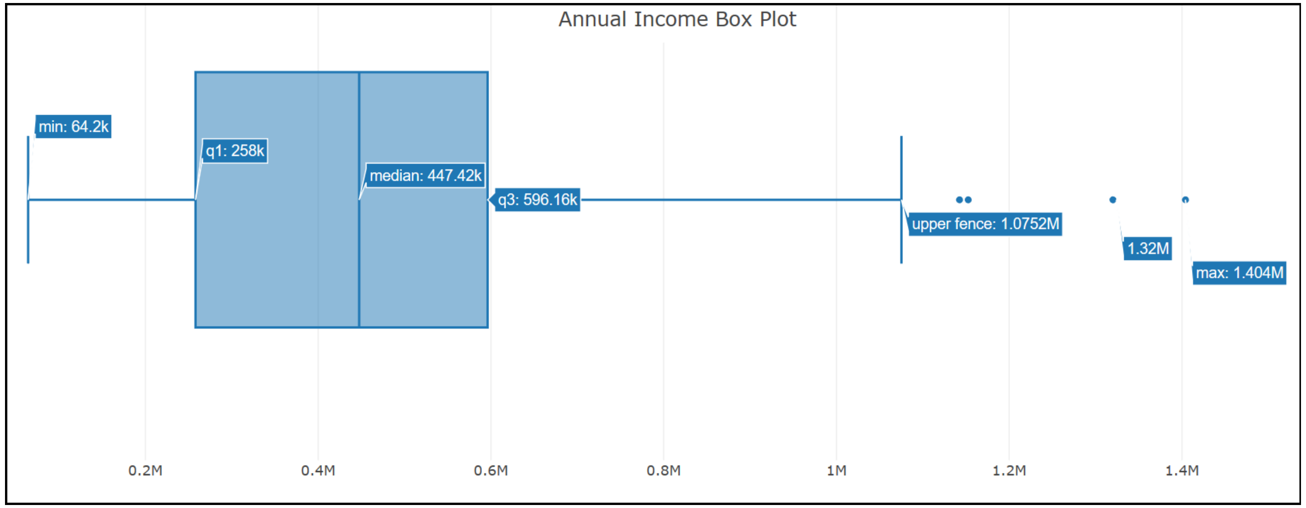
Percentile Distribution and Box Plot

```
In [16]: def percentile_distribution(df,var):
    ...:     per_distr = pd.DataFrame(df[var].describe([0,.01,.05,.1,.25,.5,.75,.9,.95,.99,1]))
    ...:     per_distr.reset_index(inplace=True)
    ...:     per_distr['var'] = per_distr.columns[1]
    ...:     per_distr = (per_distr.pivot_table(index='var', columns=['index'])).iloc[:,0:11]
    ...:     per_distr.columns = per_distr.columns.droplevel()
    ...:     per_distr = per_distr.reindex(columns=['0%','1%','5%','10%','25%','50%','75%','90%','95%','99%','100%'])
    ...:     per_distr.reset_index(inplace=True)
    ...:     return per_distr
    ...:
    ...: per_df = percentile_distribution(df = inc_exp,var = 'Annual_HH_Income')
    ...: per_df

Out[16]:
```

index	var	0%	1%	5%	10%	25%	50%	75%	90%	95%	99%	100%
0	Annual_HH_Income	64200.0	71902.8	104220.0	165360.0	258750.0	447420.0	594720.0	1036320.0	1147944.0	1362840.0	1404000.0

```
In [18]: plotly.offline.plot({
    ...:     "data": [go.Box(x=inc_exp.Annual_HH_Income)],
    ...:     "layout": go.Layout(title="Annual Income Box Plot")
    ...: }, auto_open=True)
    ...:
```

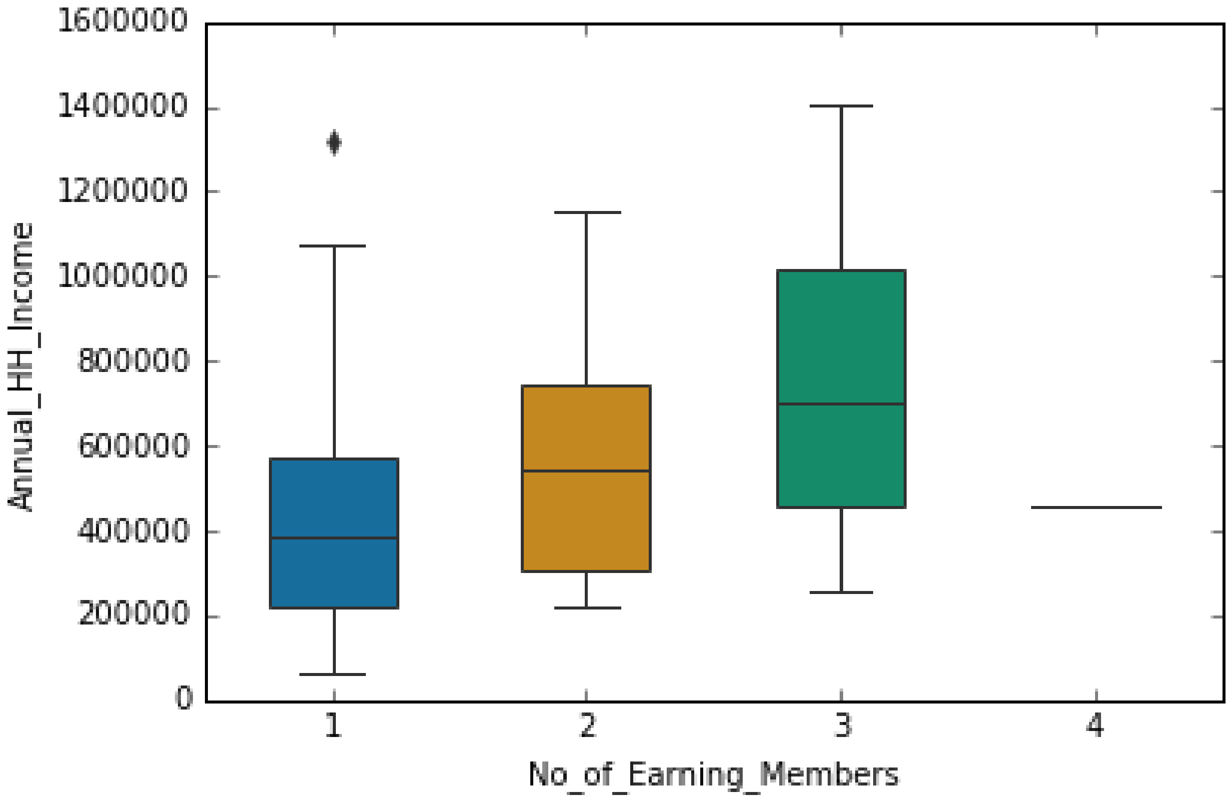


Inference

Annual Household Income of 50% of the households is less than 4.5 Lakhs

Box Plot … contd

```
In [20]: bplot = sns.boxplot(y='Annual_HH_Income', x='No_of_Earning_Members',
    ...:                     data=inc_exp,
    ...:                     width=0.5,
    ...:                     palette="colorblind")
    ...:
```



Inference

Households with more earning members have higher Annual Income

Describe function in Python

```
In [21]: inc_exp.describe()
    ...:
Out[21]:
       Mthly_HH_Income  Mthly_HH_Expense  No_of_Fly_Members  Emi_or_Rent_Amt
count        50.000000         50.000000          50.000000        50.000000
mean      41558.000000      18818.000000           4.060000      3060.000000
std       26097.908979      12090.216824           1.517382      6241.434948
min        5000.000000       2000.000000           1.000000         0.000000
25%       23550.000000      10000.000000           3.000000         0.000000
50%       35000.000000      15500.000000           4.000000         0.000000
75%       50375.000000      25000.000000           5.000000      3500.000000
max      100000.000000      50000.000000           7.000000     35000.000000

       Annual_HH_Income  No_of_Earning_Members
count      5.000000e+01              50.000000
mean       4.900190e+05               1.460000
std        3.201358e+05               0.734291
min        6.420000e+04               1.000000
25%        2.587500e+05               1.000000
50%        4.474200e+05               1.000000
75%        5.947200e+05               2.000000
max        1.404000e+06               4.000000
```

Classroom Exercise

- Perform Descriptive Analysis on the data file "LR_DF.csv"
- This file contains data of a campaign executed by MyBank
- Variable names are self-explanatory.
- Target variable captures the response of customers to marketing offer
  - Target = 1 are the customers who responded to the offer
  - Target = 0 are the customers who did not respond to the offer

The Concept of Probability

Assessing uncertainty using probability

Learning Objectives

1. What is Probability
2. Probability Terminologies (Event, Sample Space, Experiment, Outcomes)
3. Mutually Exclusive Events
4. Dependent and Independent Events
5. Marginal & Joint Probability
6. Conditional Probability & Contingency Table
7. Association of Attributes
8. Bayes’ Theorem

Probability

- Probability is used to deal with uncertainty
- Intuitively, the probability of an event is a number that measures the chance, or likelihood, that the event will occur
- Probability value ranges between 0 & 1
- Types of Probability
  - A Priori Probability
  - Empirical Probability
  - Subjective Probability (Delphi Technique)

Some dictionary definitions

a priori

adjective

1. relating to or denoting reasoning or knowledge which proceeds from theoretical deduction rather than from observation or experience.

adverb

1. in a way based on theoretical deduction rather than empirical observation.

Origin : Latin

a posteriori

adjective

1. relating to or denoting reasoning or knowledge which proceeds from observations or experiences to the deduction of probable causes.

adverb

1. in a way based on reasoning from known facts or past events rather than by making assumptions or predictions.

Origin : Latin

empirical (same as "a posteriori")

adjective

1. based on, concerned with, or verifiable by observation or experience rather than theory or pure logic.

Some terminologies

- **Probability** refers to chance or likelihood of a particular **event** taking place

- An **event** is the phenomenon or outcome of your interest in an **experiment**

- **An experiment** is a process that is performed to understand and observe **possible outcomes**

- The set of all **possible outcomes** is called the **sample space**

Sample Space e.g.

A coin is tossed three consecutive times.

1. What are the possible outcomes?
2. What is the (a priori) probability of getting at least 2 Heads?
3. What is the event in question 2

Tiny tots e.g. contd…

Sr. No	Student Name	Age in Mths	Gender
1	Chintu	24 Mths	M
2	Pintu	37 Mths	M
3	Tinku	38 Mths	M
4	Pappu	38 Mths	M
5	Munnu	36 Mths	M
6	Chunnu	39 Mths	M
7	Samy	40 Mths	M
8	Bubbly	37 Mths	F
9	Dubbly	38 Mths	F
10	Monu	41 Mths	F
11	Sonu	40 Mths	F
12	Kitu	36 Mths	F
13	Pitu	37 Mths	F
14	Guddu	37 Mths	F
15	Pinku	39 Mths	F

- Answer the below question based on data given in adjacent table:
  - What is the Sample Space for Gender?
  - What is the Sample Space for Age?
  - You are playing the Blind Fold game. What is the probability of you catching a tiny tot who is Female?
  - You are playing the game passing-the-pass game with music. What is the probability that the pass is with a tot aged above 38 when the music stops?

Mutually Exclusive Events

- In probability theory, two events (A & B) are mutually exclusive if the occurrence of one (A) implies the non-occurrence of other (B)

Eg.

- Tossing of a coin. Heads and Tails are mutually exclusive events
- Rolling of a dice
- Pulling a card from a well shuffled deck of cards and wanting to know whether it is a King or the Queen card.

Dependent & Independent Events

- Independent Events - Two events are said to be independent, if the occurrence of A is in no way influenced by the occurrence of B. Likewise occurrence of B is in no way influenced by the occurrence of A.
  - E.g. Rolling a dice and flipping a coin. The probability of getting any number on rolling of a dice does not change the probability of getting a head or tail on tossing of the coin

- Dependent Events - Two events are said to be dependent, if the occurrence of one event influences the probability of occurrence of the other
  - E.g. You draw a card from a deck. It is Ace. What is the probability of the second card being an Ace.

Rules for computing probability

- Addition Rule – Mutually Exclusive Events

  $P(A \cup B) = P(A) + P(B)$

  Symbol $A \cup B$ is called A union B

- Addition Rule – Events are not Mutually Exclusive Events

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

  Symbol $(A \cap B)$ is called A intersection B

Rules for computing probability

- Multiplication Rule – Independent Events

$$P(A \cap B) = P(A) \cdot P(B)$$

- Multiplication Rule – Events are not independent

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$

$P(B \mid A)$ is called the conditional probability of B given the fact that A has already occurred

$$P(A \cap B) = P(B) \cdot P(A \mid B)$$

$P(A \mid B)$ is called the conditional probability of A given the fact that B has already occurred

Classroom Example

- From a pack of well shuffled cards, a card is picked at random.

1) What is the probability that the selected card is a King or a Queen?
2) What is the probability that the selected card is a King or Diamond?

- You have two packs of well shuffled cards. From each pack you draw a card at random.

1) What is the probability that both the cards are Diamond?

K2 Analytics
Building Skills, Building Individuals

Marginal & Joint Probability
Conditional Probability
Contingency Table

Marginal and Joint Probability

- Let's assume you are a financial analyst and you are interested in two popular stocks: TCS and Reliance

- Since these stocks are being considered for a portfolio, you are interested in how they behave individually and as a pair

1. If TCS suffers a loss on any given day, what tends to happen to Reliance that same day?
2. If Reliance does NOT suffer a loss on any given day, what tends to happen to TCS that same day?

https://www.youtube.com/watch?v=SrEmzdOT65s
https://www.youtube.com/watch?v=DkHWKAy47X0

A Few Definitions

- For this problem we are going to analyse each trading day in a F.Y.; approximately 250 days.

- Each stock can be in one of two states:

1. **Loss**. The return from day-to-day was < 0
2. **No Loss**. The return from day-to-day was >= 0. This includes being "even"

Possible Joint Outcomes & Contingency Table

Reliance	TCS	Joint Occurrence
Loss (R’)	Loss (T’)	**R’ ∩ T’**
No Loss (R)	Loss (T’)	**R ∩ T’**
Loss (R’)	No Loss (T)	**R’ ∩ T**
No Loss (R)	No Loss (T)	**R ∩ T**

Contingency Table		**TCS**		**Row Total**
		Loss	No Loss	
Reliance	Loss	**R’ ∩ T’**	**R’ ∩ T**	**R’**
	No Loss	**R ∩ T’**	**R ∩ T**	**R**
Col. Total		**T’**	**T**	

Contingency Table

Joint Occurrence Table		TCS		Row Total
		Loss	No Loss	
Reliance	Loss	65	55	120
	No Loss	50	80	130
Col. Total		115	135	250

Joint & Marginal Probabilities		TCS		Row Total
		Loss	No Loss	
Reliance	Loss	65 / 250 = 0.26	55 / 250 = 0.22	120 /250 = 0.48
	No Loss	50 / 250 = 0.20	80 / 250 = 0.32	130 /250 = 0.52
Col. Total		115 /250 = 0.46	135 /250 = 0.54	250 /250 = 1

Joint & Marginal Probabilities Table

Joint Occurrence Table		**TCS** Loss	**TCS** No Loss	**Row Total**
Reliance	Loss	0.26	0.22	
	No Loss	0.20	0.32	
	Col. Total			

Joint Probabilities

Joint & Marginal Probabilities		**TCS** Loss	**TCS** No Loss	**Row Total**
Reliance	Loss			120 /250 = 0.48
	No Loss			130 /250 = 0.52
	Col. Total	115 /250 = 0.46	135 /250 = 0.54	

Marginal Probabilities

*Marginal Probability is also called Simple Probability

Marginal Probability ...e.g

Joint & Marginal Probabilities		TCS		Row Prob
		Loss	No Loss	
Reliance	Loss	0.26	0.22	0.48
	No Loss	0.20	0.32	0.52
	Col. Prob	0.46	0.54	250 /250 = 1

- What is the Probability of TCS stock giving LOSS tomorrow?
- What is the Probability of Reliance stock NOT giving LOSS tomorrow?

Joint Probability …e.g

Joint & Marginal Probabilities		TCS Loss	TCS No Loss	Row Prob
Reliance	Loss	0.26	0.22	0.48
	No Loss	0.20	0.32	0.52
	Col. Prob	0.46	0.54	250 /250 = 1

- What is the Probability of TCS **AND** Reliance stock giving LOSS?
- What is the Probability of TCS giving NO LOSS **AND** Reliance giving LOSS?
- What is the Probability of TCS **OR** Reliance giving NO LOSS?

Conditional Probability… e.g

Joint & Marginal Probabilities		TCS		Row Prob
		Loss	No Loss	
Reliance	Loss	0.26	0.22	0.48
	No Loss	0.20	0.32	0.52
	Col. Prob	0.46	0.54	250 /250 = 1

- TCS stock has given LOSS today. What is the probability that Reliance stock has also given LOSS today?
- TCS stock has given LOSS today. What is the probability that Reliance stock has also given NO LOSS today?
- Hint: $P(A \mid B) = P(A \cap B) / P(B)$

Classroom e.g | Contingency Table

Of the cars on a used car lot, 70% have Air Conditioner (AC) and 40% have a CD Player (CD). 20% cars have both AC and CD

- Create the Contingency Table

- What is the probability that a car has a CD player, given that it has AC

$$P\,(CD \mid AC) = \ ?$$

Concepts of Conditional Probability are used in Market Basket Analysis (Association Rules)

- Let us assume you have the Transactions for a Retail Outlet
- **Transaction Summary**

  # Invoices = 10000

  # Invoices has Product A in the item set = 900

  # Invoices has Product B in the item set = 500

  # Invoice has Product A & B in the item set = 350
- **Support Computation**

  Support of Product A = 900 / 10000 = 9%

  Support of Product B = 500 / 10000 = 5%
- **Rule A -> B (Customer who buy A also buys B)**

  Support of Product A & B = 350 / 10000 = 3.5%

  Confidence of Rule A -> B = 350 / 900 = 38.9%

  (%of customers who bought B from those who bought A)

  Lift = Confidence / Support of Product B = 38.9 / 5 = 7.77

  (Likelihood of customer purchasing product B is 7.77 times higher if the customer has purchased A)

- Items are the objects that we are identifying association between
- Association Rules a relation of the form X -> Y
  - If you have the item / items in the items set on the LHS then customer will be interested in the item Y on the RHS
- Support is the fraction of transactions in the dataset that contain the item or item set
- Confidence is the proportion of times the customer has taken the item Y given she has also taken X
- Lift is ratio of Confidence of the Rule divided by support of Product Y alone

K2 Analytics
Building Skills, Building Individuals

Association of Attributes
Bayes' Theorem

Association of Attributes

- Of 37 men and 33 women, 36 are teetotallers (completely abstain from alcoholic beverages). Nine of the women are non-smokers and 18 of the men smoke but do not drink. 13 of the men and 7 of the women drink but do not smoke.

- How many, both drink and smoke? What is the associated probability?

Bayes' Theorem

- Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event

- Bayes' Theorem is an extension of Conditional Probability

$$P\ (A \mid B) = \frac{P\ (B \mid A)\ .\ P\ (A)}{P\ (B)}$$

where $P\ (B) \neq 0$

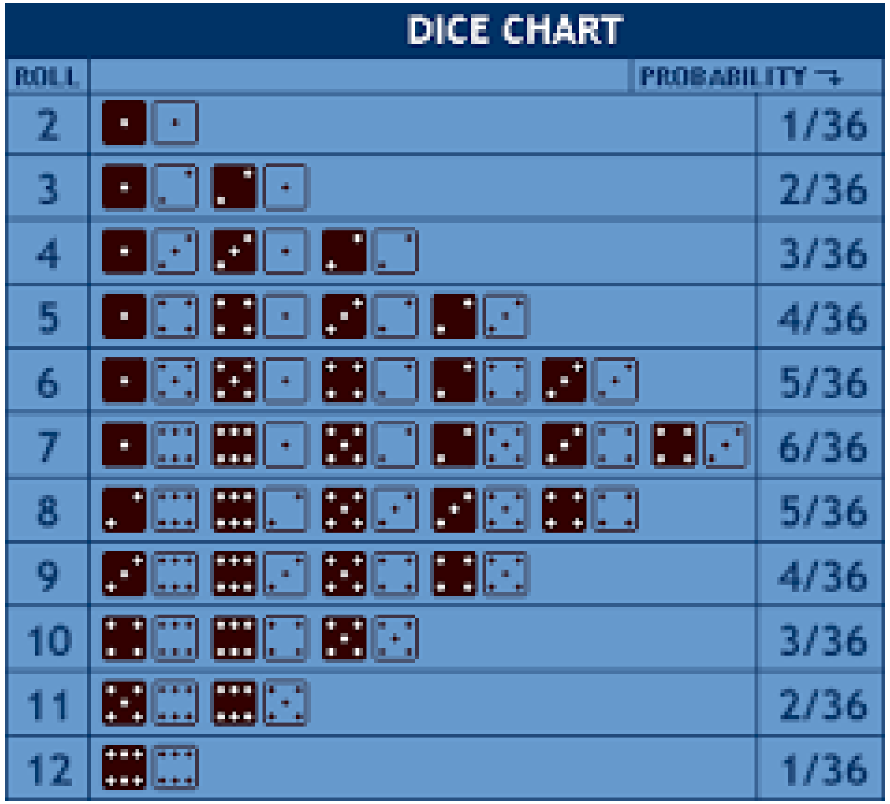
Bayes' Theorem Discussion Problem

- A drilling company has estimated a 40% chance of striking oil for their new well
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests
- Given that this well has been scheduled for a detailed test, what is the probability that the well be successful?

Probability Distribution
Binomial Distribution
Poisson Distribution

Probability Distribution

- Suppose you are playing the game of Ludo with two dice
- The sum of the value on the face of the two dice can take any number between 2 & 12. The sum of the value in this e.g. will be referred as **Random Variable**



A **Probability Distribution** is a total listing of the various values the random variable can take along with the corresponding probability of each value

From Dice to Coins… Binomial Distribution

- Assume you flip a coin 10 times.
- What is the probability that you will get Head all the 10 times?
- What is the probability that you get exactly 6 Head and 4 Tail?

Binomial Distribution (Bernoulli trials)

- The Binomial Distribution is a widely used probability distribution of a discrete random variable

- Conditions for applying Binomial Distribution
  - Trials are independent and random
  - There are fixed number of trials (n trials)
  - There are only two outcomes of the trial designated as success or failure
  - The probability of success is uniform throughout the n trials

Binomial Probability Function

- The probability of getting X successes of n trials is indeed the definition of a Binomial Distribution. The Binomial Probability Function is given by the following expression

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$$

Where x can take the values 0, 1, 2, …n

*P (X = x) is the probability of getting **x** success in **n** trials*

p is the probability of success which is the same throughout the n trials

p is the parameter of the Binomial Distribution

$\binom{n}{x}$ *is the number of ways in which **x** success can take place out of **n** trials and this is equal to*

$$\frac{n!}{x!\cdot(n-x)!}$$

Binomial Distribution e.g.

- MyBank has a large Credit Card portfolio. Based on empirical data, they have found that 60% of the customers pay their bill on time. If a sample of 10 accounts is selected from the current database, construct the Probability Distribution of accounts paying on time.

x	p	cum prob
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.832710
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1.000000

Function in Excel
BINOM.DIST(nSuccess, Trials, Prob, Cum)

Function in R
dbinom(x, size, prob) ## Probabailty
pbinom(x, size, prob) ## Cum. Probability

Mean and Standard Deviation of Binomial Distribution

- Mean of Binomial Distribution

$$\mu = \text{Exp}(x) = n \, . \, p = n \, . \, \pi$$

- Standard Deviation of Binomial Distribution

$$\sigma = \sqrt{n.p.q} = \sqrt{n.\pi \, . \, (1 - \pi)}$$

Poisson Distribution

Poisson Distribution

- The Poisson distribution is the discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time.
  - Occurrences of event can be over time, distance, area or volume

- **Conditions for Poisson Distribution:**
  - An event can occur any number of times during a time period.
  - Events occur independently. In other words, if an event occurs, it does not affect the probability of another event occurring in the same time period.
  - The rate of occurrence is constant; that is, the rate does not change based on time.
  - The probability of an event occurring is proportional to the length of the time period. For example, it should be twice as likely for an event to occur in a 2 hour time period than it is for an event to occur in a 1 hour period.

https://brilliant.org/wiki/poisson-distribution/

Poisson E.g.

- The number of car accidents in a day
- The number of customers visiting a Customer Service Center every hour
- Number of calls you receive in a day
- Number of defects per 100 Sq mtr of cloth

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Mean of Poisson Distribution

μ = λ

Standard Deviation of Poisson Distribution

σ = λ

Derivation of Poisson Distribution from Binomial Distribution
https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239

Poisson Distribution | Call Centre E.g.

- A call centre receives an average of 4.5 calls every 5 minutes. Each agent can handle one of these calls over the 5 minute period. If a call is received, but no agent is available to take it, then that caller will be placed on hold. Assuming that the calls follow a Poisson distribution, what is the minimum number of agents needed on duty so that calls are placed on hold at most 10% of the time?

Solution:

In order for all calls to be taken, the number of agents on duty should be greater than or equal to the number of calls received. If "**X**" is the number of calls received and "**k**" is the number of agents, then "**k**" should be set such that P (X > k) <= 0.1 or equivalently P (X <= k) > 0.9

https://brilliant.org/wiki/poisson-distribution/

...contd

The average number of calls is 4.5, so $\lambda = 4.5$

$$P(X=0) = \frac{4.5^0 e^{-4.5}}{0!} \approx 0.011$$

$$P(X=1) = \frac{4.5^1 e^{-4.5}}{1!} \approx 0.050 \quad \Longrightarrow \quad P(X \leq 1) \approx 0.061$$

$$P(X=2) = \frac{4.5^2 e^{-4.5}}{2!} \approx 0.112 \quad \Longrightarrow \quad P(X \leq 2) \approx 0.173$$

$$P(X=3) = \frac{4.5^3 e^{-4.5}}{3!} \approx 0.169 \quad \Longrightarrow \quad P(X \leq 3) \approx 0.342$$

$$P(X=4) = \frac{4.5^4 e^{-4.5}}{4!} \approx 0.190 \quad \Longrightarrow \quad P(X \leq 4) \approx 0.532$$

$$P(X=5) = \frac{4.5^5 e^{-4.5}}{5!} \approx 0.171 \quad \Longrightarrow \quad P(X \leq 5) \approx 0.703$$

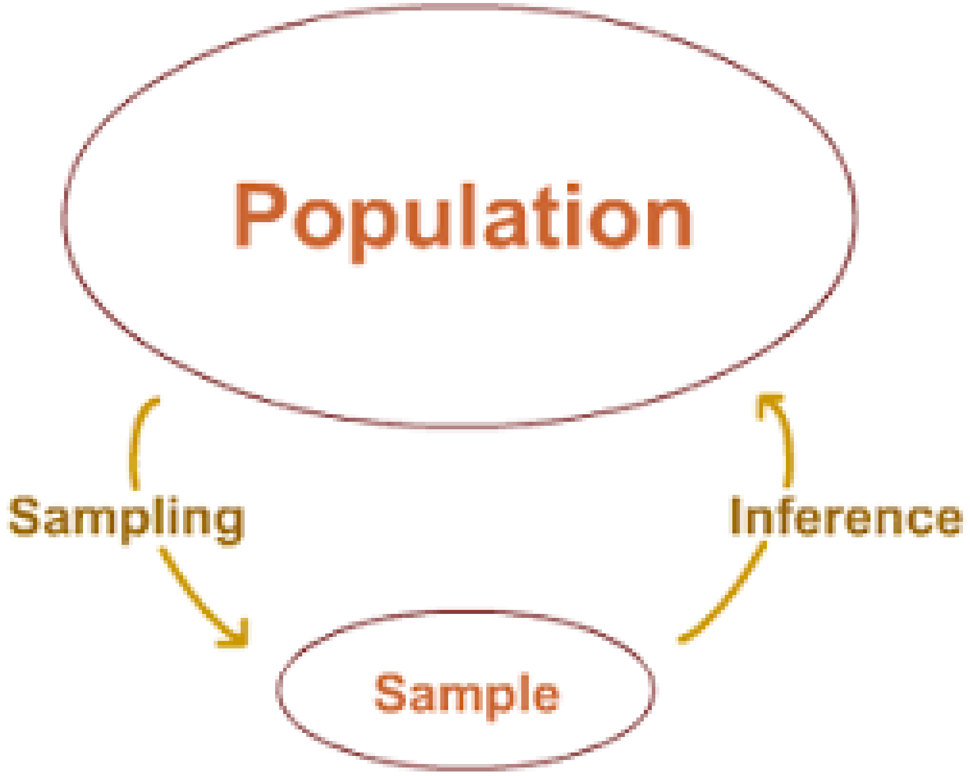
$$P(X=6) = \frac{4.5^6 e^{-4.5}}{6!} \approx 0.128 \quad \Longrightarrow \quad P(X \leq 6) \approx 0.831$$

$$P(X=7) = \frac{4.5^7 e^{-4.5}}{7!} \approx 0.082 \quad \Longrightarrow \quad P(X \leq 7) \approx 0.913.$$

- No of Agents required is 7

Sampling Distribution

Sampling Objective



- ***Objective of Sampling is to derive inference from Sample about the Population***

- ***Why not derive inference directly from the population????***

Point Estimates as Population Parameter

- Point Estimates of a population:
  - Mean
  - Proportions

Population Parameter Estimate	Sample	Population	Description
Mean	$\bar{x}$	μ	The sample mean $\bar{X}$ of a sample is an estimator of the population mean μ
Proportion	$\hat{p}$	π	The sample proportion $\hat{p}$ is an estimator of the population proportion π

- The sample mean $\bar{X}$ is as an **unbiased estimator** of the population mean μ because $E(\bar{X}) = \mu$
- An estimator (as opposed to an estimate) is a **sample statistic** that predicts a value of a parameter

Sampling Distribution – A Conceptual Framework

- The probability distribution of all the possible values a **sample statistic** can take is called the **Sampling Distribution** of the statistic.
  - Take many samples from population
  - Compute mean $\overline{x}$ of each sample
  - The distribution of mean taken over many sample is Sampling Distribution

- **Sampling Error =** $|\ \overline{x} - \mu\ |$

- Sample Size – To reduce the sample error we will have to take large sample sizes

Central Limit Theorem

Central Limit Theorem

- **Central Limit Theorem** states that ***irrespective of the shape of the distribution of the original population***, the sampling distribution of the mean will approach a ***normal distribution*** as the size of the sample increases and becomes large

- Famous **Lindeberg–Lévy** made the discovery of the this landmark, hallmark theory of CLT

Law of Large Numbers & Central Limit Theorem…

- Based on Law of Large Numbers and CLT:
  - The mean of the sampling distribution (i.e. mean of mean) will approach population mean (μ) with large number of trials,
  - the sampling **distribution of the mean** approaches a **normal distribution**
  - and variance of the sampling distribution = σ^2/N as N, the sample size, increases.
- How large should be N?
  - Thumb rule N > 30

Let's prove Central Limit Theorem through simulation

Creating Thousand Samples each having 30 observations out of 50 records

```
In [23]: sample_dst = pd.DataFrame()
    ...: for i in range(1,1001):
    ...:     temp = inc_exp.iloc[np.random.randint(0, len(inc_exp), size=30)]
    ...:     temp['sample_no'] = i
    ...:     sample_dst = sample_dst.append(temp)
    ...:     del temp
    ...:
    ...:
In [24]: sample_dst.head()
    ...:
Out[24]:
    Mthly_HH_Income  Mthly_HH_Expense  No_of_Fly_Members  Emi_or_Rent_Amt
34            46000             25000                  5             3500
32            45000             10000                  2             1000
18            29000              6600                  2             2000
11            22000             25000                  6            12000
7             18000             20000                  5             8000

    Annual_HH_Income Highest_Qualified_Member  No_of_Earning_Members  sample_no
34            596160                 Graduate                      1          1
32            437400            Post-Graduate                      1          1
18            348000                 Graduate                      1          1
11            279840               Illiterate                      1          1
7             216000                 Graduate                      1          1
```

...contd

```
In [27]: sample_mean = sample_dst.groupby('sample_no', as_index=False).agg({
    ...:             "Mthly_HH_Income": "mean", "Mthly_HH_Expense": "mean",
    ...:             "No_of_Fly_Members": "mean","Emi_or_Rent_Amt": "mean",
    ...:             "Annual_HH_Income": "mean"
    ...:             })
    ...: ###Rearrange the columns
    ...: sample_mean = sample_mean.reindex(columns=['sample_no', 'Mthly_HH_Income', 'Mthly_HH_Expense',
    ...:                                  'No_of_Fly_Members', 'Emi_or_Rent_Amt', 'Annual_HH_Income',
    ...:                                  'Highest_Qualified_Member', 'No_of_Earning_Members'])
    ...: ###Sample Mean
    ...: smean = pd.DataFrame(sample_mean.iloc[:,1:6].mean().to_frame())
    ...: smean.reset_index(inplace=True)
    ...: smean.columns = ['s_vars','smean']
    ...: ###Population Mean
    ...: pmean = pd.DataFrame(inc_exp.iloc[:,0:5].mean().to_frame())
    ...: pmean.reset_index(inplace=True)
    ...: pmean.columns = ['p_vars','pmean']
    ...:
    ...: ###cbind sample_mean and population_mean
    ...: spmean = pd.concat([smean.reset_index(drop=True), pmean], axis=1)
    ...: ### Ratio of sample_mean and population_mean
    ...: spmean['ratio'] = spmean.smean / spmean.pmean
    ...: del spmean['p_vars']
    ...: spmean
    ...:
Out[27]:
```

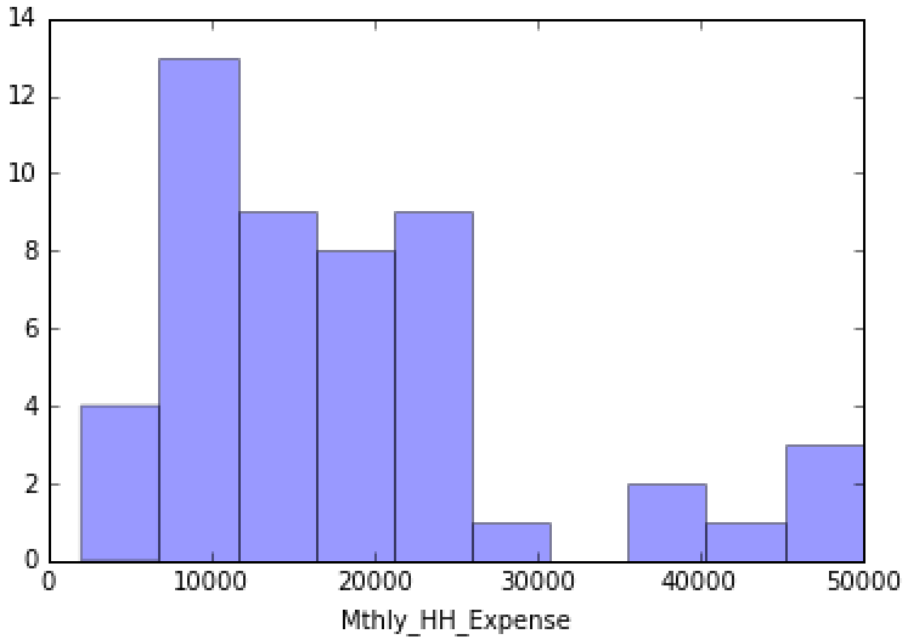
	s_vars	smean	pmean	ratio
0	Mthly_HH_Income	41262.470000	41558.00	0.992889
1	Mthly_HH_Expense	18792.463333	18818.00	0.998643
2	No_of_Fly_Members	4.057167	4.06	0.999302
3	Emi_or_Rent_Amt	3116.616667	3060.00	1.018502
4	Annual_HH_Income	486267.129600	490019.04	0.992343

Sampling Distribution of Mean follows a Normal Distribution

- 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 $\bar{x}$ follows a Normal Distribution with Mean = μ and Standard Deviation = $\sigma / \sqrt{n}$
- $\sigma / \sqrt{n}$ is known by the term Standard Error
- A standard error is the standard deviation of the sampling distribution of a statistic

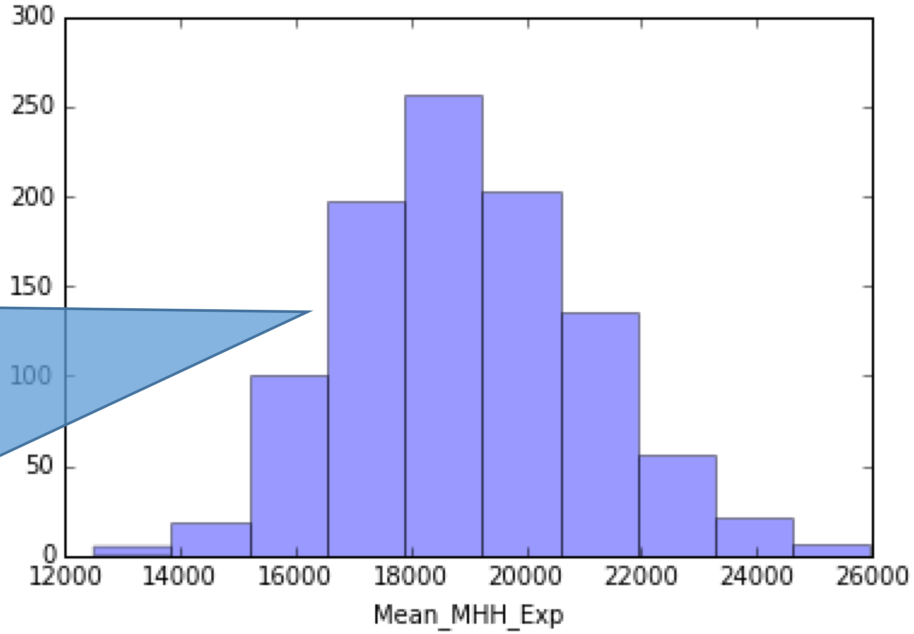
Histogram

```
In [33]: sns.distplot(inc_exp.Mthly_HH_Expense,kde=False, bins=10)
In [34]: sns.distplot(sample_mean.Mean_MHH_Exp,kde=False, bins=10)
```



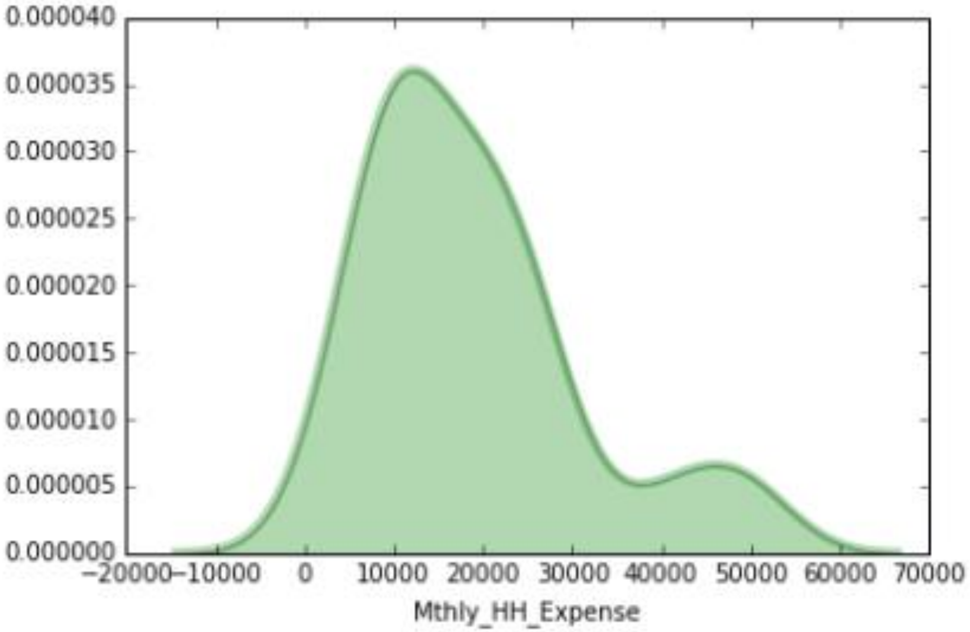
Central Limit Theorem:

the sampling ***distribution of the mean*** *approaches a* ***normal distribution*** *even if the original variables themselves are not normally* distributed.



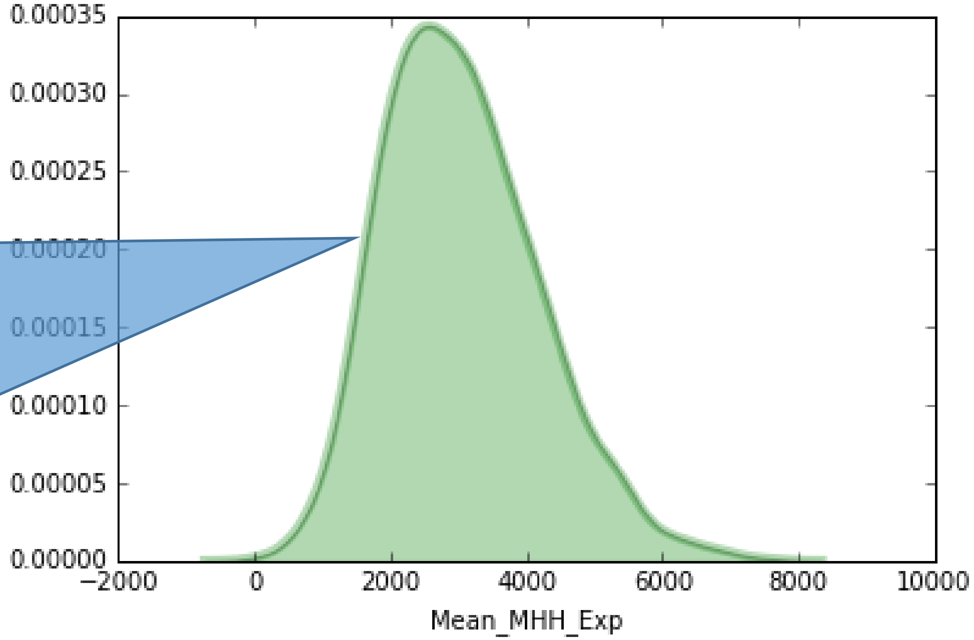
Normal Distribution | Continuous Probability Density Function

- In probability theory, the **normal** (or **Gaussian**) **distribution** is a very common continuous probability distribution



Central Limit Theorem:

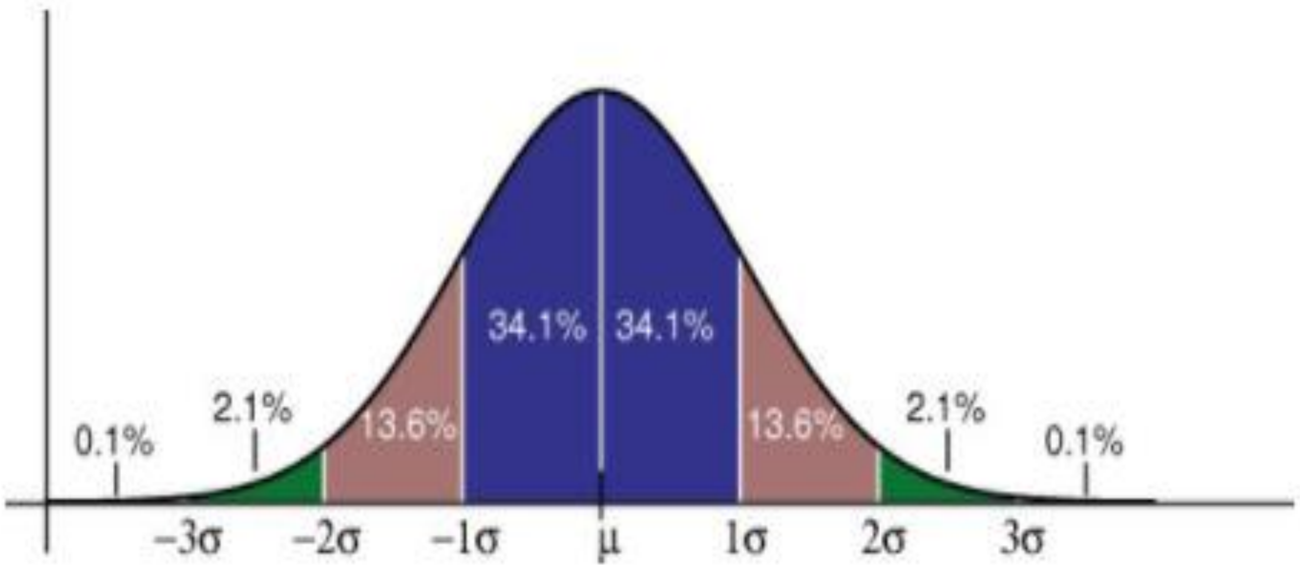
the sampling ***distribution of the mean*** *approaches a* ***normal distribution*** *even if the original variables themselves are not normally* distributed.

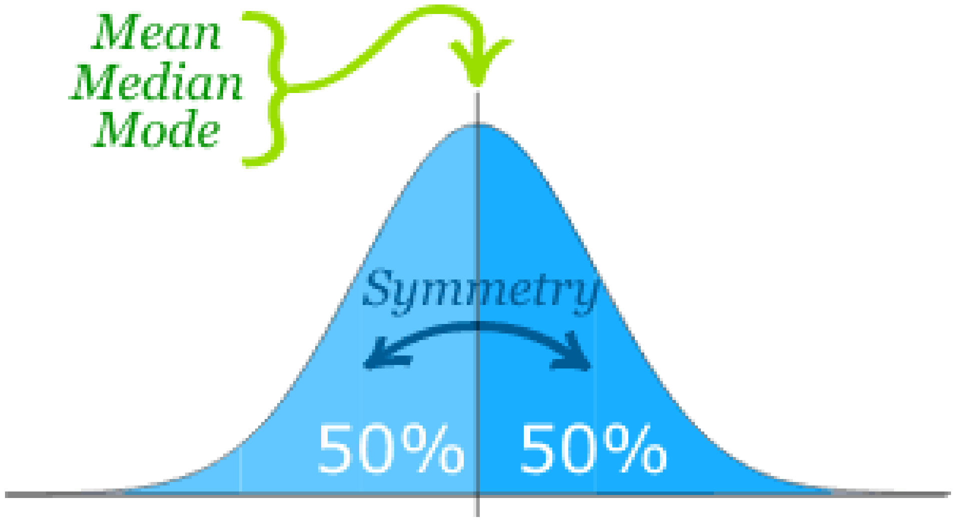


https://en.wikipedia.org/wiki/Normal_distribution

Normal Distribution
Standard Normal Distribution
(Z Transformation)

Properties of Normal Distribution





- **Normal Distribution** - a function that represents the distribution of many random variables as a symmetrical bell-shaped graph.
- A normal distribution, sometimes called the **bell curve**, is a distribution that occurs naturally in many situations
- For a perfect normal distribution – the mean, median, and mode are all equal
- If the tails of the normal distribution are extended, they will extend to the horizontal axis without actually touching it (asymptotic to X-Axis)
- The normal distribution has two properties namely μ (Mean) and σ (Standard Deviation)
- Total Area under the curve is equal to 1

http://www.statisticshowto.com/probability-and-statistics/normal-distributions/

Probability Density function of the Normal Distribution

The probability density of the normal distribution is

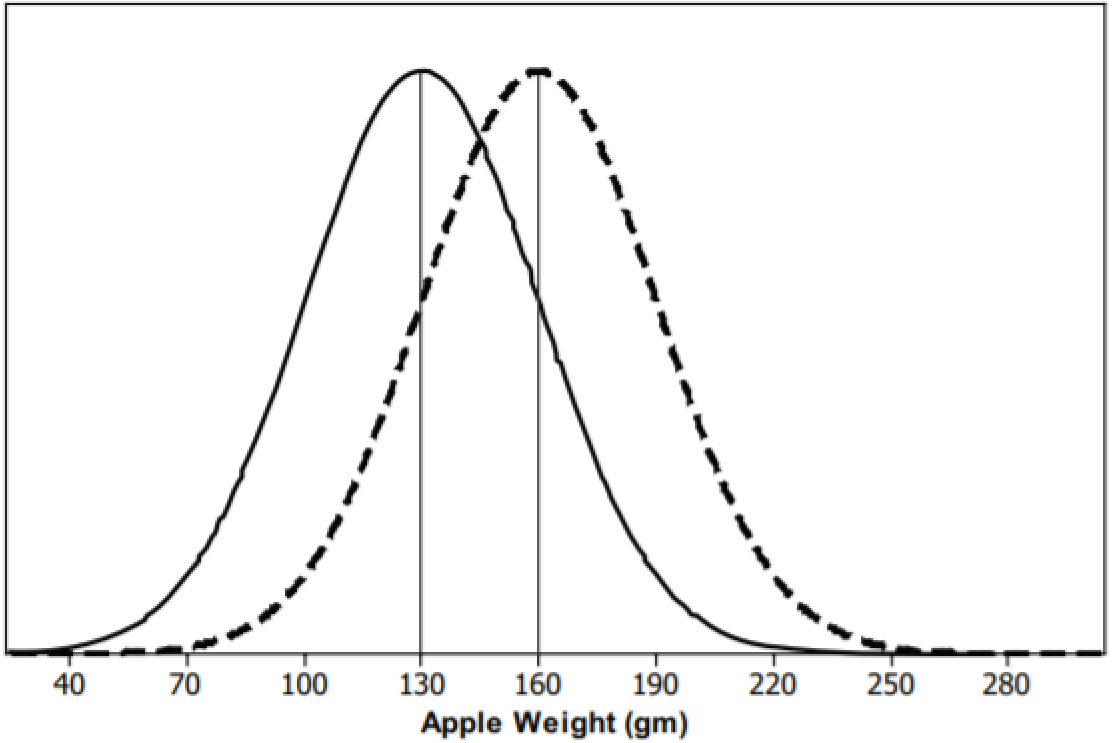
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

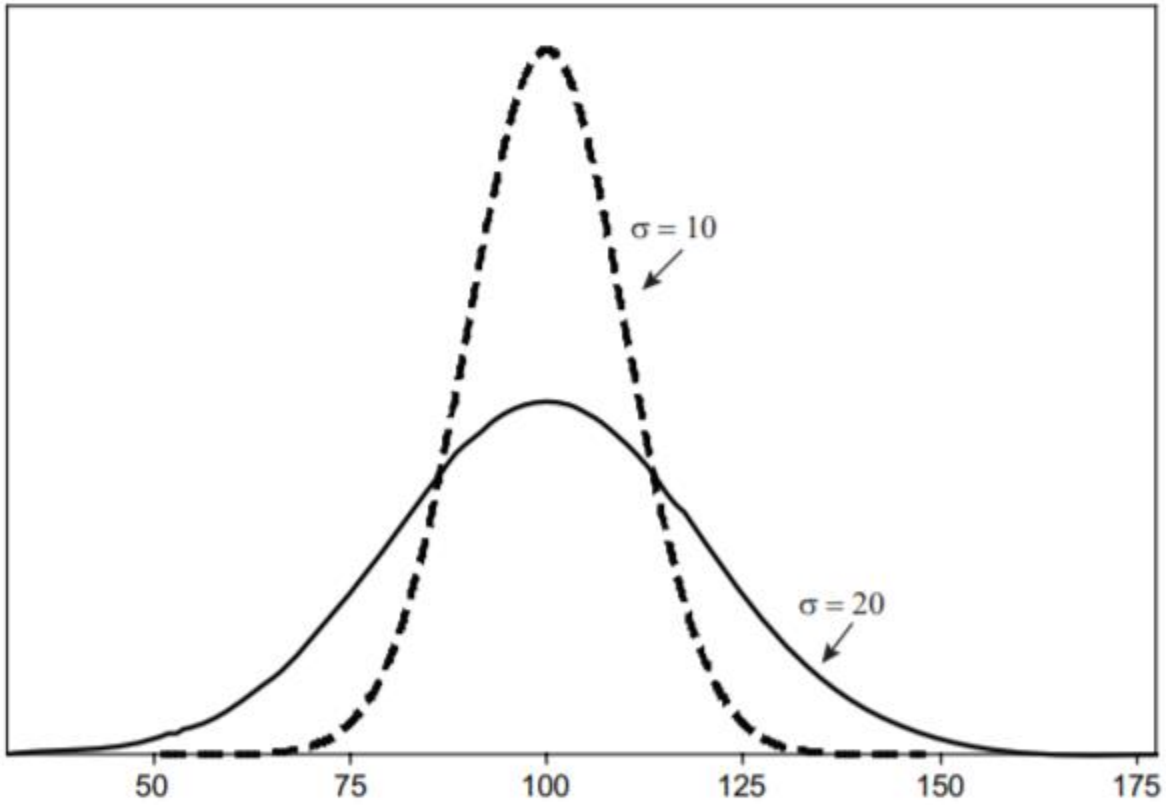
where

- μ is the mean or expectation of the distribution (and also its median and mode),
- σ is the standard deviation, and
- σ^2 is the variance.

https://en.wikipedia.org/wiki/Normal_distribution

Normal Distribution Curves

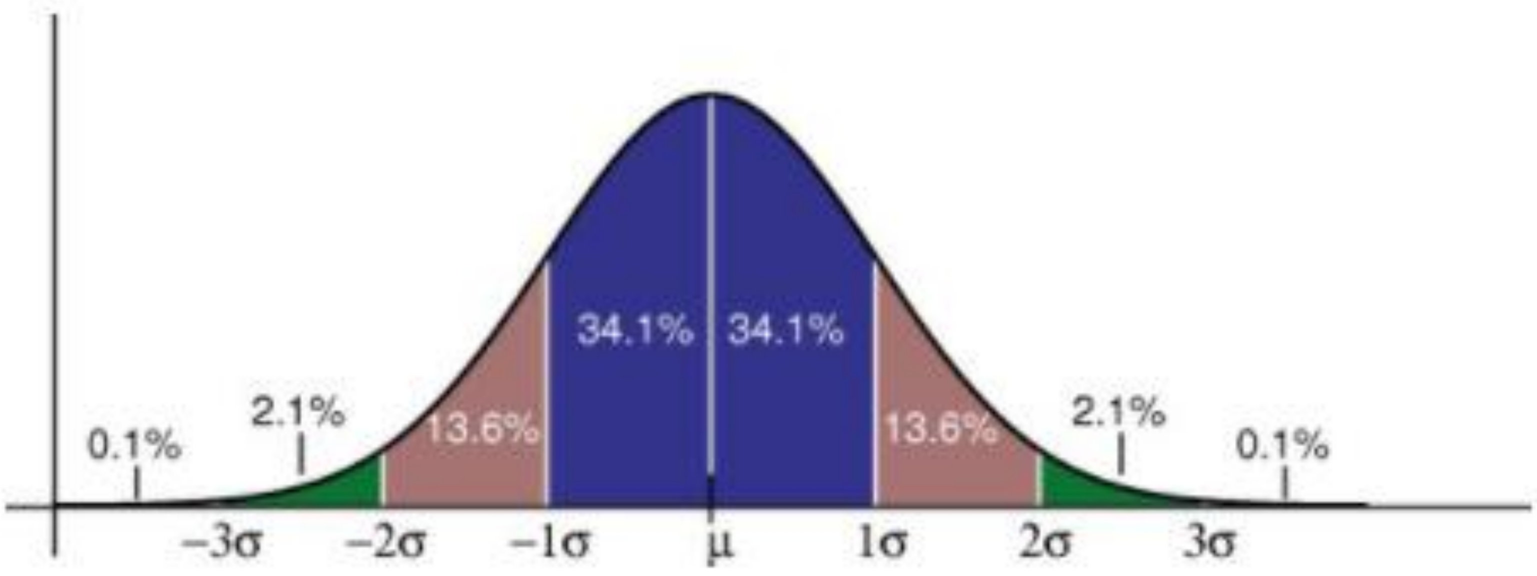




- We can have innumerable Normal Distribution Curves with different μ & σ
- Irrespective of the μ & σ, the total area under the curve is always 1
- And Mathematically the empirical relationship between the μ & σ (shown next slide) will hold

Normal Distribution **Empirical rule**

- **Empirical rule** between **Standard Deviation and Mean** for Normally Distributed Data
  - **68%** of the data falls within **one** standard deviation of the mean.
  - **95%** of the data falls within **two** standard deviations of the mean.
  - **99.7%** of the data falls within **three** standard deviations of the mean.



http://www.statisticshowto.com/probability-and-statistics/normal-distributions/

Validating SD & Mean Empirical Rule….

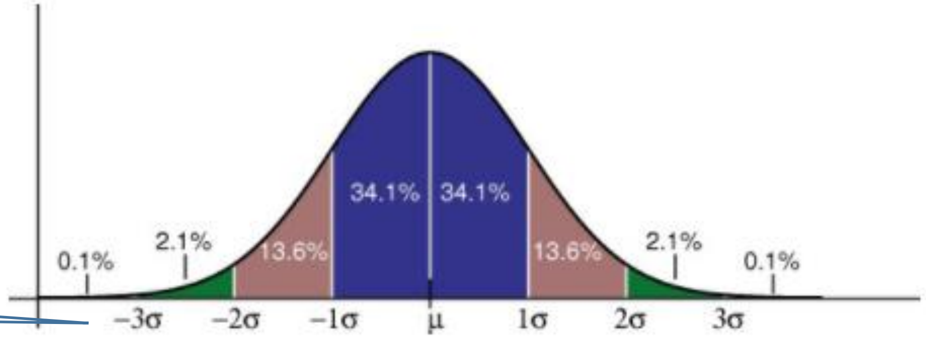
```
In [40]: sample_mean.head()
    ...:
Out[40]:
   sample_no  Mean_MHH_Inc  Mean_MHH_Exp  Mean_Fly_Mem  Mean_EMI_Rent  Mean_Ann_Inc
0          1  42846.666667  20280.000000   4083.333333       4.433333        520446
1          2  56733.333333  22643.333333   1566.666667       4.466667        664564
2          3  43383.333333  20606.666667   4216.666667       3.733333        496116
3          4  38730.000000  16903.333333   1616.666667       3.633333        478042
4          5  49416.666667  20753.333333   1616.666667       4.300000        587288


In [41]: inc_Mean = round(sample_mean.Mean_MHH_Inc.mean(),2)
    ...: inc_Mean
    ...:
Out[41]: 41698.75

In [42]: inc_SD = round(sample_mean.Mean_MHH_Inc.std(),2)
    ...: inc_SD
    ...:
Out[42]: 4711.41
```

Validating SD & Mean Empirical Rule….contd

```
In [43]: def mean_sd_fun(df, sd, inc_Mean, inc_SD):
    ...:     df = df[(df['Mean_MHH_Inc'] >= inc_Mean - sd * inc_SD) & (df['Mean_MHH_Inc'] <= inc_Mean + sd * inc_SD)]
    ...:     return df
    ...:
    ...: sample_mean_1SD_subset = mean_sd_fun(df = sample_mean, sd = 1, inc_Mean = inc_Mean, inc_SD = inc_SD)
    ...:
    ...: sample_mean_2SD_subset = mean_sd_fun(df = sample_mean, sd = 2, inc_Mean = inc_Mean, inc_SD = inc_SD)
    ...:
    ...: sample_mean_3SD_subset = mean_sd_fun(df = sample_mean, sd = 3, inc_Mean = inc_Mean, inc_SD = inc_SD)
    ...:
    ...: print('Tot_Cnt =', len(sample_mean))
    ...: print('SD1_Cnt =', len(sample_mean_1SD_subset))
    ...: print('SD2_Cnt =', len(sample_mean_2SD_subset))
    ...: print('SD3_Cnt =', len(sample_mean_3SD_subset))
    ...:
Tot_Cnt = 1000
SD1_Cnt = 673
SD2_Cnt = 963
SD3_Cnt = 997
```



Standard Normal Distribution

- Standard Normal Distribution is a Normal Curve with μ = 0 and σ = 1

  The standardized value of a normally distributed random variable is called a *Z* score and is calculated using the following formula

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

x = the value that is being standardized
μ = the mean of the distribution
σ = standard deviation of the distribution

- Why the need of Standardization? Why do we us Z instead of "the Number of Standard Deviations"?
  - Normally Distributed random variable take on many different units of measure: rupees, cms, inches, Kg, minutes.
  - By standardizing, you remove the units as such we do not require separate Normal Distribution table for each variable and one Standardized Distribution Table can be used for any random variable

Example Problem

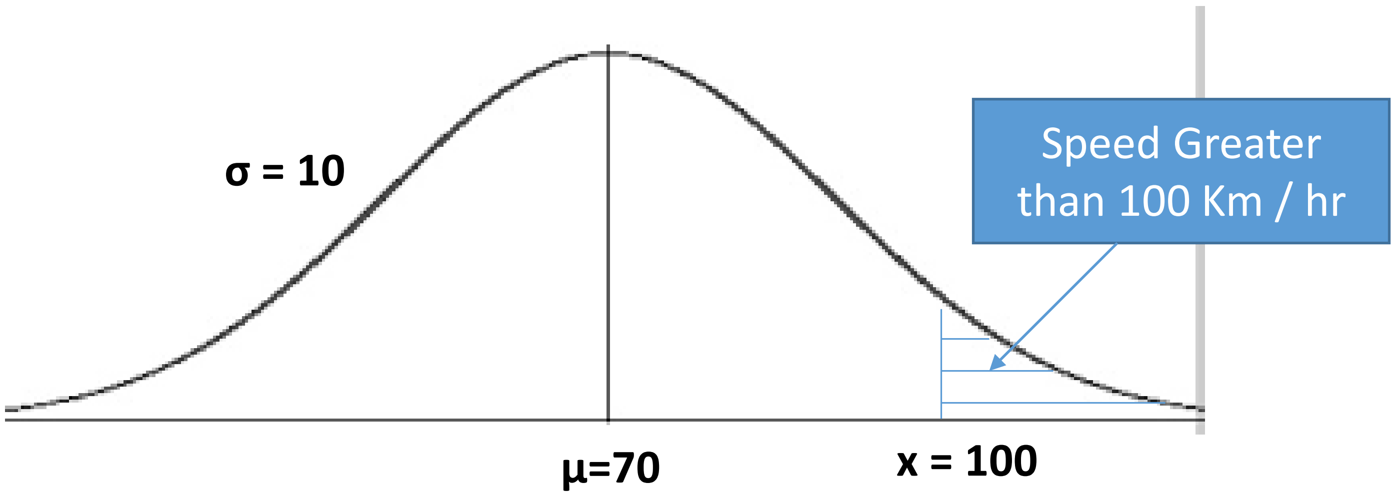
- A radar unit is used to measure speeds of cars on a Mumbai – Pune Highway. The speeds are normally distributed with a mean of 70 km/hr and a standard deviation of 10 km/hr.

A) What is the probability that a car picked at random is travelling at more than 100 km/hr?

B) What percentage of cars would be travelling at a speed less than 80 Km / hr

C) What is the probability that the car speed is between 80 Km / hr and 100 Km / hr

Solution A



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

x = the value that is being standardized
μ = the mean of the distribution
σ = standard deviation of the distribution

Solution:

Z = (100 – 70) / 10 = 3

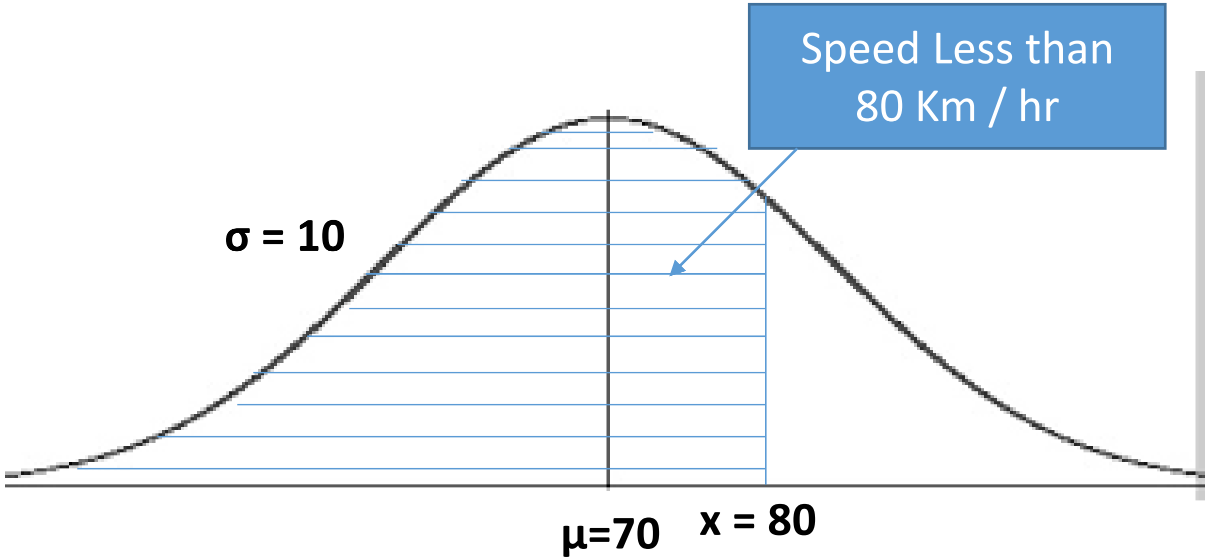
Using Excel function

= Norm.Dist(100, 70, 10, 1) = 0.99865

= Norm.S.Dist(3,1) = 0.99865

Probability of a random car picked speeding at 100 Km / Hr or more will be = 1 – 0.99865 = 0.00135

Solution B



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

x = the value that is being standardized
μ = the mean of the distribution
σ = standard deviation of the distribution

Solution:

Z = (80 – 70) / 10 = 1

Using Excel function

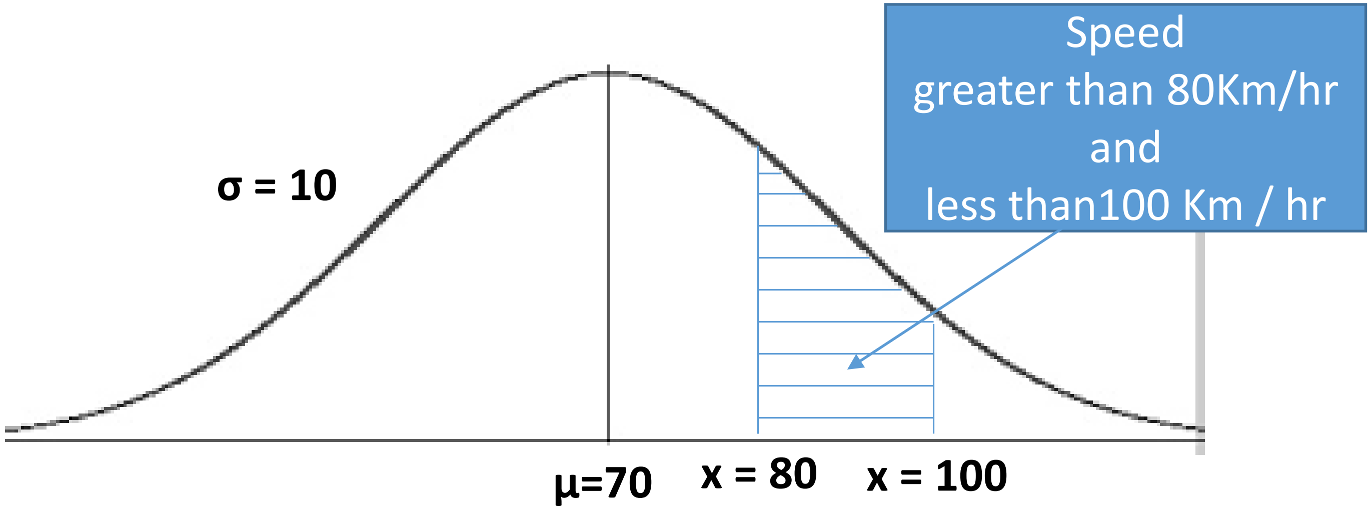
= Norm.Dist(80, 70, 10, 1) = 0.841345

= Norm.S.Dist(1,1) = 0.841345

Percentage of car traveling at speed less than 80 Km / Hr

= 0.841345 = 84.13%

Solution C



Solution:

Z = (100 – 70) / 10 = 3

Z = (80 – 70) / 10 = 1

Using Excel function

Norm.Dist(100, 70, 10, 1) = 0.998650

Norm.Dist(80, 70, 10, 1) = 0.841345

Probability of a random car picked having a speed between 80 Km / Hr and 100 Km / Hr

= 0.998650 – 0.841345 = 0.157305

This means that 15.73% of the cars are traveling between 80 Km / Hr and 100 Km / Hr

Output using Python

```
In [157]: mu = 70
     ...: sigma = 10
     ...:
     ...: def normcdf(x, mu, sigma):
     ...:     t = x-mu;
     ...:     y = 0.5*erfc(-t/(sigma*sqrt(2.0)));
     ...:     return round(1-y,10)
     ...:
     ...: normcdf(x = 100, mu = mu, sigma = sigma)
     ...:
Out[157]: 0.001349898
```

```
In [159]: def normcdf(x, mu, sigma):
     ...:     t = x-mu;
     ...:     y = 0.5*erfc(-t/(sigma*sqrt(2.0)));
     ...:     return round(y,10)
     ...:
     ...: normcdf(x = 80, mu = mu, sigma = sigma)
     ...:
Out[159]: 0.8413447461
```

#Using Z-Score

```
In [162]: round(1 - st.norm.cdf(3),10)
     ...:
     ...:
Out[162]: 0.001349898

In [163]: round(st.norm.cdf(1),7)
     ...:
Out[163]: 0.8413447
```

Thinking Problem

What is the minimum speed of the top 10% of the Fast Drivers on Mumbai – Pune Express Highway?

Thinking Problem

What is the minimum speed of the top 10% of the Fast Drivers on Mumbai – Pune Express Highway?

<u>Solution in Excel</u>
=NORM.INV(0.9,70,10)

<u>Solution in Python</u>
norm.ppf(0.9, loc=70, scale=10)

Additional Gyan | Standardization & Normalization

- Standardization & Normalization are 2 commonly used method for rescaling
- *Normalization*, which scales all numeric variables in the range.
  - One possible formula for scaling in the range [0,1]is given below:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standardization transforms data to mean zero and unit variance

$$x_{new} = \frac{x - \mu}{\sigma}$$

Hypothesis Testing

Dependent & Independent Variables
Hypothesis Creation

Hypothesis Validation

Hypothesis – Dictionary Definition

- hypothesis (noun); hypotheses (plural noun)

  a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

  a proposition made as a basis for reasoning, without any assumption of its truth

What is Statistical Hypothesis?

- Hypothesis is an assumption
- Hypothesis is a conjecture (an opinion)
- Hypothesis may be true or not true… It has to be proven

- A **Statistical Hypothesis** is a **statement** about a **population parameter**. It may or may not be true. You have to ascertain the truth of the hypothesis using Hypothesis Testing

Hypothesis

- Null Hypothesis $\boldsymbol{H_0}$:
  - A null hypothesis is status quo.
  - A general statement or default position that there is no relationship between two measured phenomena or no association among groups

- Alternate Hypothesis $\boldsymbol{H_1}$:
  - The alternative hypothesis is the hypothesis contrary to Null Hypothesis
  - It is usually taken to be that the observations are the result of a real effect

Null & Alternate Hypothesis Examples

Industry	Null Hypothesis	Alternate Hypothesis
Process Industry	Shop Floor Manager in Dairy Company feels that the Milk Packaging Process unit for 1 Litre Packs is working fine and does not need any calibration. SD = 10 ml Null Hypothesis : $\mu = 1$	Alternate Hypothesis : $\mu \neq 1$
Credit Risk	Credit Team of a Bank has been taking lending decisions based on in-house developed Credit Scorecard. Their claim to fame in the organisation is their scorecard has helped reduce NPAs by at least 0.25% Null Hypothesis: Scorecard has not helped in reducing NPA level π (scorecard NPA) - π (No scorecard NPA) = 0.25%	Alternate Hypothesis : π (scorecard NPA) - π (No scorecard NPA) > 0.25%
Motor Industry	An Electric Car manufacturer claims their newly launched eCar gives average mileage of 125 MPGe (Miles per Gasoline Equivalent) Null Hypothesis : $\mu = 125$	Alternate Hypothesis : $\mu < 125$

Type I and Type II Error

Null Hypothesis	True	False
Reject	Type I Error (α)	No Error
Accept	No Error	Type II Error (β)

- I reject the Null Hypothesis when it is True. This is Type I Error
- E.g. A manufacturer’s Quality Control department rejects a lot when it has actually met the market acceptable quality level. This is Producer’s Risk

Type I and Type II Error

Null Hypothesis	True	False
Reject	Type I Error (α)	No Error
Accept	No Error	Type II Error (β)

- I do not reject (Accept) the Null Hypothesis when it is False. This is Type II Error
- E.g. A Consumer accepts a lot when it is actually faulty. This is Conumer's Risk

Type I and Type II Error

Type I Error α probability is called **the Level of Significance** of the test

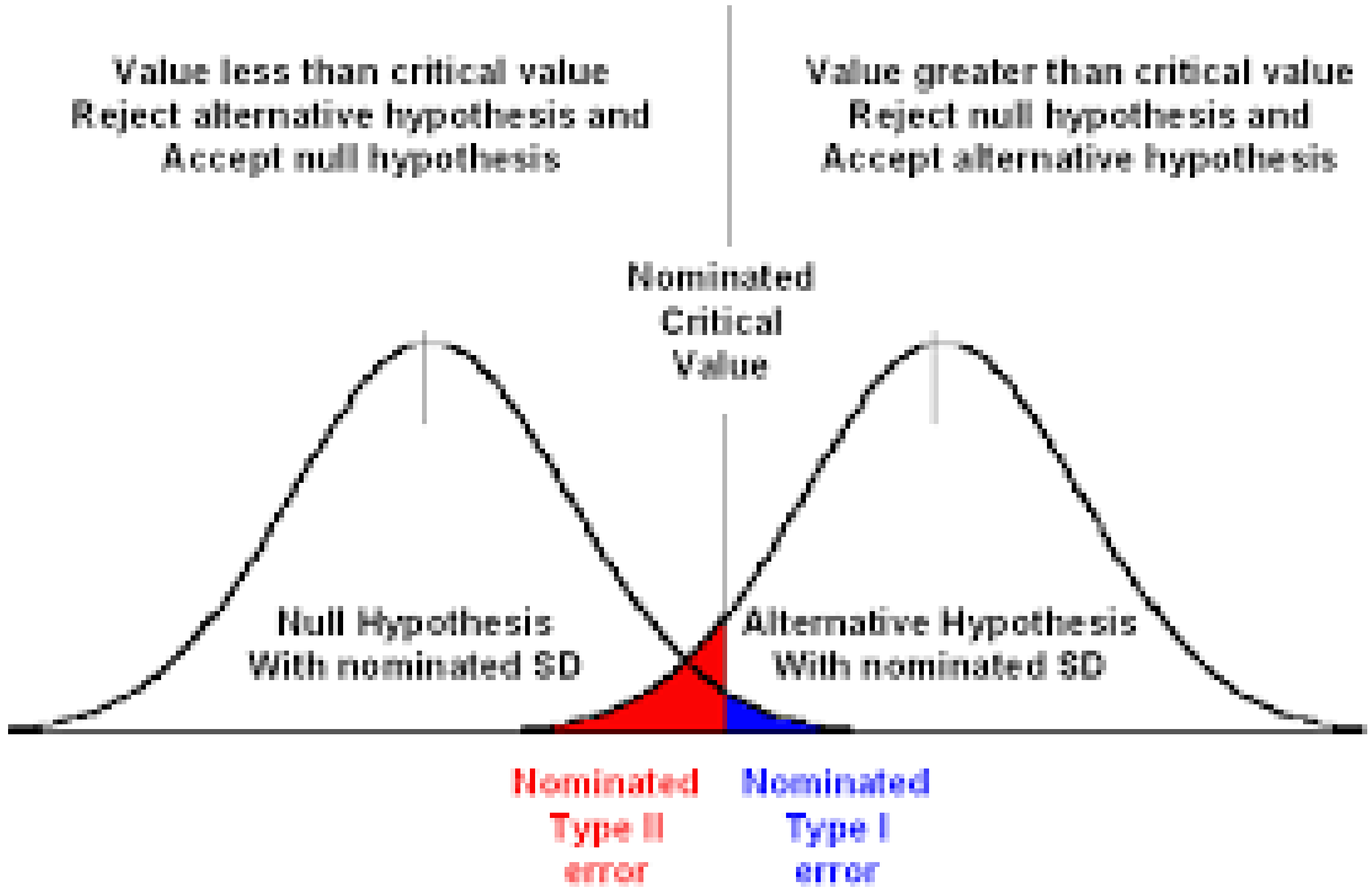
Often, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

Null Hypothesis	True	False
Reject	Type I Error (α)	No Error
Accept	No Error	Type II Error (β)

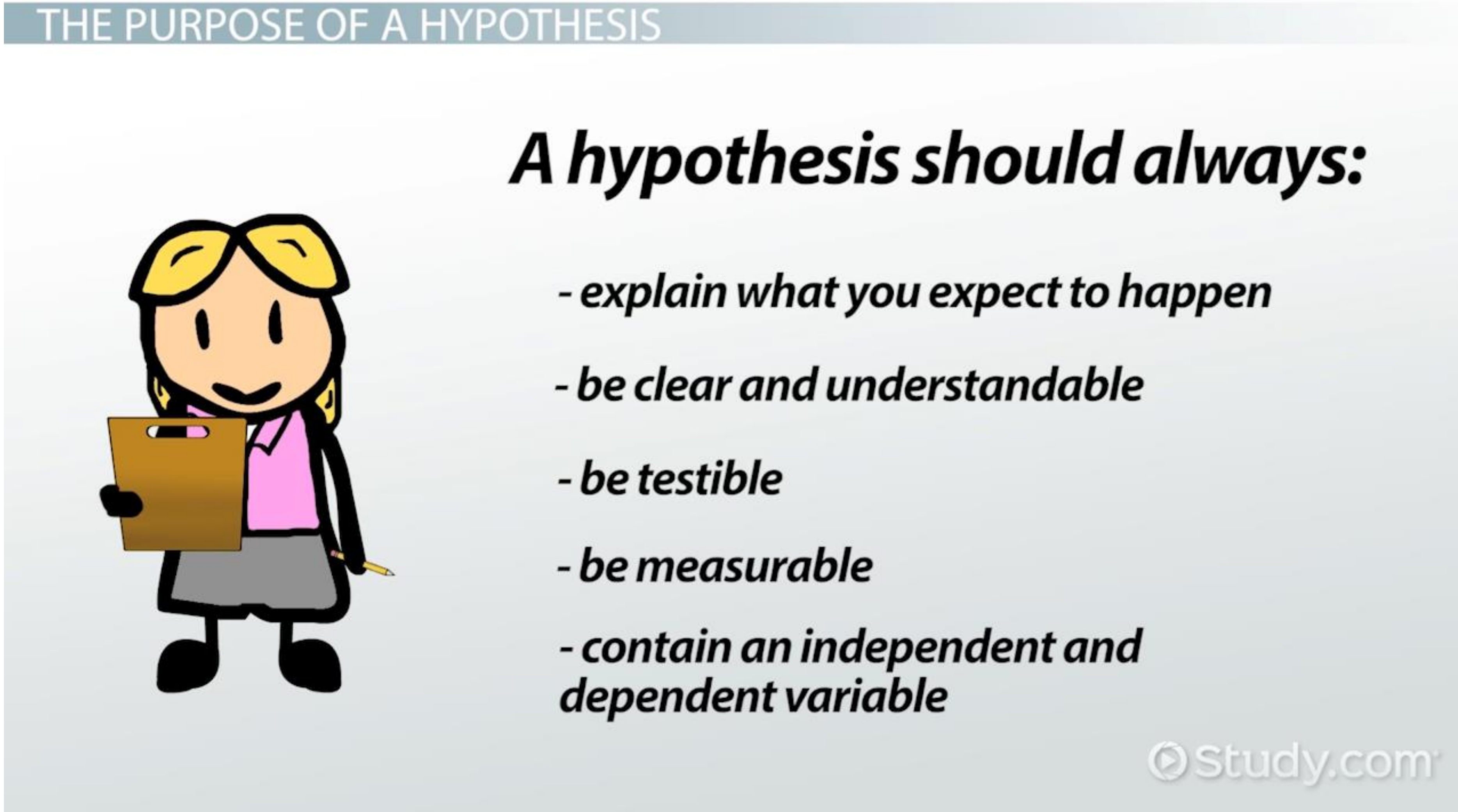
(1 – α) is called the **confidence level** of the test

A **type II error** β occurs when the null hypothesis is false, but erroneously fails to be rejected

(1 – β) is called the **power of a test**

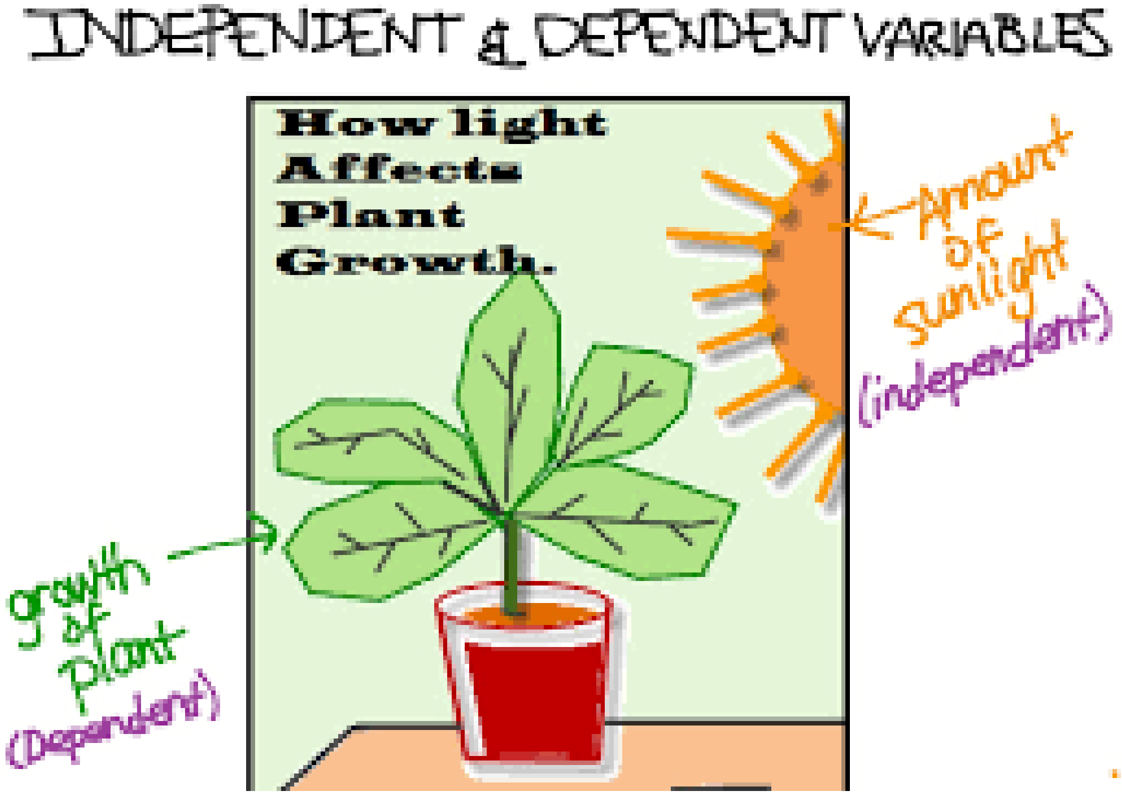

Value less than critical value
Reject alternative hypothesis and
Accept null hypothesis
Value greater than critical value
Reject null hypothesis and
Accept alternative hypothesis
Nominated
Critical
Value
Null Hypothesis
With nominated SD
Alternative Hypothesis
With nominated SD
Nominated
Type II
error
Nominated
Type I
error

Hypothesis Creation



http://study.com/academy/lesson/what-is-a-hypothesis-definition-lesson-quiz.html

Independent and Dependent Variables



- **Dependent Variable**: The variable that depends on other factors

- **Independent Variable**: The variable which is **experimented** in order to observe its effect on the Dependent Variable

- Independent Variable is also often referred as Predictor Variable

http://www.showme.com/sh/?h=9WsGXQG

Hypothesis Testing
Z Test

Launching a Product Line into a New Market Area

- Samy, Product Manager of K2 Jeans, wants to Launch a Product Line into a New Market Area
- Survey of random sample of 400 households in that market showed a mean income per household of ₹30,000. Standard Deviation based on earlier pilot study of households is ₹8,000
- Samy strongly believes the product line will be adequately profitable only in markets where the mean household income is greater than ₹29,000.
- **Samy wants our help in deciding whether the Product Line should be introduced in the New Market? Based on statistical analysis would will be your recommendation**

Practical Significance and Statistical Significance

- Practical Significance – Average income based on sample is ₹30,000 is greater than ₹29,000. Logically I can infer that it would be profitable to introduce the Product Line. **Intuition**

- **Remember… there is something called Sampling Error**

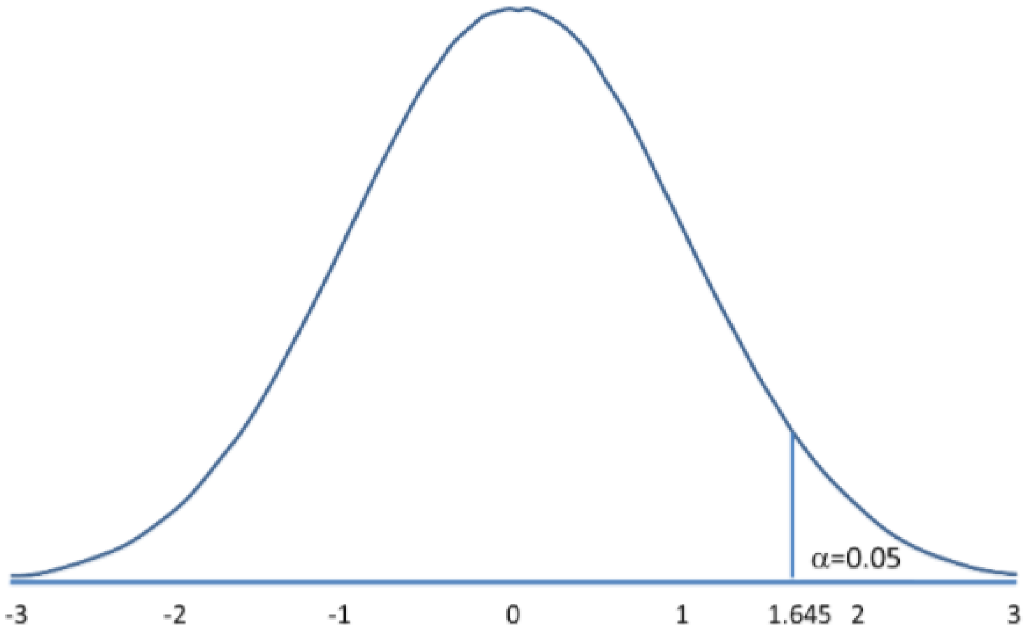
- Statistical Significance – Can I say at 95% confidence interval say that the income in the market is ₹29,000
- ….. Statistical test is to confirm your prima facie observation or intuition

Null and Alternate Hypothesis for Samy's Problem

- Null Hypothesis : Mean Income of Household is ₹29,000
  - $\boldsymbol{H_0}$ **:** $\mu = 29000$

- Alternate Hypothesis : Mean Income of Household is greater than ₹29,000
  - $\boldsymbol{H_1}$ **:** $\mu > 29000$

Upper Tail Test or Right Tail Test

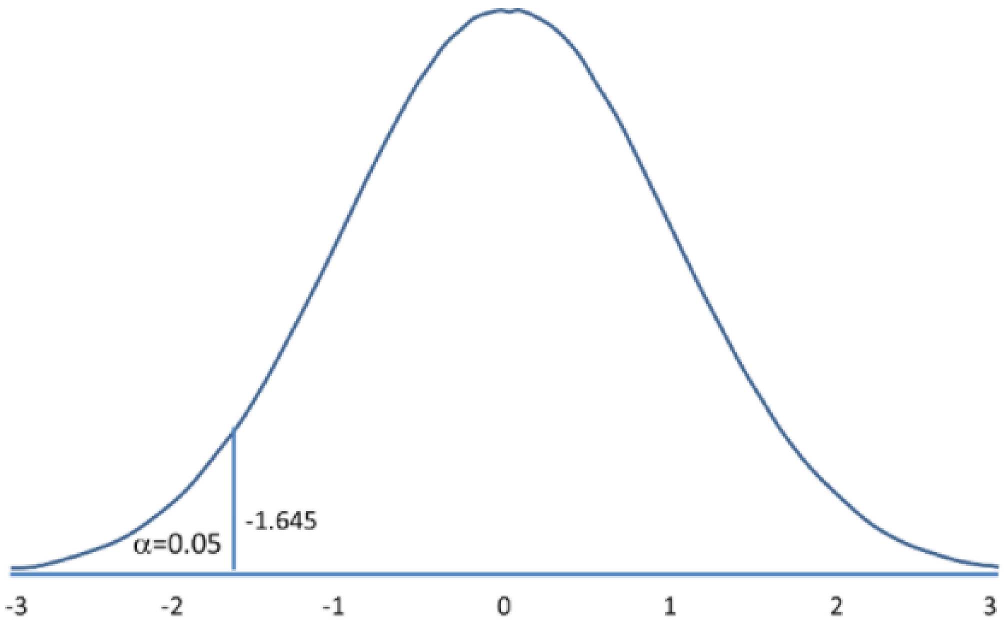


Rejection Region for Upper-Tailed Z Test (H_1: $\mu > \mu_0$) with α=0.05

The decision rule is: Reject H_0 if $Z \geq 1.645$.

Upper-Tailed Test	
α	Z
0.10	1.282
0.05	1.645
0.025	1.960
0.010	2.326
0.005	2.576
0.001	3.090
0.0001	3.719

Lower Tail Test – Left Tail Test

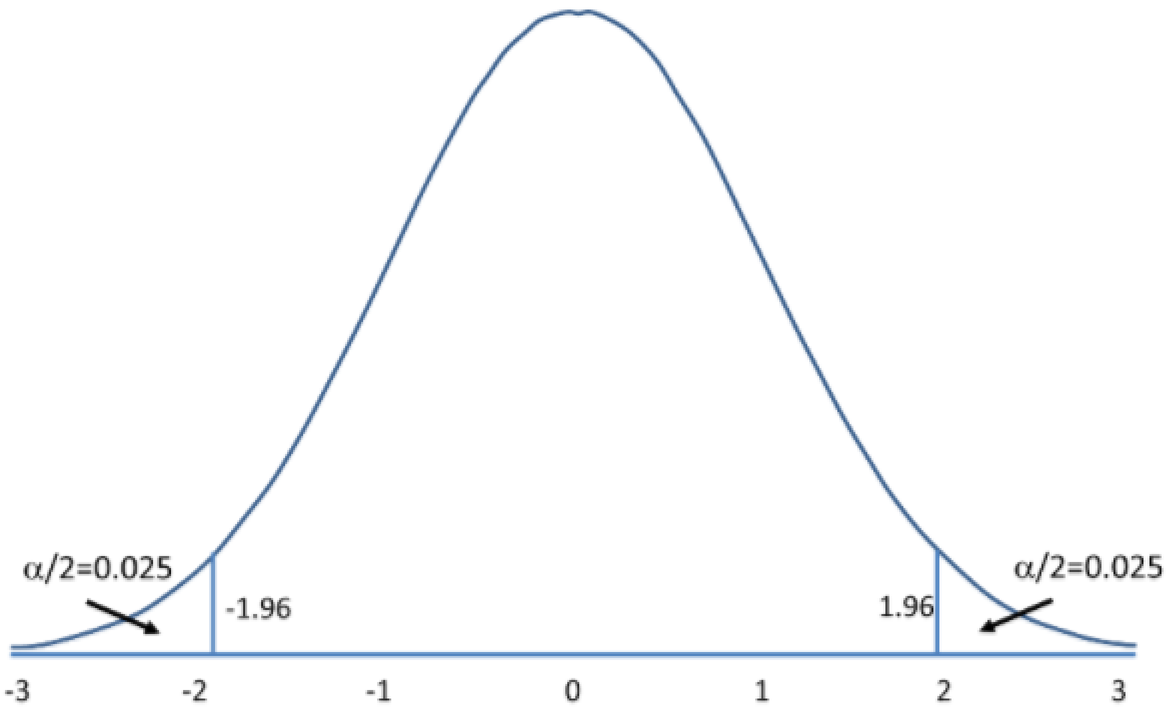


Rejection Region for Lower-Tailed Z Test (H_1: $\mu < \mu_0$) with α =0.05

The decision rule is: Reject H_0 if Z $\leq$-1.645.

Lower-Tailed Test	
a	Z
0.10	-1.282
0.05	-1.645
0.025	-1.960
0.010	-2.326
0.005	-2.576
0.001	-3.090
0.0001	-3.719

Two Tail Test



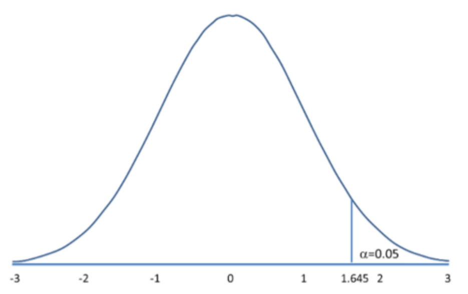
Rejection Region for Two-Tailed Z Test (H_1: $\mu \neq \mu_0$) with α =0.05

The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Two-Tailed Test	
α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

Problem Solution

- $\bar{x} = 30000$
- $\mu = 29000$ (based on null hypothesis)
- $\sigma = 8000$
- $n = 400$



Rejection Region for Upper-Tailed Z Test (H_1: $\mu > \mu_0$) with α=0.05
The decision rule is: Reject H_0 if $Z \geq 1.645$.

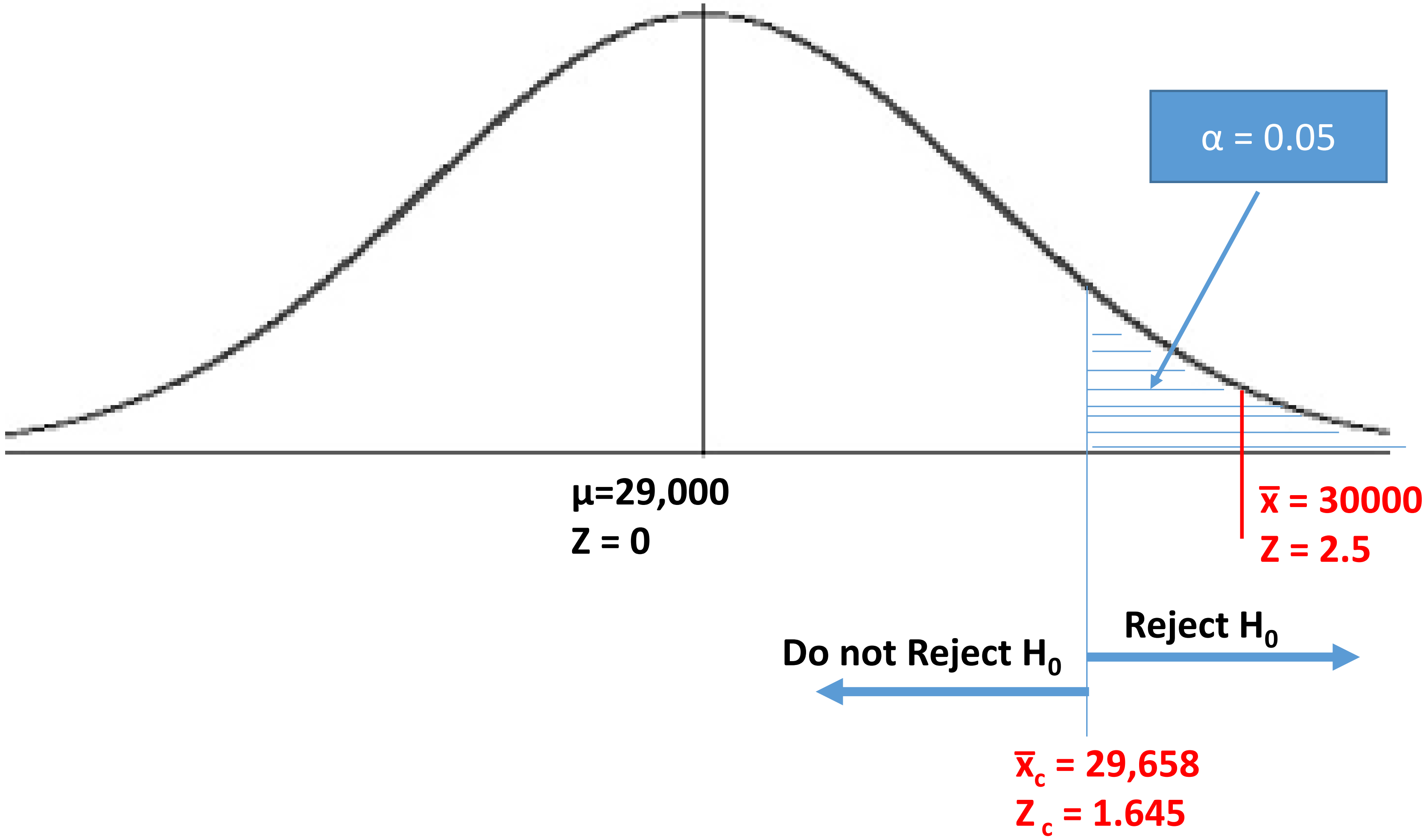
- $Z = \dfrac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \dfrac{(30000 - 29000))}{\frac{8000}{\sqrt{400}}} = 2.5$

```
## using Z Score for Samy's problem
In [164]: round(1-norm.cdf(2.5),9)
     ...:
Out[164]: 0.006209665
```

<u>Interpretation of p value:</u> The risk of rejecting the null hypothesis when it is true is 0.0062; That means at 99.38% confidence level, I can say the mean income is more than 29000

<u>Decision rule:</u> the p-value has to be compared with your desired α. When α is not specified, it is assumed as 0.05. With α = 0.05 and p-value as 0.0062, the Null Hypothesis is overwhelmingly rejected and Alternate Hypothesis may be accepted

Critical value for rejecting the Null Hypothesis



Final Recommendation to Samy – Go ahead and introduce the Product Line in the New Market

Sample Size

Sample Size

- Sample Size : The part of the population selected for analysis or experiment

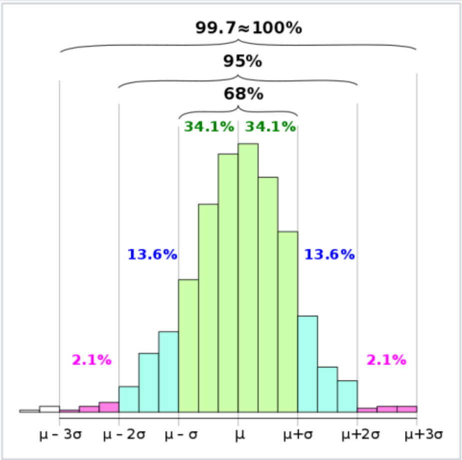
- Sampling Error : Sample Estimate may not be 100% accurate estimate of the population. The difference between the Sample Estimate and Population Estimate is the Sampling Error.

- In other words whenever we take the sample there is an uncertainty of how accurate is the sample estimate with respect to the true population estimate and this uncertainty is **Sampling Error** and is measured in terms of **Confidence Interval**
- The maximum difference between the observed sample mean $\bar{x}$ and the population mean μ is called **Margin of Error**

- The larger the **sample size** lower will be the **Margin of Error**

Samy's Product Line Example … contd

- $\bar{x}$ = 30000
- μ = 29000
- σ = 8000
- n = 400



Sample SD = s = $\frac{\sigma}{\sqrt{n}}$

s = $\frac{8000}{\sqrt{400}}$ = 400

Confidence Interval = $\bar{x}$ ± 2 s = 30000 ± 2 * 400 = (29200, 30800)

What should be my Sample Size if I don't want the Sample Estimate to differ from the Actual by more than 400 at 95% confidence level?

Hypothesis Testing

Summing up the understanding

Hypothesis Testing

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps.

1. Formulate the **null hypothesis** (commonly, that the observations are the result of pure chance) and the **alternate hypothesis** (commonly, that the observations show a real effect combined with a component of chance variation).
2. Identify a **test statistic** that can be used to assess the truth of the **null hypothesis**
3. Compute the **p-value**, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the **null hypothesis** were true. The smaller the p-value, the stronger the evidence against the null hypothesis.
4. Compare the p-value to an acceptable significance value α (sometimes called an **alpha value**). If $p <= \alpha$, then the observed effect is statistically significant, i.e., the null hypothesis is ruled out, and the alternative hypothesis is valid.

Classroom Exercise

- A retailer is weighing strawberries to sell as 250gm punnets. A customer has complained that strawberries he had bought previously weighed under 250gm. The retailer decides to check the weight of 36 punnets. He finds the average weight is 248.5gm with standard deviation of 4.8gm. In using significance test to judge whether he is selling under-weight punnets, which of the following conclusion is correct.

A) At 5% level he is selling under weight

B) At 5% level he is not selling under weight

C) At 5% level the test is inconclusive

D) A significance test is

Home Exercise

- In a Professional Examination conducted globally the marks follow a Normal Distribution. 10% of the candidates got distinction marks (i.e. >=85%), 22% of the candidates failed in the examination (<37%)

- Find out the mean and standard deviation of the marks

Student's t Test

Student’s t Test

- **William Sealy Gosset** (13 June 1876 – 16 October 1937) was an English statistician.
- He published under the pen name **Student**, and developed the Student's t-distribution

- Student’s t Test Application
  - Specifically useful for small samples; can be useful for large samples also
  - You do not know the population variance
  - Population is normally distributed

Types of t Test

- One sample t test
  - **one-sample t-test** is used to compare the mean of a population to a specified **theoretical mean** μ

- Unpaired two sample t test (Independent t-test)
  - **Independent (or unpaired two sample) t-test** is used to compare the means of two unrelated groups of samples.

- Paired t test
  - **Paired Student's t-test** is used to compare the means of two related samples. That is when you have two values (pair of values) for the same samples.

t Test Formula

- One sample t test

$$t = \frac{\bar{X} - X_0}{\frac{s_X}{\sqrt{n}}}$$

$$\mathrm{df} = n - 1$$

- Paired t test

$$t = \frac{m}{s/\sqrt{n}}$$

m and **s** are the **mean** and the **standard deviation** of the difference (d : d represents the difference between each pair), respectively. **n** is the size of d

- Independent t test

$$t = \frac{\mu_A - \mu_B}{\sqrt{\left[\frac{\left(\Sigma A^2 - \frac{(\Sigma A)^2}{n_A}\right) + \left(\Sigma B^2 - \frac{(\Sigma B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

$(\Sigma A)^2$: Sum of data set A, squared

$(\Sigma B)^2$: Sum of data set B, squared

μ_A: Mean of data set A

μ_B: Mean of data set B

ΣA^2: Sum of the squares of data set A

ΣB^2: Sum of the squares of data set B

n^A: Number of items in data set A

n^B: Number of items in data set B

http://www.statisticshowto.com/independent-samples-t-test/

t Test Application One Sample

- Experience Marketing Services reported that the typical American spends a mean of 144 minutes (2.4 hours) per day accessing the Internet via a mobile deice. (Source: The 2014 Digital Marketer, available at ex.pn/1kXJifX.) In order to test the validity of this statement, you select a sample of 30 friends and family. The result for the time spent per day accessing the Internet via mobile device (in minutes) are stored in **Internet_Mobile_Time.csv** file.

- 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 test in A?

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

Independent t-Test Two 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 each from Wing A and Wing B of the hotel. The data collated is given in **Luggage.csv** file. Analyze the data and determine whether there is difference in the mean delivery times in the two wings of the hotel. (use alpha = 0.05).

- Problem 10.83 from the Textbook adapted for Classroom Discussion (Chapter 10 – page 387)

Is Random Sample representative of Population???

- Many people do not **believe** in the concept of random sampling….
- Despite having learnt Central Limit Theorem
- … let us validate whether **sample mean** is close to **population mean** at **95% confidence level**

- Null Hypothesis : Sample Mean = Population Mean
- Alternative Hypothesis: Sample Mean ≠ Population Mean

… creating a sample from population

```
In [55]: df = pd.read_csv("hypothesis_test.csv")
    ...: popln = df[["Age", "Balance", "No_OF_CR_TXNS", "SCR"]]
    ...: popln['random'] = np.random.random(len(popln))
    ...: sample_dst = popln[popln['random'] <= 0.1]
    ...:

In [56]: len(popln)
Out[56]: 20000

In [57]: len(sample_dst)
    ...:
Out[57]: 1985
```

Index	Age	Balance	No_OF_CR_TXNS	SCR	random
13	42	2.81e+03	34	950	0.087
33	28	6.42e+03	32	130	0.082
38	29	1.69e+04	14	839	0.00549
40	24	1.84e+05	0	266	0.0906
66	32	1.11e+05	13	792	0.0192
73	55	3.88e+04	41	572	0.0235
76	34	9.27e+03	0	255	0.0169

Applying 2 tail test…

```
In [67]: one_sample = stats.ttest_1samp(sample_dst.Age, popln.Age.mean())
    ...: print("\n The t-statistic is %.3f and the p-value is %.3f." % one_sample)
    ...:
    ...: print("\n Sample Mean :", round(sample_dst.Age.mean(),4),",",
    ...:       " Population Mean :",round(popln.Age.mean(),4))
    ...:

 The t-statistic is -0.457 and the p-value is 0.648.

 Sample Mean : 38.2993 ,  Population Mean : 38.3962
```

```
In [68]: one_sample = stats.ttest_1samp(sample_dst.Balance, popln.Balance.mean())
    ...: print("\n The t-statistic is %.3f and the p-value is %.3f." % one_sample)
    ...:
    ...: print("\n Sample Mean :", round(sample_dst.Balance.mean(),4),",",
    ...:       " Population Mean :",round(popln.Balance.mean(),4))
    ...:

 The t-statistic is 1.235 and the p-value is 0.217.

 Sample Mean : 151025.7395 ,  Population Mean : 146181.3056
```

...contd

```
In [69]: one_sample = stats.ttest_1samp(sample_dst.No_OF_CR_TXNS,
popln.No_OF_CR_TXNS.mean())
    ...: print("The t-statistic is %.3f and the p-value is %.3f." % one_sample)
    ...:
    ...: print("\n Sample Mean :", round(sample_dst.No_OF_CR_TXNS.mean(),4),",",
    ...:       " Population Mean :",round(popln.No_OF_CR_TXNS.mean(),4))
    ...:
The t-statistic is 0.533 and the p-value is 0.594.

 Sample Mean : 16.81 ,  Population Mean : 16.6531
```

```
In [70]: one_sample = stats.ttest_1samp(sample_dst.SCR, popln.SCR.mean())
    ...: print("The t-statistic is %.3f and the p-value is %.3f." % one_sample)
    ...:
    ...: print("\n Sample Mean :", round(sample_dst.SCR.mean(),4),",",
    ...:       " Population Mean :",round(popln.SCR.mean(),4))
    ...:
The t-statistic is -0.855 and the p-value is 0.393.

 Sample Mean : 552.1855 ,  Population Mean : 557.136
```

Classroom Exercise

- Perform a t Test that the Average Miles clocked by Men is significantly different from Women.

- File to be used for analysis is : **“CardioGoodFitness.csv”**

- **What is the Null Hypothesis here?**
- **What is the Alternate Hypothesis?**

Paired t Test

- The file Concrete.csv contains the compressive strength, in thousands of pounds per square inch (psi), of 40 samples of concrete taken two and seven days after pouring. (Data extracted from O. Carrillo-Gamboa and R. F. Gunst, “Measurement – Error – Model Collinearities”, Technometrics, 34 (1992): 454 – 464.)

- At the 0.01 level of significance, is there evidence that the means strength is lower at two days than at seven days?

- Problem 10.26 from the Textbook adapted for Classroom Discussion (Chapter 10 – page 353)

Paired t Test for Concrete…

```
In [86]: two_sample = stats.ttest_rel(concrete_dst.SevenDays, concrete_dst.TwoDays)
    ...: print("\n The t-statistic is %.3f and the p-value is %.3f." % two_sample)
    ...:

 The t-statistic is 9.372 and the p-value is 0.000.

In [86]:

In [87]: round(concrete_dst['SevenDays'].mean(),3)
Out[87]: 3.544

In [88]: round(concrete_dst['TwoDays'].mean(),4)
    ...:
Out[88]: 2.991
```

CHI-SQ Test

Chi-Squared Test

- CHI-SQ test is a Test of Independence; The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables

- The data between the two categorical variables can be represented in Cross-table / Contingency table format

Y / X	x_1	x_l	x_L	Σ
y_1		$\vdots$		
y_k	$\cdots$	n_{kl}	$\cdots$	$n_{k.}$
		$\vdots$		
y_K				
Σ		$n_{.l}$		n

- A **chi-squared test**, also written as ***χ*² test** is then computed as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = the frequencies observed

E = the frequencies expected

$\sum$ *= the 'sum of'*

Chi-Squared Test Hypothesis

- Null Hypothesis:
  - There is no relationship between the two nominal (categorical) variables
  - There is no significant difference between the observed and expected frequencies

- Alternate Hypothesis:
  - There is a relationship between the nominal variables
  - There is significant difference between the observed and expected frequencies

- Note: Chi-Sq test is a Non-Parmeteric test

Chi-Sq Application

- A company is considering organisational change involving the use of self-managed work teams. To assess the attitudes of employees of the company toward this change, a sample of 400 employees is selected and asked whether they favour the institution of self-managed work teams in the organization. Three responses are permitted: favour, neutral or oppose. The results of the survey, cross-classified by type of job and attitude toward self-managed work teams, are summarized as follows:

	Self Managed Work Teams			
Type of Job	**Favour**	**Neutral**	**Oppose**	**Total**
Hourly Worker	**108**	**46**	**71**	**225**
Supervisor	**18**	**12**	**30**	**60**
Middle Management	**35**	**14**	**26**	**75**
Upper Management	**24**	**7**	**9**	**40**
Total	**185**	**79**	**136**	**400**

At the 0.05 level of significance, is there evidence of a relationship between attitude toward self-managed work teams and type of job?

Problem 11.32, Chapter 11, Pages 424-425 of the Textbook

...solution

```
In [100]: sm_wt = pd.read_pickle("Self_Managed_Work_Teams.pkl")
     ...: sm_wt
     ...:
Out[100]:
                   favour  neutral  oppose
Hourly Worker         108       46      71
Supervisor             18       12      30
Middle Management      35       14      26
Upper Managment        24        7       9

In [101]: from scipy.stats import chi2_contingency
     ...: chi2, p, dof, expected = chi2_contingency(sm_wt)
     ...: print("\n\t Pearson's Chi-Squared test \n",
     ...:       "Chi2 :",round(chi2,4),",","dof :",dof,",",
     ...:       " P Value :",round(p,4))
     ...:

         Pearson's Chi-Squared test
 Chi2 : 11.8953 , dof : 6 ,  P Value : 0.0643

In [102]: expected = pd.DataFrame(expected, columns = ['favour','neutral','oppose'],
     ...:                   index = ["Hourly Worker", "Supervisor",
     ...:                            "Middle Management", "Upper Managment"])
     ...: expected
     ...:
Out[102]:
                     favour  neutral  oppose
Hourly Worker      104.0625  44.4375    76.5
Supervisor          27.7500  11.8500    20.4
Middle Management   34.6875  14.8125    25.5
Upper Managment     18.5000   7.9000    13.6
```

Classroom Exercise

- Perform a t Test that the Average Miles clocked by Men is significantly different from Women.

- File to be used for analysis is : **"CardioGoodFitness.csv"**

- **What is the Null Hypothesis here?**
- **What is the Alternate Hypothesis?**

Classroom Exercise

- In Hypothesis_test.csv file assume the “Target” column is capturing the response of the customer to a Marketing Offer. Using Chi-Sq test tell whether there is any relationship between Occupation and Response of the customers

- **The crosstab between the two columns is given below**

```
In [103]: pd.crosstab(df.Occupation, df.Target,margins =True)
     ...:
Out[103]:
Target          0     1    All
Occupation
PROF         5028   435   5463
SAL          5426   413   5839
SELF-EMP     2858   508   3366
SENP         4955   377   5332
All         18267  1733  20000
```

Classroom Exercise

- Assume you have built an **Artificial Intelligence Model** to predict the occurrence of Fraud. The frequency table of Observed Fraud and Model Classified Fraud is given below.
- Null Hypothesis: There is no difference between occurrence of Observed Fraud and Model Predicted Fraud

	Actual Observed		Model Classified	
Model Risk Level	Fraud	No Fraud	Fraud	No Fraud
Very High	25	500	22	503
High	18	600	22	596
Moderate	10	600	8	602
Low	4	1000	8	996
Very Low	3	2000	0	2003
Total	60	4700	60	4700

At 99% confidence level will you recommend to use the Artificial Intelligence Model to predict the occurrence of fraud?

K2 Analytics
Building Skills, Building Individuals

Thank you

Contact us:
ar.jakhotia@k2analytics.co.in