Long questions:

Q1: What is a histogram and box-plot? Give a situation where you would prefer a histogram to box-plot and vice-versa.

Ans:

A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal).

The only difference between a histogram and a bar chart is that a histogram displays frequencies for a group of data, rather than an individual data point; therefore, no spaces are present between the bars. Typically, a histogram groups data into small chunks (four to eight values per bar on the horizontal axis), unless the range of data is so great that it easier to identify general distribution trends with larger groupings.

A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value. When graphing this five-number summary, only the horizontal axis displays values. Within the quadrant, a vertical line is placed above each of the summary numbers. A box is drawn around the middle three lines (first quartile, median, and third quartile) and two lines are drawn from the box’s edges to the two endpoints (minimum and maximum).

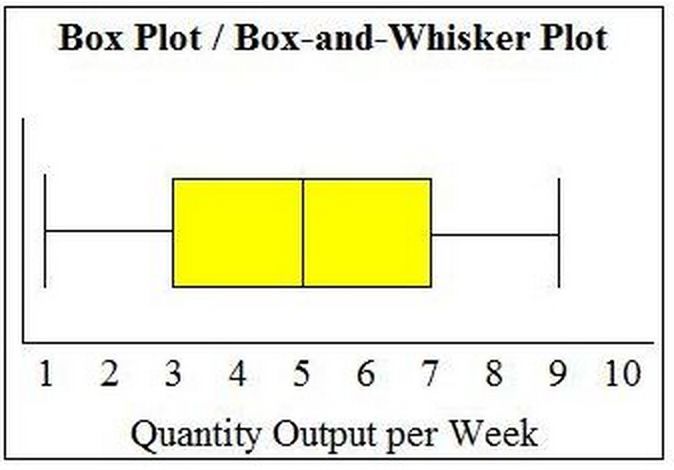
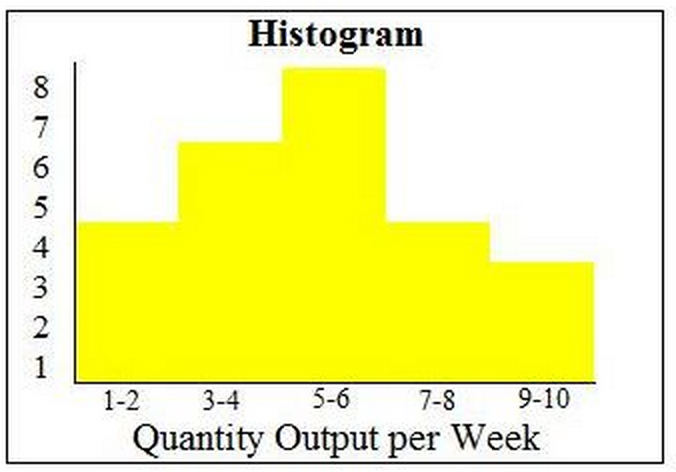
Histogram is helpful when you want to compare frequencies between different variables. A box-plot is helpful when you want to measure and compare different descriptive values of a set (minimum, first quartile, median, third quartile and maximum). A histogram is useful when you have either wide or low range of variance between the sets. The histogram helps us to determine the peaks within the data and make decision based on that. A simple box plot averages the values and the data appears to be normal. This feature of box plot is helpful to see if the data is normally distributed and not skewed. So we can plot both types of graph and see the different types of result that these present.

For eg:

Cases where Histogram might be better.

|  |  |
| --- | --- |
| Macintosh HD:Users:pratyush:Desktop:Screen Shot.pngMacintosh HD:Users:pratyush:Desktop:Screen Shot 1.png | Macintosh HD:Users:pratyush:Desktop:Screen Shot 2.pngMacintosh HD:Users:pratyush:Desktop:Screen Shot 3.png |
| Case 1 | Case 2 |
| The histogram is more helpful because it shows the variance between the different sets which is not visible in the box plot. | |

Case where box plot might be useful.



The Histogram shows that the data might be skewed to the left but the box plot helps us visualize the even distribution.

Image source: http://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn\_5

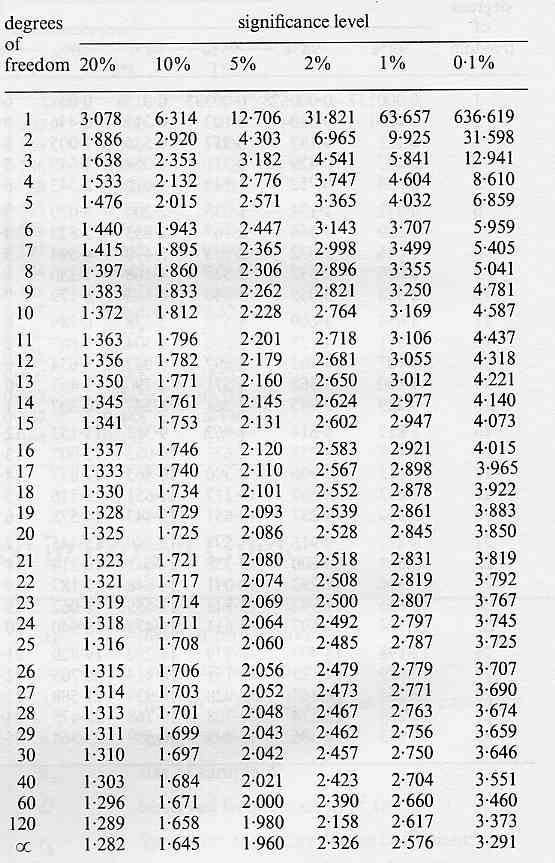
Q2: Given the following statistical data for a sample, state the null hypothesis and find out if we are 95% confident that the mean age of the population is 55 years. Use the data mentioned below:

1. Mean age of the sample: 53

2. Standard error: 0.92

3. Sample size: 30

4. The table of critical values below.



Ans:

α **=** .05

H0 = Mean age of the population is 55 years with 95% confidence.

t = = (55 – 53)/.92 = 2.17

From the table we can find the critical value of t.05 for (N-1) i.e. for 29 = 2.043

Since |t| >|tα| 🡪 H0 is not true.

Hence we cannot say that the mean age of the population is 55 years with 95% confidence.

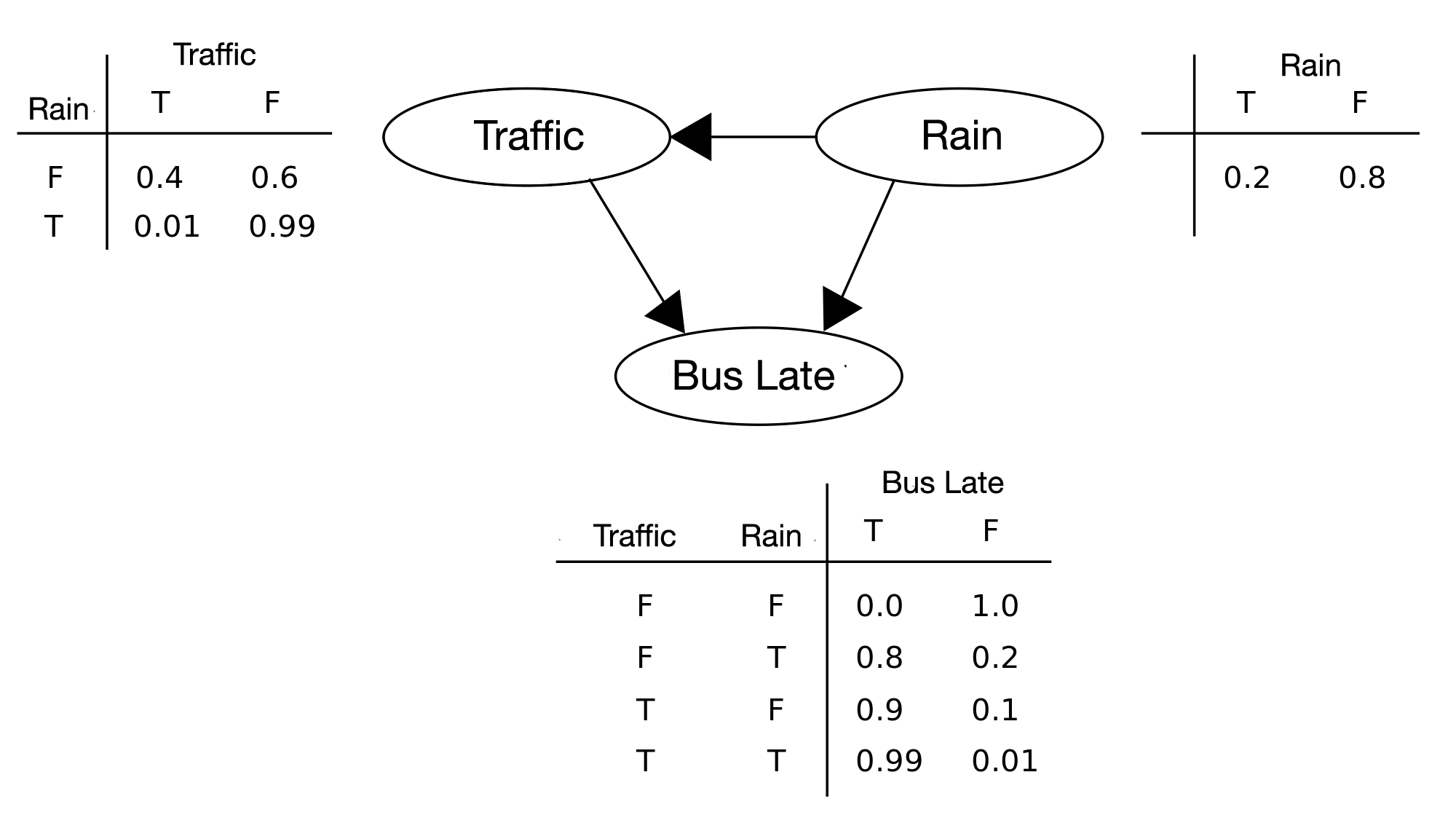
Short Questions:

Q3: What is Pruning and what are the different ways of pruning?

Ans:

Pruning is a method to check overfitting in a decision tree. Pruning helps us to remove the unnecessary nodes and reduce the depth of the nodes, which increase the noise in the decision trees. We can do prepruning or postpruning on the tree. It is more accurate to do postpruning as you can see the effect of pruning at each step. When you perform the reduced error pruning, you basically start with the leaf node, replace it with the most popular class and then see the affect on the accuracy. If the accuracy is not affected then we can keep the change otherwise we revert back.

Q4: Consider the following Bayesian network



Use the network to calculate the probability that if the bus is late then it must be raining?

Ans:

P(Rain) = 0.2

P(Late|Rain) = Probability of the bus being late when rain is true = 0.8\*0.2 + 0.99\*0.2

P(Late) = Probability that the bus is late = Bus late = true from table 3 and combination of rain and traffic from table 1 = (0.0\*0.6 + 0.8\*0.99 + 0.9\*0.4 + 0.99\*0.01)

P(Rain|Late) = P(Late|Rain)\*P(Rain) / P(Late)

= (0.8\*0.2 + 0.99\*0.2)\*0.2 / (0.0\*0.6 + 0.8\*0.99 + 0.9\*0.4 + 0.99\*0.01)

= 0.06

Q5: What is significance test of correlation and why do you need it?

Ans:

Pearson correlation represents the strength of a relationship between two independent variables. It can be misleading and cause incorrect inferences. For example there could be a very high correlation between rain and foreign exchange rate, but that does not mean that one affects the other. Therefore we should test the significance of the correlation as well. We can do this by doing a t-test for two variables. A t-test will measure how likely is this correlation to exist. We state the confidence level (say 95%) and test if the data can predict such level of probability for that correlation. Pearson correlation coefficient measures strength of linear relationship. Scatterplots are useful for checking whether the relationship is linear. The correlation coefficient is the average product of departures of two variables from their respective means divided by the product of the standard deviations of those variables.

In order to test the correlation our variables should satisfy the following assumptions:

The assumptions are:

1. The samples, x and y, are drawn from populations that follow a bivariate normal distribution

2. The samples are random samples from the population

3. The population correlation coefficient is zero.

Some caveats to interpretation of the correlation coefficient are

 Relationship in question should be linear

 Correlation does not imply causality

 Correlation does not address lagged relationships

 Statistical significance may not imply practical significance

Causality

A significant correlation coefficient between x and y does not necessarily imply that variations in x ”cause” variations in y, or vice versa. Sometimes the correlation results simply from both x and y being associated with a third variable that might have some causative influence on both x and y. For example, a positive correlation between tree-growth variations in Arizona and Colorado in no way implies that tree-growth in one region affects tree-growth in the other, but merely that trees in both regions are influenced by a third variable – precipitation – which is spatially coherent across the Southwest.

Source: http://www.ltrr.arizona.edu/~dmeko/notes\_9.pdf