STAT 100 Chance and Data Analysis Sept. 25, 2002(rev)

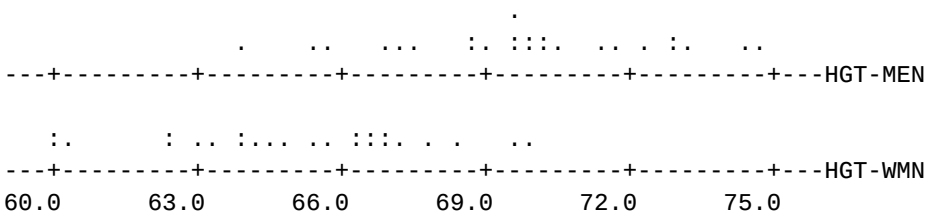
Notices:

1. Office Hours Change MF1430-1520 W1130-1220
2. Midterm Dates Oct 11 and Nov 8.

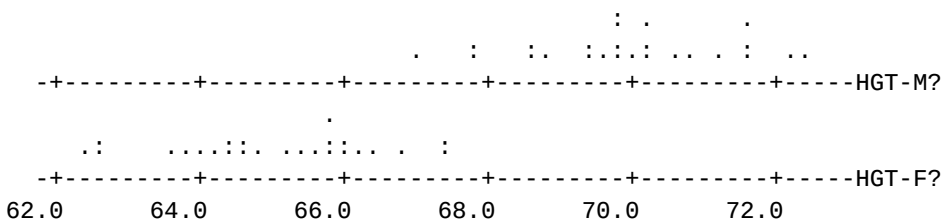
Today: More about **variability of averages**:

The usual way to summarize numerical data is to report an average: batting averages, average commuting times, average salaries, etc. When we make comparisons between groups, we often compare averages: when it is said that that men are taller than women, it is not claimed that all men are taller than all women, but rather a statement about averages. But a difference of averages does not tell the whole story - the variability in each group is obviously relevant to the degree of overlap.

Consider, for example, the height distributions of 25 men and 25 women.



In this data, the average height for man is about 70 inches while for women it is about 65 inches. The considerable overlap does not contradict the statement that men tend to be taller than women. But note that the SD of about 2.5 is important in summarizing the degree of overlap. If we shrink this data toward the means so that the SD is one-half of its former value, the graph would look like this:



Much less overlap and something more worth reporting. **SD is important in making comparisons of groups.**

Now we have been looking at the SD of raw measurements. A more subtle effect has to do with the **variability of averages** of raw measurements. Our risky company example demonstrates this, but needs a bit more work ….

Everyone take a coin and toss two times. Interpret the result as follows:

You are investing $1.00 in our "risky" company. The amount of this investment you get back at the end of the year is simulated by your coin tosses:

H,H means return of $0
H,T means return of $0.50
T,H means return of $1.00
T,T means return of $4.00

Do this five times and record the result. We are simulating a portfolio of five risky companies, to see if this is a good investment. That is, each student has five companies like the one we defined. Once you have done this, compute the mean of your five numbers. (and later the SD). I will summarize in class the resulting distribution of means. We can compare that to the class distribution of returns for the simulated companies.

Here is an example: Suppose I toss HH, HH, HT, TT, HT. My simulated company returns are:

```
 0.0    0.0    0.5    4.0    0.5

The mean is (0.0 + 0.0 + 0.5 + 4.0 + 0.5)/5 = 1.0.

The SD is the square root of

[(0.0-1.0)²+(0.0-1.0)²+(0.5-1.0)²+(4.0-1.0)²+(0.5-1.0)²]/5

= square root of [1+1+.25+9+.25]/5

= square root of 2.3 = 1.52  (See Row 5 below)
```

This should produce 5 numbers, something like this if there were 100 students doing the simulation …

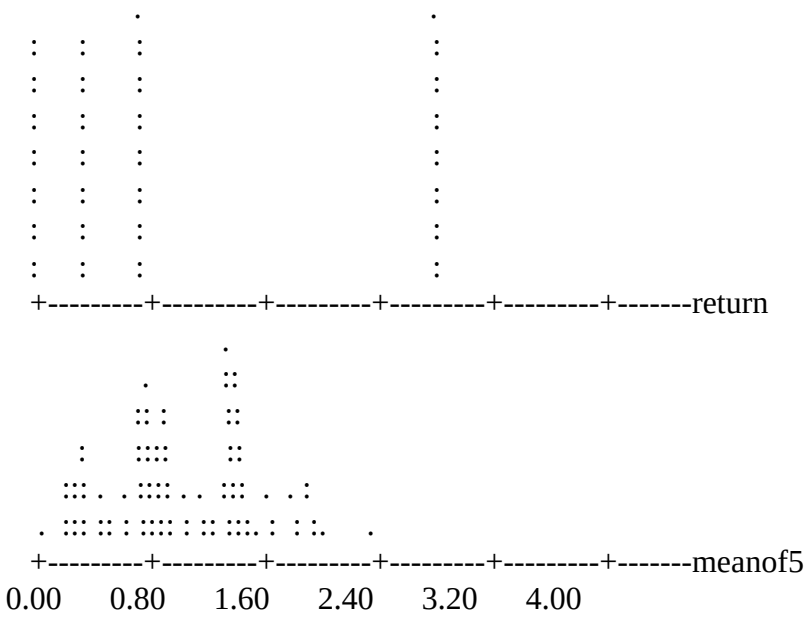
Row	sim1	sim2	sim3	sim4	sim5	mean	SD
1	0.5	0.0	0.5	1.0	0.5	0.5	0.32
2	1.0	4.0	4.0	1.0	1.0	2.2	1.47
3	1.0	1.0	1.0	4.0	4.0	2.2	1.47
4	1.0	0.5	0.0	0.0	0.5	0.4	0.37
5	0.0	0.0	0.5	4.0	0.5	1.0	1.52
6	4.0	4.0	0.0	0.5	0.5	1.8	1.81
7	4.0	0.0	4.0	4.0	1.0	2.6	1.74
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
93	1.0	0.0	0.0	0.5	0.5	0.4	0.37
94	1.0	0.5	0.0	4.0	4.0	1.9	1.74
95	1.0	4.0	0.5	0.5	0.0	1.2	1.44
96	1.0	0.5	0.0	1.0	0.5	0.6	0.37
97	0.0	0.5	4.0	0.5	0.0	1.0	1.52
98	0.5	0.5	0.0	0.0	1.0	0.4	0.37
99	4.0	0.0	0.0	0.0	1.0	1.0	1.55
100	1.0	4.0	0.0	1.0	0.0	1.2	1.47

We computed the SD of the numbers {0,0.5,1,4} to be 1.56, but we can see that for a random five simulations, the SD varies above and below this. However, if we were to pool all these SDs, we would find the 1.56 was an accurate estimate of the 500 simulated experiences (In fact it is 1.557 in this case, just by computing the SD of the 500 numbers). So, the SD of returns for this company is $1.56.

But what is the SD of the portfolio of five companies? First we can just do the calculation since we have everyone's simulated data. Based on my simulation, the mean and SD of the 100 averages (each average is based on one student's five company simulations) turns out to be $1.40 and $0.68 respectively. In other words, students have experienced returns on the portfolio that averaged $1.40 for the $1.00 investment and have a SD of $0.68. But this is not unexpected, given the theory …

The two distributions to compare are given here:

Each dot represents 9 points



Note that the portfolio mean returns ('meanof5' above) is more tightly clustered than the raw returns themselves. In fact the SD of the two simulated distributions are 1.56 and 0.68. The average portfolio return had a spread that was a fraction .68/1.56 of the spread of the raw company returns (using SD as our measure of spread). This factor is approximately $1/\sqrt{5}$.

To see it another way: we had computed the true mean to be \$1.38 and the true SD of the individual company returns to be 1.56. The average return in a portfolio of 5 individual companies should have an SD, according to our square-root law, of $1.56/\sqrt{5} = .70$. It turned out from the simulation to be .68. Pretty close.

The advantage of using the formula is that no simulation is needed. The mean 1.38 and the SD 1.56 were computed directly from the numbers {0,0.5,1,4}. In this case the simulation is used just to clarify the meaning of the formula.

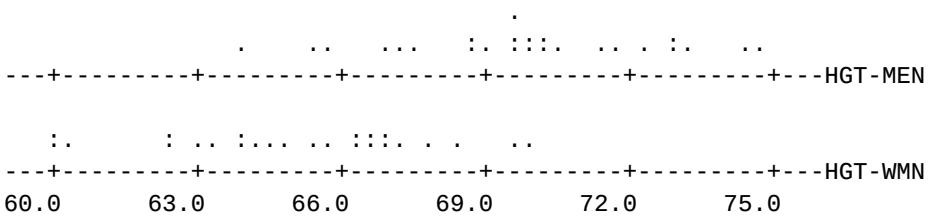
The main result of all this:

The observed mean return of five companies is more stable that the observed return of one company. The factor of SD reduction is the square root of 5. Our theory states that the mean return of five companies has an SD of \$0.70 which is $1/\sqrt{5}$ times the SD of the return of one company, \$1.56. (Our simulation produced empirical evidence for this - we got SDs that varied around \$1.56 for the individual companies in a portfolio and around \$0.68 for the average returns.)

In general, averages of n observations have an SD that is 1/√n of the SD of the observations themselves.

How is this result used?

Go back to the example:



IF we were trying to determine from this data whether it was true than mean were tallerthan women, on average, would this data convince us? For above data, means are about 65 and 70, SDs are about 2.7. How variable are the means if we were to take another random group of 25 men and 25 women? SD of means is 2.7/√25 = .54. Note that 65 and 70 are quite far apart compared with an SD of 0.54, so it is unlikely that the population averages could have been equal. Conclude that men are taller than women, on average!

(You will have more chances to understand this kind of application).

Next few lectures will be based on Tanur articles:
pp 178-187 Study Design. Test Scores
pp 3-14 Study Design. Salk Polio Vaccine Experiment.
pp 31-40 Study Design. Health Insurance.

Assignment #3 (Due Wed. Oct 2, 4:30 pm, in STAT 100 boxes outside K 9510).

1. Answer question 5 on page 169 of the Tanur book, concerning the food study.

2. With reference to the "risky" company described in class:

a) Use a coin-simulation to determine the returns of a portfolio of five "risky" companies.

b) Explain, using what you know about means and standard deviations, why an investment in a portfolio of five of these companies ($1 in each one) has less risk than an investment of $5 in just one of the companies.

3. Using whatever paper money of Canada you have available, examine the numeric portion of the serial numbers. From five different bills, compute the SD of the seven serial number digits. Also, compute the SD of the five averages of the serial number digits. Compare the original 5 SDs with the SD of the averages. How did your result compare with the theory? (i.e. say what the theory predicts, and then comment on what your calculations showed in comparison to this theory.)