STATISTICS WORKSHEET-1 (Internship-23, Konatala Mohit ID-34)

1. Bernoulli random variables take (only) the values 1 and 0.

Answer: a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized,becomes that of a standard normal as the sample size increases?

Answer: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Answer: b) Modeling bounded count data

4. Point out the correct statement

Answer: d) All of the mentioned

5. __________ random variables are used to model rates.

Answer: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

Answer: b) False

7. Which of the following testing is concerned with making decisions using data?

Answer: b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

Answer: a) 0

9. Which of the following statement is incorrect with respect to outliers ?

Answer: c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

Normal Distribution is used to study any database on a particular property which varies with each and every specimen compared (eg: marks of student database in a particular subject).

When we plot the data on a histogram it will look like a bell curve (symmetric) which is known as normal distribution.

Normal Distributions are always centred on the average value. It is also known as Gaussian or gauss or Laplace-Gauss. It is a continuous probability distribution for a real valued random variable. The general form of its probability density function is

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

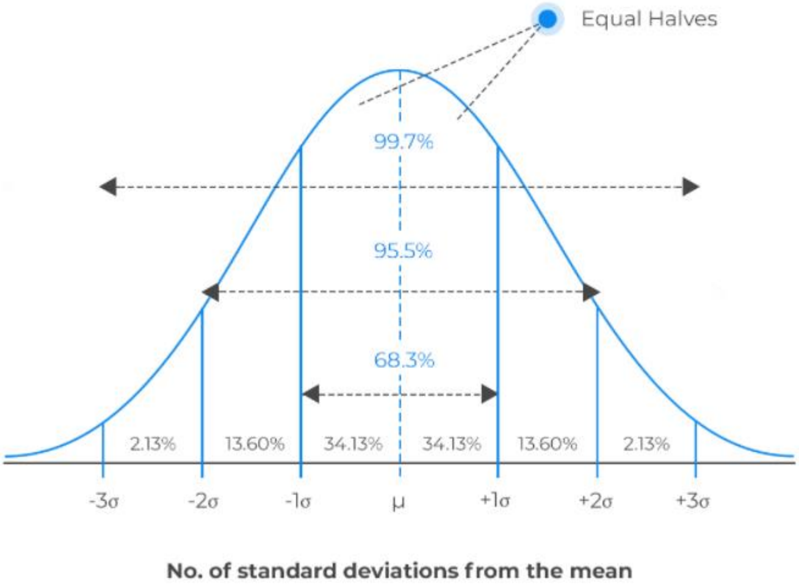
Where:

f(x) = probability density function

μ   = mean

σ   = standard deviation

The parameter μ is the mean or expectation of the distribution, while the parameter σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.



Note: 68, 95, 99.7 Law :- Outliers are extremely rare. It also suggest how much we know about the data set

68% fall between μ - σ ~ μ + σ

95% fall between μ - 2σ ~ μ +2 σ

99.7% fall between μ - 3σ ~ μ + 3σ

Knowing the standard deviation is helpful because normal curves are drawn such that 95% of the measurement falls between +/- 2 standard deviations around the mean.

To draw a normal distribution we need to know:

1. The average measurement. This tells us where the centre of the curve goes.

2. The standard deviation of the measurement, this tells us how wide the curve should be. And the width of the curve determines how tall it is. The wider the curve, the short it is and the narrower the curve, the taller it is.

In essence the curve tells us that there is a high probability of measuring i.e. +/- σ of the mean for narrow curve and vice-versa.

11. How do you handle missing data? What imputation techniques do you recommend?

Handling missing data is a vital part during the preliminary processing of the dataset as many machine learning algorithms do not support missing values.

The following are imputation techniques to handle missing values in the dataset:

1. Removal of Rows with missing values
2. Impute missing values for continuous/categorical variable
3. Other Imputation Methods
4. Using Algorithms that support missing values
5. Prediction of missing values

- **Dropping/Removal Rows with Missing Values**:

Missing values can be handled by deleting the entire row or column with contains the null values i.e. if the values of the particular are very less or doesn't affect the outcome in anyway i.e. impact on solution is very minute. By removing the null data the data-set much more robust but **if the data is large it is not recommended** as it can collapse the outcome.

- **Impute missing values with Mean/Median**:

The missing data can be filled with the mean/median of the row or column. Mean imputation works well with the small dataset. Even so mean imputation is **highly not recommended** as it does not preserve the relationship among variables and leads to miscalculation of standard errors

- **Other Imputation Methods**:

Few imputation techniques can be used handle the missing data depending on the type of data we are handling. In other words, for the data variable having longitudinal behaviour, it might make sense to use the last valid observation to fill the missing value. This is known as the Last observation carried forward (LOCF) method. For the time-series dataset variable, it makes sense to use the interpolation of the variable before and after a timestamp for a missing value.

- **Using Algorithms that support missing values**:

All the machine learning algorithms don't support missing values but few machine learning algorithms are strong enough to with stand missing values in the dataset. The K-nearest neighbor algorithm can ignore a column from a distance measure when a value is missing. Naive Bayes can also support missing values when making a prediction. These algorithms can be used when the dataset contains null or missing values. Another algorithm that can be used here is RandomForest that works well on non-linear and categorical data. It adapts to the data structure taking into consideration the high variance or the bias, producing better results on large datasets. This technique is **highly recommended as there is need to handle missing values** in each column as ML algorithms will handle them efficiently.

- **Prediction of missing values**:

In the earlier methods to handle missing values, we do not use the correlation advantage of the variable containing the missing value and other variables. Using the other features which don't have nulls can be used to predict missing values. The regression or classification model can be used for the prediction of missing values depending on the nature (categorical or continuous) of the feature having missing value. This technique is **recommended** because it gives a better result than earlier methods and it takes into account the covariance between the missing value column and other columns.

Every dataset has missing values that need to be handled intelligently to create a robust model. There is no thump rule to handle missing values in a particular manner, the method which gets a robust model with the best performance. One can use various methods on different features depending on how and what the data is about. Having domain knowledge about the dataset is important, which can give an insight into how to preprocess the data and handle missing values

12. What is A/B testing?

A/B testing also known as split or bucket testing is an experimental process used to test two variants based on the user experience research data. At a high level, A/B testing is a statistical way of comparing two or more versions i.e. to determine not only which version performs better but also to understand if a difference between two version is statistically significant.

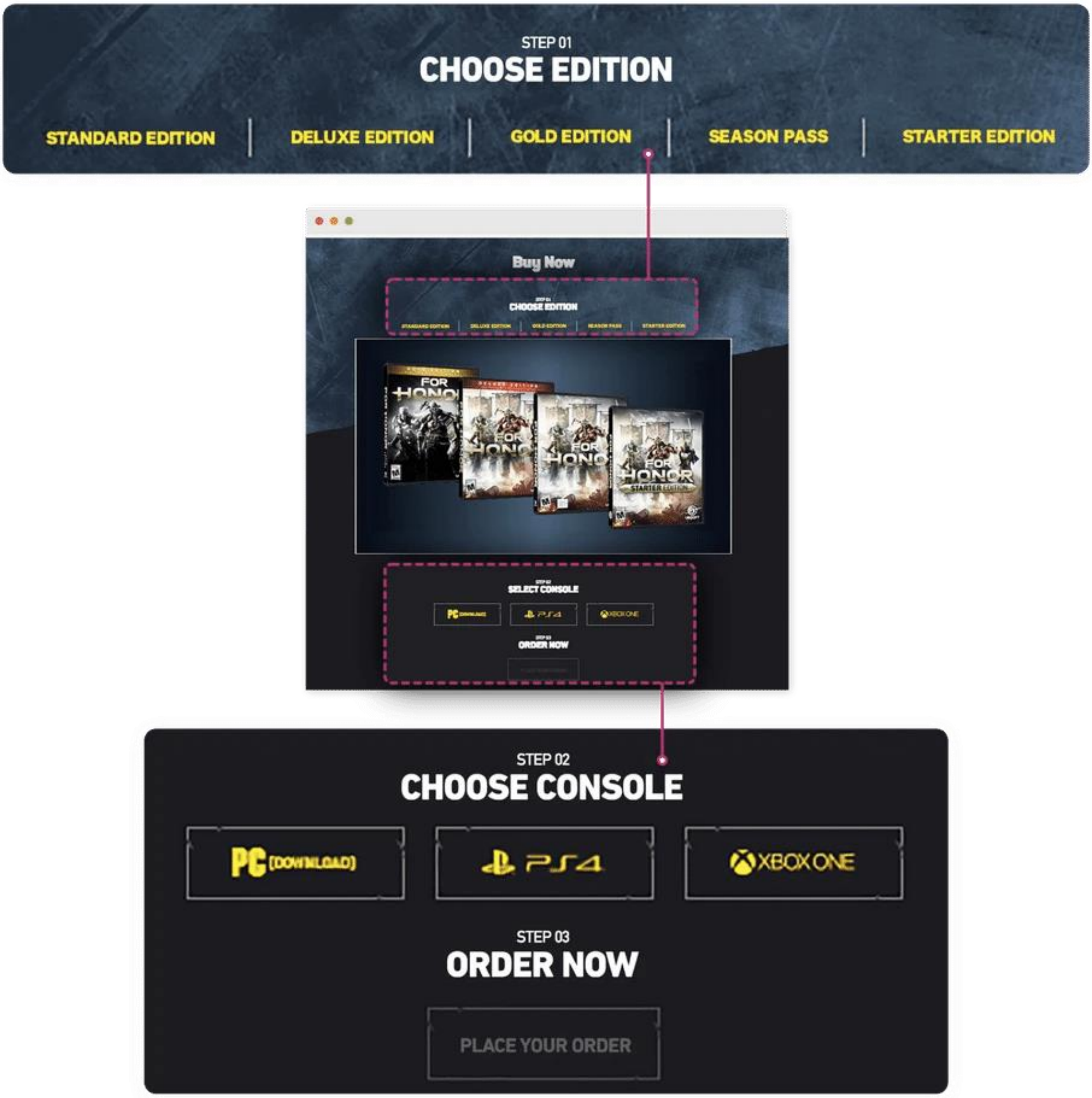
Most of the businesses in today era are all opting for data-driven approach. A common quandary that most of the companies face understands of the customers, as there is vast deviation in the behaviour it is difficult to undermine the each and every customer that is what they would think consciously or subconsciously. Most of the times the customer doesn't even notice what choices they make, they just do it, but when the experiment or A/B test is conducted we might find out otherwise and the results can often be very overwhelming and customers can behave much differently than we would think so it's best to conduct tests rather than relying on instinct.

Case study on A/B testing:

Ubisoft used A/B testing to increase its lead generation by 12%

Ubisoft Entertainment is one of the main French online game companies. It's maximum acknowledged for publishing video games for numerous fairly famed online game franchises which include For Honor, Tom Clancy's, Assassin's Creed, Just Dance, etc., and handing over memorable gaming experiences. For Ubisoft, lead era and conversion fee are key metrics to research its average performance.

While a number of its pages have been acting properly in phrases of lead era and conversion fee, its ‘Buy Now’ web page committed to the ‘For Honor’ emblem wasn’t yielding the first-rate of results. Ubisoft’s group investigated the matter, accumulated tourist information the use of clickmaps, scrollmaps, heatmaps, and surveys, and analyzed that their shopping for technique become too tedious. The organization determined to overtake For Honor’s Buy Now web page completely – lessen the up and down web page scroll and simplify the whole shopping for technique. Here’s now the manage and version seemed like:



STEP 01

CHOOSE EDITION

STANDARD EDITION | DELUXE EDITION | GOLD EDITION | SEASON PASS | STARTER EDITION

STEP 02

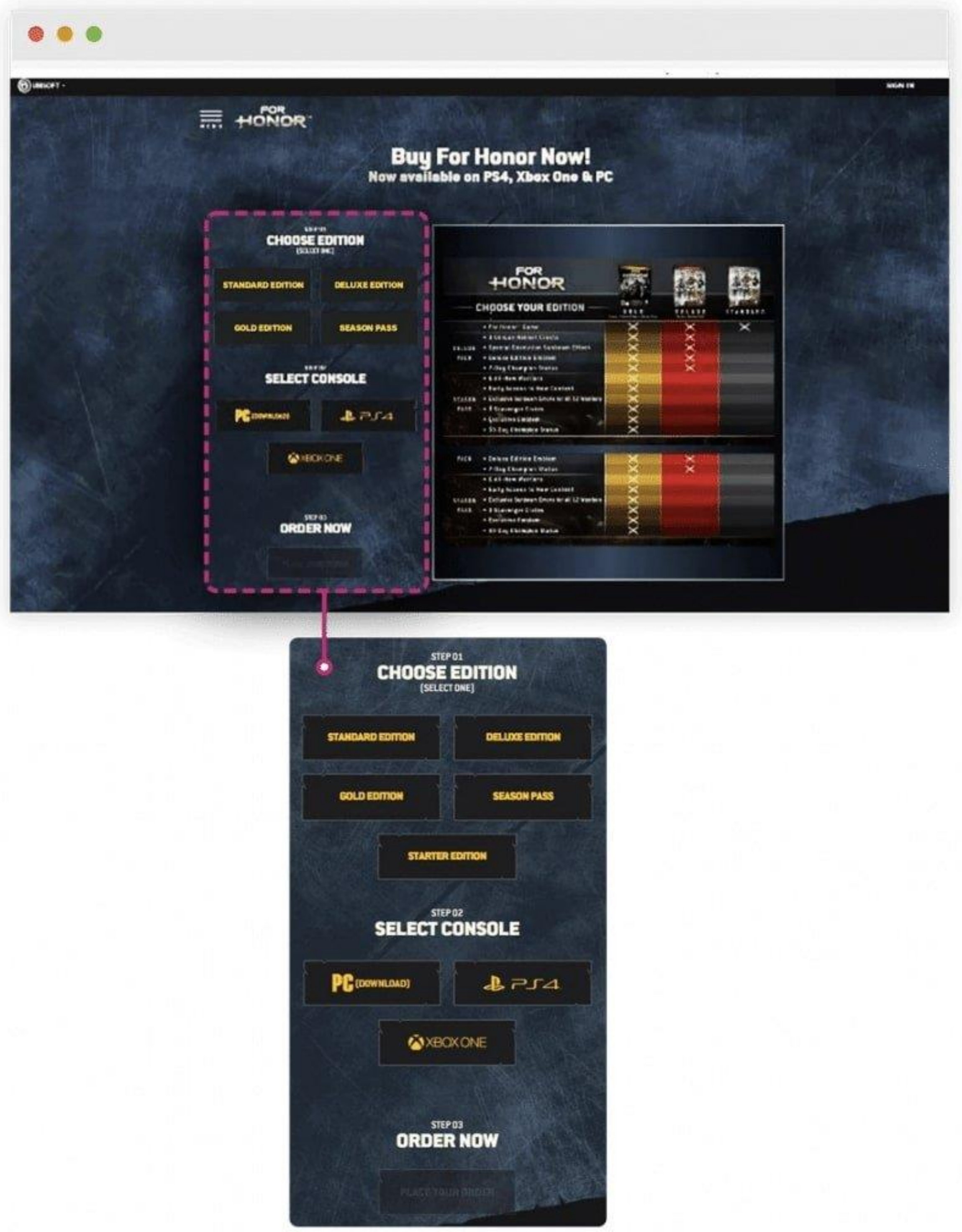
CHOOSE CONSOLE

PC (DOWNLOAD) | PS4 | XBOX ONE

STEP 03

ORDER NOW

PLACE YOUR ORDER



After running the test for about three months, Ubisoft saw that variation brought about more conversions to the company than the control. Conversions went up from 38% to 50%, and overall lead generation increased by 12%.

Criteria for success of A/B testing is one must call out the required outcome necessary prior to the testing i.e. splitting the traffic into two part not necessarily into two equal halves but we want to figure out what is the minimum number of people one need to run the A/B testing on to achieve statistically significant results.

We can do this type of test on multiple version i.e. taking to sampling with all the similarities except for the colour such type are called a multivariate test or a full factorial test since we are comparing different factors.

A/B testing is mostly advantageous as it doesn't get much effect even if something is broken in the data, or the data is messy or there is too much noise in the data or there is some sampling error. The impact can be 1~2% but one must make sure that A/B testing is conducted properly first by setting up an A/A test.

13. Is mean imputation of missing data acceptable practice?

Imputation methods are used to handle the real world data which has missing values in the dataset. A method that is often utilized is mean imputation also known as mean substitution where missing values are substituted with the variable mean value acquired from the data sample.

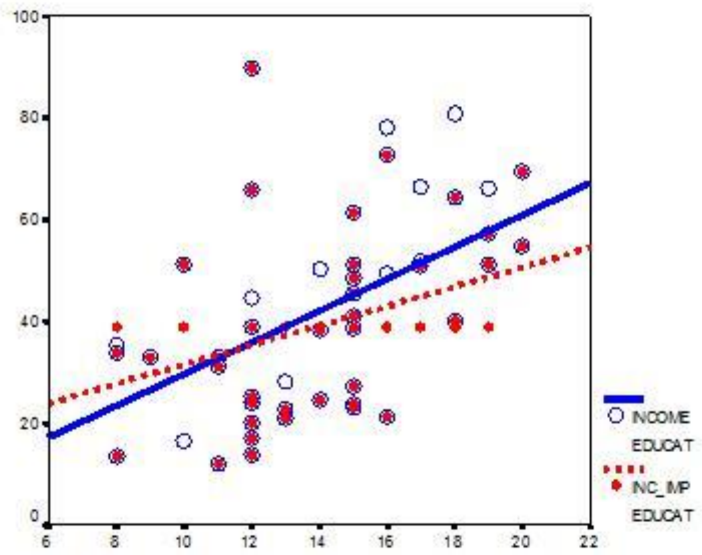
Generally mean imputation conserves the mean of the sample data, i.e. if the data is missing at any random interval the estimated of the mean value acquired is to take its place and remain unbiased. But most often one is not interested in filling the null data rather one is interested in seeking the relationship between the variables of data sample. Therefore it isn't an acceptable data imputation practice.

Major two problems in mean imputation which makes it an unacceptable practice:

I) It does not preserve the relationship among variables:

Mean imputation does keep mean variable i.e. achieved unbiased, in-addition to this it also keeps the sample size fully intact, But all we are doing is to find the average variable data and substitute it at all the random intervals were the data is missing. Due to this it induces standard error.

The below graph demonstrates why mean imputation isn't an ideal approach:



This graph illustrates hypothetical records among x=years of training and y=annual profits in thousands with n=50. The blue circles indicate the authentic data, and the striking blue straight line shows the best fit regression line for the in general data sample. The correlation among x and y is measured r = .53.

To illustrate why mean imputation isn't a good practice I then aimlessly removed 12 observed values in annual profits (y) and substituted them with the mean of the sample data.
The crimson dots are the replaced mean data of the sample set. The blue line circles with crimson filling in it are representation of the non-missing records. The empty blue circles indicate the lacking

records. If we once take a glance throughout the graph at Y = 39, one can see there is row of crimson filled circles. These circles represent the imputed values. The dotted crimson line is generated after the random values were replaced by the mean which altered the best fit regression line. As one can notice there is drastic change in the regression line when compared to the original data regression line. Replacing the random values with mean value of the dataset pulled it down. The new correlation is r = .39. That's lots smaller when compared to the original data's .53.

Due to inducing the mean substitution the relationship between the original data is underestimated. Of course, in an actual data set, you wouldn't observe so without difficulty the unfairness you're introducing. This is major drawback of this solution as when we looking to clear up the missing data with we tend to introduced another problem with create an impact in our desired output.

An important point to note is that if x have been lacking in preference to y, mean imputation could falsely blow up the correlation.

Simply put, we think by filling the void with mean value we have established a strong relation but isn't the case and to top it off it is no longer exact either. It's now not even reproducible and also we are never required an over exaggerating result. This type of solution is so exact at keeping impartial estimates for the mean that it isn't so exact for impartial estimates of relationships.

II) Mean imputation leads to miscalculation of Standard Errors:

The second cause applies to any form of single imputation. Any statistical data that uses the imputed information most of the times has minimum standard error. To say it simple words, yes we get the same mean data when compared with the mean imputed data which we got without any imputation. And yes, there are instances in which that mean value is impartial. But due to this the standard error of that mean can be very least (minimum) in some cases to be actually real and convincing. And matter of facts the imputations are themselves estimate, due to which there are some error related to them. But the statistical programming language doesn't understand that. It thinks it as actual information. In the end, due to the fact that the standard errors are very undersized, so are the p-values. Now it is making Type I errors without realizing it.

14. What is linear regression in statistics?

The primary aim in statistics is to find solution for the variable X associated with the variable Y i.e. to find the relation between variable X and Y which is used to predict Y. The procedure of training a model on a real data where there is a known outcome and the following application to data where the outcome is not known is called supervised learning. Regression is one of the sub-categories of supervised learning.

In statistics, Regression analysis is used in determining / estimating a relationship between a dependent variable Y and an independent variable X.

Simple linear regression provides a model of the relationship between the magnitudes of two variables. Unlike co-relation which strength the relation between two variable, Regression quantifies the nature of the relationship between the two variables.

Key terms used in Regression:

The variable that we try to predict is known as response.

The variable that is used to predict the response is known as independent variable.

The vector of predictor and outcome value for a specific individual or case is known as record.

The intercept of the regression line i.e. the predicted value when X=0 is called intercept (b_0, β_0).

The slope of the regression line is known as Regression coefficient.

The estimated $\hat{Y}$ obtained from the regression lines is called fitted values.

The difference between observed and fitted values is known as residuals.

Least squares is the method of fitting a regression by minimizing the sum of squared residuals.

Linear Regression Equation:

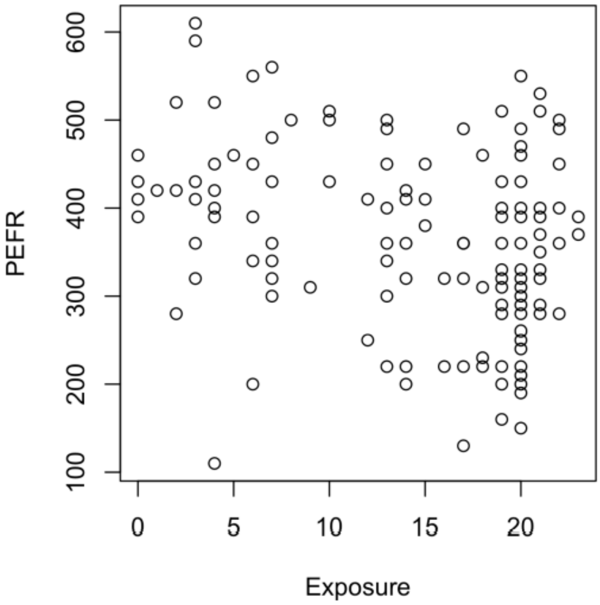
$$\boldsymbol{Y = b_0 + b_1 X}$$

Where,

Y is response and X is predictor

b_0 is constant and b_1 is intercept

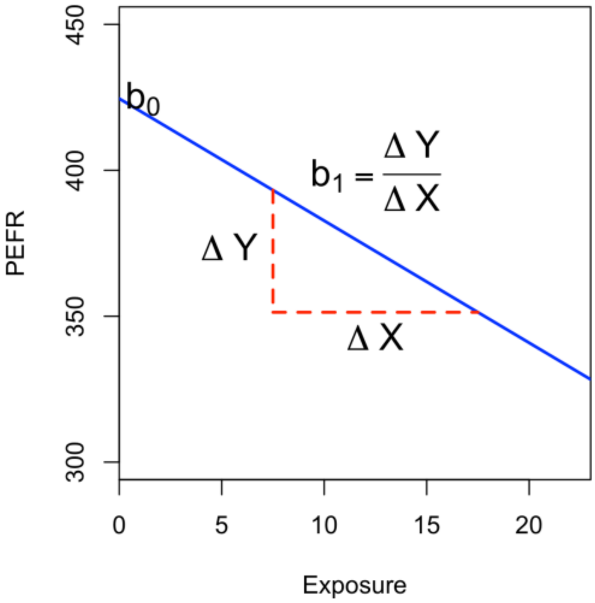
Demonstration of Linear Regression with an example:



The graph illustrates the scatter-plot of number of years a work was exposed to cotton dust (X-axis) versus measure of lung capacity i.e. peak expiratory flow rate – PEFR (Y-axis). By take a glance at the scatter-plot it is difficult to know i.e. is there any relation between X-variable and Y-variable.

Simple Linear Regression can help in finding the relation between the two variables i.e. finding the best line to the predict the response (PEFR) which is a function of predictor variable (Exposure)

$$PEFR = b_0 + b_1(Exposure) \quad \text{- I}$$



The above linear graph is the best line achieved using equation-I

The important concepts in regression are fitted values (prediction value) and the residual (prediction error). In general the data point doesn't exactly fall on the best line. So as to correct the equation we include explicit error e_i.

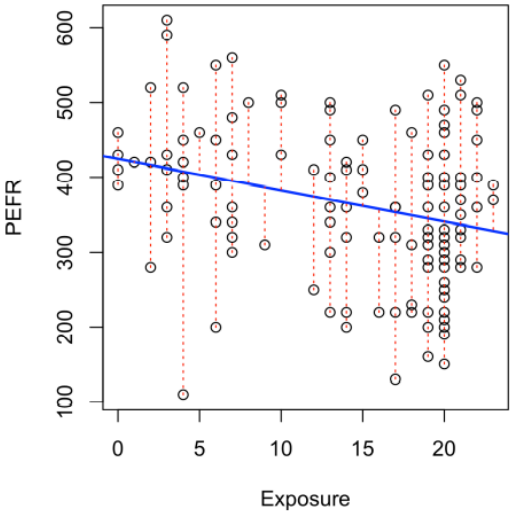
$$Y_i = b_0 + b_1X_i + e_i \quad \text{- II}$$

The fitted values (predicted values) is usually denoted with $\hat{Y}$

$$\hat{Y_i} = \hat{b}_0 + \hat{b}_1X_i \quad \text{- III}$$

The notations $\hat{b}_0$ and $\hat{b}_1$ indicate coefficients of estimated and known respectively

We can obtain the residuals $\hat{e_i}$ by substituting equation II equals to equation III



In the above graphical figure the lengths of the vertical red dashed lines are indicating the residual.

To determine if the model is best fit to the data we use Least squares i.e. the regression line is the estimate that minimizes the sum of square residual values

$$\text{RSS}= \sum_{i=1}^{n}\left(Y_i - \widehat{Y_\imath}\right)^2$$

$$\text{RSS}=\sum_{i=1}^{n}\left(Y_i - \ \hat{b}_0 - \hat{b}_1 X_i\right)^2$$

The estimate $\hat{b}_0$ and $\hat{b}_1$are the values that minimize RSS.

15. What are the various branches of Statistics?

Statistics is a branch of Mathematics that deals with data. There are four branches in Statistics but mainly divided into two branches:

- Mathematical / Theoretical Statistics
- Statistical methods or function

Main Branches of Statistics:

- Descriptive Statistics
- Inferential Statistics

Mathematical Statistics: In mathematical statistics we bring in probability theories to resolve the collection of data to achieve the final outcome. Mathematical approach such probability, algebra, differential equations etc. are some of the used techniques.

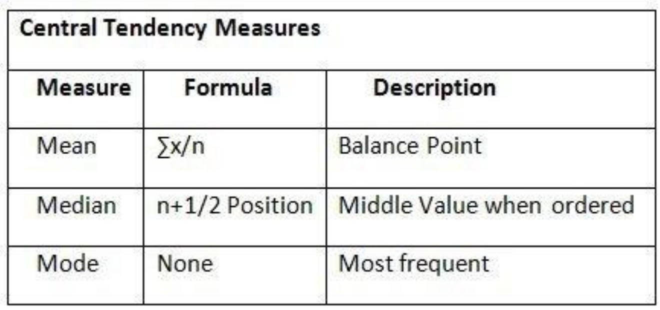
Statistical Method: It helps in collection, tabulation and interpretation of the data. It helps in analyzing the data and return insight from the data.

Descriptive Statistics: In Descriptive we present / organize the data and summarize the data. Descriptive Statistics helps in describing the characteristics of the data-set sample.

In descriptive statistics we determine if the data-set sample is normally distributed or not. Most statistical test requires the sample data-set to have a normal distribution, we also determine if the sample can be compared to the larger population (data-set). Descriptive Statistics are displayed as tables, charts, percentages, frequency distribution, and as reported measures of central tendency

Descriptive Statistics include the following information about the sample:

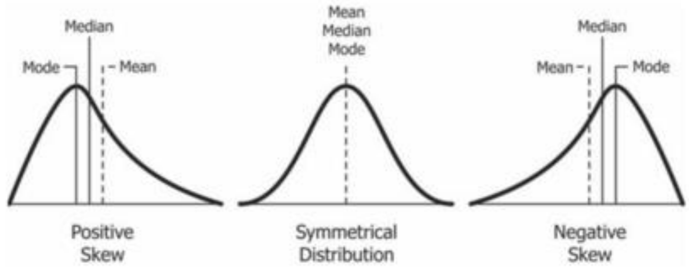
Central tendency: The sample mean (average), median (midpoint), mode (most frequently occurring number).



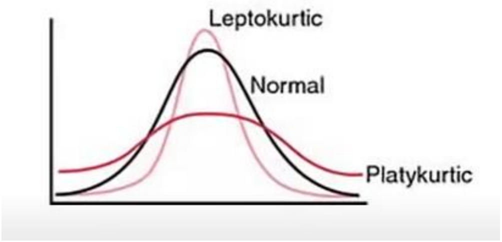
Central Tendency Measures		
Measure	Formula	Description
Mean	∑x/n	Balance Point
Median	n+1/2 Position	Middle Value when ordered
Mode	None	Most frequent

Measures of variability: range (the difference between largest and smallest variable), variance (how far the numbers are spread out), and the standard deviation (i.e. how much variation exists from the mean).

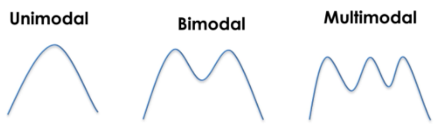
Skewness: How symmetric the distribution of variables is.



Kurtosis: Peakedness or flatness of a distribution



Shape: this includes modality and outliers.



Inferential Statistics: In Inferential we run various tests to draw conclusions about the data-set based on the data observed in the sample. To put it in simple words in inferential statistics the values that infer results of a sample to population from which the sample is drawn.

Inferential Statistics also known as null hypothesis testing uses probability to determine whether it is likely that a particular sample (outcome) is a representative of the population i.e. we test to see if the sample mean is same as the population mean and if its not its due to one of the two reasons i.e. chance or due to an effect caused due to the experiment we have done on the sample.

When we draw a sample from a population we expect the mean of the sample to the same as the mean of the population but when we do an experiment we tend to draw a large sample which we then randomly assign into two groups Experimental and Control group

Experimental group: The group that is given a treatment (testing).

Control group: This group is identical to the experimental group that is not given the treatment (no testing is done).

After the testing mean of both the control and experimental is compared to see if the testing has affected the results. If the mean is the same even after the testing the intervention which implies that the intervention hasn't worked but if the values differ then we can suspect that the intervention had an effect and this is called hypothesis testing.

Hypothesis is a testable prediction about real world phenomenon. Experimental hypothesis is a starting point that will be accepted or rejected by observing the evidence that supports (good) or contradicts it (rejected).

There are two types of Hypothesis:

Null Hypothesis: the sample mean is the same as the population mean. Any difference we observe is due to a chance.

Alterative Hypothesis: The sample mean is not the same as the population mean. Any difference we observe is due to an effect.