A quick trip through the text book shows that during our work in chapters 1, 2, and 3.1 that we have discussed the following:

- types of variables: categorical, quantitative;
- center, spread, shape
- stemplot
- histogram
- boxplot
- density plot
- modes, symmetry, skew
- outliers (suspected outliers 1.5IQR)
- mean, median
- quartiles, IQR, standard deviation
- five-number summary
- normal: rules of thumb, z-scores, standard normal, inverse values, qqplot
- lurking variables, confounding variables,
- types of experiments
- comparative experiment
- placebo effect
- bias: lack of realism

Here are some sample problems to work on during class. These are not meant to exhaust the full range of questions I may ask.

1. Recalling the rules of thumb for a normal, if math SAT scores for CSI students are normally distributed with a mean of 500 and standard deviation of 100, find the probability a randomly chosen student has an SAT score exceeding 600? Between 500 and 700?

   ANS: 1) the zscore is 1, the area is to the right, so so, $(1 - .68)/2$. 2) The zscore is -1 to 1, so just .68.

2. A data set consists of values 1,1,2,3,4,5,8,15.

   (a) What is n

   (b) What is $\bar{x}$?

(c) What is s

(d) What is the median?

(e) What is the IQR

ANS: n is 8. As for $\bar{x}$:

```
x = c(1,1,2,3,4,5,8,15)
n = length(x)
xbar = sum(x)/n
xbar

## [1] 4.875
```

The value of the standard deviation is

```
squares = (x - xbar)^2
squares

## [1]  15.015625  15.015625   8.265625   3.515625   0.765625   0.01562
## [8] 102.515625

s2 = sum(squares) / (n-1)
s = sqrt(s2)
s

## [1] 4.703722
```

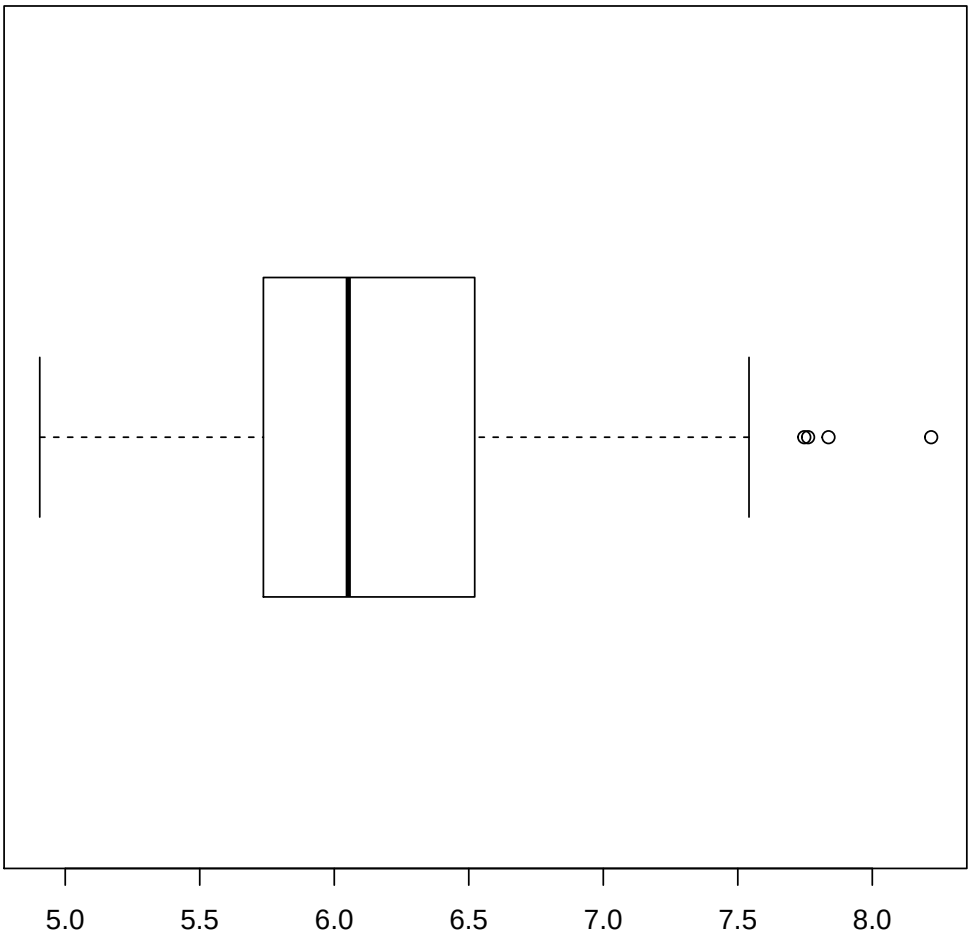
The median in the "middle" value, between 3 and 4, or 3.5.

The hinges are between 1 and 2, or $Q_1 = 1.5$ and 5 and 8, so $Q_3 = 6.5$, giving $IQR = Q_3 - Q_1 = 6.5 - 1.5 = 5$.

3. A boxplot of a data set is shown. Based on this identify

   Uhh, where is the boxplot!!! (It was down below). Here is a similar one:

```
boxplot(log(rivers), horizontal=TRUE)
```



(a) the median

(b) The IQR

(c) The max.

(d) Will the mean or median be greater? Why

The median is the heavy bar, 6.1 say; the IQR, the distance of the box, 0.75 say; the max, the rightmost, value, 8.5 say; and as this is skewed right, the mean is expected to be larger than the median.

4. The paired data $((x,y))$: $(1,6),(5,8),(7,9),(8,0),(4,1)$ is considered.

   (a) Find the pearson correlation coefficient

(b) Find the slope and intercept of the regression coeffients.

(c) What percent of the variation in y is described by variation in x?

The peason correlation is found through:

```
x = c(1,5,7,8,4)
y = c(6,8,9,0,1)
n = 5 # length(x)
xbar = mean(x)
ybar = mean(y)
sx = sd(x)
sy = sd(y)
zx = (x - xbar)/sx
zy = (y - ybar)/sy
r = sum(zx * zy) / (n-1)
```

The slope of the regression line is rs_y/s_x:

```
b1 = r * sy / sx
b1
```

```
## [1] -0.2333333
```

The intercept is found from

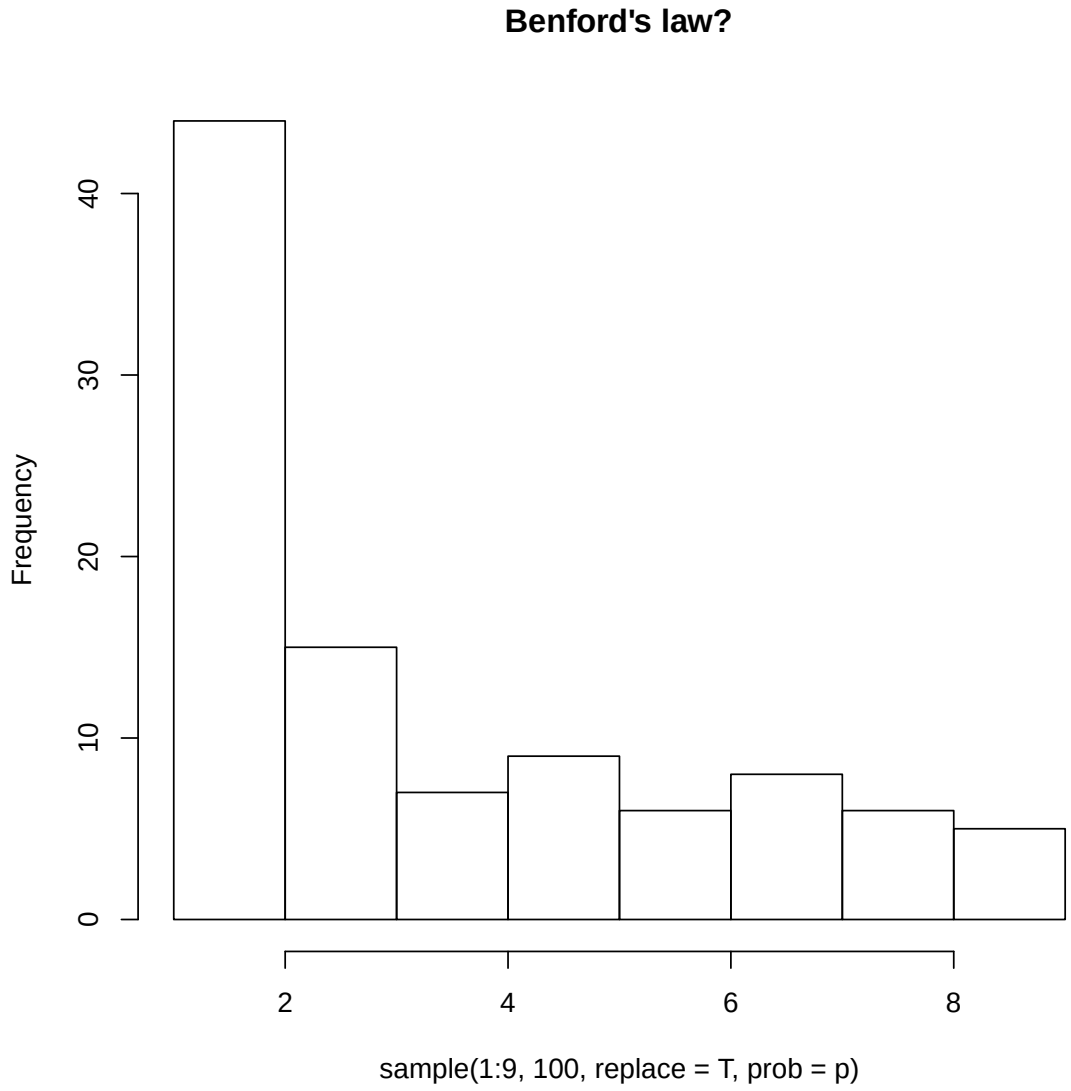
```
b0 = ybar - b1*xbar
b0
```

```
## [1] 5.966667
```

The value of $r^2 \cdot 100\%$ explains the percent of variation in the y values due to the linear model (assuming it applies):

```
r^2
```

```
## [1] 0.0244511
```



5.

A statistics teacher scrupolously records the first digit of each of receipts they receive. A histogrram is shown.

(a) Is this data set skewed? symmetric? neither?

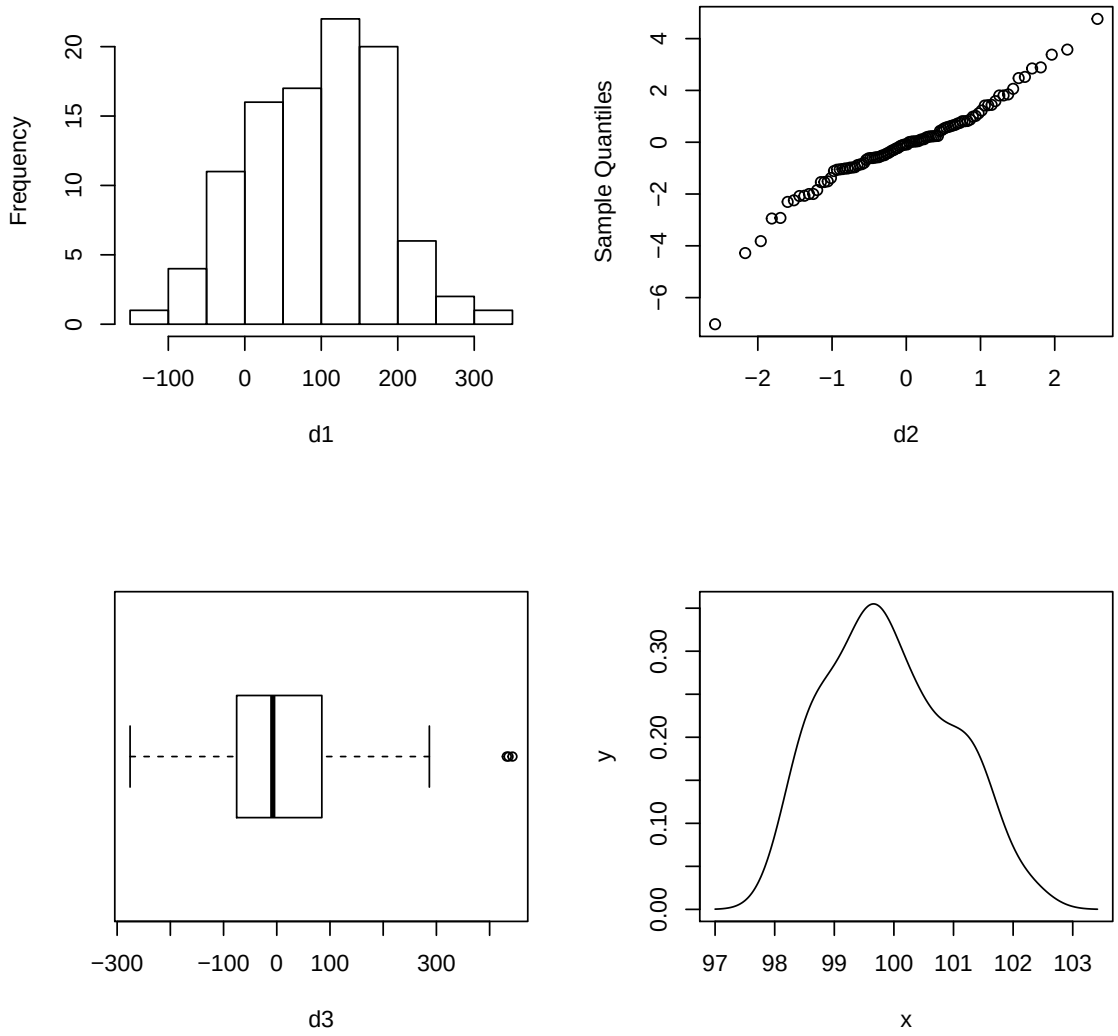
(b) Is this data set unimodal, bimodal, multimodal?

(c) Estimate the mean value for this data set.

(d) there is something funny about this graphic in terms of representing the data. Can you tell what it is?

ANS: Skewed right; unimodal, mean is about 4, finally, 0 and 1 get lumped together into one bar.

6. Which of these data sets appears to be normally distributed?



The histogram and density are, the quantile plot and boxplot have longer tails.

7. Some statistics are not resistant to a single outlier. An example would be the mean value, as one extremely large of small value can tip the scales. Which of these statistics is also not resistant to outliers: median, standard deviation, IQR, range, Pearson correlation, linear regression coefficients?

   ANS: standard deviation, range, correlation, and the regression coefficients are all not resistant.

8. For the regression model, what is the difference between an *outlier* and a *influential* observation?

   ANS: An outlier is far from the pattern, an influential value may also be an outlier (leaving it out dramatically effects the estimates), but outliers need not be influential.

9. The following are all described by the books as cautions when using the language of correlation and regression to a bivariate set of data:

   (a) Correlation only measures linear association, so results do not apply to non-linear associations

   (b) Extrapolation of the linear regression line for prediction purposes outside the range of the data can be problematic

   (c) Correlation and least squares regression are not resistant

   (d) Association does not imply causation

   Can you provide scenarios or examples illustrating why caution is necessary?

   ANS: Well, a U-shaped data set may have low correlation; but be associated. If you modle miles per gallon by weight of car using smaller cars, the model may estimate heavey SUVs to have negative mileage; A least squares fit of brain size by body mass is influenced by dinosaurs (we saw this in class); finally, soda sales and ice cream sales my be related in the summer, but likely due to hotter days influencing the sale of each.

10. Principles of experimental design are "Compare", "Randomize", and "Repeat." Match these terms with the following reasons:

    (a) This controls the bias due to cohort selection

    (b) This reduces chance variation in the results

    (c) This controls the effects of lurking variables

    These are compare; repeat; randomize

11. The book has an example of study where a treatment group was given asked to smoke marijuana cigarettes and a control group was asked to smoke non-marijuana cigarettes. This was an example of:

    - A double-blind study
    - An example where the placebo was recognized
    - An example of the failure of randomization

    ANS: The latter; kids are just too smart these days...

12. Thirty students are asked to usa an online homework system which randomizes each question asked, so that students do not see the exact same question. After a month they are surveyed as to their satisfaction on a 5-point scale.

    - What are the experimental units?
    - What is the treatment(s)?
    - Is the set of students who did not log on to the system a control group?

- Is this a comparative study?
- Could this study have been improved by randomization?

ANS: The students; the using of the system; no they are likely biased in some way; no, there is no comparison group. Well, maybe if two treatments were used, e.g., one group uses online HW, the other paper and pencil.