STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

   Ans : 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?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned

   Ans : a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a)Modeling event/time data
   b)Modeling bounded count data
   c)Modeling contingency tables
   d)All of the mentioned

   Ans : b) 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.

   a)Empirical
   b)Binomial
   c)Poisson
   d)All of the mentioned

   Ans: c)Piosson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

   Ans: b) False

STATISTICS WORKSHEET-1

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

   Ans : b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

Ans : a) 0

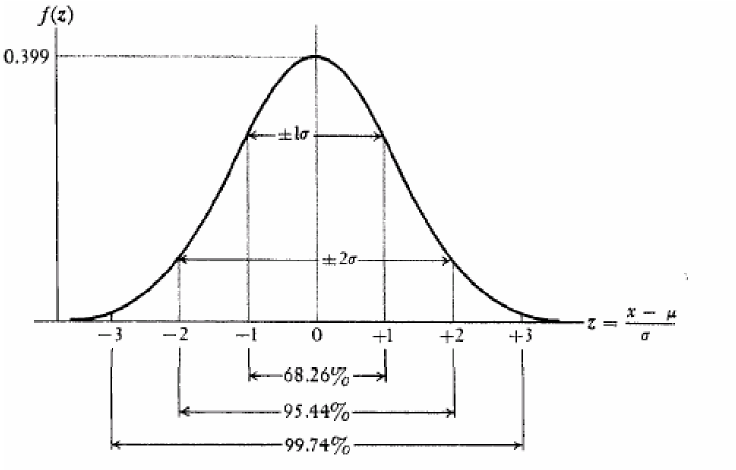
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

   Ans: c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.



Characteristics of the Normal distribution

- Symmetric, bell shaped
- <u>Continuous</u> for all values of X between -∞ and ∞ so that each conceivable interval of real

numbers has a probability other than zero.

$-\infty \leq X \leq \infty$

- Two parameters, μ and σ. Note that the normal distribution is actually a family of distributions, since μ and σ determine the shape of the distribution.

- The rule for a normal density function is

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

- The notation $N(\mu, \sigma^2)$ means normally distributed with mean μ and variance σ^2. If we say $X \sim N(\mu, \sigma^2)$ we mean that X is distributed $N(\mu, \sigma^2)$.
- About 2/3 of all cases fall within one standard deviation of the mean, that is

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = .6826.$$

- About 95% of cases lie within 2 standard deviations of the mean, that is

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = .9544$$

Why is the normal distribution useful?

- Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution
- The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.
- There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

STATISTICS WORKSHEET-1

The standardized normal distribution.

a. **General Procedure**. As you might suspect from the formula for the normal density function, it would be difficult and tedious to do the calculus every time we had a new set of parameters for μ and σ. So instead, we usually work with the standardized normal distribution, where μ = 0 and σ = 1, i.e. N(0,1). That is, rather than directly solve a problem involving a normally distributed variable X with mean μ and standard deviation σ, an indirect approach is used.

1. We first convert the problem into an equivalent one dealing with a normal variable measured in standardized deviation units, called a standardized normal variable. To do this, if X ~ N(μ, σ5), then

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

2. A table of standardized normal values can then be used to obtain an answer in terms of the converted problem.

3. If necessary, we can then convert back to the original units of measurement. To do this, simply note that, if we take the formula for Z, multiply both sides by σ, and then add μ to both sides, we get

$$X = Z\sigma + \mu$$

4. The interpetation of Z values is straightforward. Since σ = 1, if Z = 2, the corresponding X value is exactly 2 standard deviations above the mean. If Z = -1, the corresponding X value is one standard deviation below the mean. If Z = 0, X = the mean, i.e. μ.

b. **Rules for using the standardized normal distribution.** It is very important to understand how the standardized normal distribution works, so we will spend some time here going over it. Recall that, for a random variable X,

$$F(x) = P(X \le x)$$

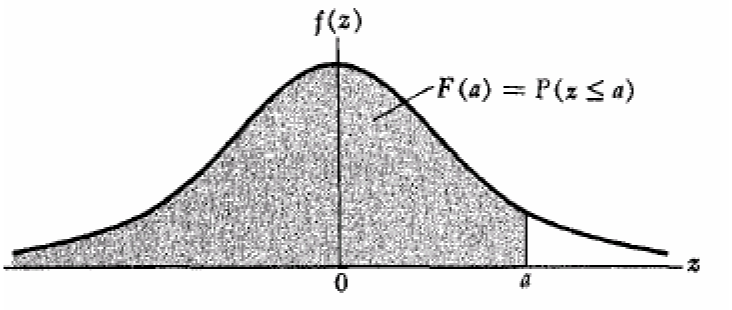
STATISTICS WORKSHEET-1

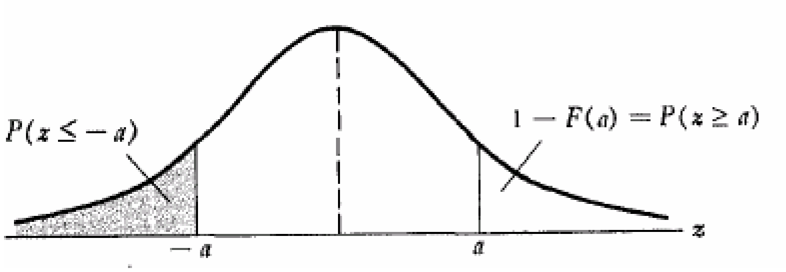
RULES:

1. **$P(Z \le a)$**

$= F(a)$ (use when a is positive)

$= 1 - F(-a)$ (use when a is negative)





EX: Find $P(Z \le a)$ for a = 1.65, -1.65, 1.0, -1.0

To solve: for positive values of a, look up and report the value for F(a) given in Appendix E, Table I. For negative values of a, look up the value for F(-a) (i.e. F(absolute value of a)) and report 1 - F(-a).

$P(Z \le 1.65) = F(1.65) = .95$

$P(Z \le -1.65) = F(-1.65) = 1 - F(1.65) = .5$

$P(Z \le 1.0) = F(1.0) = .84$

$P(Z \le -1.0) = F(-1.0) = 1 - F(1.0) = .16$

You can also easily work in the other direction, and determine what a is given $P(Z \leq a)$ EX:

Find a for $P(Z \leq a) = .6026, .9750, .3446$

To solve: for $p \geq .5$, find the probability value in Table I, and report the corresponding value for Z. For $p < .5$, compute $1 - p$, find the corresponding Z value, and report the negative of that value, i.e. $-Z$.

$P(Z \leq .26) = .6026$

$P(Z \leq 1.96) = .9750$

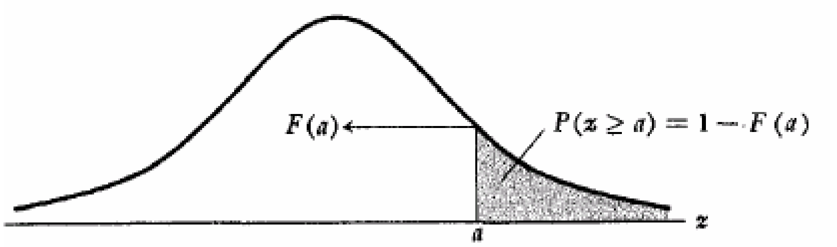
$P(Z \leq -.40) = .3446$ (since $1 - .3446 = .6554 = F(.40)$)

NOTE: It may be useful to keep in mind that $F(a) + F(-a) = 1$.

2. **$P(Z \geq a)$**

$= 1 - F(a)$ (use when a is positive)

$= F(-a)$ (use when a is negative)



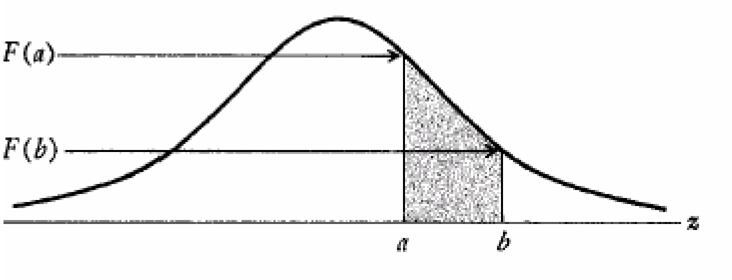
EX: Find $P(Z \geq a)$ for $a = 1.5, -1.5$

To solve: for a positive, look up $F(a)$, as before, and subtract $F(a)$ from 1. For a negative, just report $F(-a)$.

$P(Z \geq 1.5) = 1 - F(1.5) = 1 - .9332 = .0668$

$P(Z \geq -1.5) = F(1.5) = .9332$

3. **$P(a \le Z \le b) = F(b) - F(a)$**



EX: Find $P(a \le Z \le b)$ for $a = -1$ and $b = 1.5$

To solve: determine F(b) and F(a), and subtract.

$P(-1 \le Z \le 1.5) = F(1.5) - F(-1) = F(1.5) - (1 - F(1)) = .9332 - 1 + .8413 = .7745$

4. **For a positive, $P(-a \le Z \le a) = 2F(a) - 1$**

PROOF:

$P(-a \le Z \le a)$

$= F(a) - F(-a)$ (by rule 3)

$= F(a) - (1 - F(a))$ (by rule 1)

$= F(a) - 1 + F(a)$

$= 2F(a) - 1$

EX: find $P(-a \le Z \le a)$ for $a = 1.96$, $a = 2.58$

$P(-1.96 \le Z \le 1.96) = 2F(1.96) - 1 = (2 * .975) - 1 = .95$

$P(-2.58 \le Z \le 2.58) = 2F(2.58) - 1 = (2 * .995) - 1 = .99$

4B. **For a positive, $F(a) = [1 + P(-a \le Z \le a)] / 2$**

EX: find a for $P(-a \le Z \le a) = .90$, .975 $F(a) = (1 + .90)/2 = .95$, implying $a = 1.65$. For $P(-a \le Z \le a) = .975$,

$F(a) = (1 + .975)/2 = .9875$, implying $a = 2.24$

NOTE: Suppose we were asked to find a and b for $P(a \le Z \le b) = .90$. There are an infinite number of values that we could use; for example, we could have a = negative infinity and b = 1.28, or a = -1.28 and b = positive infinity, or a = -1.34 and b = 2.32, etc. The smallest interval between a and b will always be found by choosing values for a and b such that a = -b. For example, for $P(a \le Z \le b) = .90$, a = -1.65 and b = 1.65 are the "best" values to choose, since they yield the smallest possible value for b - a.

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

An analysis is only as good as its data, and every researcher has struggled with dubious results because of missing data .there are 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.

Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.It is typically safe to remove MCAR data because the results will be unbiased. The test may not be as powerful, but the results will be reliable

- **Missing At Random (MAR):**

Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, it is missing given the observed data. The missing data can be predicted based on the complete observed data

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):

- The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown.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.

Missing Not at Random (MNAR)

- The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing. In a survey, perhaps a specific group of people – say women ages 45 to 55 – did not answer a question. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model. other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.Before deciding which approach to employ, data scientists must understand why the data is missing.

Delection

There are two primary methods for deleting data when dealing with missing data: listwise and dropping variables.

❖ **Listwise**

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data are not missing completely at random (MCAR). Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis

- **Pairwise**

Pairwise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pairwise deletion allows data scientists to use more of the data. However, the resulting statistics may vary because they are based on different data sets. The results may be impossible to duplicate with a complete set of data.

Dropping Variables

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

Imputation

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

We use imputation because Missing data can cause the below issues: –

1. **Incompatible with most of the Python libraries used in Machine Learning:-** Yes, you read it right. While using the libraries for ML(the most common is skLearn), they don’t have a provision to automatically handle these missing data and can lead to errors.
2. **Distortion in Dataset:-** A huge amount of missing data can cause distortions in the variable distribution i.e it can increase or decrease the value of a particular category in the dataset.
3. **Affects the Final Model:-** the missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.

Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

STATISTICS WORKSHEET-1

- **Mean, Median and Mode**

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

- **Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**

These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

- **Linear Interpolation**

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

- **Seasonal Adjustment with Linear Interpolation**

When dealing with data that exhibits both trend and seasonality characteristics, use seasonal adjustment with linear interpolation. First you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages – say, two one-year averages – that are offset by one period relative to another. You can then complete data smoothing with linear interpolation as discussed above.

- **Multiple imputation**

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.

The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

- **K Nearest Neighbors**

In this method, data scientists choose a distance measure for k neighbors, and the average is used to impute an estimate. The data scientist must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

12. What is A/B testing?

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

Using the visual above as an example, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It’s important to note that all other variables need to be held constant when performing an A/B test.

STATISTICS WORKSHEET-1

Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. **Statistical hypothesis testing** is a method in which a sample dataset is compared against the population data. **Two-sample hypothesis testing** is a method in determining whether the differences between the two samples are statistically significant or not.

How to conduct a standard A/B test

1. Formulate your hypothesis

Before conducting an A/B testing, you want to state your null hypothesis and alternative hypothesis:

The **null hypothesis** is one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is **no** difference between the control and variant group.

The **alternative hypothesis** is one that states that sample observations are influenced by some non-random cause. From an A/B test perspective, the alternative hypothesis states that there **is** a difference between the control and variant group.

When developing your null and alternative hypotheses, it's recommended that you follow a PICOT format. Picot stands for:

- **P**opulation: the group of people that participate in the experiment
- **I**ntervention: refers to the new variant in the study
- **C**omparison: refers to what you plan on using as a reference group to compare against your intervention
- **O**utcome: represents what result you plan on measuring
- **T**ime: refers to the duration of the experience (when and how long the data is collected)

Example: "Intervention A will improve anxiety (as measured by the mean change from baseline in the HADS anxiety subscale) in cancer patients with clinical levels of anxiety at 3 months compared to the control intervention."

Does it follow the PICOT criteria?

- Population: Cancer patients with clinical levels of anxiety
- Intervention: Intervention A
- Comparison: the control intervention
- Outcome: improve anxiety as measured by the mean change from baseline in the HADS anxiety subscale
- Time: at 3 months compared to the control intervention.

Yes it does therefore, this is an example of strong hypothesis test.

STATISTICS WORKSHEET-1

2.Creat your control group and test group

Once you determine your null and alternative hypothesis, the next step is to create your control and test (variant) group. There are two important concepts to consider in this step, random samplings and sample size

Random sampling is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it's important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.**

SampleSize

It's essential that you determine the minimum sample size for your A/B test prior to conducting it so that you can eliminate **under coverage bias**, bias from sampling too few observations. There are plenty of online calculators that you can use to calculate the sample size given these three inputs.

3. Conduct the test, compare the results, and reject or do not reject the null hypothesis

$$T-statistic=\frac{Observed\ value-hypothesized\ value}{Standard\ Error}$$

$$Stamdard\ Error=\sqrt{\frac{2*Variance(sample)}{N}}$$

Once you conduct your experiment and collect your data, you want to determine if the difference between your control group and variant group is statistically significant. There are a few steps in determining this:

- First, you want to set your **alpha**, the probability of making a type 1 error. Typically the alpha is set at 5% or 0.05
- Next, you want to determine the probability value (p-value) by first calculating the t-statistic using the formula above.
- Lastly, compare the p-value to the alpha. If the p-value is greater than the alpha, do not reject the null!

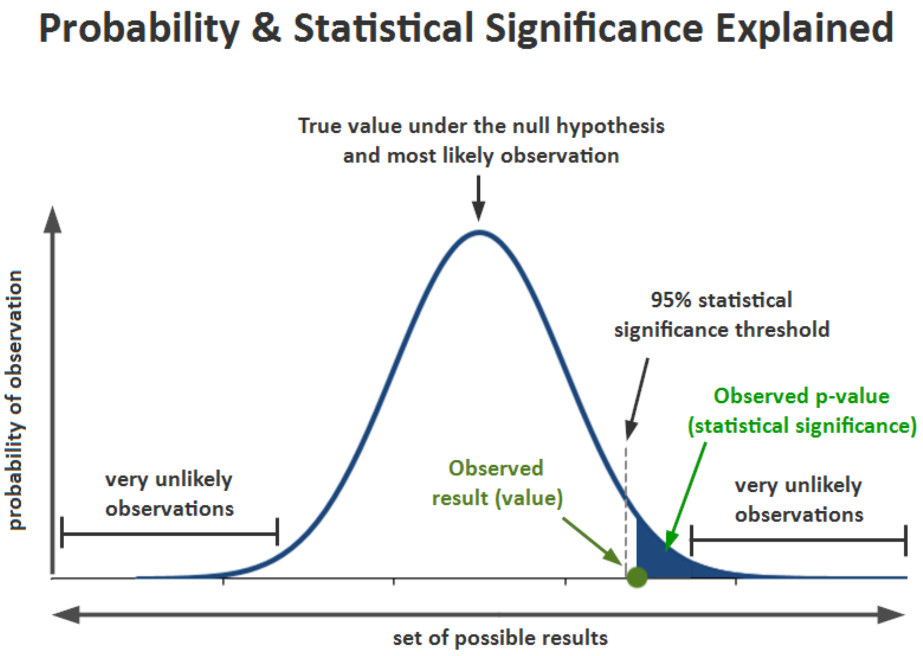
Statistical significance of the Test

There are two types of errors that may occur in our hypothesis testing:

1. **Type I error**: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A
2. **Type II error**: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A

To avoid these errors we must calculate the statistical significance of our test.An experiment is considered to be statistically significant when we have enough evidence to prove that the result we see in the sample also exists in the population.That means the difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a two-sample T-test.

The **two–sample t–test** is one of the most commonly **used** hypothesis **tests**. It is applied to compare whether averge distance between two number.


Probability & Statistical Significance Explained
True value under the null hypothesis and most likely observation
probability of observation
95% statistical significance threshold
Observed p-value (statistical significance)
very unlikely observations
Observed result (value)
very unlikely observations
set of possible results

To understand this, we must be familiar with a few terms:

1. **Significance level (alpha):** The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05
2. **P-Value:** It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value stronger the chances to reject the H_0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis
3. **Confidence interval:** The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

Here, our p-value is less than the significance level i.e 0.05. Hence, we can reject the null hypothesis. This means that in our A/B testing, newsletter B is performing better than newsletter A. So our recommendation would be to replace our current newsletter with B to bring more traffic on our website.

STATISTICS WORKSHEET-1

13. Is mean imputation of missing data acceptable practice?

Mean imputation (or mean substitution) replaces missing values of a certain variable by the mean of non-missing cases of that variable**.**

1. Missing values in your data **do not reduce your sample size**, as it would be the case with listwise deletion (the default of many statistical software packages, e.g. R, Stata, SAS or SPSS). Since mean imputation replaces all missing values, you can keep your whole database.
2. Mean imputation is very **simple to understand and to apply** (more on that later in the R and SPSS examples). You can explain the imputation method easily to your audience and everybody with basic knowledge in statistics will get what you've done.

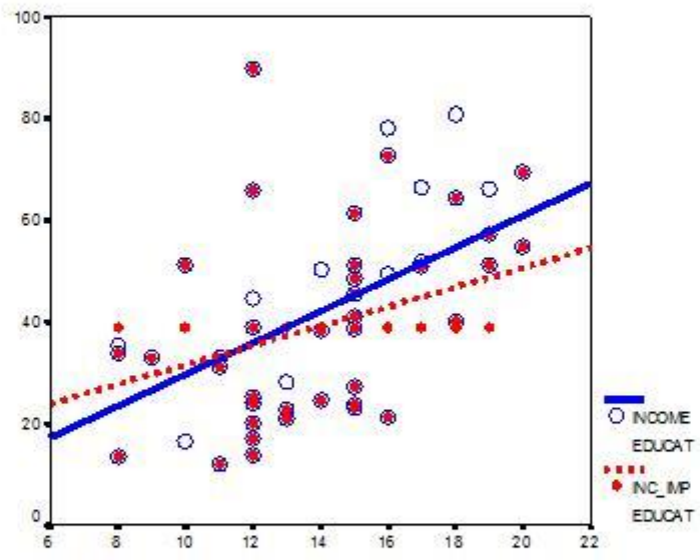
If the response mechanism is MCAR, the **sample mean of your variable is not biased**. Mean substitution might be a valid approach, in case that the univariate average of your variables is the only metric your are interested in. perhaps that's a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort.

It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it.This post is the first explaining the many reasons not to use mean imputation (and to be fair, its advantages).First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

Problem 1: Mean imputation does not preserve the relationships among variables.

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. That's a good thing.Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.This is the original logic involved in mean imputation.If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.It *will* still bias your standard error, but I will get to that in another post.Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. The following graph illustrates this well:

STATISTICS WORKSHEET-1



This graph illustrates hypothetical data between X=years of education and Y=annual income in thousands with n=50. The blue circles are the original data, and the solid blue line indicates the best fit regression line for the full data set. The correlation between X and Y is r = .53.I then randomly deleted 12 observations of income (Y) and substituted the mean. The red dots are the mean-imputed data.Blue circles with red dots inside them represent non-missing data. Empty Blue circles represent the missing data. If you look across the graph at Y = 39, you will see a row of red dots without blue circles. These represent the imputed values.

The dotted red line is the new best fit regression line with the imputed data. As you can see, it is less steep than the original line. Adding in those red dots pulled it down.The new correlation is r = .39. That’s a lot smaller that .53.The real relationship is quite underestimated.Of course, in a real data set, you wouldn’t notice so easily the bias you’re introducing. This is one of those situations where in trying to solve the lowered sample size, you create a bigger problem.One note: if X were missing instead of Y, mean substitution would artificially *inflate* the correlation.

In other words, you’ll think there is a stronger relationship than there really is. That’s not good either. It’s not reproducible and you don’t want to be overstating real results.This solution that is so good at preserving unbiased estimates for the mean isn’t so good for unbiased estimates of relationships.

Problem 2: Mean Imputation Leads to An Underestimate of Standard Errors

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that’s too low.In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small.Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.Ultimately, because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.That's not good.

We learned some reasons why mean imputation is so popular among data users. However, let's move on to the more important part – the **drawbacks of mean imputation**:

1. Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. **Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.
3. If the response mechanism is MAR or MNAR, even the **sample mean of your variable is biased** (compare that with point 3 above). Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

There are a few advantages, but many serious drawbacks. On top of that, we can also benefit from the advantages with more advanced imputation methods (e.g. predictive mean matching or stochastic regression imputation). To make it short, there is basically no excuse for using mean imputation.Mean substitution is **a method in which missing observations for a certain variable are replaced by the average of observed data for that variable in other patients**. However, both of these methods have severe drawbacks.

The process of replacing null values in a data collection with the data's mean is known as mean imputation.Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

STATISTICS WORKSHEET-1

14. What is linear regression in statistics?

Linear regression is a statistical method that tries to show a relationship between variables. It looks at different data points and plots a trend line. A simple example of linear regression is finding that the cost of repairing a piece of machinery increases with time.

More precisely, linear regression is used to determine the character and strength of the association between a dependent variable and a series of other independent variables. It helps create models to make predictions, such as predicting a company's stock price.Before trying to fit a linear model to the observed dataset, one should assess whether or not there is a relationship between the variables. Of course, this doesn't mean that one variable causes the other, but there should be some visible correlation between them.For example, higher college grades don't necessarily mean a higher salary package. But there can be an association between the two variables

$$\boxed{y = mx + b}$$

In this simple linear regression equation

- **y** is the estimated dependant variable (or the output)
- **m** is the regression coefficient (or the slope)
- **x** is the independent variable (or the input)
- **b** is the constant (or the y-intercept)

Finding the relationship between variables makes it possible to predict values or outcomes. In other words, linear regression makes it possible to predict new values based on existing data.

Types of linear regression

There are two types of linear regression: **simple linear regression** and **multiple linear regression**.

- The **simple linear regression** method tries to find the relationship between a single independent variable and a corresponding dependent variable. The independent variable is the input, and the corresponding dependent variable is the output.

STATISTICS WORKSHEET-1

- The **multiple linear regression** method tries to find the relationship between two or more independent variables and the corresponding dependent variable. There's also a special case of multiple linear regression called **polynomial regression**.

Simply put, a simple linear regression model has only a single independent variable, whereas a multiple linear regression model will have two or more independent variables. And yes, there are other non-linear regression methods used for highly complicated data analysis.

regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation *solves for*) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.In linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable. In this simple model , a straight line approximates the relationship between the dependent variable and the independent variable.When two or more independent variables are used in regression analysis, the model is no longer a simple linear one. This is known as multiple regression.

Formula For a Simple Linear Regression Model

The two factors that are involved in simple linear regression analysis are designated ***x*** and ***y***. The equation that describes how ***y*** is related to ***x*** is known as the **regression model**.

The simple linear regression model is represented by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The linear regression model contains an error term that is represented by ε. The error term is used to account for the variability in ***y*** that cannot be explained by the linear relationship between ***x*** and ***y***. If ε were not present, that would mean that knowing ***x*** would provide enough information to determine the value of ***y***.

There also parameters that represent the population being studied. These parameters of the model are represented by β_0 and β_1.

The simple linear regression equation is graphed as a straight line, where:

1. β_0 is the y-intercept of the regression line.
2. β_1 is the slope.
3. ***E***(***y***) is the mean or expected value of ***y*** for a given value of ***x***.

STATISTICS WORKSHEET-1

A regression line can show a positive linear relationship, a negative linear relationship, or no relationship[3].

1. **No relationship:** The graphed line in a simple linear regression is flat (not sloped). There is no relationship between the two variables.
2. **Positive relationship:** The regression line slopes upward with the lower end of the line at the y-intercept (axis) of the graph and the upper end of the line extending upward into the graph field, away from the x-intercept (axis). There is a positive linear relationship between the two variables: as the value of one increases, the value of the other also increases.
3. **Negative relationship:** The regression line slopes downward with the upper end of the line at the y-intercept (axis) of the graph and the lower end of the line extending downward into the graph field, toward the x-intercept (axis). There is a negative linear relationship between the two variables: as the value of one increases, the value of the other decreases.[4]

The Estimated Linear Regression Equation

If the parameters of the population were known, the simple linear regression equation (shown below) could be used to compute the mean value of ***y*** for a known value of ***x***.

$\boldsymbol{E}(\boldsymbol{y}) = \beta_0 + \beta_1 \boldsymbol{x} + \varepsilon$

In practice, however, parameter values generally are not known so they must be estimated by using data from a sample of the population. The population parameters are estimated by using sample statistics . The sample statistics are represented by β_0 and β_1. When the sample statistics are substituted for the population parameters, the estimated regression equation is formed.[3]

The estimated regression equation is:

$(\hat{\boldsymbol{y}}) = \beta_0 + \beta_1 \boldsymbol{x} + \varepsilon$

The graph of the estimated simple regression equation is called the estimated regression line.

1. $\boldsymbol{\beta_0}$ is the y-intercept of the regression line.
2. $\boldsymbol{\beta_1}$ is the slope.
3. $(\hat{\boldsymbol{y}})$ is the estimated value of ***y*** for a given value of ***x***.

Limits of Simple Linear Regression

Regression analysis is commonly used in research to establish that a correlation exists between variables. But correlation is not the same as causation: a relationship between two variables does not mean one causes the other to happen. Even a line in a simple linear regression that fits the data points well may not guarantee a cause-and-effect relationship.

Using a linear regression model will allow you to discover whether a relationship between variables exists at all. To understand exactly what that relationship is, and whether one variable causes another, you will need additional research and statistical analysis.Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation *solves for*) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.

In linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable. In this simple model, a straight line approximates the relationship between the dependent variable and the independent variable.When two or more independent variables are used in regression analysis, the model is no longer a simple linear one. This is known as multiple regression.

15. What are the various branches of statistics?

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 statistics.

- Descriptive statistics
- Inferential statistics

Descriptive Statistics

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. Measures of Central Tendency

2. Measures of Dispersion (or Variability)

3. **Measures of Central Tendency**

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types.

STATISTICS WORKSHEET-1

1. **Mean :** Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

2. **Median :** Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

- If the number of observations are odd, median is given by the middle observation in the sorted form.

- If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.An important point to note that the order of the data (ascending or descending) does not effect the median.

3. Mode : Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.

- If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.

- If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.

- If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal.**

Example to compute the Measures of Central Tendency

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

- Mean — Mean is calculated as

$$Mean=\frac{17+16+21+18+15+17+21+19+11+23}{10}=\frac{178}{10}=17.8$$

- Median — To calculate Median, lets arrange the data in ascending order.

STATISTICS WORKSHEET-1

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17+18}{2} = 17.5$$

- Mode — Mode is given by the number that occurs maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Note-

1. Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.

2. At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.

3. If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1. **Absolute Deviation from Mean** — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set. It is calculated as

$$Mean\ Absolute\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

2. Variance — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as

$$Variance = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

STATISTICS WORKSHEET-1

3. Standard Deviation — The square root of Variance is called the Standard Deviation. It is calculated as

$$Std\ Deviation = \sqrt{Variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

4. Range — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

$$Range = Maximum - Minimum$$

5. Quartiles — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

- 25% of the data points lie below Q1 and 75% lie above it.
- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
- 75% of the data points lie below Q3 and 25% lie above it

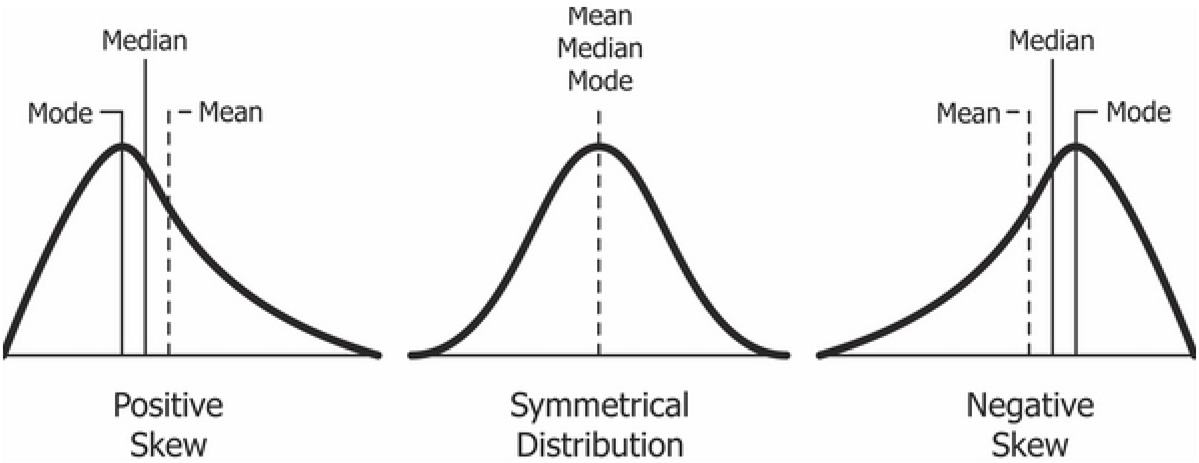
6. Skewness — The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

- Positive Skew — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.
- Negative Skew — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.
- The most commonly used method of calculating Skewness is

$$Skewness = \frac{3\,(Mean - Median)}{Std\ \ Deviation}$$

If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.

STATISTICS WORKSHEET-1

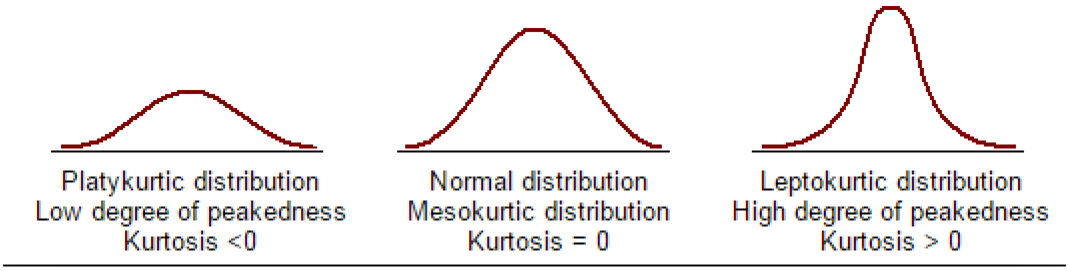


7. Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are three kinds of Kurtosis:

- Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.

- Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.

- Platykurtic — This is when the tail of the distribution is light(no outlier) and kurtosis is lesser than that of the normal distribution.



Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information.

Descriptive statistics is the type of statistics that probably springs to most people’s minds when they hear the word “statistics.” In this branch of statistics, the goal is to describe. Numerical measures are used to tell about features of a set of data. There are a number of items that belong in this portion of statistics, such as:

STATISTICS WORKSHEET-1

These measures are important and useful because they allow scientists to see patterns among data, and thus to make sense of that data. Descriptive statistics can only be used to describe the population or data set under study: The results cannot be generalized to any other group or population.Inferential statistics are produced through complex mathematical calculations that allow scientists to infer trends about a larger population based on a study of a sample taken from it. Scientists use inferential statistics to examine the relationships between variables within a sample and then make generalizations or predictions about how those variables will relate to a larger population.

It is usually impossible to examine each member of the population individually. So scientists choose a representative subset of the population, called a statistical sample, and from this analysis, they are able to say something about the population from which the sample came. There are two major divisions of inferential statistics:

- A confidence interval gives a range of values for an unknown parameter of the population by measuring a statistical sample. This is expressed in terms of an interval and the degree of confidence that the parameter is within the interval.
- Tests of significance or hypothesis testing where scientists make a claim about the population by analyzing a statistical sample. By design, there is some uncertainty in this process. This can be expressed in terms of a level of significance.

Techniques that social scientists use to examine the relationships between variables, and thereby to create inferential statistics, include linear regression analyses, logistic regression analyses, ANOVA, correlation analyses, structural equation modeling, and survival analysis. When conducting research using inferential statistics, scientists conduct a test of significance to determine whether they can generalize their results to a larger population. Common tests of significance include the chi-square and t-test. These tell scientists the probability that the results of their analysis of the sample are representative of the population as a whole.

Hypothesis Testing

Hypothesis testing makes use of inferential statistics and is used to analyze relationships between variables and make population comparisons through the use of sample data. The steps for hypothesis testing include having a stated research hypothesis (null and alternate), data collection per the hypothesis test requirements, data analysis through the appropriate test, a decision to reject or accept the null hypothesis, and finally, a presentation and discussion of findings made.

Hypothesis testing falls under the “statistical tests” category. Statistical tests account for sampling errors and can either be parametric (includes assumptions made regarding population distribution parameters) or non-parametric (does not include assumptions made regarding population distribution parameters).

STATISTICS WORKSHEET-1

Parametric tests tend to be more trusted and reliable because they enable the detection of potential effects. Parametric tests assume that the population from which sample data is derived is normally distributed. The sample size provides an adequate representation of the population from which it was derived. The groups, variances, and measures of spread are comparable.

Other Testing Methods

There are other testing methods, including correlation tests and comparison tests. Correlation tests examine the association between two variables and estimate the extent of the relationship. Examples of correlation tests are the Pearson’s r test, Spearman’s r test, and the Chi-square test of independence