Math 203
Topics for the Chapter 9 quiz:
Building data

To do:

Choose a sample from a population, using simple random sampling, independent sampling, and systematic sampling.

Understand the advantages and disadvantages of a range of sampling methods.

Find the mean, median, standard deviation, and quartiles of a data set; construct a box plot.

Describe the characteristics of a data set, and compare data sets, using the above measures.

Statistics is all about drawing conclusions about the opinions/behavior/structure of large *populations* based on information about a relatively small part of that population, called a *sample*. The population is the collection of individuals we would like to measure; the sample consists of the individuals that we actually measure. The level of confidence we have in how accurately the conclusions drawn from a sample reflect the opinions/behavior/structure of the whole population rests in part on how 'good' our sampling procedure is!

Samples can be gathered in a variety of ways:

Convenience sample: interview people as they walk by

Voluntary response sample (=call-in poll): the subject contacts the interviewer

Random sample: subjects are chosen 'randomly' from the target population

A sampling method is *biased* if it systematically favors one outcome over another (i.e., there are systematic reasons why the sample will ňot reflect the true characteristics of the population). The idea: if the sample accurately represents the population (i.e., is not biased), then the measurement of the sample has a better chance of being close to what we would have obtained if we had measured the entire population.

Biased: in a convenience poll, ask people at a mall 'Do you like malls?'.

Random sample: every sample of the same size is as likely to be chosen as any other. Basic technique: simple random sampling (SRS). Assign every subject in the population a different number with the same number of digits (pad with 0's), and then choose your sample by picking strings of numbers from a *table of random digits*.

Basic procedure: First, determine the size N of the population, and a size n for your sample, and choose a number of digits so that each individual can be assigned a distinct number with the same number of digits. Decide how numbers from the table will be chosen (e.g., "last 3 digits in each group of 5, reading down the page, starting with the second group on the 14th line"). Then assign each individual in the population a number of the required size. Then, following your procedure, extract the first n distinct numbers which have been assigned to individuals in the population; these individuals are your sample.

SRS is as random as your table is. It can be time-consuming; you need to be able to identify your entire poplulation before starting, and it requires you to assign a number to each. Its main advantage is that you get the exact sample size you want.

In *independent sampling* (think: sampling the people walking in a door), you decide as you ad individuals to the population whether they will be in your sample. For a $k\%$ sample, you flip a '$k\%$ coin'; using a random number table, you read the next 2-digit number from the table, and place the individual in the sample if the number read is between 00 and $k-1$ (or is one of k numbers you had previously agreed to use).

Advantages: you don't need to idenitfy the population ahead of time, and it saves the time of numberng the population. Disadvantage: you don't know ahead of time the size of your sample. It will be approximately $k\%$ of the eventual population, but a fair (50%) coin desn't come up heads exactly 50 out of every 100 flips!

In *systematic sampling*, we decide precisely what fraction of the population we want in the sample (k out of every n). We choose k distinct numbers from 1 to n, $a_1, \ldots, a_k$ and as we identify the individuals of the population, one at a time, we take for our sample the individuals in each group of n.

Advantages: same as independent sampling. Also, the sample size will be very nearly k out of n. Disadvantages: We don't know in advance the size of the sample. Also, if the line of individuals (think: parts off of an assembly line) has a charcteristic which follows a pattern that falls in sync with the sampling procedure (repeating every or nearly every (or a multiple or factor of every) n), then our sample is likely to be populated by only one or a few of the possible characteristics (biasing the sample).

If there are some characteristics of the individuals in the population (gender, education, age?) that you feel are important, and we want to make sure that your sample mimics the range of values that your population takes on those characteristics, we can use *stratified sampling*. The population is divided into non-overlapping *strata*, each stratum containing the individuals taking on a given range of values for the characteristics. Our sample is then chosen by SRS on each stratum, choosing the percentage of the target sample size equal to the perentage of the population that is in that stratum.

(One idea: each stratum is more homogeneous, so there may be less variability in our measurements within each stratum. Later we will see: this means that a smaller sample size could be used, without sacrificing how accurately the sample measurement can predict the behavior of the entire population.)

Another approach: *quota sampling*. As we identify individuals in the population, using independent or systematic sampling, we identify which stratum each lies in. We take the individual for the sample if the sampling procedure would pick them <u>and</u> we have not filled our quota for that stratum, i.e., we do not choose more from a stratum than it represents as a fraction of the population. (This tacitly assumes that we know the size of the population, so that we know what size our sample should have.)

In *cluster sampling* the goal is to lower the cost of sampling (e.g., traveling to an individual's home to conduct interviews) by making sure that our random sample falls in clumps. We divide the population into non-overlapping clusters, and then use a sampling method to randomly choose some of the clusters for further consideration. Those not in clusters chosen will not be chosen for the sample. Within each chosen cluster, we apply a further (possibly different!) sampling method to randomly choose individuals within the cluster for our sample.

The idea: chosen individuals within a cluster are 'close to' one another, lowering the cost of interviewing each of them. Unlike stratified samplng (which also divides up the population) individuals within a cluster need not share any common charateristics (other than location?).

Measures of 'central tendency' and variability.

Having collected the data, we can ask what 'shape' the data has. (The last chapter gave some approaches to this.) Where is the 'middle' of the data? What is the 'typical' value? How much does the data concentrate around the middle? How evenly distributed around the middle is it? There are two basic kinds of measures that prove useful in addressing this.

Median/Quartiles: list the numbers in increasing order. The *median* m is the middle number on this list (or the average of the two middle-most numbers, if the number of data points is even). The fist quartile q_1 is the median of the bottom half (i.e., one-fourth of the way along the list), and the third quartile q_3 is the median of the top half. The median describes the center; the quartiles describe the spread.

Mean/standard deviation: the *mean* μ is the 'average'; add all of the numbers up, and divide by how many numbers you have. This gives a (usually different) description of the center of the data.
The *standard deviation* σ describes the spread; it is the **square root** of the variance σ^2, which is calculated as (here $x_1, \ldots, x_n$ are our data points)

$$\sigma^2 = \frac{(\mu - x_1)^2 + (\mu - x_2)^2 + \cdots + (\mu - x_n)^2)}{(n-1)}$$

Each number in the sum measures how far from the middle each data point is, so we are taking a sort of 'average of the deviations'. The square root makes direct comparisons among data sets possible; if every measurement is multiplied by c, then both the mean and the standard deviation are multiplied by c. The larger σ is, the more spread out around the middle our data is (.i.e., the larger the 'average deviation' is).

The distinction between these two ways of measuring center and spread is basically that μ and σ are much more sensitive to outliers, while M and the quartiles are almost completely ignorant of them.

In the case of M and the quartiles, we can include the highest and lowest values to create a *5-number sumary* and *boxplot*; draw a horizontal line at M, a rectangle on top whose top and bottom are the quartiles, and finally a vertical line on top of this whose top and bottom are the highest and lowest values in the data set. This gives a very quick representation of the center, spread, and outliers in the data, especially good for making quick comparisions between different data sets.

We say that a data set is *skewed left* (i.e., low) if $\mu < M$; this means that most of the data lies to the right of the mean, so there are unusually small values that pull the mean down. Similarly, the data is *skewed right* (i.e., high) if $\mu > M$, since there are unusually large values that pull the mean to the right.

The *mode* is the most popular value; it is the value (or values) that ocur most often. In a histogram, they represent the highest peaks in the data; they are the most typical value in the data. They can easily have no relation to the mean and median, although certainly a popular value pulls the mean toward it (the same number is averaged into the mean many times).