Solution for HW1(38 points total)

STA113, ISDS

1 2.48(13points)

Refer to the Lab2 for programming. Some hints are provided below.

- hint(a): median is the average of the middle two observations, once the observations have been arranged in order.

- hint(c):The number of observations that appear in $\overline{y} \pm s$ is 31. This is close to the 68 % or $0.68 * (50) \approx 34$ given by the Empirical Rule.
  The number of observations that appear in $\overline{y} \pm 2s$ is 49. This is close to the 95 % or $0.95 * (50) \approx 48$ given by the Empirical Rule.
  The number of observations that appear in $\overline{y} \pm s$ is 50. This agrees with ≈ 1 given by the Empirical Rule.

- hint(d): Inputting boxplot(filename) in matlab, you can easily get the figure. However, we will take a second to review how the box plot works. You first compute Q_L and Q_U. $Q_L$is the data point with the rank of $(n+1)/4 \approx 13$.The 13th ranked data point is 109. $Q_U = 3(n+1)/4 \approx 38$. The 38th ranked data point is 131. The interquartile range, IQR, is $Q_U - Q_L = 22$. The inner fences are located $1.5(IQR) = 33$ below Q_L and above Q_U. The inner fences are 76 and 164. The outer fences are located $3(IQR) = 66$ below Q_L and above Q_U.The outer fences are 43 and 197. There is no suspect outliers because no data points lie outside the inner fence. A picture of this box plot has been attached to the end of this file(See figure1). Please check if you have any questions.

- hint(e):The 70 % percentile is the data point that has a rank of $.70(n) = 35$. The 35th data point is 128. Seventy percent of the data points have a value less than or equal to 128.

2 2.54(10points)

- a. Using 'boxplot' command and 'hist' command, we get a box plot and histogram respectively. These figures have been attached to our solution in case you may want to have a check with yours(See figure2 and figure3).

- b. You can use commands in Lab2 to generate descriptive statistics in Matlab. Be sure to generate the following descriptive statistics: N, Mean, Std Dev, Max, Q_U, Med, Q_L, and Min. The results means the average of the 164 observations is 63.7. Half of the 164 observations have a value below 4, and we expect most of the values to fall in the interval $63.7 \pm 3(219.9)$.

- c. The textbook discussed two rules for identifying outliers: 1.5IQR (suspect) and 3IQR (highly suspect). When you do problem 2.54(c), you can choose either of the rules. We use the 3IQR method here.
  A good boxplot will help you to solve this problem. Using $IQR = 11.5$, $Q_U = 12.5$, $Q_L = 1$, any observation falling above $Q_U + 3IQR = 47$ would represent an outlier. You do not need to consider observation falling below $Q_L - 3IQR$ because no one falls in this interval.

- d. Matlab code is provided in lab2, see the part 'Empirical Rule and Outliers'. Ch . eck your figures with figure4 and figure5.

- e. We see that the mean and the standard deviation were greatly affected by the extreme values in the data set. The median, however, did not change very much. This lead us to conclude that the median is not affected very much by those outliers.

3 3.14(3points)

hint, $P(A|B) = P(AB)/P(B)$.

4 3.24(2points)

- a. Events are considered mutually exclusive if they cannot occur at the same time.

- b. Mutually exclusive events are not independent. Check whether $P(X|Y) = P(X)$ for independence.

5 Additional Problem(2points)

5.1 Solution1

$$\begin{aligned} \sum(y_i - a)^2 &= n(a - \overline{y})^2 + \sum y_i^2 - n\overline{y}^2 \\ &\geq \sum y_i^2 - n\overline{y}^2 \end{aligned}$$

So when $a = \overline{y}$, the sum reaches its minimum.

5.2 Solution2

Suppose $f(a) = sum(y_i - a)^2$; then take derivative of f(a) with respect to 'a' to find the extreme value at $a = \overline{y}$. Be sure to take 2nd derivative of $f(a)$ to make sure that $f(a)$ is a convex function instead of concave, that is, $f(\overline{y})$ is a min, not a max.

PS: questions 3.3, 3.8 are assigned 3 points and 5 points respectively.

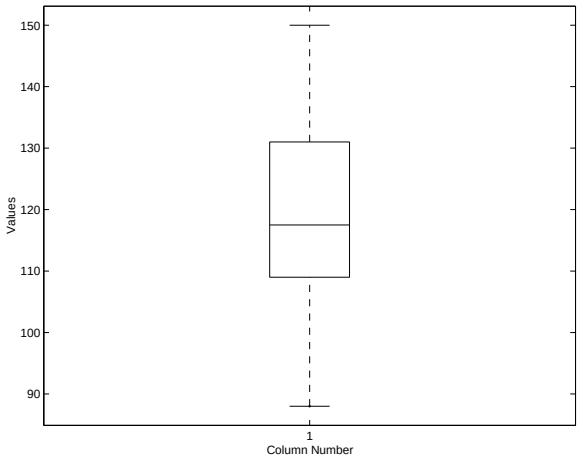

Figure 1: Boxplot for Q2.48

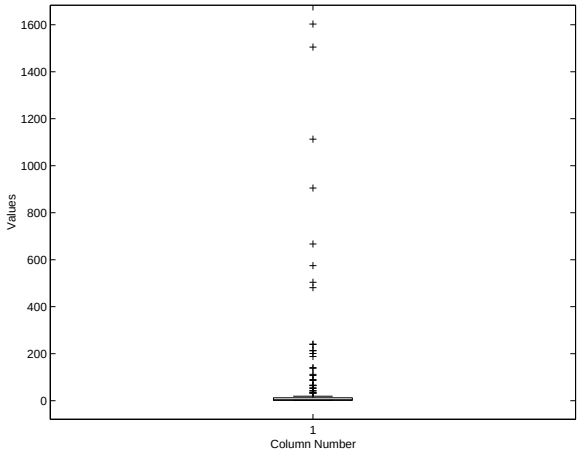

Figure 2: Boxplot for Q2.54

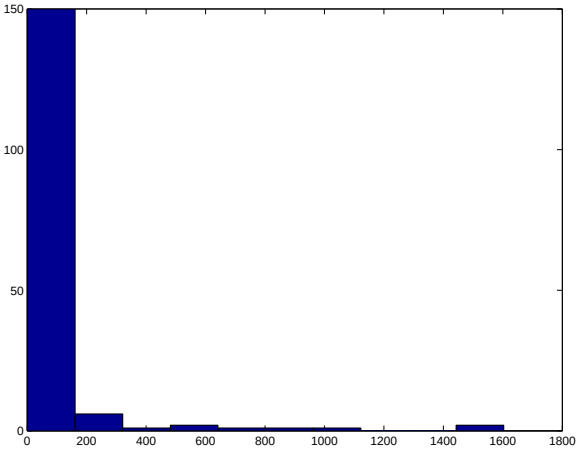

Figure 3: Histogram for Q2.54

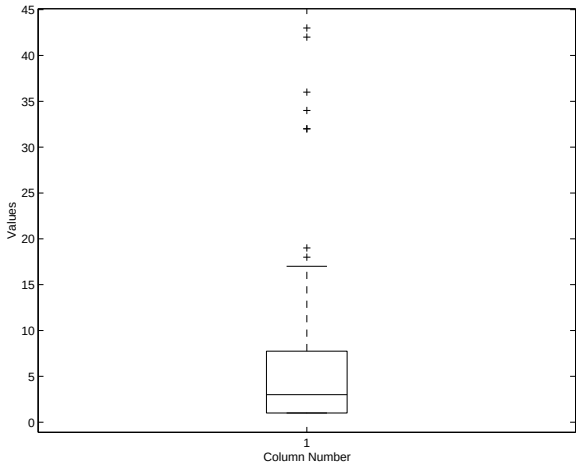

Figure 4: Boxplot for newQ2.54

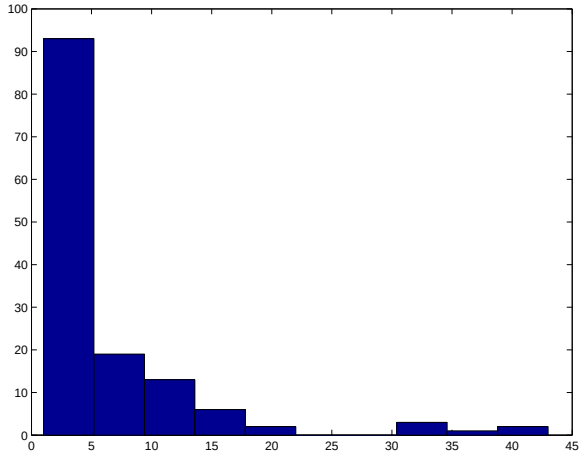

Figure 5: Histogram for newQ2.48