Select Reading Question Responses (9/1/2021)

> Could you clarify how to find the Q1, Q3 and the IQR? I was confused on how to solve 2.11 on the homework fully.

> I really struggled with question 2.11. I did not know the equation/ what numbers to use to get the Q1, Q3, or IQR.

Q1 is the point which is above 25% of the data. The idea behind calculating Q1 in 2.11 is the start from the left, and start adding up the heights of the bars. When you hit roughly 0.25, you've found the first quartile. The heights of the first three bars starting from the left of the histogram are roughly: 0.025, 0.05, 0.2. The sum of those is 0.275, so we're past the Q1. This means that Q1 should be somewhere in that first tall bar, ie, somewhere between 15 and 20. Let's say Q1 is approximately 17.5.

To find the median, we keep going until we hit 0.5. The height of the next (short) bar is 0.05, and then we have 0.2. The sum of all the heights up till the second tall bar is 0.525, so we've overshot the median again and the median must be somewhere in that second tall bar. In other words, it's between 25 and 30. Let's say it's at 27.5.

To find Q3, we do the same thing, adding up till we hit 0.75. This happens somewhere between 40 and 45, so maybe we say Q3 is about 42.5.

Finally, IQR is Q3—Q2 $\approx 42.5 - 17.5 = 25$.

> The question I had today is for 2.11(b). Why would the mean be larger than the median in this case?

The data depicted has a long tail to the right. The mean is more heavily impacted by outliers than the median, so that long tail "pulls" the mean towards it more than it pulls the median. In other words, the mean will be higher than the median.

> In the reading there was discussion of modes as being the "peaks" of a distribution. Is there a specific term for the lower spikes or points in a distribution?

You could use the word "valley." But usually people just talk about "peaks" of distributions.

> In an instance where a researcher wanted to look at three categorical variables, do people ever use three-dimensional mosaic plots, or does that overcomplicate things?

You *could* make three-dimensional mosiac plots, theoretically, but I personally feel like it's not a great idea. The problem is that translating between a three-dimensional object and a two-dimensional representation can be challenging. Even if you produce an excellent two-dimensional representation, it might be hard for many viewers to parse the three-dimensional object from its two-dimensional representation. When possible, it's probably best to stick to two-dimensional representations to keep things as easy for viewers to parse as possible.

There are versions of two-dimensional mosaic plots that work for three categorical variables (see here, for instance). But they're definitely more confusing than two-variable mosaic plots, so you might ask yourself if it might be better to just have a few mosaic plots for visualizing each pair of variables instead of having just a single plot that tries to visualize all three.

How do you know what your null hypothesis should be?

We'll discuss this later as well, but you want your null hypothesis to be something *precise* and *skeptical*. "Precise" means that assuming your null hypothesis should result in a complete description of the sampling distribution; in less theoretical and more heuristic terms, this often means that your null hypothesis will say that some population parameter is *equal* to something on the nose (as opposed to saying it's *greater than* something, or something like that). "Skeptical" means that your null hypothesis should say something like "it makes no difference" or "both are the same" or something along those lines. These are both heuristic principles and you have to figure out how to apply them in practice.

On page 47, it says "Like the mean, the population values for variance and standard deviation have special symbols: σ^2 for the variance and σ for the standard deviation." What exactly is a population value and why does it warrant its own special symbol?

If I have a list of numbers, I can compute a mean and a standard deviation of that list of numbers.

Suppose I'm interested in the weights of adult male housecats in the US. If I was able to collect a list of the weights of all adult male housecats in the US, I could compute the mean and standard deviation of those numbers. Those are "population values" (also called "parameters"). That mean would be denoted μ and that standard deviation would be denoted σ.

Of course, it would be very difficult in practice to acquire a list of all adult male housecats in the US. In practice, the best you can do is take a sample and compute the mean and standard deviation of that sample. Depending on the size and randomness of your sample, this may (or may not) be a good stand-in for the population values you're ultimately interested in. The mean calculated from a sample (sometimes called "statistics") is often denoted $\bar{x}$, and the standard deviation is s.

The two sets of quantities have different symbols because they could have different values. One might hope that the sample values are the same as the population ones, but this may not be the case. So it's convenient to have a different symbol so you can clearly distinguish the sample values, which you have access to, from the population values, which you're trying to estimate.

> I am confused about observational studies and the conclusions that can be drawn from them. It seems that a lot of studies have to be completed observationally because it is unethical to do them experimentally, but yesterday's reading essentially said observational studies can't prove correlation because of too many confounding variables and biases. Is this true all the time or can extensive data analysis help prove correlation? If not, is apparent correlation as long as extensive analysis has been done good enough to justify what ever outcomes come from the findings?

Correction: Yesterday's reading said that observational studies cannot provide evidence for *causation*.

Observational studies can in fact provide evidence for correlation. But this is still useful. Seeing a correlation in an observational study might lead you to conduct a (more expensive) experiment that can provide evidence of causation. And even if you can't conduct an experiment for ethical reasons, you could at least go in and conduct futher studies to see if you can tease out the effects of some confounding variables.

One of the key ideas you should take away from statistics is that *you never know anything for certain*. Every study gives you some data, and that data might point one way or another, but nothing is ever established in stone, even if it's coming from experimental data. Observational data might very well be weaker than experimental data, but it's still better than nothing if there isn't a practical or ethical alternative.

> In question 2.23, I am confused why it would be dependent, I would have answered it independent because both political ideologies seem to support and not support DREAM, wouldn't that mean it is independent of each other and the political ideologies do not affect it? or am I confusing the concept of independent and dependent?

> How do you tell if a relationship between variables is independent or dependent when looking at a mosaic plot?

"Independent" in this context would mean that all ideologies support DREAM *in equal percentages*. In other words, "independence" would mean that the percentage of conservatives who support DREAM is equal to the percentage of moderates who support DREAM is equal to the percentage of liberals who support DREAM. Similarly with percentages of "not support" and percentages of "not sure." This is not the case (as we can tell from the mosaic plot, since the three columns of the mosaic plot don't have boxes of the same height) so the variables are *dependent*.

We can generalize this as follows. If the "gaps between boxes" of the two-variable mosaic plot all line up, your two categorical variables are independent. In other words, if you can start at one end of your mosaic plot, and travel all the way through to the other end by going through the "white space" between boxes without ever making a turn, the variables are independent. On the other hand, if the gaps don't line up (ie, if you have to make a turn somewhere), the variables are dependent.

> How do you determine when a difference is large enough to reject the independence model/ accept the alternative model?

This is really very subjective! There's a "standard answer" (reject H_0 when the "p-value" is less than 0.05, as we'll talk about), but people sometimes follow this "standard answer" without thinking about it and end up misunderstanding their results. There are two articles I'll ask you to read towards the end of this class about this issue. You might decide to read them sooner to get a better sense of this issue.

> Can transformations of a data set distort the data being displayed in a harmful manner that prevents us from understanding something important about the data? For example, the textbook says that transforming data can 'reduce skew' for a data set. Isn't it possible that reducing the data set's skew harmfully alters our ability to understand the data set?

It can be harmful, but really only if you forget that you've done the transformation! On the other hand, if you do a transformation and see a pattern you didn't see before, you can use that to write down a model for the original "untransformed" data.

The answer to this question and the previous one are sort of related. The thing that I think is important to keep in mind is that statistics provides you with some very powerful tools for understanding the world around you, but *you should not apply these tools blindly*. Doing so will inevitably lead you astray. If, on the other hand, you think clearly about what you're doing and why it makes sense, you'll be okay.