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.

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?

ANS: - a) Central Limit Theorem

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

ANS: - b) Modeling bounded count data

4. Point out the correct statement.

ANS: - d) All of the mentioned

5. _____ random variables are used to model rates.

ANS: - c) Poisson

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

ANS: - b) False

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

ANS: - b) Hypothesis

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

ANS: - a) 0

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

ANS: - c) Outliers cannot conform to the regression relationship

Q10 to Q15 are subjective answer type questions, Answer them in your own words briefly.

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

ANS: - A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

Here's an example of a normal distribution curve:

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

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

ANS: -
Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programmed will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values.

And how would you choose that estimate? The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10. Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilizing the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value. This has all of the benefits of regression imputation plus the random component's benefits. The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

Furthermore, standard errors are underestimated by all single imputation approaches. Because the imputed observations are estimates, their values have a random error associated with them. However, your programmed is unaware of this when you enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values.

And, while imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances.

As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above–hot deck and stochastic regression–work as the imputation method.

The multiple estimates varied significantly because these two approaches contain a random component. This reintroduces some variance that your program can account for in order to provide reliable standard error estimates for your model.

About 20 years ago, multiple imputation was a big advance in statistics. It eliminates many (but not all) difficulties with missing data and, when done correctly, leads to unbiased parameter estimations and accurate standard errors

12. What is A/B testing?

ANS: - It's a method to compare two things and, in a nutshell, find out which one is preferred by people who are using your website or app. To make the comparison, you'll need two different designs, or word choices or images.

- The control. This is the one that's actually on your website, aka in production, in the wild or live.
- The experiment. This is the design, copy or image you want to test to see if it works better than what you already have on your website.

For example, you might wonder if more visitors would add items to their carts if the "buy now" button on your product pages was bigger. Your control for this A/B test is the Buy Now buttons already on your product pages. The experiment is a larger Buy Now button on some product pages.

You can A/B test any element on your website that you suspect could be better, but the most efficient way to A/B test is to look at your site analytics first.

For example, my design team and I look at our site data for areas that have a high bounce rate or low conversions—even or especially if those pages are getting a lot of traffic. In fact, if the page gets lots of visitors, that traffic volume can help you get valid A/B results faster.

You can also flip the concept around and look for pages that have low bounce rates and high conversions. Then you could A/B test to see if you get better results on lower-performing pages by applying elements from those strong pages.

If you're still not sure where to start, you can focus on elements on your checkout or signup pages, because those are typically the most important pages on a website, the ones where you want the best possible performance. However, if your checkout and signup rates are already

high, you may want to leave things as they are and focus on other elements of your site, like landing page bounce rates or buy now buttons.

13. Is mean imputation of missing data acceptable practice?

ANS: - 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.

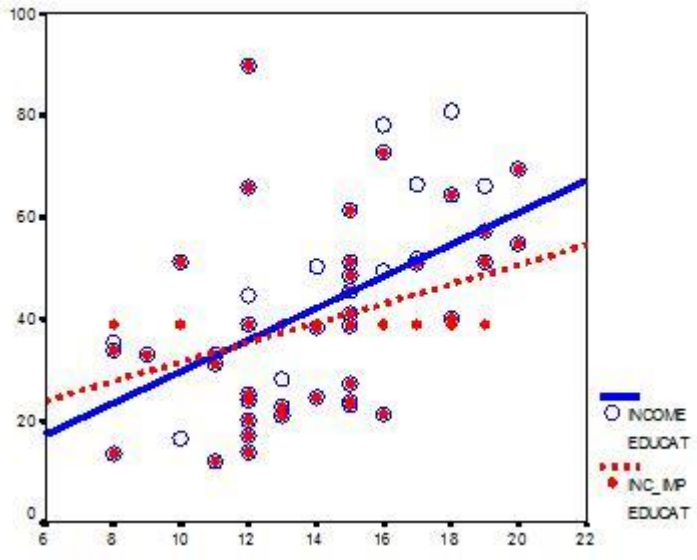
Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

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:



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.

A second reason is applying 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.

14. What is linear regression in statistics?

ANS: - Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. Linear regression is commonly used for predictive analysis and

modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression. This post will show you examples of linear regression, including an example of simple linear regression and an example of multiple linear regression.

Example of simple linear regression

The table below shows some data from the early days of the Italian clothing company Benetton. Each row in the table shows Benetton's sales for a year and the amount spent on advertising that year. In this case, our outcome of interest is sales—it is what we want to predict. If we use advertising as the predictor variable, linear regression estimates that Sales = 168 + 23 Advertising. That is, if advertising expenditure is increased by one million Euro, then sales will be expected to increase by 23 million Euros, and if there was no advertising, we would expect sales of 168 million Euros.

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

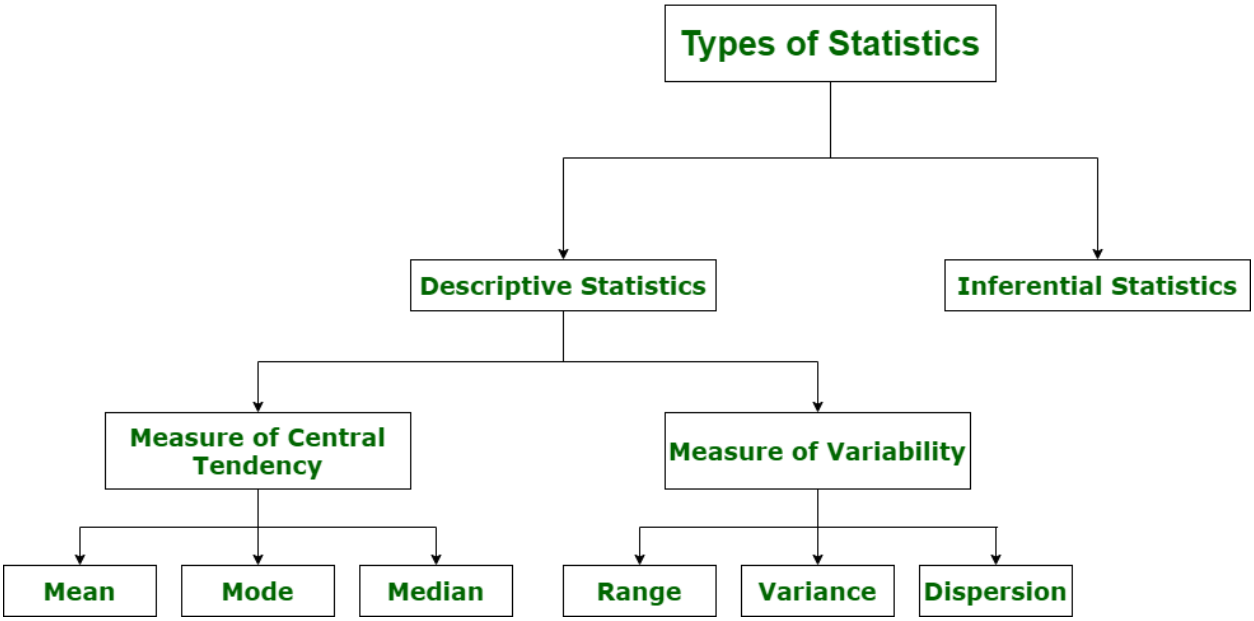
Example of multiple linear regression

Linear regression with a single predictor variable is known as simple regression. In real-world applications, there is typically more than one predictor variable. Such regressions are called multiple regression. For more information, check out this post on why you should not use multiple linear regression for Key Driver Analysis with example data for multiple linear regression examples.

Returning to the Benetton example, we can include **year** variable in the regression, which gives the result that Sales = 323 + 14 Advertising + 47 Year. The interpretation of this equation is that every extra million Euro of advertising expenditure will lead to an extra 14 million Euro of sales and that sales will grow due to non-advertising factors by 47 million Euro per year.

15. What are the various branches of statistics?

ANS: - branches of statistics: -



1. Descriptive Statistics:
Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

- (a). Measure of central tendency –
  Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.
  In statistics, there are three common measures of central tendency as shown below:
  - (i) Mean:
    It is measure of average of all value in a sample set.
    For example,

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

$$\text{Mean (m)} = \frac{\text{Sum of all the terms}}{\text{Total no. of terms}}$$

$$m = \frac{21.3 + 20.8 + 19}{3}$$

$$= 20.366$$

- (ii) Median:
  It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.
  For example,

Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i 20	15	4

Ordering the set from lowest to highest = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2}$$

$$\text{Median} = 23.5$$

- (iii) Mode:
  It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.
  For example,

  2 3 4 2 4 6 4 7 7 4 2 4

  Mode = 4

- (b). Measure of Variability –
  Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below:
  - (i) Range:
    It is given measure of how to spread apart values in sample set or data set. Range = Maximum value - Minimum value
  - (ii) Variance:
    It simply describes how much a random variable defers from expected value and it is also computed as square of deviations$^2 = \sum^{n}_{i=1} [(x_i - \overline{x})^2 \div n]$
  - In this formula, n represent total data points, $\overline{x}$ represents mean of data points and x_i represent individual data points.
  - (iii) Dispersion:
    It is measure of dispersion of set of data from its mean.$\sigma = \sqrt{} (1 \div n) \sum^{n}_{i=1} (x_i - \mu)^2$

2. Inferential Statistics:
Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive stats. It is simply used to analyze, interpret result, and draw conclusion. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis.

Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population. Inferential statistics are generally used to determine how strong relationship is within sample. But it is very difficult to obtain a population list and draw a random sample.

Inferential statistics can be done with help of various steps as given below:

1. Obtain and start with a theory.
2. Generate a research hypothesis.
3. Operationalize or use variables
4. Identify or find out population to which we can apply study material.

5. Generate or form a null hypothesis for these population.
6. Collect and gather a sample of children from population and simply run study.
7. Then, perform all tests of statistical to clarify if obtained characteristics of sample are sufficiently different from what would be expected under null hypothesis so that we can be able to find and reject null hypothesis.