1. a
2. a
3. b
4. d
5. c
6. b
7. b
8. a
9. c
10. A random variable X with probability density function f(x) = [1/{sq.rt(2*pi*)*sigma}]*e[-{x-mu}.sq/{2 * (sigma.sq)}] is a normal random variable with parameters mu and sigma where mu lies between negative and positive infinity and sigma > 0.

Normal distribution is a bell shaped curve with a mean mu and standard deviation sigma.

99% data lies in 3 standard deviations from mean

95% data lies in 2 standard deviations from mean

68% data lies in standard deviations from mean.

Standard normal distribution is a special case of normal distribution with a mean 0 and sd 1.

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

Missing data is handled in various ways.

With respect to a feature if only some rows may be like less than 1% rows are missing this feature then we can just drop such rows.

On the other hand, if more than 50% data has a feature missing, then it is better to drop the entire feature column.

Imputation techniques:

Other option is to fill the missing data with values. It is based on the type of the data that feature holds, for example if a feature is number of cars in basement, if a house has no basement then the number of cars feature can be just filled with 0 value.

For numerical columns we can fill with mean values of that column.

For categorical column, they are filled with mode of that feature.

If some numerical column is based on other feature of the data , in such cases we first group by based on the feature and this we take mean of the value and fill the missing value with the mean of that group.

12.

A/B testing refers to the experiments where two or more variations of the same webpage are compared against each other by displaying them to real-time visitors to determine which one performs better for a given goal. A/B testing is not limited by web pages only, you can A/B test your emails, popups, sign up forms, apps and more.

Different kinds of metrics can be used to measure a website efficacy. With discrete metrics, also called binomial metrics, only the two values 0 and 1 are possible. The following are examples of popular discrete metrics.

Click-through rate — if a user is shown an advertisement, do they click on it?

Conversion rate — if a user is shown an advertisement, do they convert into customers?

Bounce rate — if a user is visits a website, is the following visited page on the same website?

With continuous metrics, also called non-binomial metrics,, the metric may take continuous values that are not limited to a set two discrete states. The following are examples of popular continuous metrics.

Average revenue per user — how much revenue does a user generate in a month?

Average session duration — for how long does a user stay on a website in a session?

Average order value — what is the total value of the order of a user?

A two-sample hypothesis test is used. Null hypothesis H0 is that the two designs A and B have the same efficacy, i.e. that they produce an equivalent click-through rate, or average revenue per user, etc. The statistical significance is then measured by the p-value, i.e. the A two-sample hypothesis test is used. Null hypothesis H0 is that the two designs A and B have the same efficacy, i.e. that they produce an equivalent click-through rate, or average revenue per user, etc. The statistical significance is then measured by the p-value, i.e. the probability of observing a discrepancy between our samples at least as strong as the one that we actually observed.

13. Is mean imputation of missing data acceptable practice?

Mean imputation of missing data is not always a best practice.

1. The correlation is affected by simply filling the missing data with missing values.
2. Underestimation of Standard errors with mean imputation and it leads to type 1 errors in hypothesis testing.

14.

Linear Regression is a predictive model that assumes a linear relationship between the dependent variable and independent variables. Linear Regression can be implemented using Ordinary Least squares method, Gradient descent algorithm, Normal equation.

15.

The main branches of statistics are Descriptive statistics and inferential statistics.

Descriptive Statistics describes the important characteristic of the data using measures of central tendency like mean, median, mode.

Inferential statistics : it uses data from sample and makes inferences about larger population using methods like Hypothesis tests.