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) Centric 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) Poisson

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

   Ans. b) False

7. 1. 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. 4. 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?

Ans. Normal Distribution

The Normal distribution is also known as Gaussian or Gauss distribution.

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

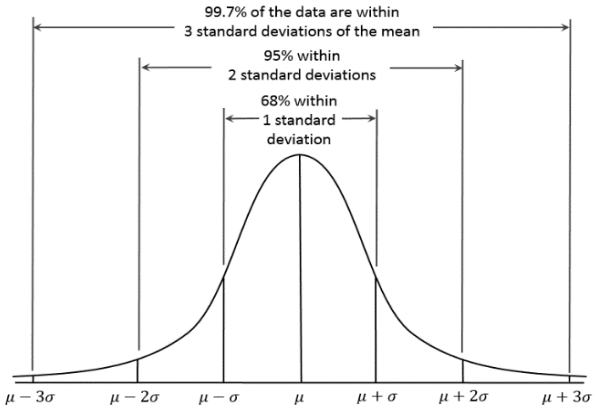


Fig. a Normally distributed curve

There are two main parameters of a normal distribution- the mean and standard deviation.

Mean

>>Researchers used the mean or average value as a measure of central tendency. It can be used to describe the distribution of variables that are measured as ratios or intervals.

>>The mean determines the location of the peak, and most of the data points are clustered around the mean in a normal distribution graph.

>>If we change the value of the mean, then the curve of normal distribution moves either to the left or right along the X-axis.

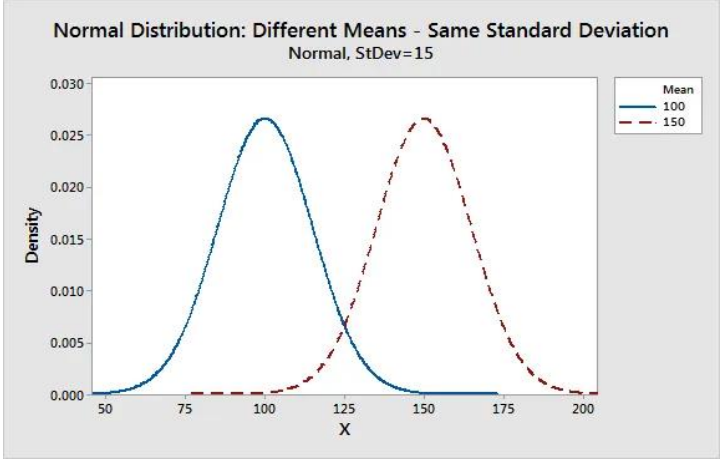


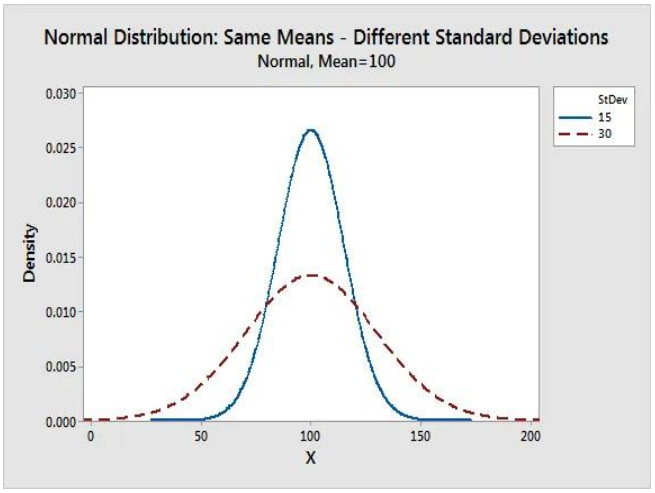
Fig.b

Standard Deviation

>> It determines how far the data points are away from the mean and represents the distance between the mean and the data points.

>> Deviation tightens or expands the width of the distribution along the x-axis.

>>Usually, a smaller standard deviation with respect to the mean results in a steep curve while a larger standard deviation results in a flatter curve.



Properties of Normal distribution

>> It is symmetrical

>>The shape of the normal distribution is perfectly symmetrical.

The mean, median, and mode are equal

>>The midpoint of normal distribution refers to the point with maximum frequency i.e., it consists of most observations of the variable.

>>The midpoint is also the point where all three measures of central tendency fall. These measures are usually equal in a perfectly shaped normal distribution.

Empirical rule

>>In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean.

>>Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.

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

Ans. **Handling missing data**

>>Real-world data is messy and usually holds a lot of missing values.

>>Missing data appear when no value is available in one or more variables of an individual.

>>Due to Missing data, the statistical power of the analysis can reduce, which can impact the Validity of the results.

>>Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it.

Imputation techniques used for missing values

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.

The following are some of the most prevalent methods:

>> Basic Imputation Techniques.

a) Imputation with a constant value.

b) Imputation using the statistics (mean, median, mode)

>>K-Nearest Neighbor Imputation.

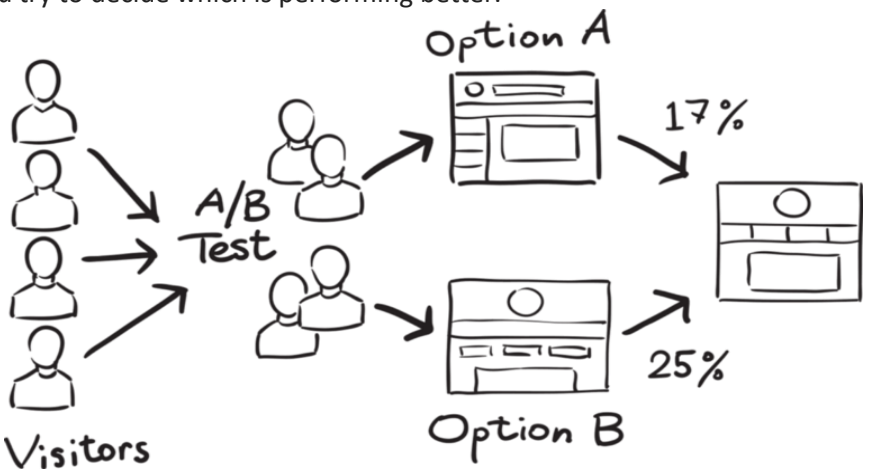
12. What is A/B testing?

Ans.

>>A/B testing is a basic randomized control experiment. It is a way to compare the two Versions of a variable to find out which performs better in a controlled environment.

>>A\B test divides the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging.

>>Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



>>It is a hypothetical testing methodology for making decisions that estimate population Parameters based on sample statistics.

>>The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

Ans.

>>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.

>>Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

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.

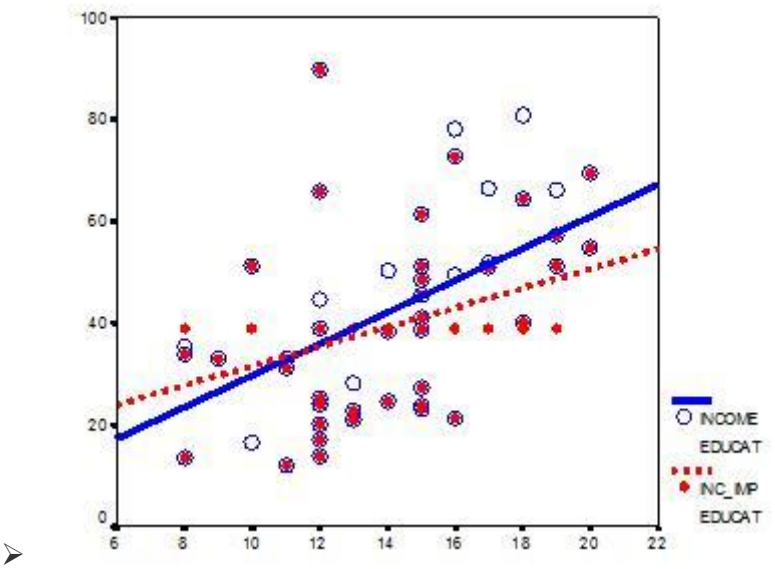
>>Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

>>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:



>>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 than .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.

>>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

14. What is linear regression in statistics?

Ans.

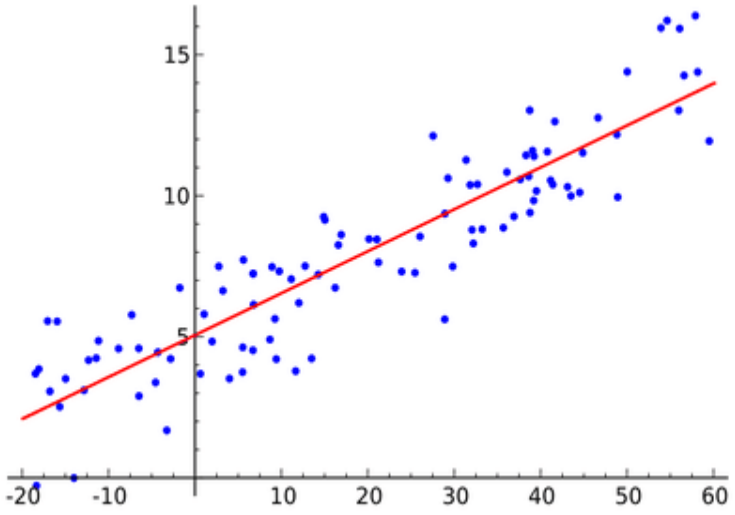
>>Linear regression analysis is used to predict the value of a variable based on the value of another variable.
>>The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
>> A linear regression line has an equation of the form $Y = a + bX$,

>> Where X is the explanatory variable and Y is the dependent variable.
The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

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

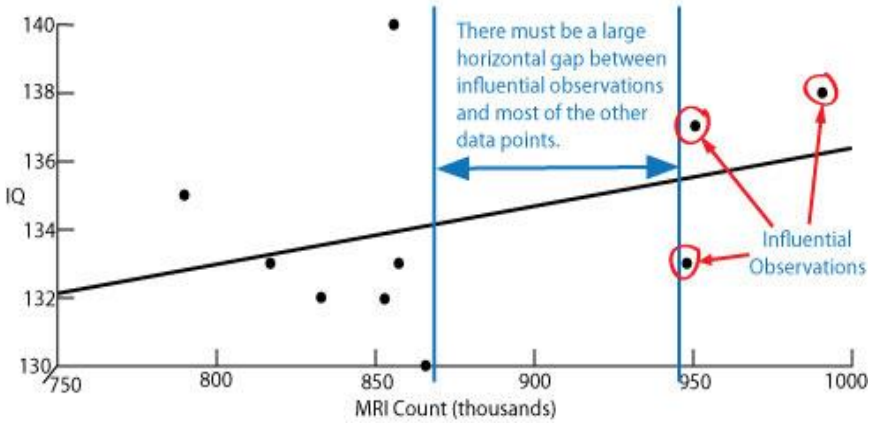
>>its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.



The fig shows Least-Squares Regression

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

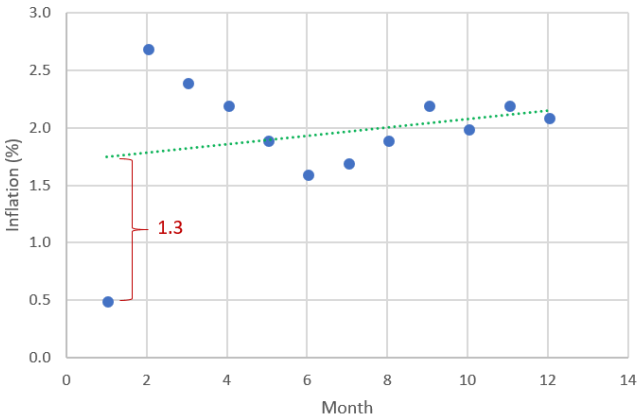


The above fig shows Outliers and Influential Observations

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.



15. What are the various branches of statistics?

Ans.

There are three real branches of statistics:

>>Data collection,
>>Descriptive statistics and
>>Inferential 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 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 through manipulate studies and

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.