Variability Exercises

Pages	Suggested Reading
73 – 81	Section 2.9

Pages	Problems
87 – 100	(***Section 2.13***) 1(a-d), 15(c) (#15 has a solution in the text, but this video might help, as well), 18(d,e,f), 22(a,b,c), 28

Addenda Videos
• A video to derive a manual formula for the standard deviation of grouped data sets! You'd use it, for example, if you ever had to use Excel to find stats of grouped data.
• More evidence that the sum of the deviations from the average is zero! (Also see extra credit below!)
• The Empirical Rule Illustrated! (Sorry about the volume when the music kicks in…)

Many students (and, news people, and scientists, and just about *anyone*, for that matter) have a hard time, when given a choice, deciding when to use the mean or the median. In class, we decided that, in the case of extremely skewed data, the median should be used as a measure of center, since it's less affected by skewed data. But how skewed is, well, ***too*** skewed?

I'd like to open a can of worms here and offer one measure of skew that's fairly easy to compute, and only slightly more controversial to interpret. It's called Pearson's second skew coefficient (PSSC):

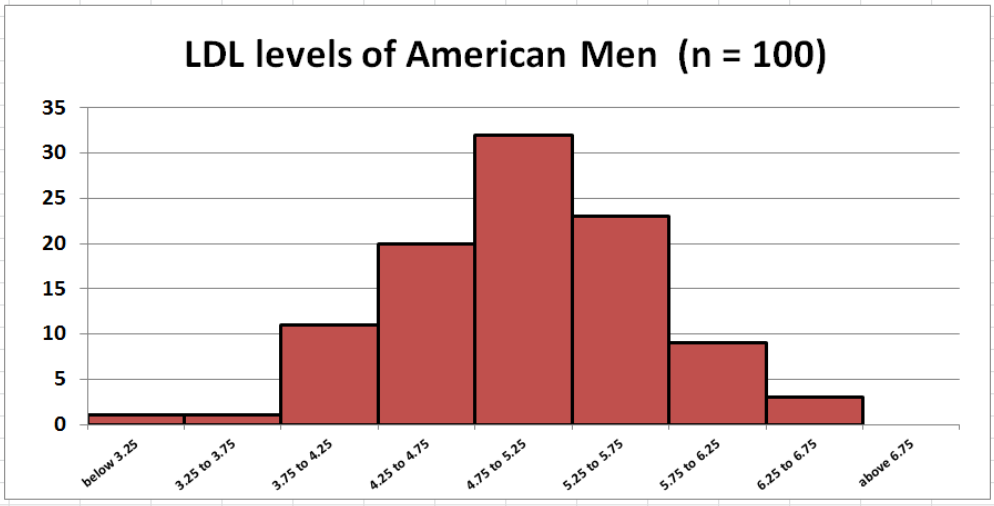
$$\text{PSSC} = \frac{\bar{x} - \text{sample median}}{s}$$

According to Hildebrand (1986), any PSSC absolute value greater than 0.2 shows "great skewness", and, as such, care should be taken when using the mean.

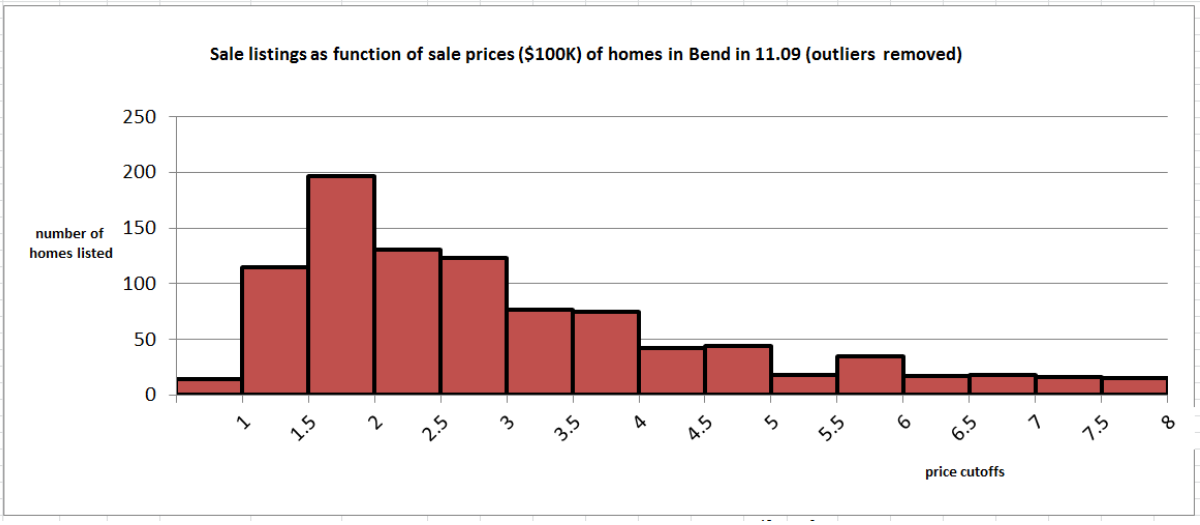
E1. Why "absolute value"? Could skew be negative? If so, what does that mean?

For the following datasets (in E2, E3, and E4), calculate the PSSC and tell whether or not the data show "great skew". You'll recognize these data sets from class (the lecture on central tendency). What measure of center would ***you*** use?

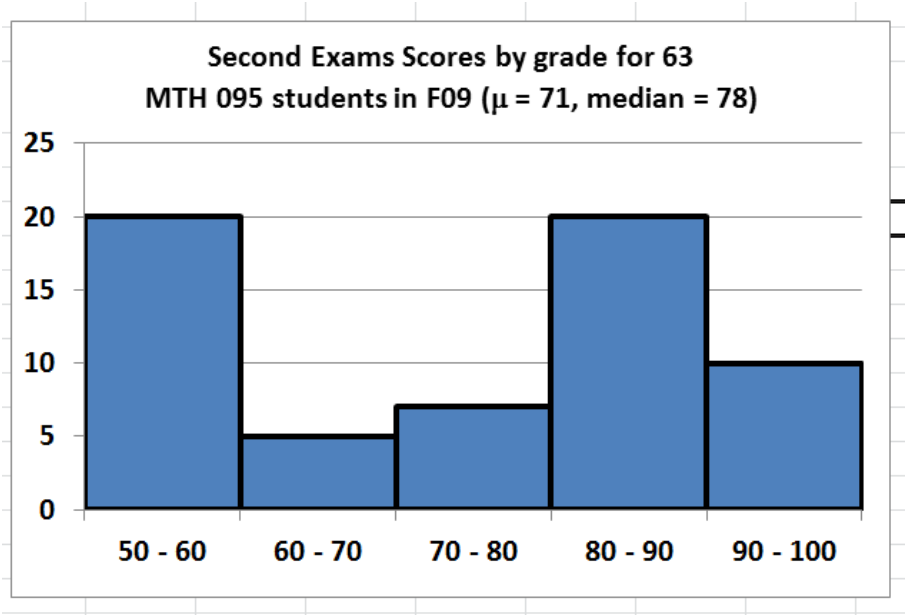
E2.



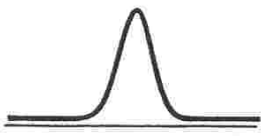
E3.

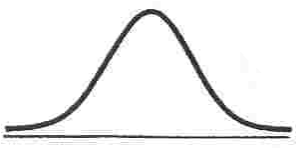


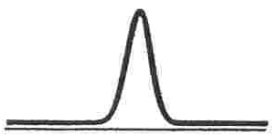
E4.



E5. Rank the following four sets of data is order of increasing variation around the mean:

a. 

b. 

c. 

d. 

A random sample of the ages of 10 consumers at a mall is taken. Their ages all fall between 17 and 81, inclusive.

E6. Suppose the sample consists of 10 consumers that are all 23 years old. Their average age, therefore, would be 23 years. What would the standard deviation of this sample be?

E7. The answer in the previous question represents the *smallest* possible standard deviation of a random sample of 10 consumers within the age range above (it's also highly unlikely to draw 10 people that all have the exact same age). Now, let's draw 10 other people from the same mall population. Assuming all of their ages still fall between 17 and 81, inclusive, what would the *largest* possible standard deviation be?

(this spreadsheet might help: http://coccweb.cocc.edu/srule/MTH243/homework/variation.xlsm)

(continued)

E8. What would the average of the ages in E7 be?

E9. Refer back to question E1 from the previous homework exercises ("Central Tendency"). How is the standard deviation affected?

E10. Refer back to question E2 from the previous homework exercise. How is the standard deviation affected?

E11 (extra points). In class, I claimed that $\Sigma(x - \bar{x}) = 0$, for any set of data. Prove this to me. Please submit your proof either in writing, or, if you like, via a demonstration at the office.

A very common variability convention (that we will investigate at depth in MTH 244) is called the ***margin of error*** about a statistic. A margin of error is (usually) expressed as 2 standard deviations (or, better, standard "errors"…we'll learn about that later). Here's an example of one in the media today (9.14.11)

Obama in Close Race Against Romney, Perry, Bachmann, Paul

Romney has slight edge over Obama, Bachmann slightly lags

by Frank Newport

PRINCETON, NJ -- President Barack Obama is closely matched against each of four possible Republican opponents when registered voters are asked whom they would support if the 2012 presidential election were held today. Mitt Romney leads Obama by two percentage points, 48% to 46%, Rick Perry and Obama are tied at 47%, and Obama edges out Ron Paul and Michele Bachmann by two and four points, respectively.

President Barack Obama vs. Potential Republican Candidates

Suppose the presidential election were held today. If Barack Obama were the Democratic Party's candidate and _____ were the Republican Party's candidate, who would you vote for -- [ROTATED: Barack Obama, the Democrat (or) _____, the Republican]?

	Registered voters
	%
Barack Obama	46
Mitt Romney	48
Other/Don't know	6
Barack Obama	47
Rick Perry	47
Other/Don't know	6
Barack Obama	47
Ron Paul	45
Other/Don't know	8
Barack Obama	48
Michele Bachmann	44
Other/Don't know	7

For results based on the total sample of national adults, one can say the maximum margin of sampling error is ±4 percentage points.

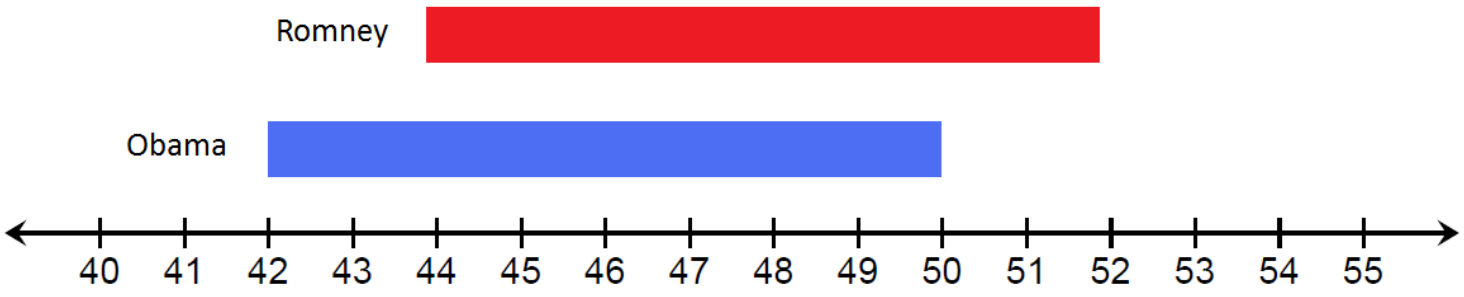
This was from Gallup (http://www.gallup.com/poll/149114/Obama-Close-Race-Against-Romney-Perry-Bachmann-Paul.aspx) . Look at the first comparison (Obama vs. Romney[a]):

	Registered voters
	%
Barack Obama	46
Mitt Romney	48
Other/Don't know	6

[a] I know this is dated. No matter…the underlying idea is timeless.

The headline of this poll makes it sound like Romney, at this point in time, is ahead of Obama in potential electability. That would only be true, however, if Gallup had asked ***every single registered voter in America***, and the percentages were 48 to 46 in favor of Romney (they would then be "parameters", remember?). In this poll, however, they randomly sampled Americans, and are using their response ("statistics") as a representation of all Americans.

Statistics (again, as you'll see more in 244) always have to carry a measure of variability; in this case, a margin of error. So, when Gallup claims that 46% of registered voters would vote for Obama, with a 4% margin of error, what they really mean to say is this: "We're sure that between 42% and 50% of voters will vote for Obama." (46% ±4%). As far as Romney goes, they should say, "We think that between 44% and 52% will vote Romney." When looked at this way, their subheadline, "Romney has slight edge over Obama" is meaningless…Obama might actually be ahead! Consider the graph below, which pictorially represents the two intervals above:



When you visualize this graph realize this: all you know about the true voter percentage is that it's ***somewhere*** within the bar. That means that Romney's could be as low as 44%, while Obama's could be as high as 50%. Thus, the subheadline is, in fact, misleading. What Romney and Obama are in at this point is called a "statistical dead heat"; that is, their statistics (46% and 48%) are within a margin of error of each other, and, therefore, are indistinguishable. In fact, their main headline ("close race") is far more truthful.

E12. Conduct a similar analysis for Obama vs. Bachmann, and comment on the subheadline, "Bachmann slightly lags".

Answers.

E1 . Yes, skew could be negative, if the sample mean is less than the sample median. This can happen, for example, with highly right skewed (that is, left – leaning) data.

E2. PSSC ≈ 0 (video to show you how I got this!). The data appears unimodal, so it's safe to use the mean ***or*** median in this case. Your choice.

E3. PSSC ≈ 0.18 (another video!). According to the 0.2 rule, we're not "greatly skewed" yet, but I think it's close enough to show concern. It's unimodal, so I'd go with the median. Also, remember that this Hildenbrand guy picked the 0.2 cutoff…it's a ***definition***, not a law.

E4. PSSC ≈ -0.05 (one more video!), which might imply that the data isn't skewed. However, the data is bimodal, which means that both the mean and median are misleading. As such, I would use the mode, and forget the mean and median (see, I ***told*** you it's a can of worms). Seriously, though…this example just reiterates how important it is to look at data graphically before blindly crunching numbers.

E5. d, c, a, b

E6. 0…there is no variation.

E7. It's about 33.7…do you see why? Use the spreadsheet to help you!

E8. 49.

E9. It's doubled. Here's proof, if you want to see why!

E10. It's not (why?). Here are two videos (one using Excel, and one using algebra) to help you understand why!

E11. Due before the end of the term, if you choose.

E12. I don't buy it. You?

Variability Quizzes

Quiz 1.

Please visit this website:

http://www.gallup.com/poll/162029/pelosi-best-known-least-liked-congressional-leaders.aspx

1. **(2 points)** What is the margin of error on the poll[b]?

   When Gallup makes the claim that *"[t]hirty-one percent of Americans view Pelosi favorably and 48% unfavorably"*, they're referring to the chart near the top of the site:

Favorable Ratings of U.S. House and Senate Leaders

Finally, we'd like to get your overall opinion of some people in the news. As I read each name, please say if you have a favorable or unfavorable opinion of these people -- or if you have never heard of them. How about [Speaker of the House, John Boehner; Senate Democratic leader, Harry Reid, House Democratic leader, Nancy Pelosi; Senate Republican leader, Mitch McConnell]?

	Nancy Pelosi	Harry Reid	John Boehner	Mitch McConnell
	%	%	%	%
Favorable	31	27	31	26
Unfavorable	48	38	41	34
Never heard of	11	21	14	22
No opinion	10	15	14	17
Net favorable	-17	-11	-10	-8

April 11-14, 2013

2. **(1 point each)** However, their claim ignores the margin of error (MOE) from question #1. Apply the MOE to each of the sample statistics to complete the following (I've done Nancy Pelosi as an example):

Nancy Pelosi's "Unfavorable" rating could be as low as 44% or as high as 52%.

Harry Reid's "Unfavorable" rating could be as low as _____ or as high as _____.

John Boehner's "Unfavorable" rating could be as low as _____ or as high as _____.

Mitch McConnell's "Unfavorable" rating could be as low as _____ or as high as _____.

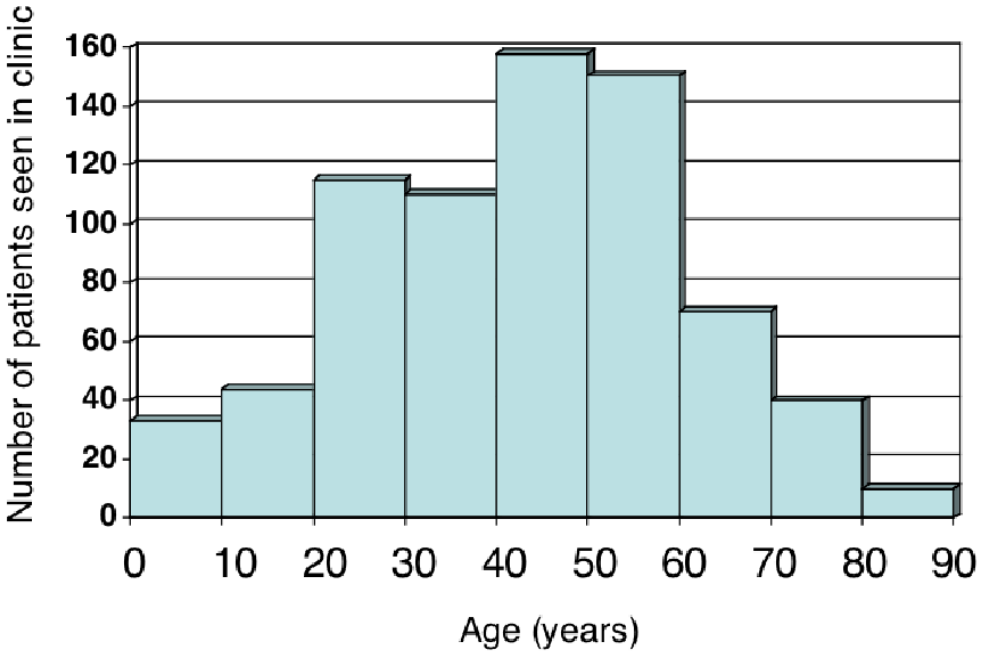
3. (**2 points**) Based on your work in 2, can you support the claim in the headline…that Pelosi is the "*Least Liked of Congressional Leaders*"? Why or why not[c]?

---

[b] You have to scroll down near the bottom. It's in, like, 4 point font. Of course.

[c] Let's define "least liked" to mean that they have the absolute highest unfavorable rating – higher than any of the other 3, even when considering the margins of error.

Quiz 2.



The roughly bell – shaped data above was collected from a free health clinic trying to track the demographics of its clients (in this case, age). On this quiz, you'll use some technology to glean some data from this graph.

a. (**1 point**) Find the average age of someone using this clinic. I know that your answers might be varied a little, since you're eyeballing frequencies. That's fine). Round to the nearest year. I'd use the Excel Calculator ("Grouped Data" sheet!).

b. (**2 points**) What's the sample standard deviation of the ages? Again, nearest year, please.

c. (**w**) (**2 points**) Calculate the PSSC for this data (if you forget what that is, look at E1 through E4 above). To the nearest tenths' place, please.

d. (**3 points**) The PSSC was almost 0, so I feel good about this data' symmetry. The "empirical rule" tells us that roughly 95% of data (in bell – shaped distributions) should fall within 2 standard deviations of the mean. What are those two ages? In other words, what is $\bar{x} + 2s$ and what is $\bar{x} - 2s$? Round these to the nearest ***tens***' place, please.

e. (**w**) (**2 points**) Approximately what percentage of your data ***actually*** lie between your $\bar{x} + 2s$ and $\bar{x} - 2s$? Not **much** work to show here – just the division.

Quiz 3.

The "empirical" (i.e., 2/3, 95%, ~100%) rule is a great way to analyze data and its spread…assuming (with a capital "A") you have data that is roughly bell – shaped. When you don't, and you can't use the percentages listed above, what can you do?

Well, a fine young Russian lad, way back in the 1800's, was bothered by this, as well. His name was Pafnuty Lvovich Chebyshev[d]. He discovered and proved a result, known mathematically conversationally as "Chebyshev's Theorem", that gives the least percentage of data points that must occur within a certain number of standard deviations from average.

(**2 points**) Google "Chebyshev's Theorem". There's an algebraic relationship (usually, the variable is called "***k***", where ***k*** is the number of standard deviations from the mean) that gives a minimum percentage of data within ***k*** standard deviations of the average. What is this relationship?

(**2 points each**) Complete each grayed box. Round each to the nearest whole percent.

Number of Standard Deviations from Mean	**Percentage** of data within that many Standard Deviations…	
	…if data is (arguably) bell – shaped:	…if data is NOT bell – shaped:
1	About 2/3	At *least* none of it[e].
2	About 95%	At *least*
3	Almost all of it	At *least*
2/3	About [f]	(n/a)
0.5	About	(n/a)

[d] A name worth 58 points in Scrabble.

[e] Pretty useful, huh? ☺ Technically not meaningful, since the number of standard deviations must be ***larger*** than 1.

[f] Use this!: http://onlinestatbook.com/2/calculators/normal.html (you'll have to write the "2/3" as a decimal – use as much precision as you can)

Quiz 4.

You might remember from class that the formula for the sample standard deviation is given by the formula at right…

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Now, in a population, the standard deviation ("little sigma", or ***σ)*** is given by…

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Do some Googling, and give me a convincing argument as to why we divide by one less than the sample size (***n***-1) for s, but the entire population size (***N***…assuming it's known) for little sigma. List sources, please!