STATISTICS Worksheet Set 1:

Q1. → a.)

Q2. → a.)

Q3. → b.)

Q4. → d.)

Q5. → c.)

Q6. → b.)

Q7. → b.)

Q8. → a.)

Q9. → c.)

Q10. →

Normally distributed variables are so common, many statistical tests are designed for normally distributed populations.Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

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.

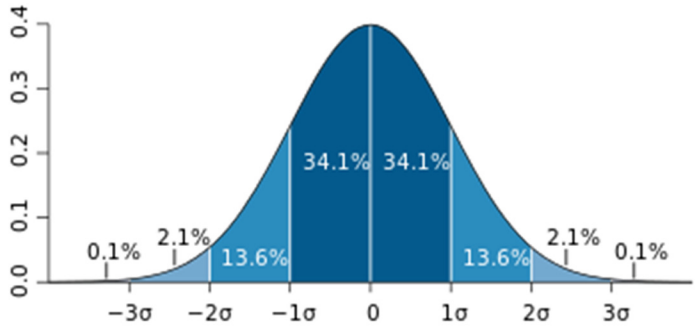
Normal distributions are symmetrical, but not all symmetrical distributions are normal.

In reality, most pricing distributions are not perfectly normal. For example, the bell curve is seen in tests like the SAT and GRE. The bulk of students will score the average (C), while smaller numbers of students will score a B or D.

It has symmetric bell shape. Half the values fall below the mean and half above the mean.Mean and Median are equal; Both located at the center of the distribution

1.) 68% of the data falls within one standard deviation of the mean.
2.)95% of the data falls within two standard deviations of the mean.
3.)99.7% of the data falls within three standard deviations of the mean.

The mean is the location parameter while the standard deviation is the scale parameter. The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.



Q11. →

In statistics, missing data, or missing values, **occur when no data value is stored for the variable in an observation**. Missing data are a common

occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Missing values can be handled by **deleting the rows or columns having** null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

There are techniques available to handle missing values.they are as follows,

1. Replacing With Mean/Median/Mode
2. Assigning An Unique Category
3. Using Algorithms Which Support Missing Values
4. Predicting The Missing Values
5.Deductive Imputation. This is an imputation rule defined by logical
6.reasoning, as opposed to a statistical rule.
7.Mean/Median/Mode Imputation. ...
8.Regression Imputation. ...
9.Stochastic Regression Imputation.
10.Listwise or case deletion. ...
11.Pairwise deletion. ...
12.Mean substitution. ...
13.Regression imputation. ...
14.Last observation carried forward. ...
15.Maximum likelihood. ...
16.Expectation-Maximization. ...
17.Multiple imputation.
18.Substitution. ...
19.Hot deck imputation. ...
20.Cold deck imputation. ...
21.Interpolation and extrapolation.

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability. Multiple imputation facilitates simple formula for variance estimation and interval estimation of the

parameter of interest.

One approach to deal with missing data is simple imputation, which is **the process whereby a single estimated value for the missing observation is obtained**, thereby enabling standard statistical methods to be applied to the augmented data set. Various methods can be implemented to impute the missing data.

In a single imputation method the **missing data are filled by some means** and the resulting completed data set is used for inference. Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean.

SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. It replaces the NaN values with a specified placeholder. fill_value : The constant value to be given to the NaN data using the constant strategy.

Mean/Median/Mode Imputation:
In this method, any missing values in a given column are replaced with the mean (or median, or mode) of that column.

The simplest imputation method is **replacing missing values** with the mean or median values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach, and one that doesn't introduce any undue bias into the dataset.

It is a popular approach because the statistic is easy to calculate using the training dataset and because it often results in good performance.

The technique which I recommend to handle missing values is

Hot-Deck Imputation:-Works by randomly choosing the missing value from a set of related and similar variables. Cold-Deck Imputation:-A systematically chosen value from an individual who has similar values on other variables. This is similar to Hot Deck in most ways, but removes the random variation.

Q12. →

A/B testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics

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.) ... In A/B testing, **A refers to 'control' or the original testing variable**. Whereas B refers to 'variation' or a new version of the original testing variable.

A/A tests, which are often used to **detect whether your testing software is working**, are also used to detect natural variability. It splits traffic between two identical pages. If you discover a statistically significant lift on one variation, you need to investigate the cause.

When do you do an AB test?

The 5 Times When You Absolutely Must do A/B Testing

- Do A/B testing when you redesign your website. ...
- Do A/B testing when you change a service, plugin, or feature. ...
- Do A/B testing when you change prices. ...
- Do A/B testing when you think your conversion rates might be screwed. ...
- Do A/B testing when you just want to raise revenue.

How to Conduct A/B Testing

1. Pick one variable to test. ...
2. Identify your goal. ...

3. Create a 'control' and a 'challenger. ...
4. Split your sample groups equally and randomly. ...
5. Determine your sample size (if applicable). ...
6. Decide how significant your results need to be. ...
7. Make sure you're only running one test at a time on any campaign.

Q13. →

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

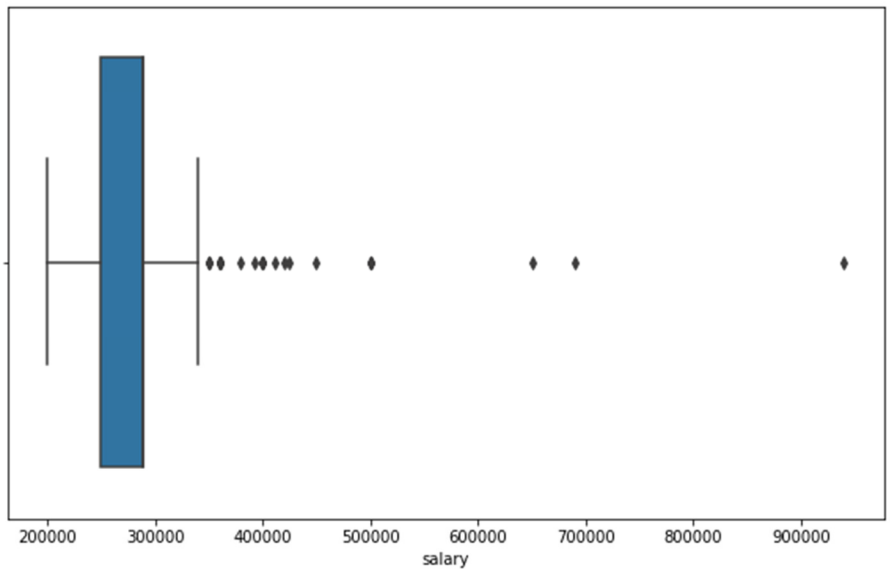
Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out.

Mean reduces a variance of the data

As we can see, the variance was reduced (that big change is because the dataset is very small) after using the Mean Imputation. Going deeper into mathematics, a smaller variance leads to the narrower confidence interval in the probability distribution[3].

```
In [118]: fig, ax = plt.subplots(figsize=(10, 6))
          sns.boxplot(df.salary)

Out[118]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8e0ac37ed0>
```



Q14. →

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

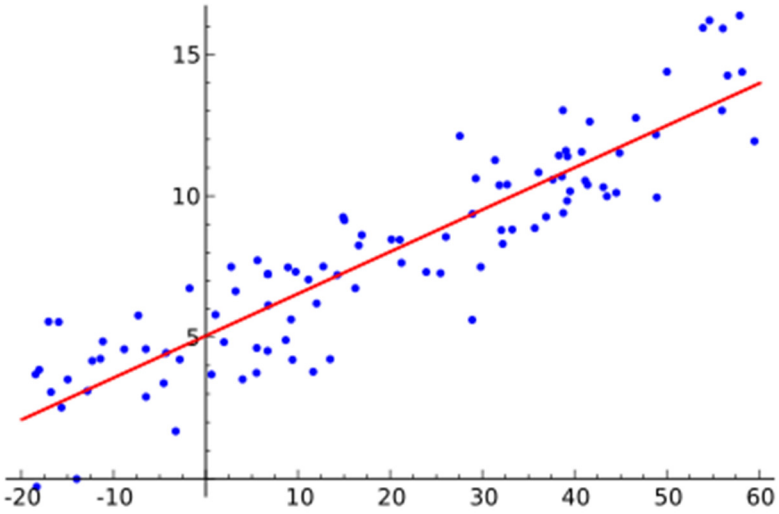
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. ... For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome. Regression can help finance and investment professionals as well as professionals in other businesses.

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables(also known as dependent and independent variables). Such models are called linear models.

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.



Q15. →

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.
There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

Descriptive statistics:
A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information, while descriptive statistics is the process of using and analysing those statistics.
Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).
Descriptive statistics can be useful for two purposes: 1) to provide basic information about variables in a dataset and 2) to highlight potential relationships between variables.
Descriptive statistics allow a researcher to quantify and describe the basic characteristics of a data set. As such, descriptive statistics serve as a

starting point for data analysis, allowing researchers to organize, simplify, and summarize data.

There are a variety of descriptive statistics. Numbers such as the mean, median, mode, skewness, kurtosis, standard deviation, first quartile and third quartile, to name a few, each tell us something about our data.

Descriptive statistics help you to simplify large amounts of data in a meaningful way. It reduces lots of data into a summary. Example 2: You've performed a survey to 40 respondents about their favorite car color.

Descriptive statistics are limited in so much that they only allow you to make summations about the people or objects that you have actually measured. You cannot use the data you have collected to generalize to other people or objects (i.e., using data from a sample to infer the properties/parameters of a population).

Inferential statistics :

Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates

Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects. There are many types of inferential statistics and each is appropriate for a specific research design and sample characteristics.

Descriptive statistics summarize the characteristics of a data set. Inferential statistics allow you to test a hypothesis or assess whether your data is generalizable to the broader population. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study.

What are the four types of inferential statistics?

- One sample test of difference/One sample hypothesis test.
- Confidence Interval.
- Contingency Tables and Chi Square Statistic.
- T-test or Anova.
- Pearson Correlation.
- Bi-variate Regression.
- Multi-variate Regression.

The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.
For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study.
How many types of inferential tests are there?

There are three basic types of t-tests: one-sample t-test, independent-samples t-test, and dependent-samples (or paired-samples) t-test. For all t-tests, you are simply looking at the difference between the means and dividing that difference by some measure of variation.
The first, and most important limitation, which is present in all inferential statistics, is that you are providing data about a population that you have not fully measured, and therefore, cannot ever be completely sure that the values/statistics you calculate are correct.