PLSC 502: "Statistical Methods for Political Research"

Measures of Variation
September 15, 2016

Introduction

The subject du jour is measures of *dispersion*; that is, indicators of how "spread out" one's data are across the range of values each variable can take on. Of course, we'll also discuss special cases, examples, software, etc., as well as talking about some other characteristics of variables that can be useful to explore.

Ranges and Percentiles

Range

The *range* is the most basic measure of dispersion; it is simply the difference between the highest and lowest values of a (interval- or ratio-level) variable:

$$\text{Range}(X) = \max(X) - \min(X) \tag{1}$$

So, for example, the range of our NFL `points` variable is $(39 - 0) = 39$. As a measure of dispersion, it has a number of things to recommend it:

- The range tells you what you need/want to know – how much variation is there in your variable – and does so in units that are "native" to the variable itself. So, for example,
    - Knowing the range of ages (in years) of survey respondents is 22 tells you that there are only 22 years of difference between the oldest and youngest respondents (and so there's probably something wrong with your survey...).
    - Knowing that the range of salaries in one political science department is \$150,000, and in another is \$75,000, tells you that there's a lot more variability in the salaries of profs in the former than in the latter.
    - It is common practice to present the range of a variable (in the form of the minimum and maximum empirical values) in tables of summary statistics; more on this below.
- The range also *scales* with the variable. That is, if you rescale the variable, the range "adjusts" as well. So, if we rescale the salary variable above to be measured in thousands of dollars (rather than in single dollars), the range would decrease by a factor of 1000 (here, from 150,000 to 150, and so forth).

There's not (too) much more to say about the range, so I won't.

Percentiles, IQR, etc.

A related way of getting a handle on the variation in a variate is through looking at *percentiles* of a variable. The kth percentile is nothing more than the value of the variable below which k percent of the observations fall. Thus,

- if we have data on $N = 100$ observations, the 50th percentile is the value of the 50th-largest-valued observation. Similarly,
- if our data have (say) $N = 6172$, then the 50th percentile is the value of the $6172 \times 0.5 = 3086$th-largest-valued observation in the data, and
- the 79th percentile of that variable is the value of the $6172 \times 0.79 = 4876$th-largest-valued observation.
- Etc.

Note that this implies a few things:

1. The 50th percentile is the same thing as the median ($\check{X}$),
2. The 0th percentile is the same thing as the minimum value of X (because it's the value below which no data fall), and
3. the 100th percentile is (equivalently) the same thing as the maximum value of X.

As a practical matter, we typically care about "round-numbered" percentiles. The two most commonly-used ones are *quartiles* and *deciles*.

Quartiles are the 25th, 50th, and 75th percentiles of a variable; that is, the values of that variable below which 25, 50, and 75 percent of the data fall. So, for example, in our NFL week-one data, the lower quartile was 16 (meaning that 25 percent of teams scored less than or equal to 16 points), while the upper quartile is 28.2 (meaning that only 25 percent of teams scored 28 points or more). If we combine this with the median (i.e., the 50th percentile), which we know to be 23, we can get a pretty good idea of what the points scored variable "looked like" that week.

A related measure is known as the "interquartile range" (often abbreviated "IQR," and sometimes known in Cleveland-speak as the "midspread"):

$$\text{IQR}(X) = 75\text{th percentile}(X) - 25\text{th percentile}(X) \tag{2}$$

The IQR for our NFL data is $(28.2 - 16) = 12.2$, which means that the "middle" 50 percent of teams scored points that fell within about a 12-point range. That range obviously includes the median (23).

The IQR can be thought of as analogous to the range,[1] but has one characteristic that researchers sometimes prefer over the range itself: it is *robust* to outlying data points. For example, if (in the first-week games) Detroit had scored 89 points instead of 39 – but all other teams had scored exactly as many as they actually did – the range would then be $(89 - 0) = 89$, rather than 39, but the IQR would be unaffected. This is usually believed to be a good thing.

Deciles are essentially percentiles by "tens" – the tenth, twentieth, thirtieth, etc. percentiles of the data. They are calculated the same way as any percentile, and have more or less the same characteristics. They can be useful, however, when the data are *skewed* – that is, when there are small numbers of relatively high or low values in the data. Deciles provide a finer-grained picture of the variation in X than quartiles; moreover, they are often used to analyze data where there are large disparities between large and small values.

Consider household income in the U.S., for example. A commonly-used measure of income inequality is the "90-10 ratio": the ratio of the 90th percentile of income to the 10th percentile of income. So, for example, in 1980 the U.S. 90-10 ratio was a bit over 4; by 2003, it had gone up to 6. The value of this measure is that it captures income variation at the "top" and the "bottom" in a compromising fashion: while the top and bottom 10 percent are certainly substantially different, the ratio is still not unduly affected by extreme outliers (here, the Bill Gates of the world).

Deviations

Percentiles and ranges are fine, but as measures of variation, they leave something to be desired: they do not make use of the information in *all* of the data. Approaches that do so are typically based on the idea of a *deviation*.

A deviation is nothing more than the extent to which an observation's value on X differs from some "benchmark" value. In considering those benchmarks, two obvious candidates arise: the median, and the mean. In either case, the *deviation* is just the (signed) difference between an observation's value and that benchmark. So, *deviations from the mean* are $(X_i - \bar{X})$, and *deviations from the median* are $(X_i - \check{X})$.

Mean Squared Deviation, Variance, and Standard Deviation

Individual deviations are (as a rule) not all that interesting; what we're more interested in is something like, "How large is a *typical* deviation from the (mean, median)?" Of course, the whole idea of "typical" was one that we discussed last time when we covered measures

[1] In fact, we see now that the range is nothing more than the 100th percentile of X minus its 0th percentile.

of central tendency. So the starting point for such a discussion might be something like an "average" deviation from (say) the mean value of X:

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}).$$

This seems plausible enough, but watch what happens when we "unpack" this number:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}) &= \frac{1}{N}\left[\left(\sum_{i=1}^{N} X_i\right) - N\bar{X}\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N} X_i - N\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)\right] \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} X_i\right) = \frac{1}{N}(0) \\
&= 0
\end{aligned}$$

This is because (as we discussed last time) the mean is the value of X that makes the sum of deviations from it equal to zero (a corollary of the fact that it minimizes the sums of squared deviations – think calculus). Thus, the "average" deviation from the mean will always be equal to zero.

Intuitively, the reason for this is that the positive and negative deviations "cancel each other out." The consequence is that, irrespective of how widely varying the data are around $\bar{X}$, the mean deviation from the mean will be zero. To eliminate this property, we have to consider something other than the simple deviation. For reasons we'll get to a bit later (and that we touched on last time), one possibility to consider is the *squared deviation* from the mean, $(X_i - \bar{X})^2$. If we consider the "average" of this value, we get the *mean squared deviation* (MSD):

$$\text{MSD} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 \tag{3}$$

The MSD is intuitive enough, in that it is the average squared deviation from the mean; it also deals with the "canceling out" effect described earlier. But, think for a minute about the idea of variability. Imagine a dataset with a single observation:

team	points
Bears	14

The mean is (obviously) 14; but what's the MSD? Of course, it's zero (because the mean and X_i are identical). In fact, from one observation, we can know something about how many points were scored *on average* in the NFL that week (because $\bar{X} = 14$), but we can't know anything about the distribution ("spread") of those points. Were all the games 14-14 ties? Were they mostly 28-0 blowouts? There's no way to know.

Now, add an observation:

team	points
Bears	14
Giants	20

The (new) mean is now 17, and the MSD is $(14-17)^2 + (20-17)^2 = 9 + 9 = 18$. At this point, we can begin to learn something not just about the mean of the data, but also about its variability. Informally, this suggests a principle that is wise always to remember:

You cannot learn about more characteristics of data than you have observations.

That is, if you have one observation, you can learn about the mean, but not the variability. With two, you can begin to learn about the mean and the variation around that mean, but not the "skewness" (see below). And so forth. The formal name for this intuitive idea is "degrees of freedom."

The relevance of degrees of freedom is that, in the simple little example above, we actually only have *one* effective observation that is telling us about the variation in X, not two. Accordingly, we should consider revising the denominator of our estimate of the MSD downwards by one; doing so gives the formula for *variance*:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2. \tag{4}$$

The variance is the most widely-used measure of variability out there; we'll see it a lot this term and next. Note that, as $N \to \infty$, $\sigma^2 \to$ MSD, but that the two can be quite different in small samples.

Another key thing about the variance (and the MAD, for that matter) is that it is expressed in terms of "squared units" rather than the units if X. A straightforward way to put σ^2 back on the same scale as X is to take its square root:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}. \tag{5}$$

The term σ is, of course, known as the "standard deviation;" it is the closest analogue to an "average deviation from the mean" that we have. It is also the standard measure of empirical variability used with interval- and ratio-level data. In fact, it's generally good practice to report σ every time you report $\bar{X}$.

Notice that, as with the mean and the other statistics we've talked about, σ is expressed in the units of the original variable X. In our 32-team NFL data, for example, the variance σ^2 is 87.4, while the standard deviation is (unsurprisingly) about 9.35. That means that an "average" (or, more accurately, "typical") team's score was about 9-10 points away from the empirical mean of (about) 22.4.

A Variant: Geometric σ

Just like with the (arithmetic) mean, there's a variant of σ that is better used when the geometric mean is the more appropriate statistic. The *geometric standard deviation* is defined as:

$$\sigma_G = \exp\left[\sqrt{\frac{\sum_{i=1}^{N}(\ln X_i - \ln \bar{X}_G)^2}{N}}\right] \tag{6}$$

σ_G is the geometric analogue to σ, and is best applied in every circumstance where the geometric mean is also more appropriately used.

Absolute Deviations and MAD

Variances and standard deviations – because they are based on means – have many of the same problems that means have. In particular, they can be drastically affected by outlier values of X, with the result that one or two small or large values can artificially distort the picture presented by σ. Again, suppose that Detroit put up 89 points instead of 39, but ll the other teams' scores were identical. The result is that the mean of `points` rises (from 22.4 to about 24), but the standard deviation rises even more dramatically (from 9.4 to about 14.8). That is, changing the points scored by a *single team* (out of 32) caused σ to increase by around 50 percent.

An alternative to the variance and standard deviation is to consider not squared deviations, but rather *absolute values* of deviations from some benchmark. Because the concern is over resistance to outliers, we typically substitute the median for the mean, in two ways:

1. We consider deviations around the median, rather than the mean, and
2. When considering a "typical" value for the deviations, we use the median value rather than the mean.

The result is the "median absolute deviation" (MAD), defined as:

$$\text{MAD} = \text{median}|X_i - \check{X}|. \tag{7}$$

Note that some textbooks refer to the "mean absolute deviation" when referring to "MAD;" the latter is:

$$\text{Mean Absolute Deviation} = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|.$$

The latter is merely a substitution of the absolute value into the standard formula for a mean deviation, but one that lacks the robustness to outliers that MAD does.

MAD is often presented where there are highly influential outlier observations – that is, in the same situations where the median is used as a measure of central tendency.

A Moment for Moments...

As an aside, means and variances are examples of "moments." Moments are typically used to characterize random variables (rather than empirical distributions of variates), but they give rise to some useful additional statistics

- Think of the moment as a description of a distribution...
- We can take a moment "around" some number; usually the mean (or zero).
- In general, the kth moment around a variable's mean is $M_k = \text{E}[(X-\mu)^k]$. This means that...
    - The mean is the first moment: $\mu = \text{E}(X)$.
    - The variance is the second: $\sigma^2 = \text{E}[(X-\mu)^2]$.
    - *Skewness* is the third moment: $M_3 = \text{E}[(X-\mu)^3]$.
        - Measure of *symmetry*... (Why? – cubic transformation preserves the directionality of the distribution).
        - Often use an alternative measure:

$$\begin{aligned}\mu_3 &= \frac{M_3^2}{\sigma^3} \\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right]^{3/2}}\end{aligned} \tag{8}$$

        - Skewness = 0 is symmetrical .

- Skewness > 0 is "positive" (tail to the right).
- Skewness < 0 is "negative" (tail to the left).

◦ *Kurtosis* is measured by the fourth moment: $M_4 = \mathrm{E}[(X - \mu)^4]$.

- Measure of how "peaked" the distribution is... (Why?)
- Also use an alternative here...:

$$\mu_4 = \frac{M_4}{\sigma^4} - 3 \tag{9}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right]^2} - 3 \tag{10}$$

◦ μ_4 is always ¿ 0; and in fact, it has a lower bound equal to the square of the skewness plus one:

$$\frac{\mu_4}{\sigma^4} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1$$

Kurtosis has no upper bound, and can (in theory, anyway) be infinite.

- Fat-tailed/"Peaked" = *leptokurtic*: μ_4 is large / $\mu_4 - 3$ is (much) greater than zero.
- Medium-tailed = *mesokurtic*: μ_4 is somewhat large / $\mu_4 - 3$ is (much) close to zero.
- Thin-tailed/"Flat" = *platykurtic*: μ_4 is small / $\mu_4 - 3$ is negative.

Moments will be come somewhat more important as we move through the term.

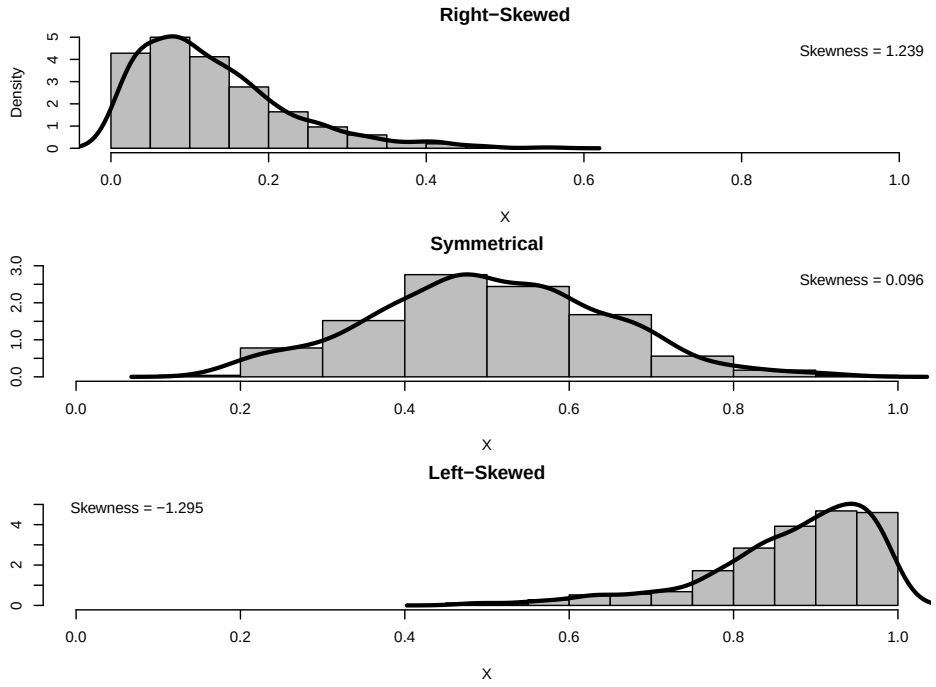
Examining our NFL Week One points data, we find get the following:

```
> library(moments)

> with(NFL, skewness(points))
[1] -0.229

> with(NFL, kurtosis(points))
[1] 2.66
```

Figure 1: Skewness Illustrated



A Few More Special Cases

Symmetrical Distributions

If the data are *symmetrical*, then a few things are true:

- The median is equal to the average (mean) of the first and third quartiles. That, in turn, means that
- Half the IQR equals the MAD (i.e., the "distance" between the median and (say) the 75th percentile is equal to the MAD).
- The skewness is zero.

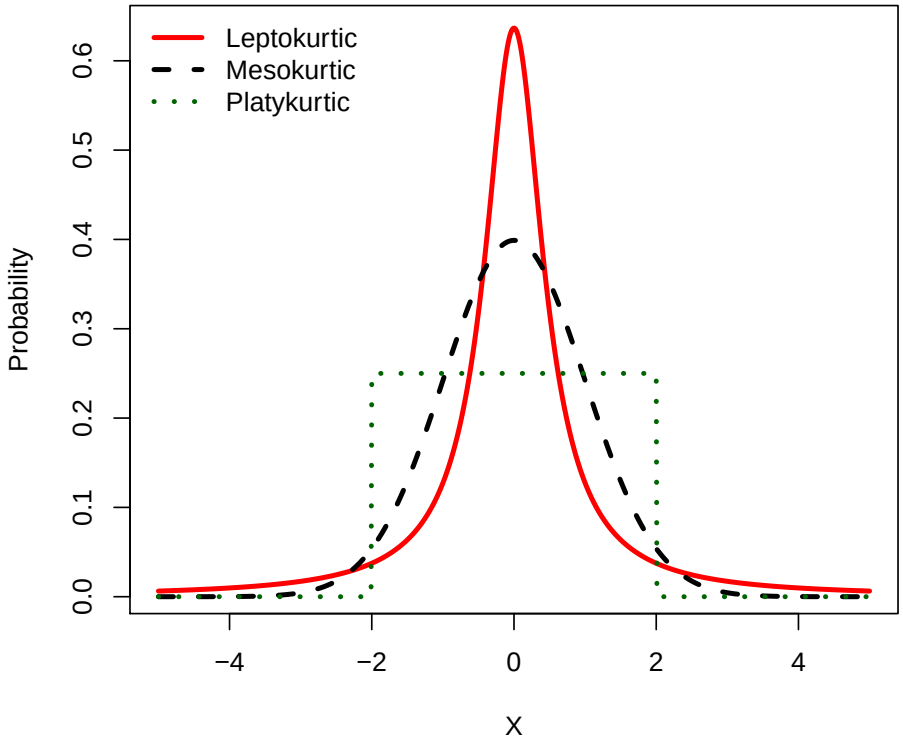
Nominal-Level Variates

There isn't much to say here. About all you can do is report category percentages. One should *not* use any of the measures discussed above with nominal-level data.

Ordinal Variates

As a rule, for ordinal-level variables, it is usually a good idea to stick with "robust" measures of variation/dispersion for ordinal-level variates. That means using MADs and the like, rather than variances and standard deviations. While this is not usually (read: almost never)

Figure 2: Kurtosis Illustrated



done in practice, it remains the right thing to do. (Changing the world one future-Ph.D. at a time, that's me...).

Dichotomous Variates

For dichotomous covariates, the measures discussed above have some special characteristics. For example, the range of a dichotomous variable is always 1.0, and the percentiles are always either values of zero or one; similarly, the IQR is always either zero (for dichotomous variables with fewer than 25 or more than 75 percent "1s") or one (for dichotomous variables with between 25 and 75 percent "1s").

Likewise, a dichotmous variate has a strict relationship between its mean and its variance. In particular, the variance of a dichotomous variable D is:

$$\sigma_D^2 = \bar{D} \times (1 - \bar{D})$$

Similarly, the standard deviation (σ_D) is simply the square root of this number. Since $\bar{D}$ is necessarily between zero and one, then σ_D^2 is as well; that means that $\sigma_D > \sigma_D^2$. This and equation 11 also suggest that the variance and standard deviation of a binary variable will always be greatest when $\bar{D} = 0.5$ (that is, equal numbers of "zeros" and "ones"), and will

decline as one moves away from this value.

Note, for example, that if we added an indicator called `NFC`[2] to our NFL Week One data, it would look like this:

```
> with(NFL, summary(NFC))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     0.5     0.5     1.0     1.0
> with(NFL, var(NFC))
[1] 0.258
> with(NFL, sd(NFC))
[1] 0.508
> with(NFL, mad(NFC))
[1] 0.741
> with(NFL, skewness(NFC))
[1] 0
> with(NFL, kurtosis(NFC))
[1] 1
```

Best Practices

Here are a few tips for summary measures of central tendency and dispersion:

1. It is customary (and useful) to present summary statistics for every variable you use in a paper/project/whatever. This is a good habit to acquire; one almost always transforms variables, rescales them, etc., and so presenting summary measures is a good way of ensuring that your reader can (better) understand what is going on.

2. Typically, one presents means, standard deviations, and minimums & maximums, as well as an indication of the total number of observations in the data. An example looks like this:

Summary Statistics

[2]half of the teams in the NFL are in the National Football Conference (NFC); the other half are in the American Football Conference (AFC).

Variable	Mean	Standard Deviation	Minimum	Maximum
Assassination	0.01	0.09	0	1
Previous Assassinations Since 1945	0.45	0.76	0	4
GDP Per Capita / 1000	5.83	6.04	0.33	46.06
Political Unrest	0.01	1.01	-1.67	20.11
Political Instability	-0.03	0.92	-4.66	10.08
Executive Selection	1.54	1.34	0	4
Executive Power	3.17	2.39	0	6
Repression	1.67	1.19	0	3

Note: $N = 5614$. Statistics are based on all non-missing observations in the model in Table X.

- For an important "dependent" variable, it's also generally a good idea to present some sort of graphical display of the distribution of the response variable (particularly if it is continuous or almost so – for a dichotomous response, it's somewhat less necessary).