**Introduction to**

**Business Analytics**

Disclaimer: This material is protected under copyright act AnalytixLabs ©, 2011-2017. Unauthorized use and/ or duplication of this material or any part of this material including data, inany form without explicit and written permission from AnalytixLabsis strictly prohibited. Any violation of this copyright will attract legal actions.

***Learn to Evolve***

**Types of Data**

**Categorical data**

Is non-numeric, can be observed but

not measured

E.g. Favorite color, Place of Birth

**Discrete**

**Quantitative data**

Is numerical data which can be measured

**Continuous**

Random variable which takes only

Random variable which takes

isolated values in its range of variation. For example number of heads in 10 tosses of a coin

**Nominal**

Values do not have ordering

Example categorical variables like color, nationality and so on

any value in its range of

variation. For example, height of a person

**Ordinal**

Values are ordered

Example Satisfaction scores

**Types of Statistics**

**Statistics** is the science of collecting, organizing,

presenting, analyzing, and interpreting numerical data to

assist in making more effective decisions.

**Descriptive Statistics** describes the data set that’s being analyzed, but doesn’t allow us to draw any conclusions or make any inferences about the data

• Measures of Central Tendency • Measures of Dispersion

• Tables and Graphs

**Inferential statistics** is a set of methods that is used to draw conclusions or inferences about the characteristics of populations based on data from a sample

• Estimation

• Hypothesis Testing

**Measures of Central Tendency**

**MEAN MEDIAN MODE**

It is just the average of the data, computed as the sum of the data points divided by the number of points

It is the easiest metric to

understand and communicate Mean is prone to presence of

outliers

Example: What is a typical student in the class doing?

It is the value in the middle of the data set, when the data points are arranged from smallest to largest.

***Tricky circumstances:***

If there is an even number of data points, you will need to take the average of the two middle values.

Median is a more “robust” to presence of outliers

It is more complicated to

communicate

Example: To compare performance of any single student against group

It is the most common value in the dataset.

***Tricky circumstances:***

If no value occurs more than once, thenthere is no mode

If two, or more, values occur as frequently as each other and more frequently than any other, then thereare two, or more, modes.

Not very practical since it is affected by skewness

Most real life distributions are multimodal

Example: A parent wanting to knowwhether their child is better or worsethan typical child at his grade level

**An example – Histogram of weekly returns of XYZ share prices** Median = -1

30

25

t

**27 272627**

Mean = 20.8

n

e

c

r

e

P

yc

n

e

u

q

e

r

F

20 15 10 5

**16**

**11**

**8**

**18**

**15**

**89**

And Mode?

**123** 0

**2**

-24 -21 -18 -15 -11 -8 -5 -1 2 5 9 12 15 18 2000 Weekly returns

**But are these sufficient?**

• There is the man who drowned crossing a stream with an average depth of six inches.

• Say you were standing with one foot in the oven and one foot in an ice bucket. According to the averages, you should be perfectly comfortable.

| Auto | Office Transport | Own Car |
| --- | --- | --- |
| 7 | 9 | 1 |
| 6 | 9 | 3 |
| 3 | 9 | 5 |
| 8 | 9 | 7 |
| 12 | 9 | 9 |
| 9 | 9 | 9 |
| 9 | 9 | 9 |
| 13 | 9 | 11 |
| 13 | 9 | 13 |
| 9 | 9 | 15 |
| 10  9 | 9  9 | 17  9 |
| 9 | 9 | 9 |

• Time taken by different modes of transport

Mean

Median

Mode 9 9 9

**Measures of Variation/Dispersion**

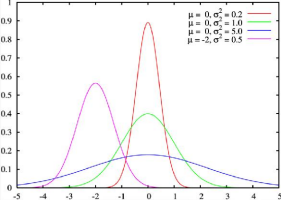
**Dispersion** refers to the spread or variability in the data. It determines how spread out are the scores around the mean.

***Why is Dispersion important?***

• It gives additional information that enables to judge the reliability of the measure of central tendency

• If data are widely spread the central location is less representative of

**Distributions with different dispersions**

data as a whole than it would be for data more closely centered around Mean 

• Since problems are peculiar to widely dispersed data, dispersion enables to identify and tackle problems accordingly

• This enables to compare dispersions of various samples • For eg. If a wide spread of values are away from center, this may be • undesirable or presents a risk, one may avoid choosing that distribution

**Common measures of dispersion**

**Standard Deviation is a measure of how spread out numbers are Variance is defined as the average of the squared differences from the Mean**

2

2

) + . . . Σ (X - μ)2

( ( )

=− μ + − μ

*x x*

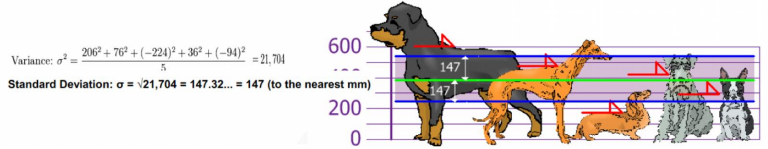
σ2=

1

2

N **Where X is the value of an observation in the population**

**µ is the arithmetic mean of the population**

Example: You have just measured the heights of your dogs (in millimeters). The heights are: 600mm, 470mm, 170mm, 430mm and 300mm. 

Mean = 394mm

Using the Standard Deviation we have a "standard" way of knowing what is normal,

and what is extra large or extra small.

**N is the number of observations in the population**

**Now have a look….**

| Auto | Office Transport | Own Car |
| --- | --- | --- |
| 7 | 9 | 1 |
| 6 | 9 | 3 |
| 3 | 9 | 5 |
| 8 | 9 | 7 |
| 12 | 9 | 9 |
| 9 | 9 | 9 |
| 9 | 9 | 9 |
| 13 | 9 | 11 |
| 13 | 9 | 13 |
| 9 | 9 | 15 |
| 10  9 | 9  9 | 17  9 |
| 9 | 9 | 9 |

Mean Median Mode 9 9 9

| **Std Dev** | **3.0** | **0.0** | **4.9** |
| --- | --- | --- | --- |
| **Variance** | **9.2** | **0.0** | **24.0** |

**Coefficient of Variation (CV)**

It is a normalized measure of dispersion of a probability distribution. It is calculated as the ratio of the standard deviation to the mean.

• Measure of relative dispersion

• Always a %

• Shows variation relative to mean

• Used to compare 2 or more groups

**Which Cricketer do you like? Who is more consistent?**

| Dravid | 150 | 150 | 130 | 125 | 145 | 110 | 100 | 152 | 120 | 50 | 128 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sehwag | 230 | 240 | 150 | 50 | 173 | 23 | 20 | 300 | 45 | 1 | 128 |

|  | **Dravid** | **Sehwag** |
| --- | --- | --- |
| **Mean** | **123.636** | **123.636** |
| **Median** | **128** | **128** |

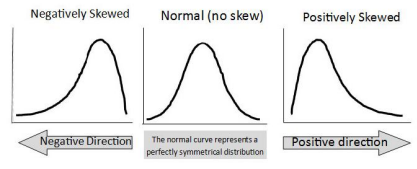
| **CV** | **24%** | **84%** |
| --- | --- | --- |

**Descriptive Statistics**

**Central Tendency:** is the middle point of distribution. Measures of Central Tendency include Mean, Median and Mode

**Dispersion:** is the spread of the data in a distribution, or the extent to which the observations are scattered.

**Skewness:** When the data is asymmetrical ie the values are not distributed equally on both sides. In this case, values are either concentrated on low end or on high end of scale on horizontal axis.

If the trail is to the right or positive 

end of the scale, the distribution is

said to be “positively skewed”.

If the distribution trails off to the left

or negative side of the scale, it is said

to be “negatively skewed”.

**Outliers**

**An outlier is an observation that is numerically distant from the rest of the data.**

**An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution.**

Example: Bill Gates makes $500 million a year. He’s in a room with 9 teachers, 4 of whom make $40k, 3 make $45k, and 2 make $55k a year. What is the mean salary of everyone in the room? What would be the mean salary if Gates wasn’t included?

Mean With Gates: **$50,040,500** Mean Without Gates: **$45,000**

A **Scatterplot** is useful

45

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

40

for "eyeballing" the

35

presence of **outliers**. **We can also use** 30

25

20

15

10

5

0

0 2 4 6 8 10

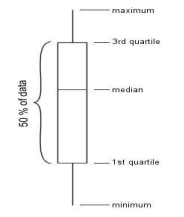
**Standard Deviation to identify Outliers!**

**Other Measures of Dispersion**

**Box-plot**

• Reveals the spread of the data

• Outliers defined using the

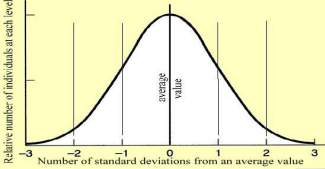
Q1 - 1.5(Q3-Q1) and Q3 + 1.5(Q3-Q1) outlier

**Normal Distribution**

Normal distribution is a pattern for the distribution of a set of data which follows a ***bell shaped curve***. This also called the *Gaussian distribution*.

The normal distribution is a theoretical ideal distribution. Real-life empirical distributions never match this model perfectly. However, many things in life do approximate the normal distribution, and are said to be “normally distributed.”

• Normal Distribution has the mean, the median, and the mode all

coinciding at its peak 

• The curve is concentrated in the center and decreases on either side ie

most observations are close to the mean

• The bell shaped curve is symmetric and Unimodal

• It can be determined entirely by the values of mean and std dev

• Area under the curve = 1

• **The empirical *68-95-99.7 rule* states that for a normal distribution:**

• **68.3% of the data will fall within 1 SD of mean**

• **95.4% of the data will fall within 2 SD’s of the mean**

• **Almost all (99.7%) of the data will fall within 3 SD’s of the mean**

**Standard Normal Distribution**

**Standard Normal distribution is a special case of the Normal distribution which has a mean of 0 and a standard deviation of 1 Any normal distribution can be converted to a Standard normal distribution through:**

Example: If X is a continuous random variable with a mean of 40 and a standard deviation of 10, what proportion of observations are a) Less than 50 b) Less than 20 c) Between 20 and 50

a) P(x<50)? 50-40

b) P(x<20)? 20-40

c) P(20<x<50)?

20-40 50-40

Z= =1 Z= = -2

< Z <

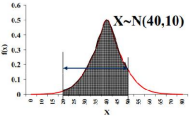
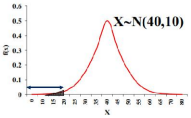
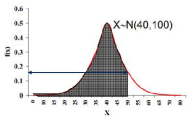
10

P(x<50) = P(z<1) = .8413

10

P(x<20) = P(x<-2) = .0228

10 10 P(-2 < Z < 1) = .8185



**Standard scores**

**A standard score (also called Z score) is the number of standard deviations that a given raw score is above or below the mean.**

• All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation.

• Standardizing variables helps to get variables to the same scale, or makes the variable unit free For eg: If we want to input Income (in INR) and Number of calling minutes into the same analysis, we will have to standardize both variables to get them to the same scale

=*X*

− **

*Z*

**

**Central Limit Theorem**

It is always not possible to get the true information about the population. In this case we have to live with samples. For eg: we don’t know the actual average income for India, but can estimate it based on a random sample picked from the Indian population

**In this case, the average we have is not the population average µ but an estimate X**

If we take a similar second sample, it is extremely unlikely that the average calculated for the second sample will be the same as the average calculated for the first sample. In fact, statisticians know that repeated samples from the same population give different sample means.

They have also proven that the distribution of these sample means will always be normally distributed, regardless of the shape of the parent population. This is known as the Central Limit Theorem.

***A distribution with a mean µ and variance σ², the sampling distribution of the mean approaches a normal distribution with a mean (µ) and a variance σ²/N as N, the sample size increases.***

The amazing and counter-intuitive thing about the central limit theorem is that *The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal distribution from which the average is computed is decidedly non-Normal.*

As the sample size n increases, the variance of the sampling distribution decreases. This is logical, because the larger the sample size, the closer we are to measuring the true population sample size, the closer we are to measuring the true population parameters.

**Standard Error**

Since all samples drawn from a population are similar BUT NOT the same as population, we calculate a Standard Error. ***Standard Error is the standard deviation of the sample means from the population mean*** Also, Standard Error ultimately converges to the Standard Deviation of the population.

**Standard Error = σ**

**N**

**Populations and Samples**

So far we have determined the results associated with individual observations or sample means when the true population parameters are known. In reality, the true population parameters are seldom known. We now learn how to infer **levels of confidence**, or a measure of accuracy on parameters, estimated using samples

***POINT ESTIMATOR***

• If we take a sample from a population, we can estimate parameters from the population, using sample statistics • Example: Sample mean (x) is our best estimate of the population mean (µ)

• Whereas, we really don’t know how close the estimate is to the true parameter

• The mean annual rainfall of Melbourne is 620mm per year

***INTERVAL ESTIMATOR***

• If we estimate a range or interval within which the true population parameter lies, then we are using an interval estimation method

• This is the most common method of estimation. We can also apply a level of how confident we are in the estimate • In 80% of all years Melbourne receives between 440 and 800 mm rain

**Sampling methodologies**

Sampling is required because it is seldom possible to measure all the individuals in a population. Researchers hence, use samples and infer their results to the population of interest

Eg: Election polls, market research surveys, etc

For a sample to be a “good sample”, it is imperative that there is a good sample size and there is no biasness in the sample.

**Simple Random Sample**

is one in which every member of the population is equally likely to be measured

Eg: Allocate a number to each member of the population and use a random number generator to determine which individuals will be measured

**Stratified Sampling**

separates the population into mutually exclusive groups and randomly samples within the groups

Eg: Randomly select a number of people within each demographic cell, while maintaining overall proportions like gender ratio, income ratio, etc

**Other methodology**

**Cluster sampling:** is used when there is a considerable variation within each group but the groups are essentially similar to each other. Here we divide the population into groups, or clusters, and then select a randomsample of these clusters.

**Confidence Intervals**

Because we know the properties of the normal distribution so well, we can use these properties to assist us in applying confidence intervals to estimates. This is essentially the interval estimator range.

For example, we know that 68.3 % of sample means lie within one standard error of the true population mean. Therefore, if we know the true population variance, we can infer the range within which we can be 68.3% confident that the

true population mean lies

Example:

X ~ N(μ, 3.62) grams, n = 36, x = 25.5 We can be 68.3% confident that μ is within the interval x ± SE

= 25.5 ± 3.6/ √ 36

= 25.5 ± 0.6

= (24.9 to 26.1) grams

**0.6 0.5 0.4 0.3 0.2 0.1 0**

****

**-se +se68.3%**

-4 -3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3 4 **Standard Errors**

**Confidence Intervals**

We can extend this principle further:

• We can be *90% confident* that the true population mean lies within *x ± 1.645(SE)* • We can be *95% confident* that the true population mean lies within *x ± 1.960(SE)* • We can be *99% confident* that the true population mean lies within *x ± 2.576(SE)*

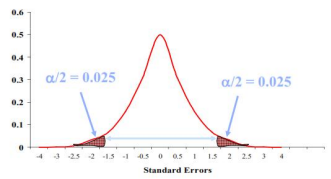
**P - value**

• Furthermore, the area outside the confidence interval is cumulatively known as α (alpha) • Confidence Interval = 1 - α

• Example: for 95% confidence interval, ∝=0.05

• α is also known as p-value.

• Hence, p-value is the probability that a randomly picked sample will have the mean **lying outside the confidence interval**.

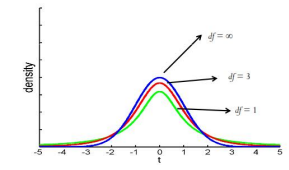


**Student’s t - distribution**

• While z-distribution is for the population, t-distribution is for the sample distribution.

• Hence, the shape of ‘t’ sampling distribution is similar to that of the ‘z’ sampling distribution in that it is a) Symmetrical

b) Centered over a mean of zero

c) Variance depends on the sample size, more specifically on the degrees of freedom (abbreviated as df) • As the number of degrees of freedom increases , variance of the t distribution approaches more closely to that of z • For n ≥ 30, shapes are almost similar 

• For n of 30 taken as dividing point between small & large samples

• t-test for population mean is:

X - µ0

s/ ~~n~~

When n < 30

**When to use**

***z-test:***

• σ is known and the population is normal

• σ is known and the sample size is at least 30. (The population need not be normal)

***t-test:***

• Whenever σ is not known

• The population is assumed to be normal

• And n<30

• The correct distribution to use is the ‘t’ distribution with n-1 df

**Hypothesis testing**

• In a Test Procedure, to start with, a hypothesis is made.

• The validity of the hypothesis is tested.

• If the hypothesis is found to be true, it is accepted.

• If it is found to be untrue, it is rejected.

• The hypothesis which is being tested for possible rejection is called null hypothesis • Null hypothesis is denoted by H0

• The hypothesis which is accepted when null hypothesis is rejected is called Alternate Hypothesis Ha • Ex. Ho: The drug works –it has a real effect.

Ha: The drug doesn’t work - Any effect you saw was due to chance.

**Hypothesis testing**

Hypothesis tests consist of the following steps: ▪ Null Hypothesis

▪ Alternative Hypothesis

▪ Confidence Level

▪ Decision Rule

▪ Test statistic

▪ Decision

**Hypothesis testing**

▪ Null hypothesis - We always assume the null hypothesis is true, or at least is the most plausible explanation before we do the test. The test can only ***disprove*** the null hypothesis. ▪ Alternative hypothesis - The alternative hypothesis is the hypothesis that we set out to test for. It is the hypothesis that we wish to ***prove***.

▪ Decision Rule - After we know the null and alternative hypotheses and the level of confidence associated with the test, we determine the points on the distribution of the test statistic where we will decide when the null hypothesis should be rejected in favor of the alternative hypothesis

▪ Use the terminology “ **Reject HO**” or “**Do not reject Ho**”. Never say “Accept Ho”

**Type I and Type II Error**

Process of testing a hypothesis indicates that there is a possibility of making an error. There are two types of errors:

Type I error: The error of rejecting the null hypothesis H0 even though H0 was true. Type II error: The error of accepting the null hypothesis H0 even though H0 was false.

**Example 1**

Suppose that we have been told that the price of petrol in Melbourne is normally distributed with a mean of 92 cents per litre, and a standard deviation of 3.1 cents/litre. To test whether this price is in fact true, we sample 50 service stations and obtain a mean of 93.6 cents/litre

**Solution:**

• Step 1: State the null and alternative hypotheses

• Ho: μ = 92

• Ha: μ ≠ 92

• Step2: Determine the appropriate test statistic and it’s distribution

Because we know the population standard deviation, we can use the z distribution

• Step3: Specify the significance level, Say α = 0.05

**Example 1 (contd.)**

• Step 4: Define the decision rule.

Using a z distribution (from tables), if α = 0.05, the rejection region is > +1.96 and < -1.96

**0.6**

**0.5**

**0.4**

i.e., if the test statistic is > 1.96 or < -1.96,

we will reject Ho and

**0.3 0.2 0.1 0**

**Rejection** 

**Region Rejection Region**

****

accept Ha

**Example 1 (contd.)**

• Step 5: Calculate the test statistic

*x ~~x~~*

** −**

−

=− *Z* = 93.6 92

= 3.653.1/ 50

*Zx o*

=

**

*SE*

*n*

• Step 6: Make a decision and answer the question: As, 3.65 > 1.96, the test statistic is in the rejection region

Reject Ho, accept Ha as a more plausible explanation **0.6**

• Step 7: Write your conclusion in the context of the aims of the study. **0.5**

**0.4**

**“The average price of petrol in Melbourne was**

**0.3**

**significantly different to 92 cents/litre”**

**0.2**

**0.1**

**Rejection Region** 

**Rejection**

**Region**

**0**

**-1.96 +1.963.65**

**Example 1 (contd.) – Importance of sample sizes**

• Consider the petrol prices in Melbourne example. If the sample size we had used was only 10, rather than 50, the test statistic would have been;

**93.6 92 Z** = **1.63**

=

− **3.1/ ~~10~~**

In which case we would not have rejected the Ho

**Example 2**

A company pays production workers $630 per week. The union claims that these workers are paid below the industry average for their work. A sample of 15 workers from other sites gives a mean wage of $670/week with a standard deviation of $58/week. Is the unions claim justified?

**Solution:**

Step 1: Ho: μ =< $630 (industry weekly average is not significantly different to $630)

Ha: μ > $630 (The industry weekly average is greater than $630)

Step2: Test Statistic - As we don’t know the population variance, and the sample size is < 30, we shall use the t test. Step3: Significance level - We will use α = 0.10 (as we want to be liberal rather than conservative)

Step 4 : Decision rule - From ‘t’ table, t **(0.1, 14df)** = 1.345**Non-Rejection**

**Region (1- =**

**0.90)**

**Rejection Region ( = 0.10)**

**1.345**

**Example 2 (contd.)**

Step 5: Calculate test statistic;

*x  x* −**

−

670 630 =−

t = 2.67*t* =

=

*x x SE ~~s~~*

58 15

*n*

Step 6: Make a decision - As 2.67 is > 1.345, we will reject the Ho

**Non-Rejection**

**Region (1- = 0.90) Rejection**

Step 7: Conclusion - “Production workers at the company earn an average of $40 per week less than the industry

standard (t = 2.67, df = 14, p < 0.1)”

**Region**

**( = 0.10)**

**1.345**

**2.67**

**P-value**

In fact, with the advent of computers it is simple to calculate the exact probability of a test statistics.

Example: P(t ≥ 2.67) = 0.00915

i.e. The area under the curve is 0.00915

** = 0.00915 2.67**

With a z test, it is possible to determine the exact probabilities using the table in the text book Ex.: What is the exact probability of obtaining a value of z ≤ 1.5?

From table, when z = 1.5, p = 0.5 + 0.4332,

P (z ≤ 1.5) = 0.9332

**A Practical Example**

****

**A Practical Example**

****

**A Practical Example**

****

**A Practical Example**

****

**A Practical Example**

****

**Comparison of two populations**

**Hypothesis testing for two samples:**

• Difference between independent samples & dependent samples • Two sample z test for means using independent samples • Two sample t test for means using independent Samples • Two sample t tests for means using dependent Samples

**Chi-square test**

Two properties are associated if the probability of having one property affects the probability of having another. Sometimes it is not known whether two properties are associated or not. What is required is a test of association, or, what is equivalent, a test of independence.

The Chi-square (χ 2) distribution can be used as a test of independence.

**Example:**

A psychologist conducted a survey into the relationship between the way in which a calculator was held and the speed with which 10 arithmetical operations were performed. The calculator could be either placed on a table or held in the hand; the sums could be performed in either less than 2 minutes, between 2 and 3 minutes or more than 3 minutes.

The following results were obtained for a sample of 150 children between 12 and 13 years old.

|  |  | **Mode of Computation** | |
| --- | --- | --- | --- |
|  |  | **On Table** | **Hand Held** |
| **Speed Of**  **Computation** | <2 | 28 | 12 |
| 2-3 | 25 | 35 |
| >3 | 21 | 29 |

**Example (contd.)**

**Solution:**

Step 1: Ho: Speed and mode are independent

Ha: Speed and mode are associated

Step2: In order to determine whether the two variables are associated it is necessary to calculate what the frequencies would be if there was absolutely no connection between them, or as we call them “Expected

Frequencies”

|  | **Table** | **Hand** |
| --- | --- | --- |
| **<2** | 40\*74/150 =  19.73 | 40\*76/150 =  20.27 |
| **2-3** | 60\*74/150 =  29.60 | 60\*76/150  =30.40 |
| **>3** | 50\*74/150  =24.67 | 50\*76/150 =  25.33 |

**Example (contd.)**

Step 3: Now we have to use the expected and observed frequencies to calculate a test statistic. ( ) ∑−

2

*O E*

The χ 2 test statistic is determined by

*i i*

*E i*

Step 4: In order to compare this with a critical value, we need to know the degrees of freedom of statistic. ν=degrees of freedom=( row number −1)⋅ (column number −1)

9.329 5.992 2 2 *Then  test* = > * critical* =

Step 5: Therefore, we reject *HO* and accept *H1*. The result is significant at the 0.05 or 5% level. This means that there is a 5 in 100 probability that the difference between the two conditions could have arisen by chance.

According to these results the way you use your calculator does affect the speed with which you do a calculation.

**ANOVA**

• Analysis of variance is a statistical technique used for comparing the means of different samples and deciding whether they are drawn from the same population or different populations. • Main Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

• The ANOVA F-statistic is a ratio of the Between Group Variation divided by the Within Group Variation: *Between F* = ~~=~~

*MSTR*

*Within*

*MSE*

• A large F is evidence against H0, since it indicates that there is more difference between groups than within groups.

**Covariance**

• A statistical measure of the joint variability of two random variables.

• An extension to *variance* but variance – *talks about the spread of a variable* and covariance– *talks about – spread of one variable wrt to other*.

• Mostly, tries to establish a linear relationship between two variable - wherein a change inone variable reciprocated by an equivalent change in another variable

• Value ranges from –Inf to Inf and large covariance denotes a strong relationship but cannotbe applicable to data with different measures and units.

• A zero value denotes no relationship.

**Correlation**

• A statistical technique that can show whether and how strongly pairs of variables arerelated or whether there exists a mutual relationship or connection between two or more things.

• Positive Correlation: If the values of two variables changing with same direction. Ex. Demand and CostPrice, Experience or Skills and CTC

• Negative Correlation: When the values of variables change with opposite direction. Ex. OperatingCost and TotalProfit



**Correlation Coefficient – Pearson’s Coefficient (r)**

• Also called Simple correlation or product moment correlation coefficient ***it measures the nature and strength between two variables of the quantitative type.*** • The value r (-1 to 1) r denotes **strength.**

• The sign of r denotes **nature** – positive, negative or no correlation strong intermediate weak weak intermediate strong

-1 0 1

-0.75 -0.25 0.25 0.75

**indirect Direct**

no relation perfect correlation

perfect correlation

**Correlation Coefficient – Spearman Rank (rs)**

• A non-parametric measure of correlation, procedure makes use of the two sets of ranksthat may be assigned to the sample values of x and Y.

• Spearman Rank correlation coefficient could be computed in the following cases: • Both variables - quantitative.

• Both variables - qualitative but ordinal.

• One variable is quantitative and the other is qualitative ordinal. • The value of rs denotes the magnitude and nature of association giving the same interpretation as simple r.

**Correlation – Examples**

| **Age (years)** | **Weight (Kg)** |
| --- | --- |
| 7 | 12 |
| 6 | 8 |
| 8 | 12 |
| 5 | 10 |
| 6 | 11 |
| 9 | 13 |





**Covariance vs Correlation**

| **BASIS FOR COMPARISON COVARIANCE CORRELATION**  Covariance is a measure indicating  Correlation is a statistical measure  Meaning  the extent to which two random  that indicates how strongly two  variables change in tandem.  variables are related.  What is it? Measure of correlation Scaled version of covarianceValues Lie between -∞ and +∞ Lie between -1 and +1 Change in scale Affects covariance Does not affects correlationUnit free measure No Yes |
| --- |

**Non - Parametric tests**

• Deals with enumerative data

• Does not deal with specific population parameters

• Does not require assumptions about specific population distributions (in particular , the assumption of normality) • Non-Parametric tests ignores the magnitude of information contained in observations

• Use either frequencies or ranks (categorical or ordinal)

• Non-parametric tests are called ''non-parametric'' because they do not make any assumption about a population parameter.

• In other words, when we apply a non-parametric test we do not have to make assumptions about mean of a population, its variance or background probability distribution.

• Thus, non-parametric tests are not as powerful as parametric tests they are of more general application and are available when the parametric tests fail.

• Basically, there is at least one non-parametric equivalent for each parametric general type of test • Non-Parametric tests broadly fall into the following categories:

• Tests of differences between independent samples: The Mann-Whitney U test (t-test for independent samples), The Kruskal- Wallis H test (ANOVA), The Kolmogorov-Smirnov test (t-test for independent samples)

• Tests of differences between dependent samples: Wilcoxon Mann-Whitney Test (t-test for independent samples)

Some Commonly Used Normal Statistical & NonParametric

**Normal theory basedtest**

t test for independent

| **Corresponding**  **nonparametric test** |
| --- |
| Mann-Whitney U test;Kolmogorov-Smirnov 2sample test |
| Wilcoxon matched  pairs signed-rank test |
| Kruskal-Wallis analysisof variance by ranks |

**Purpose of test**

Compares two

samples

independent samples

Examines a set of

Paired t test

differences

One way analysis of

Compares three or

variance (F test)

more groups

Two way analysis of variance

Compares groups

Friedman Two way

classified by two

analysis of variance

different factors

**ADDITIONAL REFERENCE SLIDES**

**OBJECTIVE**

✔ Identify situations that contain dependent or independent samples. ✔ Calculate the test statistic to test hypotheses about dependent data pairs as well as independent data pairs.

✔ Understand the shortcomings of comparing multiple means as pairs of hypotheses.

✔ Understand the steps of the ANOVA method and the method’s advantages.

✔ Understand the differences in situations that allow for ANOVA or MANOVA methods.

✔ Know the procedure of two sample t-test, ANOVA, and MANOVA and their application through technological tools.

| **Part I: Two Sample t-test** Introduction |
| --- |
| Independent Samples t-test |
| Dependent Samples t-test |
| Applications |
| Code in R |

**Introduction**

• It is used when we are to perform a statistical analysis involving two samples.

• These two samples could be independent or dependent. • When we have independent samples we assume that the scores of one sample do not affect the other.

• Two samples of data are dependent when each score in one sample is paired with a specific score in the other sample. In short, these types of samples are related to each other.

• In such a case, we would analyze both samples and the hypothesis would address the difference between two sample means.

| **Part I: Two Sample t-test** Introduction |
| --- |
| Independent Samples t-test |
| Dependent Samples t-test |
| Applications |
| Code in R |

Independent Samples t-test

• The two sample t-test is used for comparing the means of a quantitative variable (Y) in two populations.

• Our goal is comparing μ1 and μ2(which in practice is done by making inference on the difference μ1- μ2).

• The null hypotheses is Ho: μ1- μ2 = 0

• and the alternative hypothesis is one of the following (depending on the context of the problem):

• Ha: μ1- μ2 < 0

• Ha: μ1- μ2 > 0

• Ha: μ1- μ2 ≠ 0

Independent Samples t-test (2)

• The two-sample t-test can be safely used when the samples are independent and at least one of the following two conditions hold: • The variable Y is known to have a normal distribution in both populations

• The two sample sizes are large.

• When the sample sizes are not large (and we therefore need to check the normality of Y in both population), we look at the histograms of the two samples and make sure that there are no signs of non- normality such as extreme skewedness and/or outliers.

Independent Samples t-test (3)

• The test statistic is as follows and has a t distribution when the null hypothesis is true:



• P-values are obtained from the output, and conclusions are drawn as usual, comparing the p-value to the significance level alpha. • If Hois rejected, a 95% confidence interval for μ1- μ2 can be very insightful and can also be used for the two-sided test.

| **Part I: Two Sample t-test** Introduction |
| --- |
| Independent Samples t-test |
| Dependent Samples t-test |
| Applications |
| Code in R |

Dependent Samples (Paired) t-test

• The paired t-test is used to compare two population means when the two samples (drawn from the two populations) are dependent in the sense that every observation in one sample can be linked to an observation in the other sample.

• The most common case in which the matched pairs design is used is when the same subjects are measured twice, usually before and then after some kind of treatment and/or intervention.

• The idea behind the paired t-test is to reduce the data from two samples to just one sample of the differences, and use these observed differences as data for inference about a single mean—the mean of the differences, μd.

Dependent Samples (Paired) t-test (2)

• The paired t-test is therefore simply a one-sample t-test for the mean of the differences μd, where the null value is 0.

The null hypothesis is therefore:

Ho: μd = 0

and the alternative hypothesis is one of the following (depending on the context of the problem):

Ha: μd < 0

Ha: μd > 0

Ha: μd ≠ 0

Dependent Samples (Paired) t-test (3)

• The paired t-test can be safely used when one of the following two conditions hold:

• The differences have a normal distribution.

• The sample size of differences is large.

• When the sample size of difference is not large (and we therefore need to check the normality of the differences), we look at the histograms of the differences and make sure that there are no signs of non-normality like extreme skewedness and/or outliers.

Dependent Samples (Paired) t-test (4)

• The test statistic is as follows and has a t distribution when the null hypothesis is true:



• P-values are obtained from the output, and conclusions are drawn as usual, comparing the p-value to the significance level alpha. • If Hois rejected, a 95% confidence interval for μd can be very insightful and can also be used for the two-sided test.

| **Part I: Two Sample t-test** Introduction |
| --- |
| Independent Samples t-test |
| Dependent Samples t-test |
| Applications |
| Code in R |

**Applications**

• With the help of two sample t-test, we can also perform hypothesis testswith two samples. We can test two independent samples which are samples that do not affect one another or dependent samples which assume that the samples are related to each other.

• We can also test the likelihood that two dependent samples are related.

| **Part I: Two Sample t-test** Introduction |
| --- |
| Independent Samples t-test |
| Dependent Samples t-test |
| Applications |
| Code in R |

Code in R: Paired two sample t-test

• Data set of 214 undergraduate students that we surveyed both at the beginning and at the end of their semester. So among the many questions that we asked these students was this exclusive

variable. The question that they were asked is how likely they think on a scale of 1 to 10 they would be an exclusive relationship at the end of the year.

• 1 means that there’s no way they would be in an exclusive relationship. 10 means that they’re definitely going to be an exclusive relationship at the end of that year.

• Now, if we wanted to compare the mean response for the exclusive question at the beginning to the mean response at the end of the semester, we’re looking at two measurements within the same student. So to compare those means, we’re going to be doing a **paired two sample t-test.**

Code in R: Paired two sample t-test (2)

• The assumption of a paired t-test is that the distribution of the differences inmeasurements is normal. 

diff <- post$exclusive - post$post\_exclusive

**hist**(diff)

• It’s not perfectly normal, but there’s no huge outliers and there’s no skewness. Sowe are good to go on the normality assumption of our paired sample t-test.

Code in R: Paired two sample t-test (3)

• Using the **t.test()** function.

• Pass it our two variables that we want to compare separated by comma.

t.test(post$exclusive, post$post\_exclusive, paired = TRUE)

• So, this’ll tell R that these two variables that we’re giving it within the t.test function are actually from the same subjects.

Code in R: Paired two sample t-test (4)

• Output:



• The p-value is less than alpha, less than 0.05. So we would reject the null hypothesis that the mean response to the first exclusive variable is equal to the mean of the second.

Code in R: Independent Samples t-test

• Using the same data set, we’re going to look at the gender variable, where we have male or female students, and then how long they said they slept on Tuesday of that week.

• This variable is giving us an idea of how much sleep these students are getting on a typical weekday.

• So if we want to compare if the mean number of hours of sleep on Tuesday is the same for female students as it is for male students, we’re going to use the same **t.test()** function.

Code in R: Independent Samples t-test (2)

• Creating dedicated variables:

fsleep <- post$sleep\_Tues[post$gender == "Female"] msleep <- post$sleep\_Tues[post$gender == "Male"]

• This creates two vectors that contain the value of “sleep\_Tuesday” for both the female students and then the male students.

• So now I have two vectors of my two groups that I’m comparing with t-test.

Code in R: Independent Samples t-test (3)

• Checking assumptions: Are both the groups normally distributed? hist(fsleep)

hist(msleep)



• Both the distributions are roughly symmetric. There are no outliers. These are definitely approximately normal distribution.

Code in R: Independent Samples t-test (4)

• Conducting the t-test

t.test(fsleep, msleep)

• Output:



• So because our p value here is greater than 0.05, we would fail to reject the null hypothesis that the means of these two groups are equal.

| **Part II: ANOVA, MANOVA**  Introduction |
| --- |
| ANOVA |
| MANOVA |
| Applications |
| Code in R |

Introduction

• If our objective is to test the hypothesis that multiple population means and variances of scores are equal, we use ANOVA.

• Instead, we could also conduct a series of t-tests to determine if any of the sample means differ. However, this would be tedious and the probability of making one or more Type I errors multiplies exponentially.

• So, we use **Analysis of Variance (ANOVA)**, which can be used when wewant to test the means of three or more populations at once. This ANOVA is technically called “one-way” when it

has just one main grouping factor. We can also have an ANOVA with morethan one factor.