# STATISTICS WORKSHEET-1

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

Ans. 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 ?

Ans. A) Central limit theorem

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

Ans. C) Modeling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution.

b) Sums of normally distributed random variables are again normally distributed . even if the variables are dependent.

c) The square of a standard normal random variable follows what is called chi-Squared distribution

d) All of the mentioned

Ans. D) All of the mentioned

5. ______________ random variables are used to model rates.

Ans. Poisson

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

Ans. False

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

Ans. Hypothesis

8. Normalized data are centered at ______ and have units equal to standard deviations of the original data.

Ans. Zero

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

a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

Ans. C) Outliers cannot conform to the regression relationship.

10. What do you understand by the term normal distribution?

Ans. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve .

A normal distribution is the proper term for a probability bell curve

- **In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.**
- **Normal distributions are symmetrical, but not all symmetrical distributions are normal.**
- **In reality, most pricing distributions are not perfectly normal**.

## Understanding Normal Distribution

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

### Skewness and Kurtosis

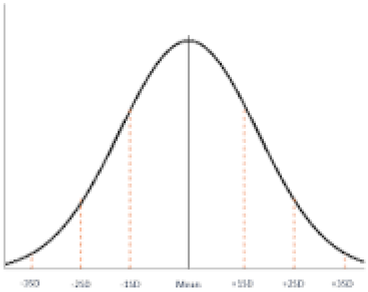
Real life data rarely, if ever, follow a perfect normal distribution.
The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution. The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail; positive skewness implies that the right tail of the distribution is longer than the left.

The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that is generally less extreme than the tails of the normal distribution. The normal distribution has a kurtosis of three, which indicates the distribution has neither fat nor thin tails. Therefore, if an observed distribution has a kurtosis greater than three, the distribution is said to have heavy tails when compared to the normal distribution. If the distribution has a kurtosis of less than three, it is said to have thin tails when compared to the normal distribution.

### Normal Distribution is Used in Finance

The assumption of a normal distribution is applied to asset prices as well as price action. Traders may plot price points over time to fit recent price action into a normal distribution. The further price action moves from the mean, in this case, the more likelihood that an asset is being over or undervalued. Traders can use the standard deviations to suggest potential trades. This type of trading is generally done on very short time frames as larger timescales make it much harder to pick entry and exit points.

Similarly, many statistical theories attempt to model asset prices under the assumption that they follow a normal distribution. In reality, price distributions tend to have fat tails and, therefore, have kurtosis greater than three. Such assets have had price movements greater than three standard deviations beyond the mean more often than would be expected under the assumption of a normal distribution. Even if an asset has went through a long period where it fits a normal distribution, there is no guarantee that the past performance truly informs the future prospects.



## Properties

- **It is symmetric. A normal distribution comes with a perfectly symmetrical shape. ...**
- **The mean, median, and mode are equal. The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. ...**
- **Empirical rule. ...**
- **Skewness and kurtosis**

**How do you handle missing data? What imputation techniques do you recommend?**

Ans. Common Methods

1. Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. ...
2. Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...
3. Random Forest.

## 3 Methods to Handle Missing Data

An analysis is only as good as its data, and every researcher has struggled with dubious results because of missing data. In this article, I will cover three ways to deal with missing data.

**Types of Missing Data**
Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

- Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

- Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
- Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

## Common Methods

**Mean or Median Imputation**
When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

**Multivariate Imputation by Chained Equations (MICE)**
MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.
To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

```
<span class="hs_cos_wrapper hs_cos_wrapper_meta_field hs_cos_wrapper_type_rich_text" data-hs-cos-general-type="meta_field" data-hs-cos-type="rich_text" id="hs_cos_wrapper_post_body" style=""><code class="language-R">Data$ID <- NULL</code></span>
```

Secondly, as mentioned above, the algorithm treats different variables differently. So, all categorical variables should be treated as factor variables before implementing MICE.

```
<span class="hs_cos_wrapper hs_cos_wrapper_meta_field hs_cos_wrapper_type_rich_text"
data-hs-cos-general-type="meta_field" data-hs-cos-type="rich_text"
id="hs_cos_wrapper_post_body" style=""><code class="language-R">Data$year <-
as.factor(Data$year)
Data$gender <- as.factor(Data$gender)</code></span>
```

Then you can implement the algorithm using the MICE library in R

```
<span class="hs_cos_wrapper hs_cos_wrapper_meta_field hs_cos_wrapper_type_rich_text"
data-hs-cos-general-type="meta_field" data-hs-cos-type="rich_text"
id="hs_cos_wrapper_post_body" style=""><code class="language-R">library(mice)

init = mice(Data, maxit=0)

method = init$method

predMat = init$predictorMatrix

set.seed(101)

imputed = mice(Data, method=method, predictorMatrix=predMat, m=5)</code></span>
```

You can also ignore some variables as predictors or skip a variable from being imputed using the MICE library in R. Additionally, the library also allows you to set a method of imputation discussed above depending upon the nature of the variable.

### 3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.

The MICE library in R also allows imputations by random forest by setting the method to “rf”. The authors of the MICE library have provided an example on how to implement the random forest method .

**Imputation Techniques**

- Complete Case Analysis(CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data that is we consider only those rows where we have complete data that is data is not missing.
- Arbitrary Value Imputation.
- Frequent Category Imputation.

- **Ignore the records with missing values.**

Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- **Substitute a value such as mean**.

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

- **Predict missing values.**

Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

- Logistic Regression
- Discriminant Regression
- Markov Chain Monte Carlo (MCMC)

- **Predict missing values - Multiple Imputation**. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required **statistical assumptions** for multiple imputation:

1. Whether the data is missing at random (MAR).
2. Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).

- The type of imputation algorithm used.
- Some justification for choosing a particular imputation method.
- The proportion of missing observations.
- The number of imputed datasets (m) created.
- The variables used in the imputation model.

**12 What is A/B testing?**

Ans A/B testing (also known as split testing or bucket testing) is **a method of comparing two versions of a webpage or app against each other to determine which one performs better**.

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

### 1. Solve visitor pain points

Visitors come to your website to achieve a specific goal that they have in mind. It may be to understand more about your product or service, buy a particular product, read/learn more about a specific topic, or simply browse. Whatever the visitor's goal may be, they may face some common pain points while achieving their goal. It can be a confusing copy or hard to find the CTA button like buy now, request a demo, etc.

Not being able to achieve their goals leads to a bad user experience. This increases friction and eventually impacts your conversion rates. Use data gathered through visitor behavior analysis tools such as heatmaps, Google Analytics, and website surveys to solve your visitors' pain points. This stands true for all businesses: eCommerce, travel, SaaS, education, media, and publishing.

### 2. Get better ROI from existing traffic

As most experience optimizers have come to realize, the cost of acquiring quality traffic on your website is huge. A/B testing lets you make the most out of your existing traffic and helps you increase conversions without having to spend additional dollars on acquiring new traffic. A/B testing can give you high ROI as sometimes, even the minutest of changes on your website can result in a significant increase in overall business conversions.

### 3. Reduce bounce rate

One of the most important metrics to track to judge your website's performance is its bounce rate. There may be many reasons behind your website's high bounce rate, such as too many options to choose from, expectations mismatch, confusing navigation, use of too much technical jargon, and so on.

Since different websites serve different goals and cater to different segments of audiences, there is no one-size-fits-all solution to reducing bounce rate. However, running an A/B test can prove beneficial. With A/B testing, you can test multiple variations of an element of your website till you find the best possible version. This not only helps you find friction and visitor pain points but helps improve your website visitors' overall experience, making them spend more time on your site and even converting into a paying customer.

### 4. Make low-risk modifications

Make minor, incremental changes to your web page with A/B testing instead of getting the entire page redesigned. This can reduce the risk of jeopardizing your current conversion rate.

A/B testing lets you target your resources for maximum output with minimal modifications, resulting in an increased ROI. An example of that could be product description changes. You can perform an A/B test when you plan to remove or update your product descriptions. You do not know how your visitors are going to react to the change. By running an A/B test, you can analyze their reaction and ascertain which side the weighing scale may tilt.

Another example of low-risk modification can be the introduction of a new feature change. Before introducing a new feature, launching it as an A/B test can help you understand whether or not the new change that you're suggesting will please your website audience.

Implementing a change on your website without testing it may or may not pay off in both the short and long run. Testing and then making changes can make the outcome more certain.

5. **Achieve statistically significant improvements**

Since A/B testing is entirely data-driven with no room for guesswork, gut feelings, or instincts, you can quickly determine a “winner” and a “loser” based on statistically significant improvements on metrics like time spent on the page, number of demo requests, cart abandonment rate, click-through rate, and so on.

6**. Redesign website to increase future business gains**

Redesigning can range from a minor CTA text or color tweak to particular web pages to completely revamping the website. The decision to implement one version or the other should always be data-driven when A/B testing. Do not quit testing with the design being finalized. As the new version goes live, test other web page elements to ensure that the most engaging version is served to the visitors.

Therefore, every piece of content that reaches your target audience via your website must be optimized to its maximum potential. This is especially true for elements that have the potential to influence the behavior of your website visitors and business conversion rate. When undertaking an optimization program, test the following key site elements (the list, however, is not exhaustive):

**Copy**

**1. Headlines and subheadlines**

A headline is practically the first thing that a visitor notices on a web page. It’s also what defines their first and last impression, filling the blanks whether or not they’ll go ahead and convert into paying customers. Hence, it’s imperative to be extra cautious about your site’s headlines and subheadlines. Ensure they’re short, to-the-point, catchy, and convey your desired message in the first stance. Try A/B testing a few copies with different fonts and writing styles, and analyze which catches your visitors’ attention the most and compels them to convert. You can also use VWO’s AI-powered text generation system to generate recommendations for the existing copy on your website.

**2. Body**

The body or main textual content of your website should clearly state what the visitor is getting – what’s in store for them. It should also resonate with your page’s headline and subheadline. A well-written body can significantly increase the chances of turning your website into a conversion magnet.

While drafting your website's content, keep the following two parameters in mind:

- **Writing style:** Use the right tonality based on your target audience. Your copy should directly address the end-user and answer all their questions. It must contain key phrases that improve usability and stylistic elements that highlight important points.
- **Formatting:** Use relevant headlines and subheadlines, break the copy into small and easy paragraphs, and format it for skimmers using bullet points or lists.

Interestingly, experience optimizers can now take advantage of artificial intelligence to create website copies. GPT-3 or Generative Pre-trained Transformer 3, is an AI-powered neural network that has the ability to produce nearly flawless text content relevant to any given context. Built by OpenAI, GPT-3 uses machine learning to predict and draft content just like a human. The best part? You can now integrate OpenAI's GPT-3 with VWO Testing account and create variations for your website copy and deploy them without the help of an expert writer or IT, respectively.

### 3. Subject lines

Email subject lines directly impact open rates. If a subscriber doesn't see anything they like, the email will likely wind up in their trash bin.

According to recent research, average open rates across more than a dozen industries ranging from 25 to 47 percent. Even if you're above average, only about half of your subscribers might open your emails.

A/B testing subject lines can increase your chances of getting people to click. Try questions versus statements, test power words against one another, and consider using subject lines with and without emojis.

## Design and layout

Because everything seems so essential, businesses sometimes struggle with finding only the most essential elements to keep on their website. With A/B testing, this problem can be solved once and for all.

For example, as an eCommerce store, your product page is extremely important from a conversion perspective. One thing for sure is that with technological progress in its current stage, customers like to see everything in high definition before buying it. Therefore, your product page must be in its most optimized form in terms of design and layout.

Along with the copy, the page's design and layout include images (product images, offer images, etc.) and videos (product videos, demo videos, advertisements, etc.). Your product page should answer all of your visitor's questions without confusing them and without getting cluttered:

- **Provide clear information:** Based on the products you sell, find creative ways to provide all necessary context and accurate product descriptions so that prospective buyers do not get overwhelmed with an unorganized copy while looking for answers to their queries. Write clear copies and provide easily noticeable size charts, color options, etc.
- **Highlight customer reviews:** Add both good and bad reviews for your products. Negative reviews add credibility to your store.
- **Write simple content:** Avoid confusing potential buyers with complicated language in the quest to decorate your content. Keep it short, simple, and fun to read.
- **Create a sense of urgency:** Add tags like 'Only 2 Left In Stock', countdowns like 'Offer Ends in 2 Hours and 15 Minutes', or highlight exclusive discounts and festive offers, etc., to nudge the prospective buyers to purchase immediately.

Other important pages whose design needs to be on point are pages like the home page and landing page. Use A/B testing to discover the most optimized version of these critical pages. Test as many ideas as you can, such as add plenty of white space and high definition images, feature product videos instead of images, and test out different layouts.

Declutter your pages using insights from heatmaps, clickmaps, and scrollmaps to analyze dead clicks and identify distractions. The less cluttered your home page and landing pages, the more likely it is for your visitors to easily and quickly .

### CTA (Call-to-action)

The CTA is where all the real action takes place – whether or not visitors finish their purchases and convert if they fill out the sign-up form or not, and more such actions that have a direct bearing on your conversion rate. A/B testing enables you to test different CTA copies, their placement across the web page, toy with their size and color scheme, and so on. Such experimentation helps understand which variation has the potential to get the most conversions.

## 13. Is mean imputation of missing data acceptable practice?

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the **mean remains unbiased**. ... Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

- **Bad practice in general**
- **If just estimating means: mean imputation preserves the mean of the observed data**
- **Leads to an underestimate of the standard deviation**
- **Distorts relationships between variables by “pulling” estimates of the correlation toward zero**

Imputing mean has long been the “usual business”, which it has no longer any real reason to be, what with ample computing power, and sophisticated multiple imputation methods being developed.

Using the mean as a substitute for missing values is used because it does not affect linear regression estimations and projections. And back in the days of hand calculations and such, substitution with mean was the easiest way around missing values short of dropping the whole case (which alters every other mean, if they only miss that one value).

Multiple imputation is the way forward, with increasing use across the board.

Using the median, though often the “better” descriptor of the data would affect the mean of the data and affect the linear regression estimates and confidence intervals

## 14. What is linear regression in statistics?

Ans. Linear regression **attempts to model the relationship between two variables by fitting a linear equation to observed data**. ... A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable.Simple linear regression is **a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line**. Both variables should be quantitative. ... Linear regression most often uses mean-square error (MSE) to calculate the error of the model.

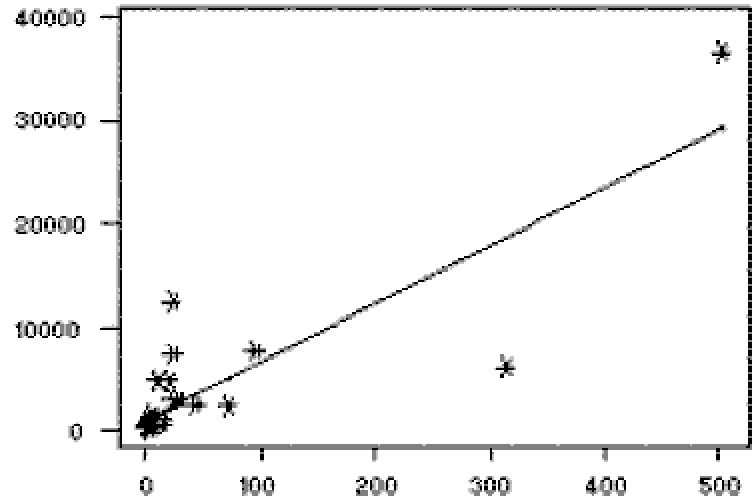
### Least-Squares Regression

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

**Example**

The dataset "Televisions, Physicians, and Life Expectancy" contains, among other variables, the number of people per television set and the number of people per physician for 40 countries. Since both variables probably reflect the level of wealth in each country, it is reasonable to assume that there is some positive association between them. After removing 8 countries with missing values from the dataset, the remaining 32 countries have a correlation coefficient of 0.852 for number of people per television set and number of people per physician. The *$r^2$* value is 0.726 (the square of the correlation coefficient), indicating that 72.6% of the variation in one variable may be explained by the other. Suppose we choose to consider number of people per television set as the explanatory variable, and number of people per physician as the dependent variable. Using the MINITAB "REGRESS" command gives the following results:

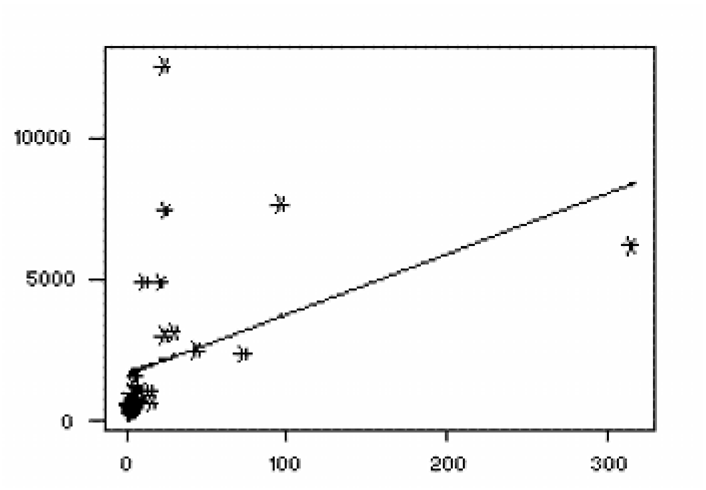
The regression equation is People.Phys. = 1019 + 56.2 People.Tel.



To view the fit of the model to the observed data, one may plot the computed regression line over the actual data points to evaluate the results. For this example, the plot appears to the right, with number of individuals per television set (the explanatory variable) on the x-axis and number of individuals per physician (the dependent variable) on the y-axis. While most of the data points are clustered towards the lower left corner of the plot (indicating relatively few individuals per television set and per physician), there are a few points which lie far away from the main cluster of the data. These points are known as ***outliers***, and depending on their location may have a major impact on the regression line .

### Outliers and Influential Observations

After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line. Notice, in the above example, the effect of removing the observation in the upper right corner of the plot:



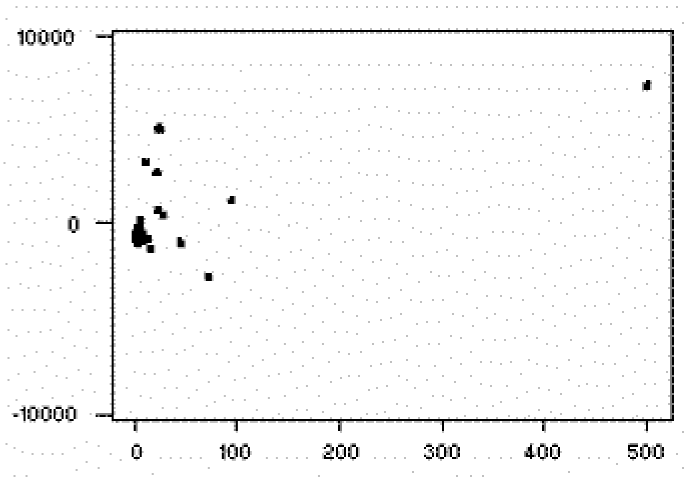
With this influential observation removed, the regression equation is now

People.Phys = 1650 + 21.3 People.Tel.

The correlation between the two variables has dropped to 0.427, which reduces the r² value to 0.182. With this influential observation removed, less that 20% of the variation in number of people per physician may be explained by the number of people per television. Influential observations are also visible in the new model, and their impact should also be investigated.

### Residuals

Once a regression model has been fit to a group of data, examination of the residuals (the deviations from the fitted line to the observed values) allows the modeler to investigate the validity of his or her assumption that a linear relationship exists. Plotting the residuals on the y-axis against the explanatory variable on the x-axis reveals any possible non-linear relationship among the variables, or might alert the modeler to investigate lurking variables. In our example, the residual plot amplifies the presence of outliers.



**Lurking Variables**
If non-linear trends are visible in the relationship between an explanatory and dependent variable, there may be other influential variables to consider. A lurking variable exists when the relationship between two variables is significantly affected by the presence of a third variable which has not been included in the modeling effort. Since such a variable might be a factor of time (for example, the effect of political or economic cycles), a time series plot of the data is often a useful tool in identifying the presence of lurking variables.

**Extrapolation**

Whenever a linear regression model is fit to a group of data, the range of the data should be carefully observed. Attempting to use a regression equation to predict values outside of this range is often inappropriate, and may yield incredible answers. This practice is known as extrapolation. Consider, for example, a linear model which relates weight gain to age for young children. Applying such a model to adults, or even teenagers, would be absurd, since the relationship between age and weight gain is not consistent for all age groups.

**15. What are the various branches of statistics?**

Ans. **Statistics:**
Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are **two main branches** of statistics
- Inferential Statistic.
- Descriptive Statistic.

**Inferential Statistics:**
Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

**Descriptive Statistics:**
Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphycal form.

Statistics plays a main role in the field of research. It helps us in the collection, analysis and presentation of data. In this blog post we will try to learn about the two main branches of statistics that is descriptive and inferential statistics.

## Branches of Statistics

Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

### Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

### Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

### Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.
Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.

## Descriptive Statistics

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations. Descriptive statistics can be categorized into

- Measures of central tendency
- Measures of variability

To easily understand the analyzed data, both measures of tendency and measures of variability use tables, general discussions, and graphs.

.

### Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

- **Mean**

This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.

- **Median**

This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score which is at the center of the arranged sample.

- **Mode**

This is the frequently occurring value in a given set of scores.

**Measures of Variability**

The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

**Inferential Statistics**

Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population.

probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.

The different types of calculation of inferential statistics include:

- **Regression analysis**
- **Analysis of variance (ANOVA)**
- **Analysis of covariance (ANCOVA)**
- **Statistical significance (t-test)**
- **Correlation analysis**