**SLIDE-1**

**Data Science using R, Minitab & XL-Miner**

R, Minitab

XL-Miner for Forecasting

****

**SLIDE-4**

**AGENDA**

Statistics Analysis

Data Mining – Supervised & Unsupervised (Machine Learning)

Text Mining & NLP

Data Visualization using Tableau

Forecasting

**SLIDE-5**

**What does it take to be a DATA SCIENTIST?**

>All Agenda Topics

>Domain Knowledge

>Practice

* Forecasting
* *Data Mining*
* *Forecasting*
* *Data Visualization*

Successful Data Scientist

**SLIDE-6**

Welcome to the Information Age …

Drowning in data and starving for Knowledge

**Data in every domain…**

* **Web** – web content, link structure, search clicks…
* **Retail** – customer details, point of sales, inventory, …
* **Census** – demographics, population indicators
* **Medical** –literature, biological data, diagnosis, drug trials,…
* **Science** – literature, scientific measurements,…
* **Remote Sensing** – optical, infrared, hyper-spectral

**SLIDE-7**

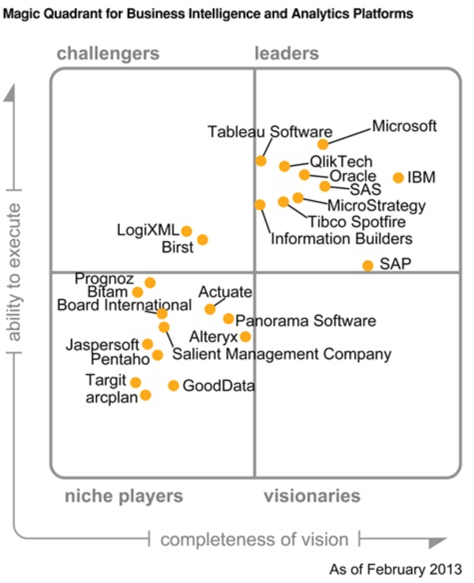
How BIG is Big Data?

* **TWITTER** 🡪 175 million tweets every day, 465 million accounts. (Early 2012)
* **YOUTUBE** 🡪 YouTube users upload 100 hours of video every minute
* **WWW** 🡪 800 new websites are created every minute
* **FACEBOOK** 🡪 100 terabytes of data uploaded daily 30 Billion Pieces of content shared every month 30+ Petabytes of user data
* **GOOGLE** 🡪Processing 20 petabytes a day (2008)
* **WALMART 🡪** more than 1 million customer transactions every hour

**SLIDE-8**

**TUCKMAN’S MODEL**

**Why Tableau?**

****

**SLIDE-9**

****

**SLIDE-10**

****

**SLIDE-11**

****

****

**SLIDE-12**

**Agenda – Basic Statistics**

1. Data Types – Continuous, Discrete, Nominal, Ordinal, Interval, Ratio, Random Variable, Probability, Probability Distribution
2. First, second, third & fourth moment business decisions
3. Graphical representation – Bar plot, Histogram, Boxplot, Scatter diagram
4. simple Linear Regression
5. Hypothesis Testing

**SLIDE-13**

**Data types:**

1. Continuous 2) Discrete

**SLIDE-14**

Data types: Preliminaries

Normal: Merely labels, no further information can be gleaned.

Ex: “coke” and “Pepsi”

Ordinal: Conveys only up to preference information. Direction alone.

Ex: “I prefer coffee to tea”

Interval: Conveys relative magnitude information, in addition to preference.

Ex: “I rate coke a 7 and Pepsi a 4 on a scale of 10.

Ratio: Conveys information on an absolute scale.

Ex: “I paid Rs11 for coke and Rs13 for Pepsi”.



**SLIDE-15**

**Random Variable**

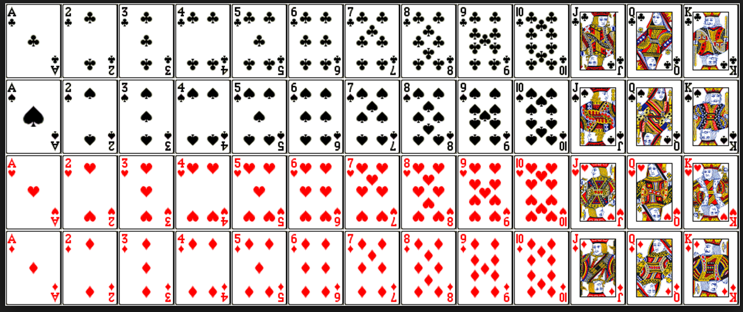
A random variable describes the probabilities for an uncertain future numerical outcome of a random process.

It is variable because it can take one of several possibilities.

It is a random because there is some chance associated with each possible value.

**SLIDE-16**

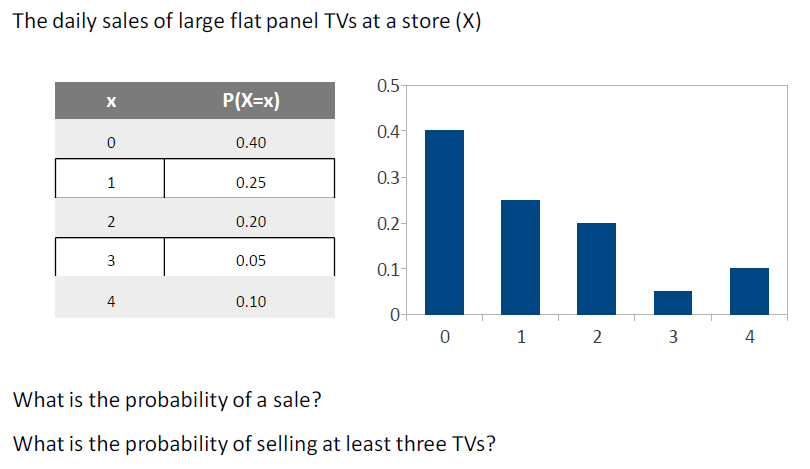
Poker cards example:



Suppose you have randomly picked a card from the card deck. What is the probability that this card will be?

* Bigger than 10?
* Equal to or Bigger than 10?
* Smaller than 3
* Greater than 4 and less than 8

**SLIDE-17**



What is probability of sale?

What is the probability of selling at least 3 tv’s?

**SLIDE-18**

**Sampling Funnel:**

1. Population 2) Sampling frame 3) SRS 4) Sample

**SLIDE-19**

**Measures of central tendency**

**First moment Business decision**:

Population –Mean or Average (µ) = (∑ (xi))/N

Sample-Mean or Average () = (∑ (xi))/n

Median - Middle value of the data

Mode - Most occurring value in the data

**SLIDE-20**

**Measures of Dispersion**

**Second moment Business decision:**

Range= Max-Min

Population variance = σ2= (∑(X-µ)2)/N

Population standard deviation =sqrt ((∑ (xi-population mean) 2)/N)

Sample variance

(∑(x-) 2)/ (n-1)

Sample standard deviation = sqrt ((∑ (xi-sample mean) 2)/(n-1))

**SLIDE-21**

**Expected Value**

For a probability distribution, the mean of the distribution is known as the expected value

The expected value intuitively refers to what one would find if they repeated the experiment an infinite number of times and took the average of all of the outcomes

Mathematically, it is calculated as the weighted average of each possible value

The formula for calculating the expected value for a discrete random variable X, denoted by μ, is:

∑ Xp(X)

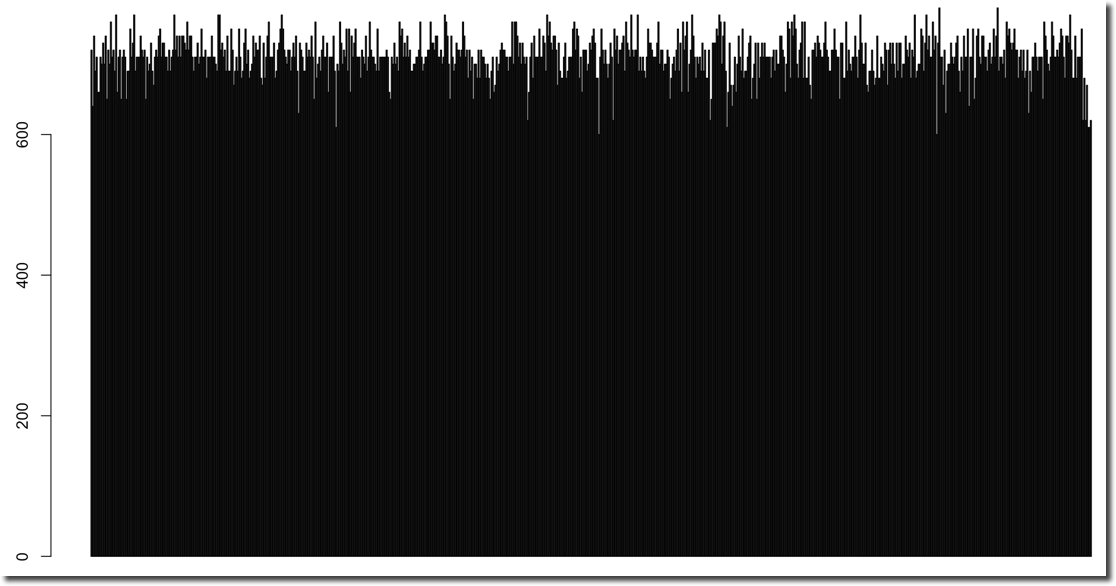
The variance of a discrete random variable X, denoted by σ2 is

σ2 = ∑ [(x-µ/σ)] 2 = ∑ (x- µ)2p(x)

**SLIDE-22**

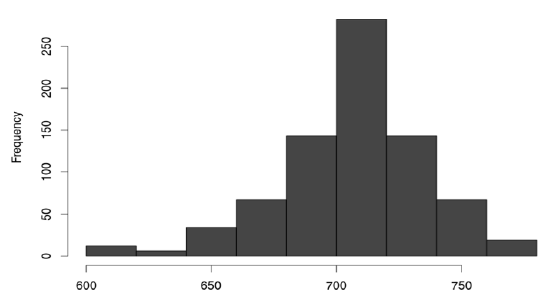
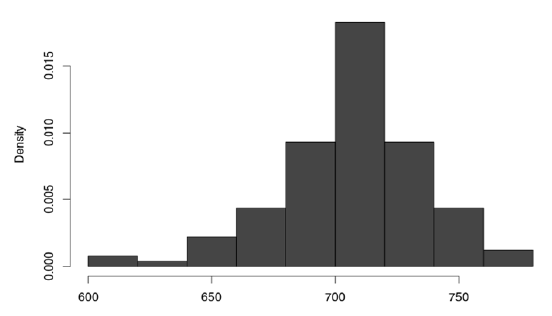
**Graphical techniques:**

1. Bar plot : plotting each point in bar shape



**SLIDE-23**

Histogram: Represents frequency distribution of data, how many observations of take the value within certain interval.

**SLIDE-24**

Third Business Moment: Skewness

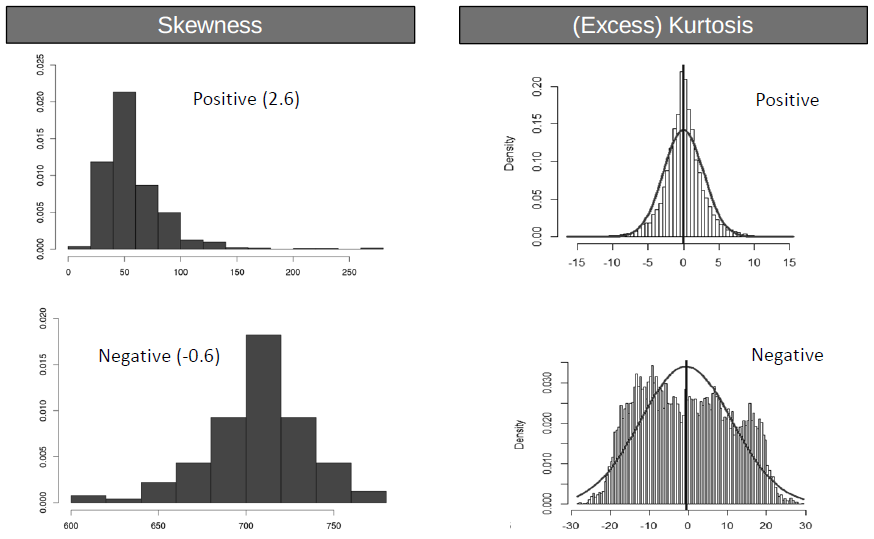
4rth Business Moment: Kurtosis

***Skewness***

* A measure of asymmetry in the distribution
* Mathematically it is given by: E [(x-µ/σ)] 3
* Negative skewness implies mass of the Distribution is concentrated on the Right

***Kurtosis***

* A measure of the “Peakedness” of the distribution
* Mathematically it is given by E[(x-µ/σ)]4 -3
* For Symmetric distributions, negative Kurtosis implies wider peak and thinner tails

****

**SLIDE-25**

**Boxplot:**

****

* Range (IQR): The middle half of a data set falls within the inter- quartile range. – Inter Quartile Range.
* **Box Plot:** This graph shows the distribution of data by dividing the data into four groups with the same number of data points in each group. The box contains the middle 50% of the data points and each of the two whiskers contain 25% of the data points. It displays two common measures of the variability or spread in a data set
* **Range:** It is represented on a box plot by the distance between the smallest value and the largest value, including any outliers. If you ignore outliers, the range is illustrated by the distance between the opposite ends of the whiskers

**SLIDE-26**

**Normal Distribution**

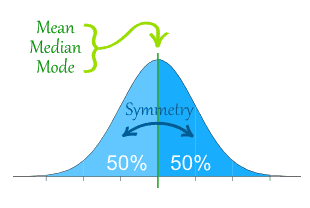
The normal random variable takes values from -∞ to +∞

The Probability associated with any single value of a random variable is always zero

Area under the entire curve is always equal to 1.

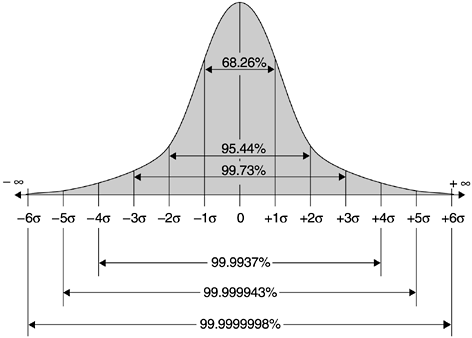
**SLIDE-27**

**. Characterized by bell shaped**

****

**Properties:**

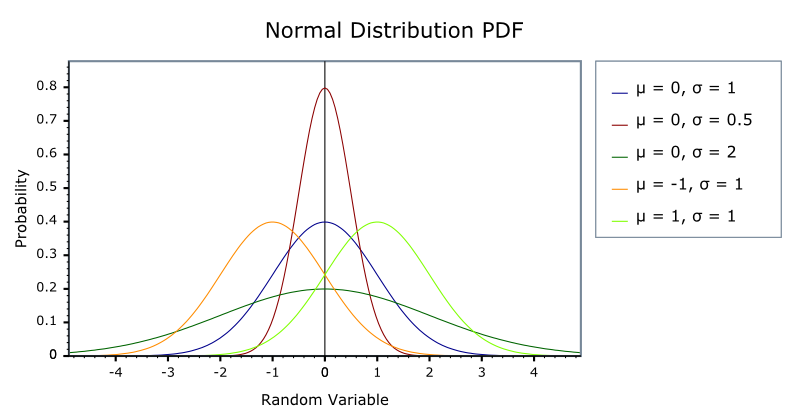
* 68.26% of values lie within ±1 σ from the mean
* 95.46% of the values lie within ±2 σ from the mean
* 99.73% of the values lie within ± 3σ from the mean



**SLIDE-28**

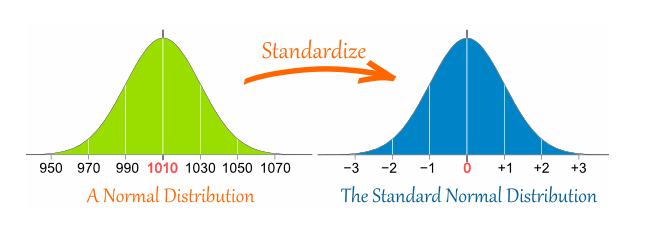
X~N(µ,σ)

**Characterized by mean, µ, and standard deviation, σ**



**SLIDE-29**

**Z scores, Standard Normal Distribution:**

****

* For every value (x) of the random variable X, we can calculate Z score **:** Z = (X-µ)/ σ
* Interpretation − How many standard deviations away is the value from the mean?

**SLIDE-30**

**Calculating Probability from Z distribution**

Suppose GMAT scores can be reasonably modelled using a normal distribution

− µ = 711 σ = 29



What is p(x ≤ 680)?

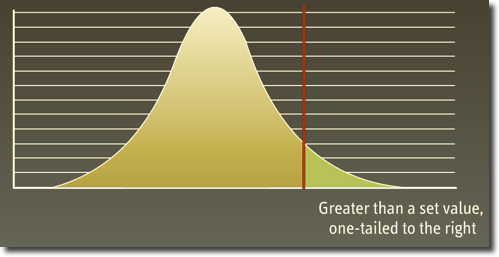
Step 1: Calculate Z score corresponding to 680

Z = (680-711)/29 = -1.06

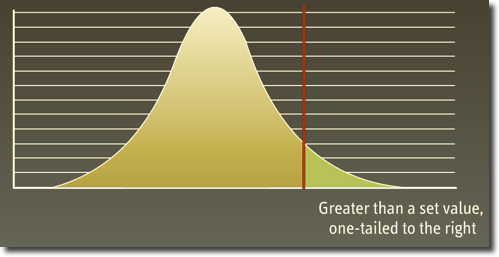
Step 2: Calculate the probabilities using Z – Tables

- P (Z ≤ -1) = 0.14





**SLIDE-31**

* What is P (697≤ X ≤ 740)?
* Step 1 : Use P(x1 ≤ X ≤ x2) = Use P( X ≤ x2) − P( X ≤ x1)
* 
* Step 2 : Calculate P( X ≤ x2) and P( X ≤ x1) as before

P(X ≤ 740) = P (Z ≤ 1) = 0.84; P(X ≤ 697) = P ( Z ≤ - 0.5) = 0.31

* Step 3 : Calculate P( 697 ≤ X ≤ 740 ) = 0.84 – 0.31 = 0.53

**SLIDE-32**

**Normal Quantile plot (Q-Q plot):**



To check whether the data is normally distributed

If plot is straight line (do not have to be absolute straight line) then we say data is normally distributed

If not then they are not normally distributed.

X-axis ->theoretical Quantiles

Y-axis ->Sample Quantiles

**SLIDE-33**

**Sampling variation**

* Sample mean varies from one sample to another.
* Sample mean can be (and most likely is) different from the population mean.
* Sample mean is a random variable.



**SLIDE-34**

**Central Limit Theorem**

The Distribution of the sample mean

- will be normal when the distribution of data in the population is normal

- will be approximately normal even if the distribution of data in the population is not normal if the “sample size” is fairly large

Mean (X) = µ (the same as the population mean of the raw data)

Standard Deviation (X) = σ /, where σ is the population standard deviation and n is the sample size

- This is referred to as standard error of mean.

The standard error of the mean estimates the variability between samples whereas the standard deviation measures the variability within a single sample.

**SLIDE-35**

**Sample Size Calculation**

A Sample Size of 30 is considered large enough, but that may /may not be adequate

More Precise conditions

- n > 10( K3 )2 , where ( K3 ) is sample skewness and

- n > 10( K4 ) , where ( K4) is sample kurtosis

**SLIDE-36**

**Confidence Interval**

* What is the Probability of tomorrow’s temperature being 42 degrees?
* Probability is ‘0’
* Can it be between [-50⁰C & 100⁰C]?

**SLIDE-37**

**Case Study: Confidence Interval**

* A University with 100,000 alumni is thinking of offering a new affinity credit card to its alumni.
* Profitability of the card depends on the average balance maintained by the card holders.
* A Market research campaign is launched, in which about 140 alumni accept the card in a pilot launch.
* Average balance maintained by these is $1990 and the standard deviation is $2833. Assume that the population standard deviation is $2500 from previous launches.
* What we can say about the average balance that will be held after a full−fledged market launch?

**SLIDE-38**

**Interval estimates of parameters**

* Based on sample data

− The point estimate for mean balance = $1990

− Can we trust this estimate?

* What do you think will happen if we took another random sample of 140 alumni?
* Because of this uncertainty, we prefer to provide the estimate as an interval (range) and associate a level of confidence with it
* Interval Estimate = Point Estimate ± Margin of Error

**SLIDE-39**

**Confidence Interval for the Population Mean**

Start by choosing a confidence level (1-α) % (e.g. 95%, 99%, 90%)

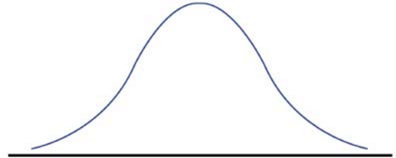
Then, the population mean will be with in

X ± Z1-ᾳ σ/ where Z1-ᾳ satisfies p (-Z1-ᾳ ≤ Z ≤ Z1-ᾳ) = 1-ᾳ

Margin of error depends on the underlying uncertainty, confidence level and sample size.

**SLIDE-40**

Calculate Z value - 90%, 95% & 99%



**SLIDE-41**

**Confidence Interval Calculation**

* Based on the survey and past data
* − n = 140; σ = $2500; X = $ 1990

σ = σ/ = 2500/ = 211.29

* Construct a 95% confidence interval for the mean card balance and interpret it?
* Construct a 90% confidence interval for the mean card balance and interpret it?

**SLIDE-42**

Confidence Interval Interpretation

Consider the 95% Confidence interval for the mean income: [$1576, $2404]

Does this mean that?

- The mean balance of the population lies in the range?

- The mean balance is in this range 95% of the time?

- 95% of the alumni have balance in this range?

Interpretation 1 : Mean of the population has a 95% chance of being in this range for a random sample

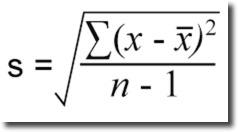
Interpretation 2 : Mean of the population will be in this range for 95% of the random samples

SLIDE-43

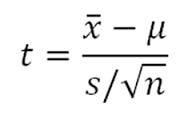
What if we don’t know Sigma?

* Suppose that the alumni of this university are very different and hence population standard deviation from previous launches cannot be used

We replace ***σ*** with our best guess (point estimate) **s**, which is the standard deviation of the sample:



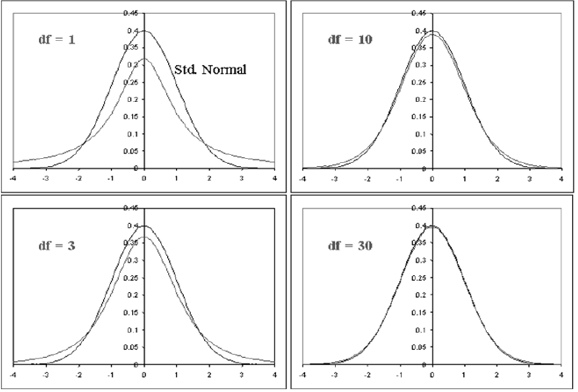
Calculate:



* If the underlying population is normally distributed , T is a random variable distributed according to a t-distribution with n-1 degrees of freedom Tn-1
* Research has shown that the t-distribution is fairly robust to deviation of the population of the normal model

**SLIDE-44**

**Student’s t-distribution**

****

As n -> ꝏ

tn -> N(0,1)

i.e., as the degrees of the freedom increase, the t-distribution approaches the standard normal distribution.

**Slide-45**

**Confidence Interval for mean with unknown Sigma**

± Z1-ᾳ σ/ where Z1-ᾳ satisfies p (-Z1-ᾳ ≤ Z ≤ Z1-ᾳ) = 1-ᾳ

Instead of above equation we can use the below t distribution equation

± t1-ᾳ, n-1 s/ where t1-ᾳ, n-1 satisfies p (-t1-ᾳ, n-1 ≤ Tn-1≤ t1-ᾳ, n-1) = 1-ᾳ

**Slide-46**

**Calculating t-value**

* Construct a 95% confidence interval for the mean card balance and interpret it?

n = 140; σ = $2500; = $ 1990

σ = 2833/sqrt(140) = 239.46

Calculate t0.95, 139 = 1.98

Then the 95% confidence interval for balance is [$1516, $2464]

**Slide-47**

**Hypothesis Testing**

Start with Hypothesis about a Population Parameter

Collect Sample Information

Reject/Do Not Reject Hypothesis

|  |  |  |
| --- | --- | --- |
|  | Ho is TRUE | H1 is TRUE |
| Fail to Reject Ho | Right Decision  Confidence  1-alpha | Type II error  beta |
| Reject Ho | Type I error  alpha | Right Decision  Power  1-beta |

The factors that affect the power of a test include sample size, effect size, population variability, and .Power and are related as increasing decreases Since power is calculated by 1 minus , if you increase ,You also increase the power of a test. The maximum power a test can have is 1, whereas the minimum value is 0.

**Slide - 48**

**Hypothesis testing case studies exercise:**

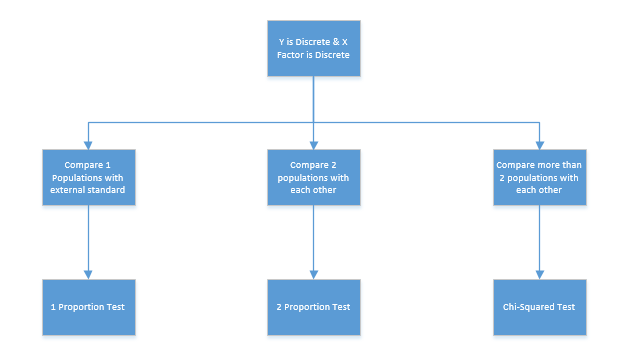
Write Null and alternate Hypothesis for the following case studies:

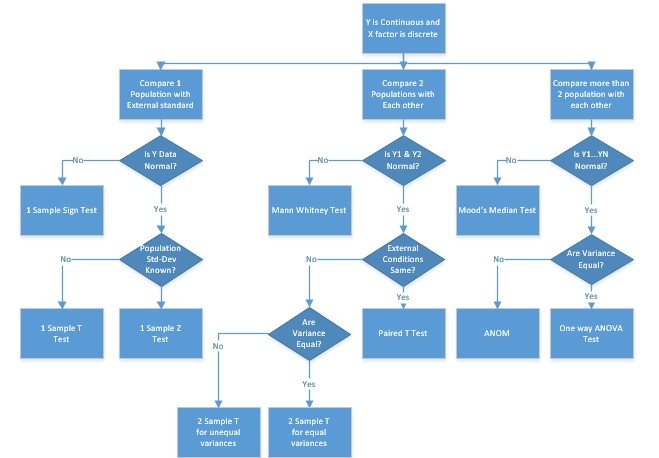
1. Our quality will not improve after the consulting project
2. The retail market will grow by 50% in the next 5 years
3. We will acquire 8,000 new customers if I open a store in this area
4. Less than 5% clients will default on their loans
5. We will need 400 more person hours to finish this project
6. Our potential customers do not spend more than 60 minutes on the web every day

**Slide-49**

Different tests based on type of input and output data types and number of data types:

|  |  |  |
| --- | --- | --- |
| Y | X | Test |
| Continuous | Discrete in 2 categories | 2 - Sample t test |
| Continuous | Discrete more than 2 categories | ANOVA – One Way |
| Discrete | Discrete in 2 categories | 2 - Proportion test |
| Discrete | Discrete more than 2 categories | Chi-square test |

****

**SLIDE-50  
**

**SLIDE-51**

**1-Sample Z test:**

Normality Test

*Stat > Basic Statistics > Graphical Summary*

Population Standard Deviation Known or Not

1 Sample Z Test

Stat > Basic Statistics > 1 Sample Z

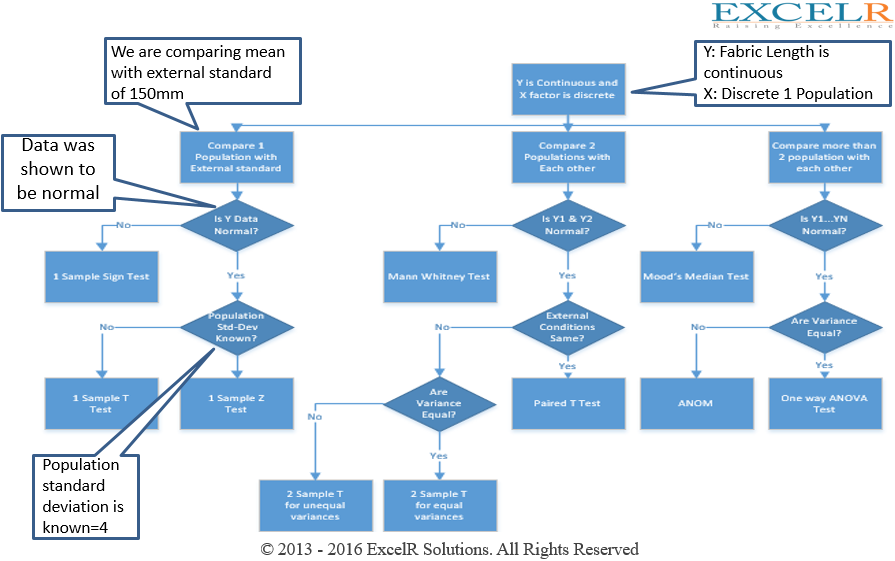
Fabric Data

The length of 25 samples of a fabric are taken at random. Mean and standard deviation from the historic 2 years study are 150 and 4 respectively. Test if the current mean is greater than the historic mean. Assume α to be 0.05

**SLIDE-52**

1-Sample Z test – Write Hypothesis

**SLIDE-53**



**SLIDE-54**

**1-Sample t Test**

* Normality Test

*Stat > Basic Statistics > Graphical Summary*

* Population Standard Deviation Known or Not
* 1 Sample t Test

*Stat > Basic Statistics > 1 Sample t*

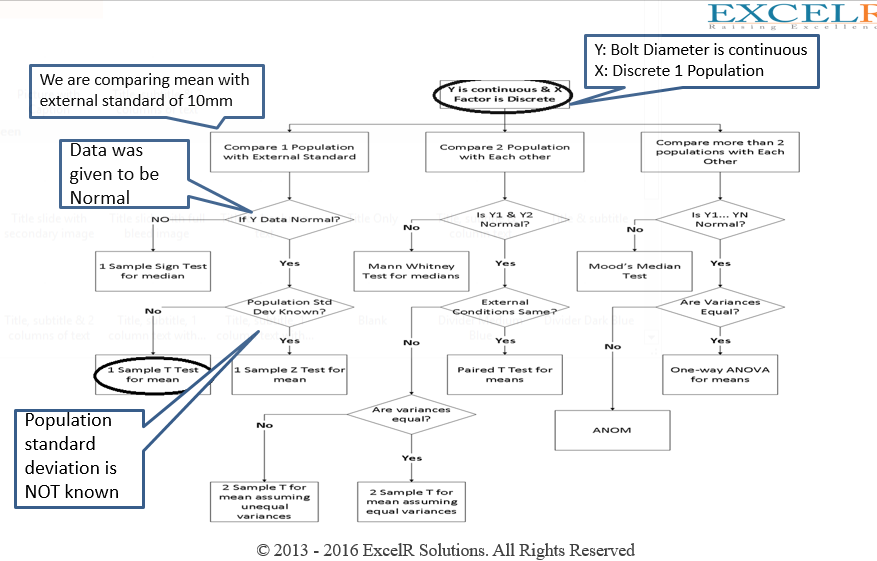
**Bolt Diameter**

The mean diameter of the bolt manufactured should be 10mm to be able to fit into the nut. 20 samples are taken at random from production line by a quality inspector. Conduct a test to check with 95% confidence that the mean is not different from the specification value.

**SLIDE-55**

1-Sample t Test – Write Hypothesis

SLIDE-56



**SLIDE-57**

1-Sample Sign Test

Normality Test

*Stat > Basic Statistics > Graphical Summary*

1 Sample Sign Test

*Stat > Non Parametric >*

*1 Sample sign*

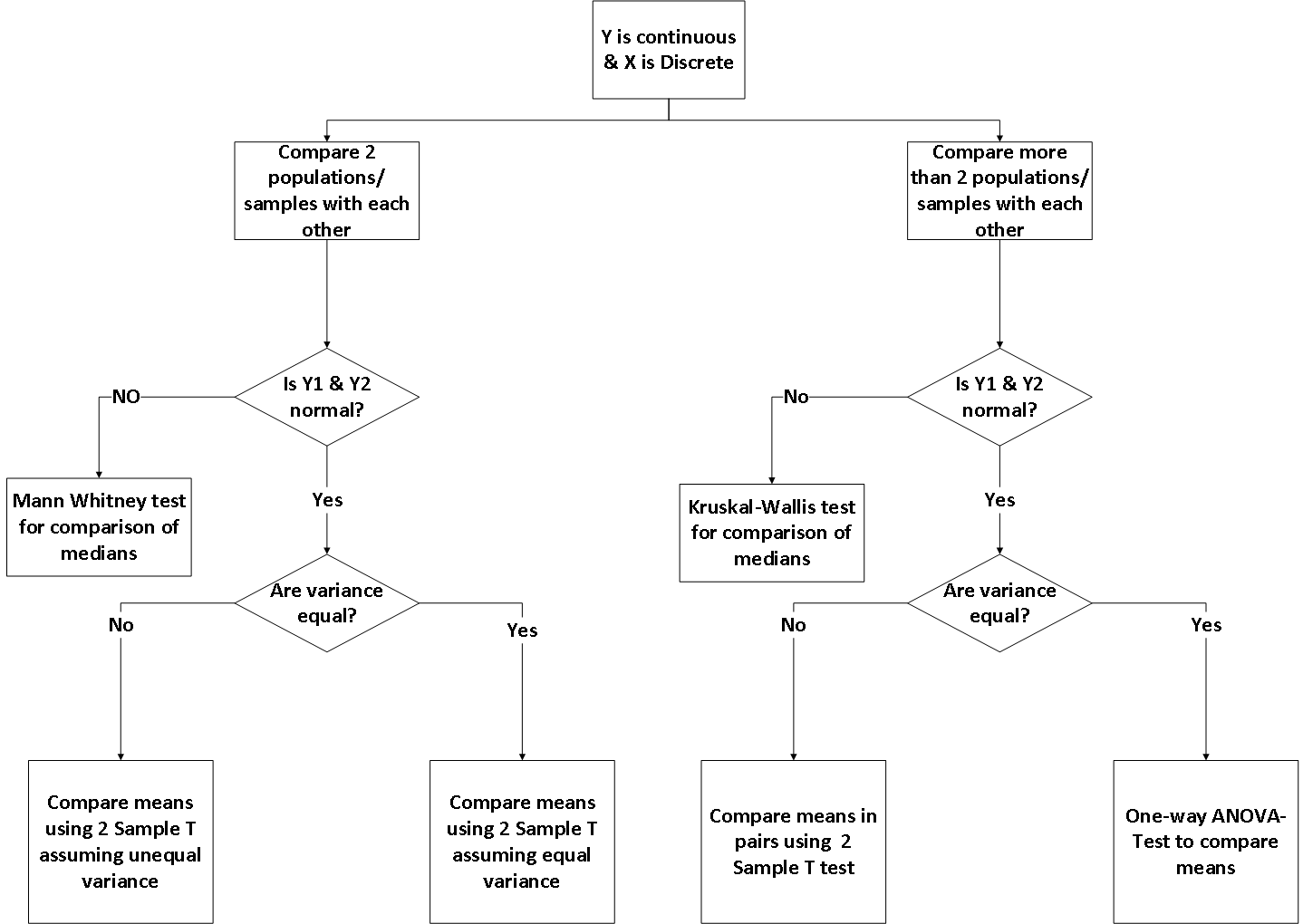
**Student Scores**

The scores of 20 students for the statistics exam are provided. Test if the current median is not equal to historic median of 82. Assume ‘α’ to be 0.05

**SLIDE-58**

1-Sample Sign Test – Write Hypothesis

**Slide-59**

****

**Slide-60**

**2-Sample t-Test**

* Normality test

*Stat > Basic Statistics > Graphical Summary*

* 2 Variance Test

*Stat > Basic Statistics > 2 Variance*

* 2 sample t- Test

*Stat > Basic Statistics > 2 sample t*

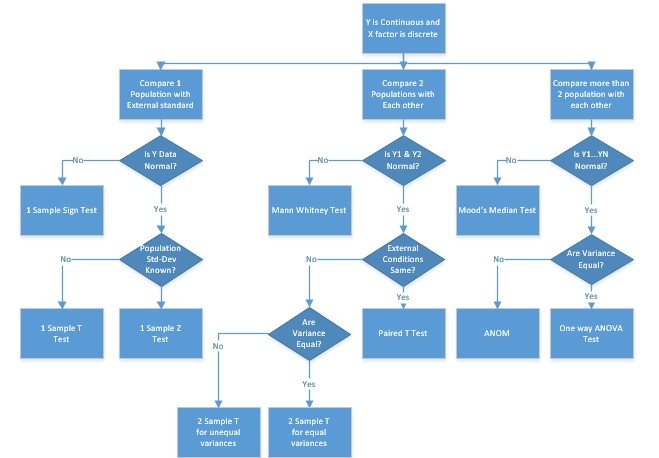
***Marketing Strategy : case study***

A financial analyst at a Financial institute wants to evaluate a recent credit card promotion. After this promotion, 450 cardholders were randomly selected. Half received an ad promoting a full waiver of interest rate on purchases made over the next three months, and half received a standard Christmas advertisement. Did the ad promoting full interest rate waiver, increase purchases?

**SLIDE-61**

**2-Sample t Test – Write Hypothesis**

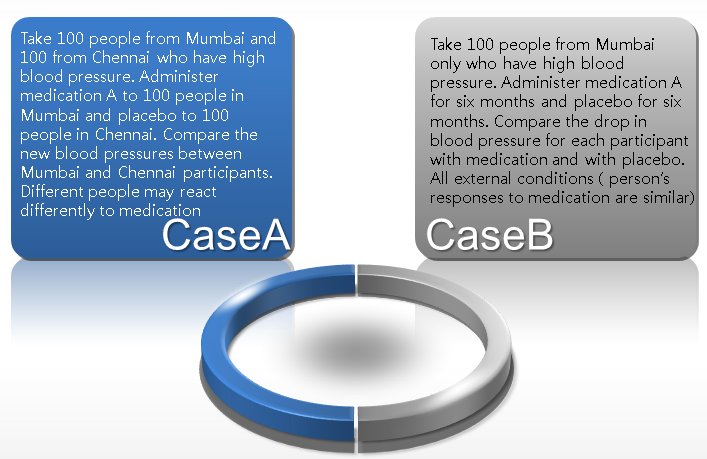
**SLIDE-62**

**Slide-63**

**Paired T Test**

* This test is used to compare the means of two sets of observations when all the other external conditions are the same
* This is a more powerful test as the variability in the observations is due to differences between the people or objects sampled is factored out

Example: To find out if medication A lowers blood pressure

****

**SLIDE-64**

**Trigger your thoughts!**

|  |  |
| --- | --- |
| Comparing the performance of machine A vs. machine B by feeding different raw materials to each machine | Compare the performance of machine A vs. machine B when the same raw material is fed to each machine |
| Compare the power output of two wind mills next to each other simultaneously when you use motor A on one wind mill and motor B on another | Compare the power output of a wind mill when you use motor A for 1 month and motor B for 1 month |
| Identifying resistor defects and capacitor defects in same PCB by collecting such data using 20 PCB units | Identifying resister defects on 20 PCB’s and capacitor defects on 20 (different) PCB’s |

**SLIDE-65**

**2-Sample t test or Paired T test**

Effect of fuel additive on vehicles is being studied. Out of a total of 20 vehicles, 10 vehicles are chosen randomly and mileage is recorded. In rest of the 10 vehicles, additive to be tested is added with the fuel and their mileage is recorded. Find if the mileage increases by adding the fuel additive.

**2-Sample t test**

Assume the same data was recorded if only 10 vehicles were chosen and mileage was recorded before and after adding the additive. What method will you choose to find the result.

**Paired T test**

**SLIDE-66**

**Mann-Whitney test**

Normality Test

*Stat > Basic Statistics > Graphical Summary*

Mann – Whitney test for Medians

*Stat > Non Parametric > Mann Whitney*

Vehicle with

& without Additives

Effect of fuel additive on vehicles is being studied. Out of a total of 20 vehicles, 10 vehicles are chosen randomly and mileage is recorded. In rest of the 10 vehicles, additive to be tested is added with the fuel and their mileage is recorded. Find if the mileage increases by adding the fuel additive.

**SLIDE-67**

**Mann-Whitney Test – Write Hypothesis**

**SLIDE-68**

**Paired T test**

Normality Test

Paired T Test

*Stat > Basic Statistic > Paired T*

* Since the data was not normal, the cause of non-normality was investigated and it was found that the first data point for “with additive” was wrongly entered. This value should have been 20. Now, proceed with the rest of the analysis.
* If the data were truly non-normal our analysis would stop here.
* Vehicle with & without Additives
* Effect of fuel additive on vehicles is being studied. Out of a total of 20 vehicles, 10 vehicles are chosen randomly and mileage is recorded. In rest of the 10 vehicles, additive to be tested is added with the fuel and theirmileage is recorded. Find if the mileage increases by adding the fuel additive. Assume the same data was recorded if only 10 vehicles were chosen and mileage was recorded before and after adding the additive.

**SLIDE-69**

**Paired T test – Write Hypothesis**

**SLIDE-70**

**One-Way ANOVA :**

Normality Test

* *Stat > Basic Statistics > Graphical Summary*

Variance Test

* *Stat > ANOVA > Test for Equal Variances*

ANOVA

*Stat > ANOVA > One-Way….*

***Contract Renewal: Case Study***

*A marketing organization outsources their back-office operations to three different suppliers. The contracts are up for renewal and the CMO wants to determine whether they should renew contracts with all suppliers or any specific supplier. CMO want to renew the contract of supplier with the least transaction time. CMO will renew all contracts if the performance of all suppliers is similar.*

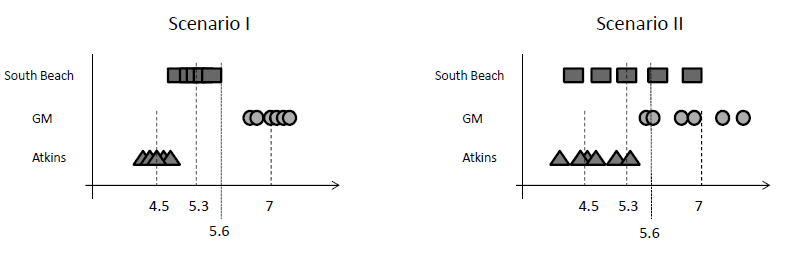
**Slide-71**

Example: More weight reduction programs

* Suppose the nutrition expert would like to do a comparative evaluation of three diet programs(Atkins, South Beach, GM)
* She randomly assigns equal number of participants to each of these programs from a common pool of volunteers
* Suppose the average weight losses in each of the groups(arms) of the experiments are 4.5kg, 7kg, 5.3kg
* What can she conclude?

**Slide-72**

**Two kinds of variation matter**

****

* Diet program having high variances for the sample weight losses
* Diet program having low variance for the sample weight losses

Not every individual in each program will respond identically to the diet program

Easier to identify variations across programs if variations within programs are smaller

Hence the method is called Analysis of Variance(ANOVA)

**Slide-73**

**Formalizing the intuition behind variations**

* It should be obvious that for every observation : Totij = ti + eij
* What is more surprising and useful is:

Sum of squares total (SST) = ij2

Sum of squares Treatment (SSTR) = iti2

Sum of squares Error (SSE) = ij2

**SST = SSTR + SSE**

**Slide-74**

**Statistically test for equality means**

* n subjects equally divided into r groups
* Hypothesis

- H0: μ1 = μ2 = μ3 = … = μr

- Not all μi are equal

**Calculate**

- Mean Square Treatment MSTR = SSTR / (r‐1)

- Mean Square Error MSE = SSE / (n‐r)

- The ratio of two squares f = MSTR/MSE

- Strength of this evidence p‐value = Pr(F(r‐1,n‐r) ≥ f)

• Reject the null hypothesis if p‐value < α

**Slide-75**

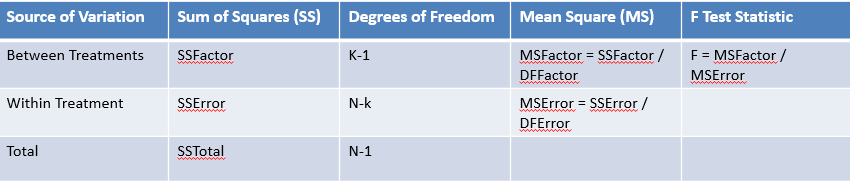
**Analysis of variance(ANOVA)**

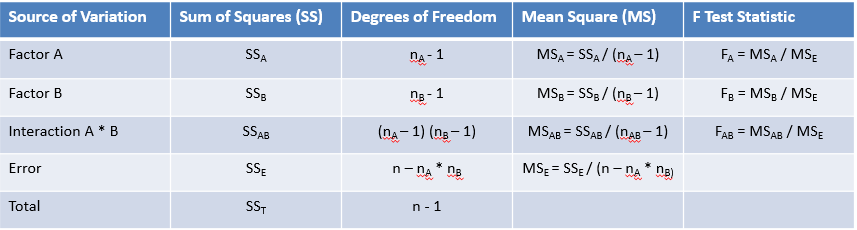
* ANOVA can be used to test equality of means when there are more then 2 populations
* ANOVA can be used with one or two factors
* If only one factor is varying, then we would use a one-way ANOVA
  + Example: We are interested in comparing the mean performance of several departments within a company. Here the only factor is the name of department
  + If there are two factors, we would use a two way ANOVA. Example: One factor is department and the second factor is the shift.(day vs. Night)

**SLIDE-76**

**Analysis of variance (ANOVA)**

**ONE-WAY ANOVA**



****

**Slide-77**

**Dichotomies**

**2-Sample t test**

Is the Transaction time dependent on whether person A or B processes the transaction?

Is medicine 1 effective or medicine 2 at reducing heart stroke?

Is the new branding program more effective in increasing profits?

**ANOVA – One-Way**

Does the productivity of employees vary depending on the three levels? (Beginner, Intermediate and Advanced)

Three different sale-closing methods were used. Which one is most effective?

Four types of machines are used. Is weight of the Rugby ball dependent on the type of machine used?

**Slide-78**

Non-Parametric equivalent to ANOVA

* When the data are not normal or if the data points are very few to figure out if the data are normal and we have more than 2 populations, we can use the Mood’s Median or Kruskal Wallis test to compare the populations

Ho: All the medians are the same

Ha: One of the medians is different

* Mood’s median assigns the data from each population that is higher than the overall median to one group, and all points that are equal or lower to another group. It then uses a Chi-Square test to check if the observed frequencies are close to expected frequencies
* Kruskal Wallis is another test that is non-parametric equivalent of ANOVA. Kruskal Wallis is the extension of Mann-Whitney test

**Slide-79**

**Mood’s Median & Kruskal Wallis**

**Mood’s Median – handles outliers well**

* *Stat > Nonparametric > Mood’s Median*

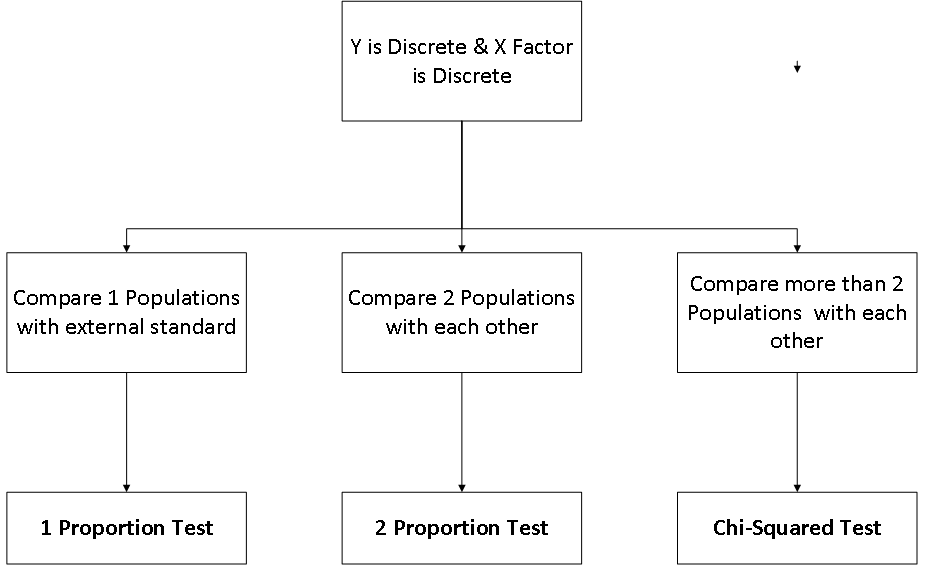
**Kruskal Wallis – more powerful than Mood’s Median**

* *Stat > Nonparametric > Kruskal Wallis*

**Height** **Growth:**

Growth is measured for three treatments as shown in the case study. Compare the effect of the three treatments on growth.

**Slide-80**

****

**Slide-81**

**1-Proportion Test**

Stat > Basic Statistics > 1-Proportion

**Football Coach**

* The people carry out a poll to find the acceptability of new football coach. It was decided that if the support rate for the coach for the entire population was truly less than 25%, the coach would be fired
* 2000 people participated and 482 people supported the new coach
* Conduct a test to check if the new coach should be fired with 95% level of confidence.

**Slide-82**

**2-Proportion Test**

|  |  |  |
| --- | --- | --- |
| HO | Proportion A = Proportion B | Check p-value |
| Ha | Proportion A NOT = Proportion B | If p-value < alpha,  we reject Ho |

**Johnnie Talkers : case study**

Johnnie Talkers soft drinks division sales manager has been planning to launch a new sales incentive program for their sales executives. The sales executives felt that adults (>40 yrs) won’t buy, children will & hence requested sales manager not to launch the program. Analyze the data & determine whether there is evidence at 5% significance level to support the hypothesis

**Slide-83**

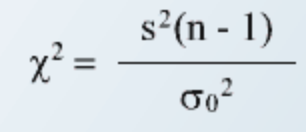
**Chi-Square Test**

* + *How can you determine whether the distribution of defects in your product or service has changed from the historic distribution over time, or exceeds an industry standard*
* Do you think mean is more significant or variance?
  + *Comparing population’s variance to a standard value involves calculating the*

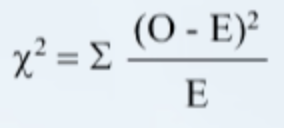
*Chi-square test statistic*

We can also:

* + *Determine whether one variable is dependent over another*



* + *Comparing observed & expected frequencies where variance is unknown.*



*This is called as goodness-of-fit test*

* + *Compare multiple proportions*

**Slide-84**

**Chi-Square Goodness-of-fit test**

Goodness-of-fit test is to test assumptions about the distributions that fit the process data

Are observed frequencies (O) same or different from historical, expected or theoretical frequencies (E)?

If there’s a difference between them, this suggests that the distribution model expressed by the expected frequencies does not fit the data

**Slide-85**

* A city has a newly opened nuclear plant, and there are families staying dangerously close to the plant. A health safety officer wants to take this case up to provide relocation for the families that live in the surrounding area. To make a strong case, he wants to prove with numbers that an exposure to radiation levels is leading to an increase in diseased population. He formulates a contingency table of exposure and disease.
* Does the data suggest an association between the disease and exposure?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Disease** | | **Total** |
| **Exposure** | Yes | No |
| Yes | 37 | 13 | 50 |
| No | 17 | 53 | 70 |
| Total | 54 | 66 | 120 |

Slide-86

Calculate the number of individuals of exposed and unexposed groups expected in each disease category (yes and no) if the probabilities were the same

If there were no effect of exposure, the probabilities should be same and the chi-squared statistic would have a very low value.

Proportion of population exposed = (50/120) = 0.42

Proportion of population not exposed = (70/120) = 0.58

Thus, expected values:

Population with disease = 54

Exposure Yes : 54 \* 0.42 = 22.5

Exposure No : 54 \* 0.58 = 31.5

Population without disease = 66

Exposure Yes : 66 \* 0.42 = 27.5

Exposure No : 66 \* 0.58 = 38.5

**Slide-87**

* Calculate the Chi-squared statistic

χ2

= (∑ (observed frequency-expected frequency) 2)

Expected frequency

= 

* Calculate the degrees of freedom :

(Number of rows – 1) X (Number of columns – 1)

df = (2 – 1) X (2 – 1) = 1

* Calculate the p-value from the Chi-squared table

For chi-squared value 29.1 and degrees of freedom = 1, from the table, p-value is < 0.001

* Interpretation: There is 0.001 chance of obtaining such discrepancies between expected and observed values if there is no association
* Conclusion : There is an association between the exposure and disease

**Slide-88**

**Chi-Square Test**

|  |  |  |
| --- | --- | --- |
| HO | All proportions are equal | Check p-value |
| Ha | Not all proportions are equal | If p-value < alpha we reject Ho |

**Bahaman Research: case study**

Baha ManTech Research Company uses 4 regional centers in South Asia (India, China, Srilanka and Bangladesh) to input data of questionnaire responses. They audit a certain % of the questionnaire responses versus data entry. Any error in data entry renders it defective. The chief data scientist wants to check whether the defective % varies by country. Analyze the data at 5% significance level and help the manager draw appropriate inferences. [‘1’ means not defectives & ‘0’ means defective]

**Slide-89**

**Non-Parametric Tests:**

* Referred to as “distribution free”, as they don’t involve making assumptions of any data
* They have lower power than the parametric tests and hence are always given the second preference after the parametric tests
* These tests are typically focused on median rather than mean
* They involve straight-forward procedures like counting and ordering

**Slide-90**

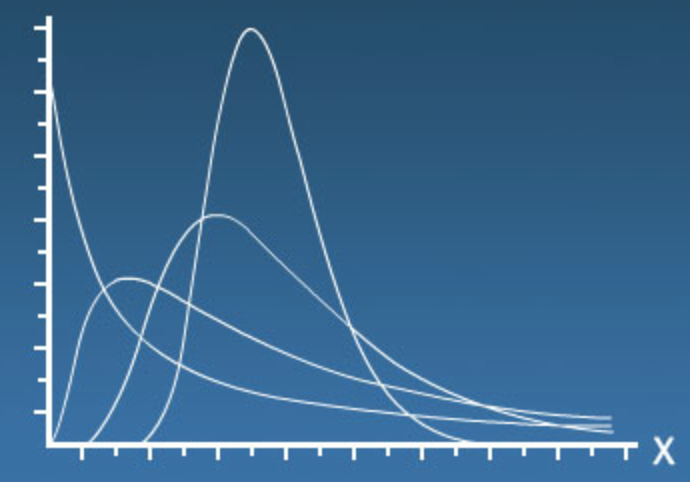
**Probability Distributions**

***Lognormal:***

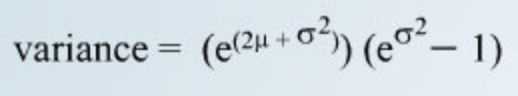
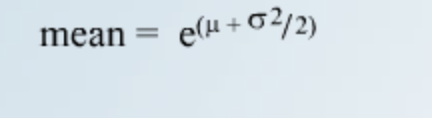
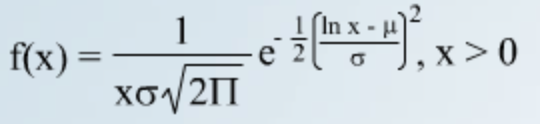
* Fits many kinds of failure data
* Used for reliability analysis, cycles-to-failure, loading variables & fatigue stress
* Tensile strength of fibers & breaking strength of concrete
* Environment data such as random quantities of pollutants in water or air

|  |  |
| --- | --- |
| Data | Log transformed |
| 12 | 2.48 |
| 28 | 3.33 |
| 87 | 4.47 |
| 143 | 4.96 |

* Economic variables such as per capita income

****

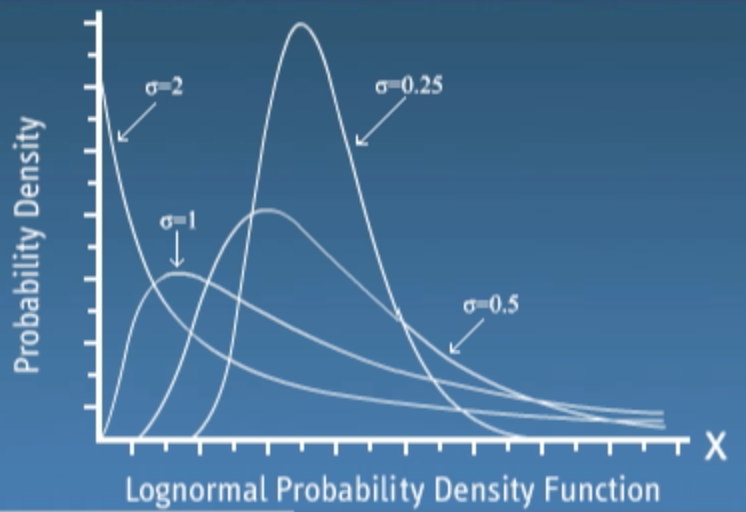
* Extreme values are well managed & makes data normal
* μ, σ are mean & standard deviation of natural logarithms

** **

**Slide-91**

**Probability Distributions**

*Lognormal:*

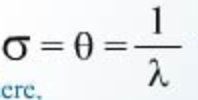
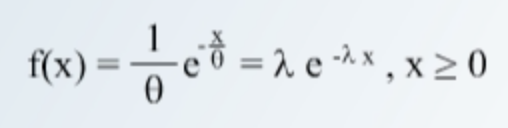
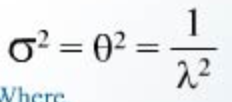
* This distribution is right skewed
* Skewness increases as value of σ increases
* Pdf starts at zero, increases to its mode, and then decreases
* If time-to-failure has a lognormal distribution, then the logarithm of time-to-failure has a normal distribution
* 

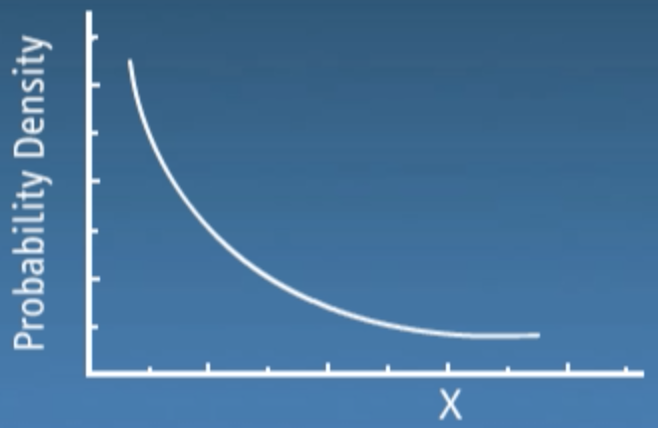
**Slide-92**

**Probability Distributions**

***Exponential:***

* Length of time between check-ins at a reception desk, calls at a call center, customers at a cashier
* Used when events occur continuously & independently at a constant average rate
* Used to model rate of change that will occur in a given amount of time
* How long equipment will keep working with proper maintenance & part replacement
* Use to model behavior of independent variables that have a constant rate
* The occurrences of variables are described by a Poisson distribution, but the times between occurrences are described by Exponential distribution
* If X is Poisson distributed then Y = 1/X will be exponentially distributed
* # of arrivals at a checkout counter, # of product failures over time – Poisson
* Length of time between events, i.e., one arrival or failure & the next – Exponential distribution
* Exponential distribution can model the interval between random events

****



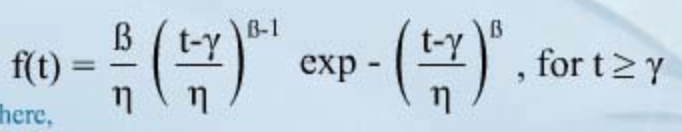
* λ = failure rate; θ = mean; x = random variable
* Used to model mean time between occurrences
* In exponential population, 37% of observations are below the mean & 63% are above
* Uses constant failure rate

**Slide-93**

**Probability Distributions**

***Weibull:***

* Model failure rate; rate is not constant



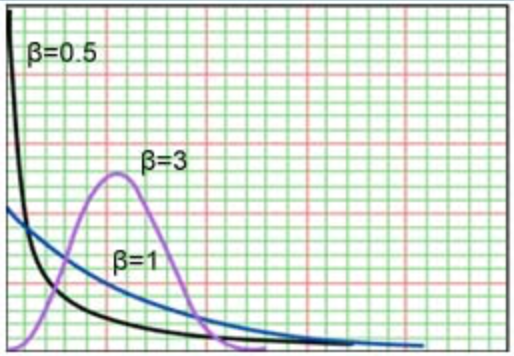
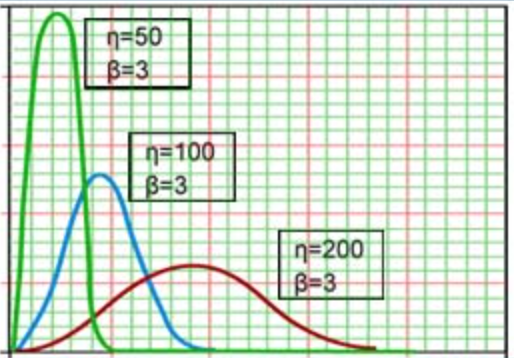
* Model time to failure, time to repair & material strength
* When system/item ages & failure rate increases/decreases
* Can model different distributions due to having parameters of shape, scale & location
* Can simulate Lognormal, Exponential & many other distributions
* Use widely in reliability & statistical applications
* Weibull & Lognormal are from same family & both can be used to assess the dataset that contains close to average values (not too high / low)
* However, Weibull is a better fit when majority of data falls to the higher side
* Lognormal is a better fit when majority of data falls to the lower side

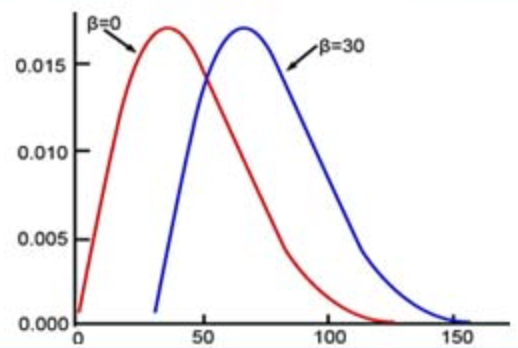
**Slide-94**

**Probability Distributions**

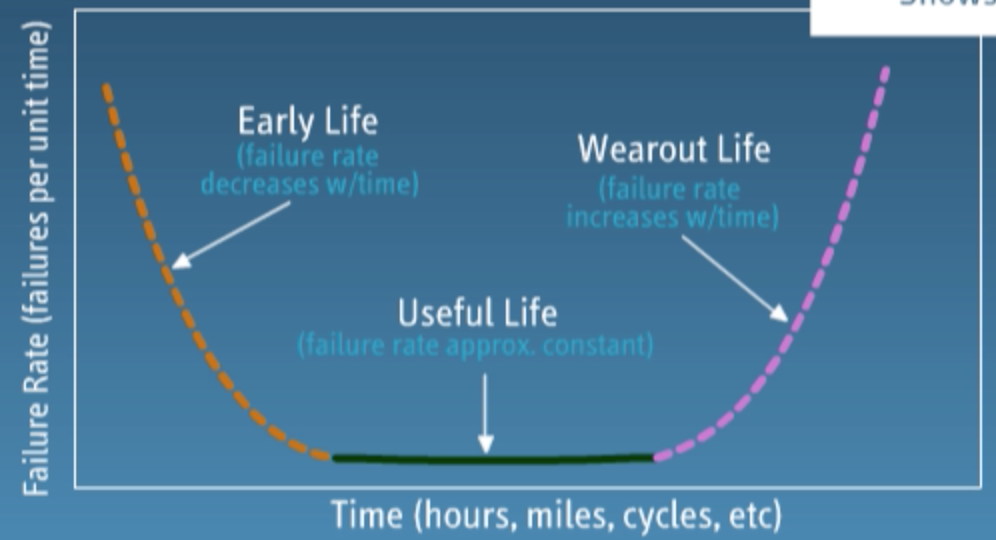
***Weibull:***

* β is shape parameter, also called as slope, determines the shape of the distribution
  + When beta = 1, shape of distribution = exponential distribution
  + When beta: 3 to 4, shape of distribution = normal distribution
  + Several beta values can approximate lognormal distribution



* η is scaled parameter (eta), determines the spread or width of distribution
* γ is non-zero location parameter, is the point, below which there are no failures, changing the value will move distribution to right or left
  + Gamma > 0, there is a period when no failures occur
  + Gamma < 0, failures have occurred before time equals zero
* e.g., defective raw materials or failure during transportation
  + When Gamma = 0, eta is called as characteristic life
* Regardless of specific value of beta, 63.2% of values fall below the characteristic life



**Slide-95**

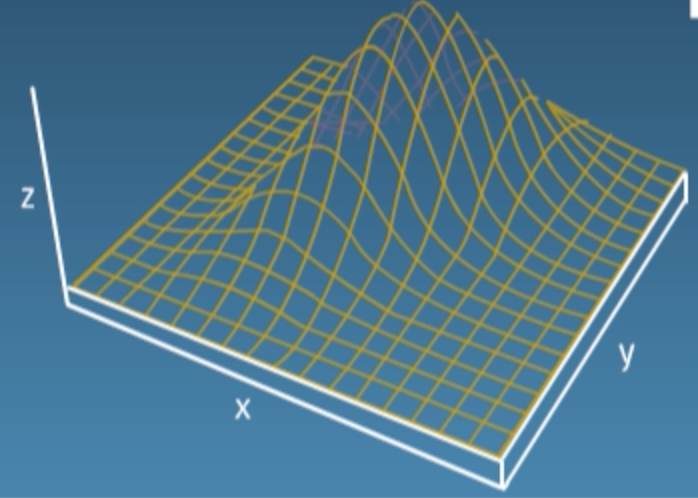
**Probability Distributions**

***Bivariate Normal Distribution:***

* Used when 2 variables that are normally distributed & may be totally independent or may be correlated to some degree
* A joint distribution of two independent variables that simultaneously

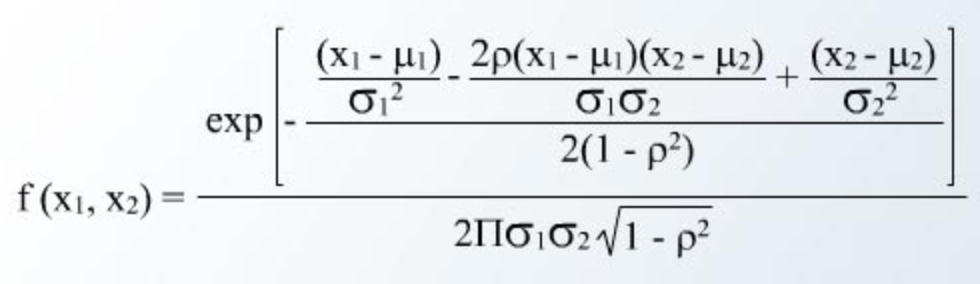
& jointly cross-classifies the data

* Can be discrete or continuous
* 3D plot like mountain terrain



* X & Y axes represent independent variables
* Z axis shows either
  + frequency for discrete data
  + probability for continuous data
* The maximum or peak occurs when X1 = Mu1 & X2 = Mu2. You can take a

“Slice” anywhere along the distribution by fixing one of the variables. This is known as a conditional distribution

****

**Slide-96**

**Probability Distributions**

***Bivariate Normal Distribution:***

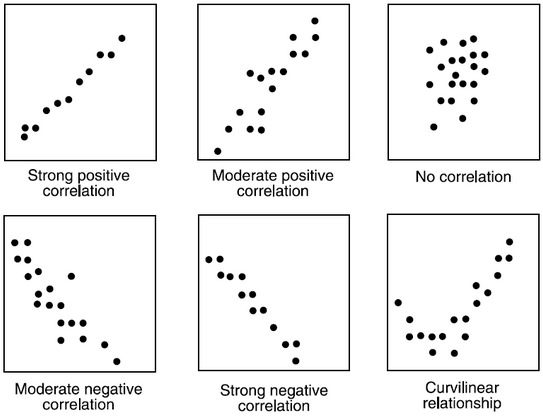
* Can help determine items of critical importance:
  + Causality – examine the joint frequencies to investigate if the second variable changes in a systematic way when the first variable changes
  + Predictions – reviewing outcomes from one variable as the other changes
  + Importance – if two variables are causally related they should have a statistically significant impact

**Slide-97**

**Scatter Diagram**

Scatter diagrams or plots provides a graphical representation of the relationship of two continuous variables

Be Careful - Correlation does not guarantee causation. Correlation by itself does not imply a cause and effect relationship!



Judge strength of relationship by width or tightness of scatter

Determine direction of the relationship, e.g. If X increases, and Y decreases; it is negative correlation, similarly if X increases, and Y increases, it is positive correlation

Scatter Plot can show Strong positive correlation, Moderate positive correlation, No correlation, Moderate negative Correlation, Strong

Negative correlation, curvilinear relation.

**Slide-98**

**Correlation Analysis**

Correlation Analysis measures the degree of linear relationship between two variables

Range of correlation coefficient -1 to +1

Perfect positive relationship +1

Perfect negative relationship -1

No Linear relationship 0

If the absolute value of the correlation coefficient is greater than 0.85, then we say there is a good relationship

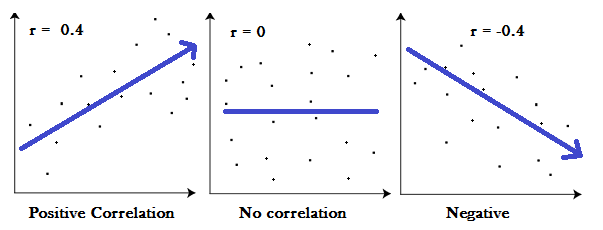
* + Example: r = 0.87, r = -0.9, r = 0.9, r = -0.87 describe good relationship
  + Example: r = 0.5, r = -0.5, r = 0.28 describe poor relationship

Correlation values of -1 or 1 imply an exact linear relationship. However, the real value of correlation is in quantifying less than perfect relationships

We can perform regression analysis, which attempts to further describe this type of relationship, if the correlation is good between the 2 variables

**Slide- 99**

**Correlation Analysis:**

****

Positive correlation: r>0

Negative correlation: r<0

No correlation: r=0

r = (n (-(∑x) (/ (sqrt ([n∑x2-(∑x) 2] [n∑y2-(∑y) 2]))

**Slide-100**

**Linear Regression Model**

The equation that represents how an independent variable is related to a dependent variable and an error term is a regression model

y = β0 + β1x + ε

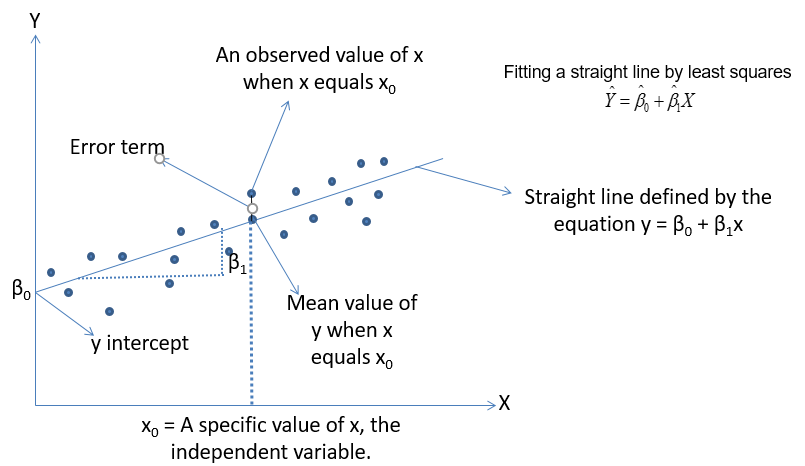
Where, β0 and β1 are called parameters of the model,

ε is a random variable called error term.

β o= ( (2)-(∑x) (/ ([n∑x2-(∑x) 2] )

β 1= ( (-(∑x) (/ ([n∑x2-(∑x) 2] )

**Slide-101**



**Slide-102**

**Regression Analysis**

R-squared-also known as Coefficient of determination, represents the % variation in output (dependent variable) explained by input variables/s or Percentage of response variable variation that is explained by its relationship with one or more predictor variables

Higher the R^2, the better the model fits your data

R^2 is always between 0 and 100%

R squared is between 0.65 and 0.8 => Moderate correlation

R squared in greater than 0.8 => Strong correlation

R2=SSR/SST= (SSR/(SSR+SSE))

0<=R2<=1

Mathematically

SSR =∑(-)2 🡪 measure of an explained variation

SSE =∑()2🡪 measure of an unexplained variation

SST = SSR+SSE =∑(y-)2 🡪 measure of total variation in y

**Slide-103**

**Regression Analysis**

Prediction and Confidence Interval are types of confidence intervals used for predictions in regression and other linear models

Prediction Interval: Represents a range that a single new observation is likely to fall given specified settings of the predictors

Confidence interval of the prediction: Represents a range that the mean response is likely to fall given specified settings of the predictors

The prediction interval is always wider than the corresponding confidence interval because of the added uncertainty involved in predicting a single response versus the mean response

**Slide-104**

**Regression Techniques – Simple Linear Regression**

Y-continuous, x – single & continuous

We apply simple linear Regression

Y-continuous, x – single & discrete

We create dummy variable for discrete component and

We then apply simple linear Regression

**Simple Linear Regression – Dummy Variable**

**Slide-105**

**Example:**

|  |  |
| --- | --- |
| **Gender** | **Dummy Variable** |
| Male | 1 |
| Female | 0 |
| Male | 1 |
| Female | 0 |
| Male | 1 |

**Slide-106**

**Simple Linear Regression – R**

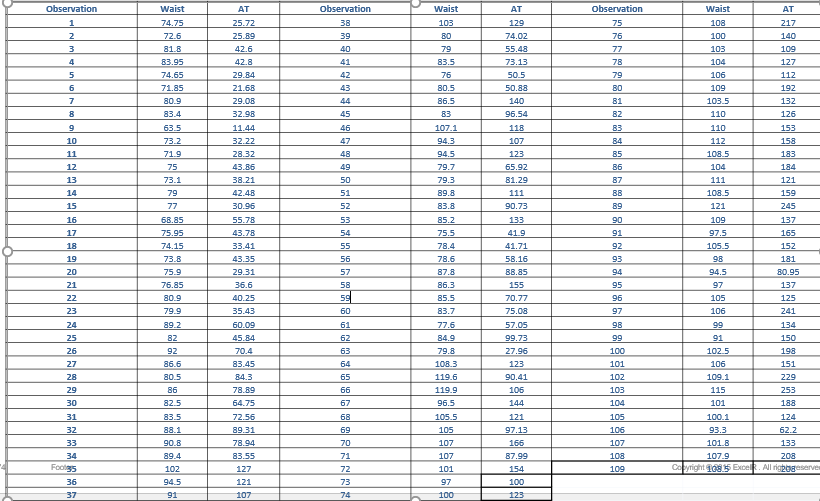
**A business problem:**

The Waist Circumference – Adipose Tissue data

* + Studies have shown that individuals with excess Adipose tissue (AT) in the abdominal region have a higher risk of cardio-vascular diseases
  + Computed Tomography, commonly called the CT Scan is the only technique that allows for the precise and reliable measurement of the AT (at any site in the body)
  + The problems with using the CT scan are:
    - Many physicians do not have access to this technology
    - Irradiation of the patient (suppresses the immune system)
    - Expensive
  + Is there a simpler yet reasonably accurate way to predict the AT area? i.e.,
    - Easily available
    - Risk free
    - Inexpensive
  + A group of researchers conducted a study with the aim of predicting abdominal AT area using simple anthropometric measurements, i.e., measurements on the human body
  + The Waist Circumference – Adipose Tissue data is a part of this study wherein the aimis to study how well waist circumference (WC) predicts the AT area

**Side-107**

**Simple Linear Regression – Data Set**



**Slide- 108**

**Simple Linear Regression – Transformation**

reg <- lm(AT ~ Waist) # Linear Regression

summary(reg)

confint(reg, level=0.95)

predict(reg, interval="predict”)

reg\_log <- lm(AT ~ log(Waist)) # Regression using Logarithmic Transformation

summary(reg\_log)

confint(reg\_log, level=0.95)

predict(reg, interval="predict”)

reg\_exp <- lm(log(AT) ~ Waist) # Regression using Exponential Transformation

summary(reg\_exp)

confint(reg\_exp, level = 0.95)

predict(reg, interval="predict”)

**Slide-109**

**Regression Techniques – Multiple Linear Regression**

Y-continuous, x – Multiple & continuous

We apply Multiple linear Regression

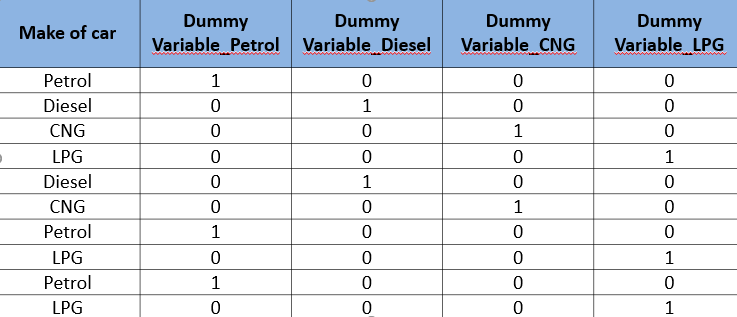
Y-continuous, x – Multiple & discrete

We create dummy variable for discrete component and

We then apply Multiple linear Regression

**Slide-110**

**Multiple Linear Regression – Dummy Variable**



**Slide- 111**

**Multiple Regression Model**

DATA : CARS, 81 observations, *“cars.csv”*

* + VOL = cubic feet of cab space

* + HP = engine horsepower
  + MPG = average miles per gallon
  + SP = top speed, miles per hour
  + WT = vehicle weight, hundreds of pounds

Our interest is to model the MPG of a car based on the other variables.

**Slide-112**

**Model and Assumptions**

**Our Model:** 

Linear

Independent

Normal

Equal Variance

Linearity (Assumptions about the form of the model):

* + - * Linear in parameters
      * Assumptions about the errors:
      * IID Normal (Independently & identically distributed)
      * Zero mean
      * Constant variance (Homoscedasticity)
      * If no constant variance (HETEROSCEDASTICITY)
      * Independent of each other. If not independent, it is called as AUTO CORRELATION problem
      * Assumptions about the predictors:
      * Non-random
      * Measured without error
      * Linearly independent of each other. If not it is called as COLLINEARITY problem
      * Assumptions about the observations:
      * Equally reliable

**Slide-113**

**Techniques used for Discrete Output**

* Logistic Analysis
* Logit Analysis
* Probit Analysis

**Slide-114**

**Regression Techniques – Simple Logistic Regression**

Y-Discrete, x – Single & continuous

We apply Simple logistic Regression

Y- Discrete, x – Single & discrete

We create dummy variable for discrete component and

We then apply Simple logistic Regression

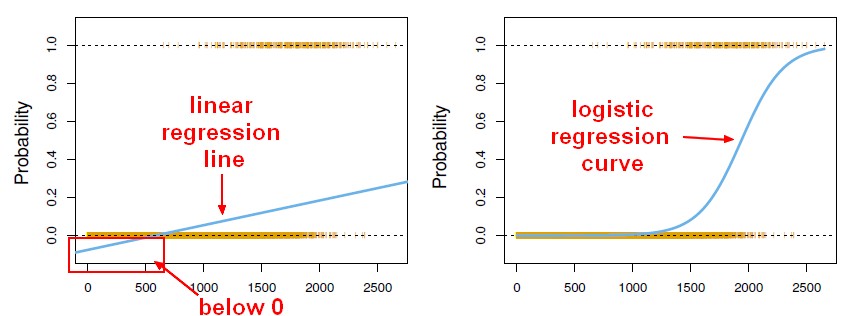
**Slide-115**

**Logistic Regression**

* + Logistic Regression model predicts the probability associated with each dependent variable Category

*How does it do this?*

* + It finds linear relationship between independent variables and a link function of these probabilities. Then the link function that provides the best goodness-of-fit for the given data is chosen.



**Slide-116**

* + Multiple Logistic Regression Model is quite similar to the Multiple Linear Regression Model, Only β coefficients vary

Multiple Linear Regression model :



Where  = Y intercept

 -> the model coefficient for the linear effect of variable i on y

e -> the random error

**Slide-117**

The probability in a logistic regression curve

P=ey/(1+ ey)

Where e is a real constant, the base of natural algorithm and equals to 2.7183

Y is the response value of the observation

**SLIDE-118**

**Logistic Regression Methods**

|  |  |  |
| --- | --- | --- |
| Method | Description of categorical response variable | Example |
| binary | 2 categories | Presence/absence of disease |
| Nominal | 3 or more categories with no natural ordering to the levels | Crunchy/mushy/crispy |
| ordinal | Three or more variables with ordering of levels | Strongly disagree/disagree/neutral/  Agree/strongly agree |

**SLIDE-119**

**Assumptions in Logistic Regression**

* Only one outcome per event – Like pass or fail
* The outcomes are statistically independent
* All relevant predictors are in the model
* One category at a time – Mutually exclusive & collectively exhaustive
* Sample sizes are larger than for linear regression

**SLIDE-120**

**Steps in Logistic Regression**

* Collect & organize sample data
* Formulate Logistic Regression Model
* Check the model’s validity
* Determine Probabilities using Probability equation
* Compile the results

**SLIDE-121**

**Logistic Regression Example**

Imagine that you are a Data Scientist at a very large scale integration circuit manufacturing company. You want to know whether or not the time spent inspecting each product impacts the quality assurance department’s ability to detect a designing error in the circuit

* *Step-1: Collect and organize the sample data*

*Number of Observations*

*Error Identification*

*Inspection Time*

Number of Observations: 55 Observations of circuits with errors, and determine whether those errors were detected by QA