Descriptive Statistics:

Now that we know what our sample looks like, we'd like to be able to describe it numerically. In other words, instead of having a lot of numbers (one for each record), we'd like to be able to describe the sample with just one or two numbers.

If we only use one number, what could we use?

Some examples:

- minimum (is this useful?)

- maximum

- third largest number?

- mode

- the number in the middle?

- mean

The first two candidates are interesting, but really don't represent the sample. The third candidate is just silly.

The fourth (the mode) we already talked about (see p. 18 **[15]** *{33}*). It's useful to describe the data, but it's not used much in more complicated analyses because it's kind of hard to work with.

We'll focus on the last two, beginning with the last:

I. Mean (see p. 32 & 33 **[26 & 27]** *{41 & 42}*)

- measures the center of our distribution. In the case of a sample, it's given by:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where n = sample size.

- this is nothing new - here is the example [2.15] *{2.2.3}*from the book (everyone should know how to calculate an average!):

weight gain in lambs over two weeks:

11, 13, 19, 2, 10, 1

thus we have 11 + 13 + 19 + 2 + 10 + 1 = 56

and we get 56/6 = 9.33 pounds.

- this is the SAMPLE mean. One can also talk about the population mean or the mean of a distribution. More on this later.

II. Median (p. 33 & 34 **[28 & 29]** *{40 & 41}*)

- the sample median is simply the value in the middle.

- if there is no "middle" number, then it's considered to be halfway between the two middle values. In other words:

  - if there are an odd number of observations, it's in the middle.

  - if there are an even number of observations, it's half way between the two middle values.

- Example (exrc. 2.14 **[2.16, p. 30]** *{2.3.3, p. 44}*):

  arranging the values from smallest to largest:

  5.9 5.9 6.3 6.9 7.0

  here the median is 6.3 nmoles/gm (the middle value)

- Example (exrc. 2.15 **[2.18, p. 30]** *{2.3.5, p. 44}*):

  again, arranging the values from smallest to largest:

  230 274 274 292 327 366

  to calculate the median, take the average of the two middle numbers: 274 + 292 = 566, and then 566/2 = 283.

so the median is 283 mg/dl

Finally, which is better? Mean or median? (See also p. 36 **[30]** *{43}*)

Depends (don't you love a vague answer like that?)

For most things (particularly in this class) the mean is probably a better indication of the "center". Why? Because it uses all of the data. The median uses only the middle or middle two numbers (though the other numbers do determine where the middle is). The mean is extensively used in statistics, particularly the kind we're going to learn.

So why bother with the median? It does better when the data are highly skewed, very spread out, or have lots of "outliers". A common example is in income. Listing the average income is very misleading. Why?

Consider Bill Gates. He pulls the average income WAY up. Also note that income usually doesn't drop below 0.

The median does much better here, since Bill Gates only moves it up half a notch, if at all.

(Lots of research going on in statistics. Some years back there was a talk in the statistics department about the median).

So now we have an idea of how to measure the center of our distribution. What about the spread? We also want to know:

- are all the observations sort of the same?

- or are they all very different from each other?

Example (draw two different normals on board)

Here we also have some candidates:

- range

- average absolute deviation

- variance

- standard deviation

Let's go through these:

I. Range (p. 48 **[p.40]** *{59}*):

- maximum value - minimum value = range.
(your book talks about interquartile ranges - ignore these references for now).

- sensitive to extremes (e.g. Bill Gates again).

II. So why not use something like "average deviation"?

- here's why, using the example from exrc. 2.15 **[2.18]** *{2.3.5}*which we talked about:

230 - 293.8333 = -63.8333
274 - 293.8333 = -19.8333
274 - 293.8333 = -19.8333
292 - 293.8333 = -1.8333
327 - 293.8333 = 33.1667
366 - 293.8333 = 72.1667

now we sum all the totals:

(-63.8333) + (-19.8333) + (-19.8333) +
(-1.8333) + (33.1667) + (72.1667) = 0 (oops!)

dividing 0 by 6 is pointless, so we can stop here.

The sum of the deviations from the mean is always 0.

III. So what can we do instead? Average absolute deviations (this one's not in the book):

- Take the absolute value of each of our numbers above.

- So we get (remember |-63.8333| = 63.8333):

63.8333 + 19.8333 + 19.8333 + 1.8333 +
33.1667 + 72.1667 = 210.6666

- And now we have 210.6666/6 = 35.1111.

- This is used, but as it turns out, is not terribly useful for us. The mathematics needed to use this for doing anything useful can be difficult (the folks using this use a computer to deal with the details).

- incidentally, there are actually several very similar measures, but we won't discuss them.

IV. Variance (& standard deviation) (p. 49 -52 **[p. 41 - 44]** *{60 - 63}*):

- The basic problem is that we need to make our "deviations" positive. So what else can we do? Square the deviations, which makes them positive, and then "take an average" (well, sort of).

- sample variance:

- take all the deviations and square them.

- sum these up (this, incidentally gives you the SUM OF SQUARES, an important quantity)

- divide by n-1. We get:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

- Here's an example, using the same set as above:

- Remember, we got -63.8333 by taking 230, one of our observations, and subtracting the average, 293.8333. (On occasion you may see the "^" symbol in the notes, which means "raised to the power, thus 2^2 would mean 2 squared, or 4). In any case, here we get:

$-63.8333^2 + 19.8333^2 + 19.8333^2 +$
$-1.8333^2 + 33.1667^2 + 72.1667^2$

= 11172.8333 = Sum of Squares = SS

And then we get the variance: 11172.8333/5 = 2234.5666

- The units on this are $(mg/dl)^2$.

- The variance is used extensively in statistics.

    - Often, statisticians don't even bother with standard deviations until they're ready to present results.

    - The problem with variance is that the units are not directly comparable to the original. Thus we use the "standard deviation", which is simply the square root of the variance.

- Here's an example of standard deviation, using exrc. 2.34 p. 58 **[2.46, p. 49]** *{2.6.7, p. 67}*:

    mean:

    6.8 + 5.3 + 6.0 + 5.9 +
    6.8 + 7.4 + 6.2 = 44.4

    and 44.4/7 = 6.343.

    variance:

    $(6.8 - 6.343)^2 = 0.20898$
    $(5.3 - 6.343)^2 = 1.08755$
    $(6.0 - 6.343)^2 = 0.11755$
    $(5.9 - 6.343)^2 = 0.19512$
    $(6.8 - 6.343)^2 = 0.20898$
    $(7.4 - 6.343)^2 = 1.11755$
    $(6.2 - 6.343)^2 = 0.02041$

    Sum of Squares = $\overline{2.9571}$

    so variance = 2.9571/6 = 0.49285

    (remember, divide by n-1; 7-1 = 6)

    standard deviation:

        This is the square root of 0.49285, which is equal to 0.70203.

Some concluding remarks about all this.

- Here is the formula for the standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

- the usual abbreviation we use for the SAMPLE standard deviation is s. The SAMPLE variance is simply s^2.

- Why on earth do we use n-1 instead of n in the denominator?

- an intuitive explanation (ex. 2.31, p. 52 **[p. 43 - 44]** *{62-63}*):

- take a sample of size 1.

- now, what is the variance?

- using the formula, one winds up with:

$$\frac{0}{0} = \textit{undefined}$$

- this makes sense, because a sample of size one can't tell us anything about the variation of a population. There ISN'T any variation in a sample of size one.

Note that it can be shown that if you use n instead of n-1 that your variance will be biased. Strangely enough, the standard deviation is always a bit biased regardless of whether or not you use n or n-1.

- is n ever appropriate? Yes, if you're really ONLY interested in the data you have, and NOT in making inferences about the population at large. This is not usually the case. We will pick up with this theme next time.