[*Hollywood-style disclaimer: The following contains fictional test questions. Any resemblance these questions have to actual test questions is purely coincidental.*]

So far in this term we have discussed data sets and how to describe them. We have learned methods to view their distribution and numeric summaries.

You should be familiar with these descriptions of a dataset

- Qualitative (categorical) data
- Quantitatitive data (both discrete and continuous)

For each, have some specific dataset in mind. Other vocabulary words in the end of chapter "Study Guide" that are covered in class will be fair game for this exam. I may ask in terms of true or false questions.

Describing data: Our primary task so far has been to describe data. We have several ways of doing so with graphs or diagrams and with numeric summaries.

For Qualitative data, we have used tables, barplots and piecharts to describe datasets

For Quantitative data we have used the following graphics

- frequency tables (just use tally marks)
- stem and leaf displays
- dot plots
- histograms
- boxplots (covered on wednesday)

Can you explain what each does? Which are effective with smaller datasets? Which are better with larger datasets? I call "classes" bins, what are they and how are they used? From a graphic, can you give a good guess what the mean and median are?

For numeric summaries of the data, you should be able to compute all of these quantities

- The mean ($(x_1 + x_2 + \cdots + x_n)/n$)
- The median (the middle data point)
- The range (Max - Min)
- The standard deviation, S. ($S^2 = 1/(n-1)\sum(x_i - \bar{x})^2$).
- Q_1 and Q_3 (covered on wednesday. Q_1 is the median of the data to the left of the median, Q_3 the median of the area to the right of the median.)

The following formulas should be known prior to the exam. As there are just a few, I expect you to know the,.

Some sample problems might involve

1. Explain the difference between a frequency and a relative frequency?
2. What is a sample and why is it different from a population?
3. For the dataset on number of hours of TV watched per day Make a barplot and piechart of the data. If a person watches TV for 90 minutes what category are they in? If they watch for 120 minutes?
4. A stem and leaf plot of monthly phone bills is

Hours	frequency
$[0, 1)$	25
$[1, 2)$	15
$[2, 3)$	20
$[3, 4)$	10
4 or more	5

Table 1: table of TV watched per day

```
The decimal point is 1 digit(s) to the right of the |

3 | 34479
4 | 44456788
5 | 6
6 | 04556
7 | 1
```

(a) What is the minimum amount spent, the maximum, the median, the mode?

(b) Make a histogram of the data. Use classes or bins of size 5. For example, $[30, 35)$, ...

5. The histogram in figure 1 is the weight distribution of 100 students.

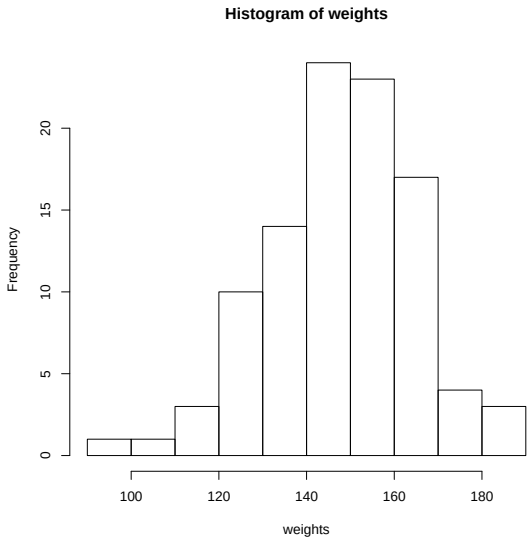

Figure 1: histogram of weights

(a) How many bin (classes) are used, what is the bin size?

(b) Which bin has the highest frequency? What is the frequency?

(c) Visually estimate the median of the distribution. What do you get?

(d) Visually estimate the mean of the distribution. Compare to the median. Is it different?

6. A dotplot of the number of cousins for 8 students is given in figure 2

(a) What is n?

(b) Find the mean number of cousins

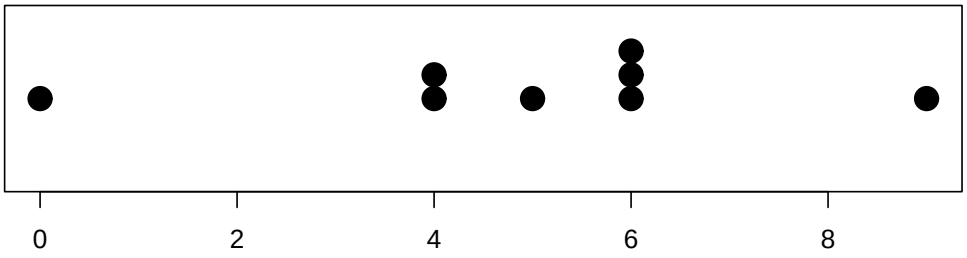

Figure 2: dotplot of number of cousins

(c) Find the median number of cousins

(d) What is the sample standard deviation?

(e) Find Q_1, Q_3

(f) What percentile rank is the value of 6?

(g) What z score is 6?

(h) What is $(1/n)\sum_{i=1}^{n} |x_i - 6|$?