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?
- 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
- 3. A boxplot of a data set is shown. Based on this identify
  - (a) the median
  - (b) The IQR
  - (c) The max.
  - (d) Will the mean or median be greater? Why
- 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 coefficients.
  - (c) What percent of the variation in y is described by variation in x?



50

40

30

20

9

0

Frequency





2 4 6 8 sample(1:9, 100, replace = T, prob = p)

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



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

- 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
- 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
- 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?