Descriptive Statistics:

One of the most important things we can do is to describe our data. Some of this can be done graphically (you should be familiar with histograms, boxplots, scatter plots and so on) or numerically. Here we'll mostly discuss numerical methods.

Describing the "center" of our data:

Here are some sample statistics that can be used to describe the "center" of our distribution (or data):

mean

weighted mean

median

trimmed mean

mode

I. Mean (p. 21)

- 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 an example from a different text (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. Weighted mean (not in text, but see p. 23)

We can use a weighted mean if we have our data grouped into various categories, or if we are trying to calculate the average of two averages (where the sample sizes are unequal).

For example, suppose we have the following data on butterfly wing lengths: (example 3.2, p. 23):

cm	frequency	frequency x cm
3.3	1	3.3
3.4	0	0
3.5	1	3.5
3.6	2	7.2
3.7	1	3.7
3.8	3	11.4
3.9	3	11.7
4.0	4	16.0
4.1	3	12.3
4.2	2	8.4
4.3	2	8.6
4.4	1	4.4
4.5	1	4.5

note that this is simply the data from example 3.1 p. 22 arranged into a frequency table.

Our total sample size here is 24 (add up the frequency column).

So what is our mean? Well, we have 1 measurement of size 3.3, 0 measurements of size 3.4, 1 measurement of size 3.5, 2 measurements of size 3.6, etc.

We could either do:

$$3.3 + 3.5 + 3.6 + 3.6 + ...+ 4.5 = 95$$

Or we could do this quicker by doing:

$$3.3\ (1) + 3.5\ (1) + 3.6\ (2) + ... + 4.5\ (1) = 95$$

In other words, multiply each of or data points by the number of data points at that measurement (the frequency x cm column above).

If we add the last column, we get 95, and 95/24 = 3.96 cm = $\bar{y}$

In other words, each measurement was multiplied by the "weight" (= frequency) of that measurement.

Now suppose we had the following data for the same exam in two different sections of the same class:

Section 1: 67 89 78 92 87 34 78 97

Section 2: 56 92 67 89 66 78

The average for section 1 is $\bar{y}_1 = 77.75$

The average for section 2 is $\bar{y}_2 = 74.67$

The goal is to combine the two sections and get an overall average. Would it be correct to do:

$$\frac{77.75 + 74.67}{2} = 76.21 \quad ??$$

No. Because section 1 has 8 students, and section 2 has 6 students. We need to "weigh" the averages from the two sections by their sample size. In other words, we do:

$$\frac{77.75(8) + 74.67(6)}{8+6} = 76.43$$

(note that this is the same as adding up both samples and dividing by 14)

Weighted means can obviously be quite useful.

III. Median (p. 24)

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.

Here's an example from a different text (numbers are already sorted) looking at 5 measurements of benzopyrene in the liver tissue of mice (in nmol/g).

5.9 5.9 6.3 6.9 7.0

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

And an example using an even number of measurements (this time cholesterol levels in 6 men). Again, already sorted.

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

Which is better? Mean or median?

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).

IV. Trimmed mean, or how to fix the mean if you have outliers (not in your text):

A trimmed mean is very similar to the regular mean, but it tries to make up the fact that the mean is very sensitive to outliers.

To do this, it excludes the 10% (or whatever % you choose) of values that are most extreme.

For example, if you had a sample size of 100, you'd remove the 5 smallest values and the 5 biggest values, and then just calculate the mean.

Obviously you're discarding data here, which is always something you should be careful with.

Often we say we want the middle 85% of the data (or whatever), instead of "we exclude the most extreme 15%").

This has the advantage of still being a "mean", but there are two problems:

1) A lot of statistical tests need to be tweaked a bit to work correctly

2) It is usually biased (that is, it doesn't estimate the population mean correctly)

(Believe it or not, "2" is not necessarily all that serious).

V. Mode (p. 27)

The mode is simply the "most frequently occurring measurement as you text puts it. It's not always a very good measurement of the middle, and can often be quite a bit different from the mean or median.

We won't be using it in here except descriptively (for some reason all introductory statistics books talk about the mode, and then never use it for anything!)

Describing the spread in our data:

Here are some measures of spread:

range

interquartile range

average absolute deviation

variance

standard deviation

I. Range (p. 33):

Simple given by:

maximum value - minimum value = range

The range is very sensitive to extremes (e.g. Bill Gates again) and isn't used much except descriptively (it *is* interesting, but not useful for statistics).

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

here's why, using the example from p. 34 (a hypothetical population of 7 insect body weights). Note that $\bar{y}$ = 1.8 g:

1.2 - 1.8 = -0.6
1.4 - 1.8 = -0.4
1.6 - 1.8 = -0.2
1.8 - 1.8 = 0.0
2.0 - 1.8 = 0.2
2.2 - 1.8 = 0.4
2.4 - 1.8 = 0.6

and you should be able to see quite quickly that the sum is 0.

dividing 0 by 7 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 (see p. 34 & 37):

Take the absolute value of each of our numbers above.

So we get:

06. + 0.4 + 0.2 + 0.0 + 0.2 + 0.4 + 0.6 = 2.4

And now we have 2.4/7 = 0.3429.

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. 34, 37, and 41):

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:

Basically, we can square each of the deviations given above:

$$-0.6^2 + (-0.4)^2 + (-0.2)^2 + 0.0^2 + 0.2^2 + 0.4^2 + 0.6^2$$

$$= 1.12 \text{ g}^2 = \text{Sum of Squares}$$

And then we get the variance: $1.12/7 = 0.1867 \text{ g}^2$

The variance is used extensively in statistics.

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

Standard deviation

The problem with variance is that the units are not directly comparable to the original.

So we use the "standard deviation", which is simply the square root of the variance.

So continuing the above example, we have:

$$\sqrt{0.1867} = 0.43\,g$$

V. 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:

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.

Note that the text gives you a computational formula for *s* and s^2. They're okay if you don't have a button on your calculator that will do *s* or s^2, but otherwise you should avoid them (see 2nd footnote on p. 39, and note that it also applies to calculators, unlike what the text implies).

Make sure you use the right button/menu on your calculator - don't use the population variance or standard deviation.

VI. Coefficient of variation:

This is a measure developed to deal with the fact that the units for variances (& standard deviations) can be very different.

Suppose you're trying to find out if mice or elephants are more variable.

Don't you expect that the units of measurement for elephants will be much higher?

It's also logical that the variance for, say, elephant length will be much bigger than the variance for mouse length.

To deal with this we simply scale the standard deviation by dividing by $\bar{y}$:

$$CV = \frac{s}{\bar{y}}$$

(your book uses *V*, and says *CV* is often used; it turns out almost everyone uses *CV*).

CV doesn't have any units! Which is exactly what we want if we want to compare mice with elephants.