**STATISTICS– WORKSHEET 6**

Solution:1 D

Solution:2 A

Solution:3 A

Solution:4 C

Solution:5 A

Solution:6 A

Solution:7 C

Solution:8 B

Solution:9 B

**Solution:10** Although histograms and box plots are collectively part of the chart aid category, they do represent very different types of charts. Both charts effectively represent different data sets; however, in certain situations, one chart may be superior to the other in achieving the goal of identifying variances among data. The type of chart aid chosen depends on the type of data collected, rough analysis of data trends, and project goals.

1. A histogram is highly useful when wide variances exist among the observed frequencies for a particular data set. This is important because to improve processes, it is critical to understand what is causing these three modes. Had this data simply been graphed using a box plot, the values would average one another out, causing the distribution to look roughly normal.
2. Another instance when a histogram is preferable over a box plot is when there is very little variance among the observed frequencies. The histogram displayed to the right shows that there is little variance across the groups of data; however, when the same data points are graphed on a box plot, the distribution looks roughly normal with a high portion of the values falling below six.
3. The final set of graphs shows how a box plot can be more useful than a histogram. This occurs when there is moderate variation among the observed frequencies, which causes the histogram to look ragged and non-symmetrical due to the way the data is grouped. This may lead one to assume the data is slightly skewed. However, when a box plot is used to graph the same data points, the chart indicates a perfect normal distribution.

**Solution:11** There isn’t a one-size-fits-all metric. The metric(s) chosen to evaluate a machine learning model depends on various factors:

Is it a regression or classification task?

What is the business objective? Eg. precision vs recall

What is the distribution of the target variable?

There are a number of metrics that can be used, including adjusted r-squared, MAE, MSE, accuracy, recall, precision, f1 score, and the list goes on.

**Solution:12** You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

**Solution:13** Any type of categorical data won’t have a gaussian distribution or lognormal distribution.

Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

**Solution:14** When there are a number of outliers that positively or negatively skew the data.

**Solution:15** Likelihood is a measure of the extent to which a sample provides support for particular values of a parameter in a parametric model.