WORKSHEET_SET_6

1) d

2) a

3) a

4) c

5) a,b

6) d

7) c

8) b

9) b

10) Below is the difference between Box Plot and Histogram:

While boxplots and histograms are visualizations used to show the distribution of the data, they communicate information differently.

Histograms are bar charts that show the frequency of a numerical variable's values and are used to approximate the probability distribution of the given variable. It allows you to quickly understand the shape of the distribution, the variation, and potential outliers.

Boxplots communicate different aspects of the distribution of data. While you can't see the shape of the distribution through a box plot, you can gather other information like the quartiles, the range, and outliers. Boxplots are especially useful when you want to compare multiple charts at the same time because they take up less space than histograms.

12) Statistical significance can be accessed using hypothesis testing:
– Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
– Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
– Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
– We calculate the observed test statistics from the data and check whether it lies in the critical region
Common tests:
– One sample Z test
– Two-sample Z test
– One sample t-test

– paired t-test
– Two sample pooled equal variances t-test
– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
– Chi-squared test for variances
– Chi-squared test for goodness of fit
– Anova (for instance: are the two regression models equals? F-test)
– Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

13) On the right, I tallied the measurements in a histogram. This can help us check if a variable is Gaussian or not.

Non-Gaussian distributed time series data arise when the mean or noise statistics vary with time.

If the mean varies with time, the variable could be non-stationary / time-varying (its trend changes with time), auto- or cross-correlated (it changes depending on its previous value or the values of other variables), or its value is computed from the values of other Gaussian variables but in a nonlinear way.

14)

Let's say you run a customer satisfaction survey with a sample of 9 and rate their overall satisfaction scores on a scale of 1 to 10. You get an average of 5.22. You know that in general, you tend to retain customers with a score over 3, so you're satisfied, because this indicates that you're still above where you want to be. But then, suddenly, you lose 6 of those 9 customers. You go back to look at your data, and you find these scores:

1, 3, 3, 3, 3, 5, 9, 10, 10

The median of this group is a 3, indicating that at least half of your customers or more were unhappy. The scores became lopsided because of the unexpected 10's, and you missed out on an important part of your data – the midpoint that indicated that as many as half of your customers or more were dissatisfied with your company.

Median can play a major role in things like income level research as well, because a few millionaires may make it look like the socio-economic status of your sample is higher than it really is.

Whenever a graph falls on a normal distribution, using the mean is a good choice. But if your data has extreme scores (such as the difference between a millionaire and someone making 30,000 a year.

15) Likelihood functions are an approach to statistical inference (along with Frequentist and Bayesian). Likelihoods are functions of a data distribution parameter. For example, the binomial likelihood function is

$$L(\theta)=n!x!(n-x)!\cdot \theta_x \cdot (1-\theta)_{n-x} L(\theta)=n!x!(n-x)!\cdot \theta x \cdot (1-\theta)n-x$$

You can use the binomial likelihood function to assess the likelihoods of various hypothesized population probabilities, $\theta\theta$. Suppose you sample n = 10 coin flips and observe x = 8 successful events (heads) for an estimated heads probability of .8. The likelihood of a fair coin, $\theta\theta$ = .05 given the evidence is only 0.

11)

User average metrics

User level conversion metrics

Page view level conversion metrics

Percentile metrics

Sum metric