PLSC 502: "Statistical Methods for Political Research"

Measures of Central Tendency
September 13, 2016

Introduction

We'll focus today on measures of the *central tendency* of a variable; these are sometimes also called *location* statistics. There are three such measures in wide use: the *mean*, the *median*, and the *mode*; we'll discuss each in turn, compare their relative merits, and talk a bit about special cases of each.

By way of example, we'll use as a single variable X the number of points scored by each of the 32 NFL teams in week one of the 2016 season:

Redskins	16	Giants	20	Bengals	23	Ravens	13
Jets	22	Dolphins	10	Chiefs	33	Patriots	23
Texans	23	Steelers	38	Jaguars	23	Titans	16
Lions	39	Falcons	24	Seahawks	12	Bills	7
Rams	0	Eagles	29	Buccaneers	31	Saints	34
Bears	14	Colts	35	Panthers	20	Chargers	27
Cardinals	21	49ers	28	Cowboys	19	Browns	10
Vikings	25	Packers	27	Broncos	21	Raiders	35

The histogram for these data is in Figure 1.

The Mean

We all know what a mean is; it's also known as an "expected value,"[1] or colloquially as an "average." It's usually denoted with a "bar" over the variable (a la $\bar{X}$), and is calculated as:

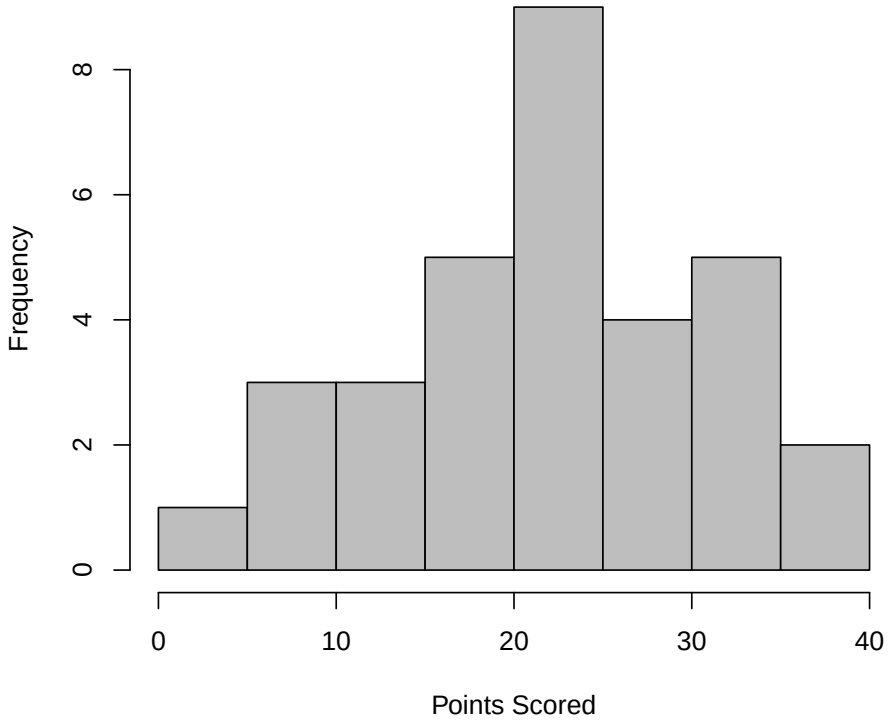
$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{1}$$

where N is the number of observations in the data. We also occasionally use the Greek letter "mu" (μ) to denote the mean.

Note at the outset that calculating a mean only makes sense of the data are *numeric*, and (in general) if the variable in question is measured at least at the *interval* level. The mean

[1] Formally, the expected value is a probability weighted mean; thus, for discrete variables, it is typically written as $\mathrm{E}(X) = \sum xf(x)$, while for continuous variates the formula is $\mathrm{E}(X) = \int xf(x)dx$. We'll return to this formulation a bit later in the course.

Figure 1: Points Scored in Week One of the 2016 NFL Season



of an ordinal variable is a quantity of some significant disagreement: while it might provide some useful information, the fact that the data are only ordinal means that the assumption necessary for summation (i.e., that the values of the variable are "equally spaced") is highly questionable. It *never* makes sense to calculate the mean of a nominal-level variate.

It's common to think of the mean as the "balance point" of the data – that is, the point in X at which, if every value of X was given "weight" according to its size, the data would "balance." This suggests that, if we add a single data point (call it X_{N+1}) to an existing variable X, the "new" mean (using $N+1$ observations) will:

- be greater than the old one if $X_{N+1} > \bar{X}$,
- be less than the old one if $X_{N+1} < \bar{X}$, and
- be the same as the old one iff $X_{N+1} = \bar{X}$.

A mean can also be thought of as the value of X that *minimizes the squared deviations between itself and every value of* X_i. That is, suppose we are interested in choosing a value for X (call it μ) that minimizes:

$$\begin{aligned} f(X) &= \sum_{i=1}^{N}(X_i - \mu)^2 \\ &= \sum_{i=1}^{N}(X_i^2 + \mu^2 - 2\mu X_i) \end{aligned}$$

To find the minimum of this function, we need to calculate $\frac{\partial f(X)}{\partial X}$, set that equal to zero, and solve. Doing so yields:

$$\frac{\partial f(X)}{\partial X} = \sum_{i=1}^{N}(2\mu - 2X_i)$$

$$\begin{aligned} \sum_{i=1}^{N}(2\mu - 2X_i) &= 0 \\ 2N\mu - 2\sum_{i=1}^{N} X_i &= 0 \\ 2N\mu &= 2\sum_{i=1}^{N} X_i \\ \mu &= \frac{1}{N}\sum_{i=1}^{N} X_i \equiv \bar{X} \end{aligned} \tag{2}$$

Frequency Data

Note also that the fact that we're summing means that it's possible to calculate the mean by examining combinations of "frequency" data. Suppose that instead of the NFL data above, we were given a set of scores and the numbers (frequencies) of teams that scored that number of points:

Points	Frequency
3	1
7	1
10	5
13	2
⋮	⋮
41	1

For such frequency data, we can calculate the mean by summing across the frequency-weighted values of X:

$$\bar{X} = \frac{1}{N} \sum_{j=1}^{J} f_j X_j \tag{3}$$

where j now indexes distinct values of X, and f_j are the frequencies (counts) of each of the J distinct values of X. Our example would then have a mean equal to:

$$\bar{X} = [(1)3 + (1)7 + (5)10 + (2)13 + ... + (1)41]/32$$

Note also that a mean is very susceptible to the effect of *outliers* in the data. Single, exceptionally large values in X will tend to "pull the mean in their direction," and so can provide a misleading picture of the actual central "location" of the variable in question. We'll talk more about this below.

Other Flavors of Means: Geometric

Formally, the mean we're used to is the "arithmetic mean." Two other potentially useful variants on the mean are the *geometric* mean and the *harmonic* mean.

The geometric mean is defined as:

$$\bar{X}_G = \left(\prod_{i=1}^{N} X_i \right)^{\frac{1}{N}}. \tag{4}$$

which can also be written:

$$\bar{X}_G = \sqrt[N]{X_1 \cdot X_2 \cdot ... \cdot X_N}.$$

It's also interesting to note that we can rewrite (4) using logarithms as:

$$\left(\prod_{i=1}^{N} X_i \right)^{\frac{1}{N}} = \exp \left[\frac{1}{N} \sum_{i=1}^{N} \ln X_i \right]. \tag{5}$$

That is, the geometric mean of a variable X is equal to the exponent of the arithmetic mean of the natural logarithm of that variable.

As the name suggests, the geometric mean has a geometric interpretation. You can think of the geometric mean of two values X_1 and X_2 as the answer to the question:

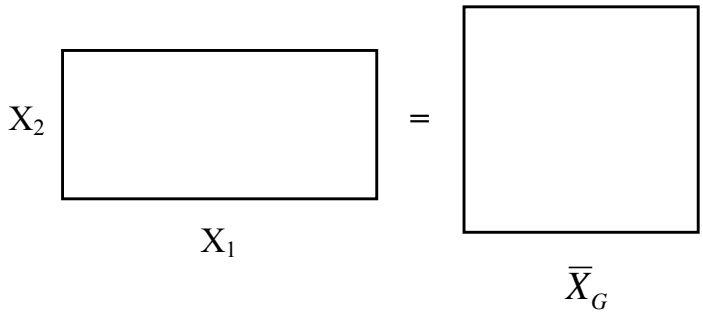
> *What is the length of one side of a square that has an area equal to that of a rectangle with width X_1 and height X_2?*

Similarly, for three values X_1, X_2, and X_3, one could ask:

> *What is the length of one side of a cube that has a volume equal to that of a box with width X_1, height X_2, and depth X_3?*

This idea is represented graphically (for the two-value case) in Figure 2.

Figure 2: The Geometric Mean: A (Geometric!) Interpretation



Finally, note a few things:

- As the geometric interpretation suggests, the geometric mean is only appropriate for variables with *positive* values. Variables that have either negative values, or values of zero, have an undefined geometric mean. Thera are various ways of dealing with zeros and negative values in calculating geometric means;
- It can be shown (though I'm not going to do it here) that the geometric mean is always less than or equal to the arithmetic mean:

$$\bar{X} \geq \bar{X}_G.$$

  In fact, the two are only equal if the values of X are the same for all N observations.

- While the geometric mean is not seen much, it is actually the *more* appropriate measure of central tendency to use for phenomena (such as percentages) that are more accurately multiplied rather than summed. For example, if we say the price of something doubled (that is, went up to 200 percent of the original price) in 2015, then decreased by 50 percent (back to its original price) in 2016, the actual "average" change in the price across the two years is *not* $(200 + 50)/2 = 125$ percent, but rather $\sqrt{200 \times 50} = \sqrt{10000} = 100$, or zero percent average (annualized) net change.

Other Flavors of Means: Harmonic

The harmonic mean is defined as:

$$\bar{X}_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}; \tag{6}$$

that is, it is the product of N and the reciprocal of the sum of the reciprocals of the X_is. Equivalently:

$$\bar{X}_H = \frac{1}{\left(\bar{\frac{1}{X}}\right)},$$

that is, the harmonic mean is the reciprocal of the (arithmetic) mean of the reciprocals of X.

Because it considers reciprocals, the harmonic mean is the friendliest toward small values in the data, and the least friendly to large values. This means, among other things, that:

- the harmonic mean will always be the smallest of the means in value (the arithmetic mean will be the biggest, and the geometric will be "in between" the other two),[2] and
- the harmonic mean tends to limit the impact of large outliers, and increase the "weight" given to small values of X.
- Finally, like the geometric mean, the harmonic mean is undefined if there are zero or negative values in the data; this follows pretty directly from Equation (6).

To be honest, there are not a lot of instances where one is likely to use the harmonic mean. But, what the heck...

The Median

The *median* of a variable X – sometimes denoted with a "check" ($\check{}$) – can be defined as:

$$\begin{aligned} \check{X} &= \text{"middle observation" of } X \\ &= \text{50th percentile of } X. \end{aligned} \tag{7}$$

Note that this isn't a very formal definition; we'll come back to this when we discuss quantiles a bit later. Practically speaking, the median is the value of:

- the $\left(\frac{(N-1)}{2} + 1\right)$th-largest value of X when N is odd, and

[2] Note again, however, that all three are the same if all the values of X are identical.

- the mean of the $\left(\frac{N}{2}\right)$th- and $\left(\frac{N+2}{2}\right)$th- largest values of X when N is even.

For our NFL opening day data, there are 32 teams; the median is therefore the average of the numbers of points scored by the 16th- and 17th-most point-scoring teams. At it happens, both of those values are 23, making the median 23. (This also happens to be the *mode*; see below.)

As with the mean, the median is typically only calculated for ordinal-, interval- or ratio-level data. Unlike the mean, however, there is no particular problem with using it for ordinal variates, since it just reflects a "middle" value (and an ordinal variable orders the observations). It is *not*, however, calculated for nominal data.

The median has a number of interesting qualities. For example, while the mean is the number that minimizes the squared distance to the data, the median is the value of c that minimizes the *absolute value* of the distance to the data:

$$\check{X} = \min\left(\sum_{i=1}^{N} |X_i - c|\right). \tag{8}$$

This latter fact is important in social science applications, e.g., in the application of the "median voter theorem."[3] As we'll see in a minute, a median is also (relatively) unaffected by "outliers;" for this reason, the median is often known as a "robust" statistic.

The Mode

The *mode* (which – creatively – I'll just denote by mode(X)) is nothing more than the most commonly-occurring value of X. In our NFL data, for example, the most common number of points scored was ten (by no less than five teams), so 10 is the mode for that variable.

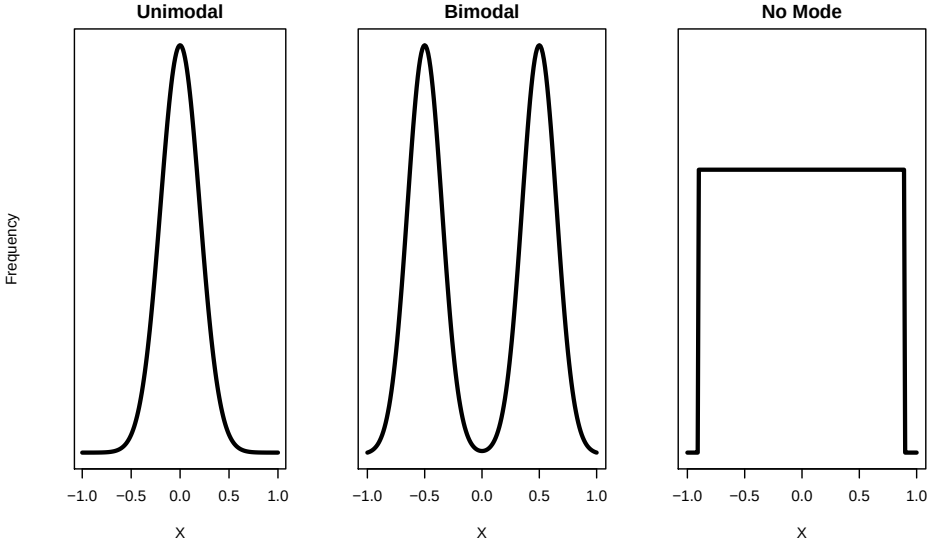
While a mode doesn't sound like much, it's actually a very useful quantity, for a host of reasons:

- The mode is the *only* measure of central tendency appropriate for use with data measured at any level, including nominal. That means that the model is also the only summary statistic appropriate for nominal variates.

[3]Note, interestingly, that when N is even, any value of $\check{X}$ between the $N/2$th and $(N+2)/2$th values will satisfy (8); formally, this means that the median is indeterminate when N is even. Also, note that both (2) and (8) can be thought of as special cases of minimizing the generalized Minkowski distance $D = \left(\sum_{i=1}^{N} |X_i - c|^p\right)^{\frac{1}{p}}$ for $p = 1$ (for the median) or $p = 2$ (for the mean).

- At the same time, the mode is (technically) undefined for any variable that has equal maximum frequencies for two or more variable values. This can happen in a number of ways:
    - Discrete data may have two (or more) categories with equal numbers of maximum-frequency categories. Such data are said to be *bimodal* (or *multimodal*). Strictly speaking, a bimodal distribution must have two categories with identical frequencies of data; more generally, however, the term is used to describe any data with two large "lumps" in the frequency distribution (like the Muslim population data we saw last week).
    - With continuous (and fine-grained discrete) variables, it is often the case that there is only one of every value of X in the data. In this case, the data are said to have *no mode* at all.

Figure 3: Modes Illustrated



Other Things to Know

Binary Variables

Think for a minute about a dichotomous (binary) variable; call it D. Note a few things:

- $\bar{D} = \frac{1}{N}\sum_{i=1}^{N} D_i \in [0, 1]$ is equal to the proportion of "1s" in the data.
- $\check{D} \in \{0, 1\}$, depending on whether there are more "0s" or "1s" in the data.
- $\text{mode}(D) = \check{D}$.

The first of these means that the mean of a binary variable tells us a lot about it: not just its mean, but (as we'll see next week) its variance as well. At the same time, because it is never the case that a binary variable's mean is equal to a value actually present in the data,[4] it is common to use the median/mode as a measure of central tendency for binary variables as well.

Relationships

For reasons we'll delve into at greater length next class, it is the case that (in general):

- In a perfectly symmetrical continuous variable, the mean and median are identical.
- If the same variable is unimodal, then the mode is also equal to the mean and the median.
- If a continuous variable Z is *right-skewed*, then it is generally the case that

$$\text{mode}(Z) \leq \check{Z} \leq \bar{Z} \tag{9}$$

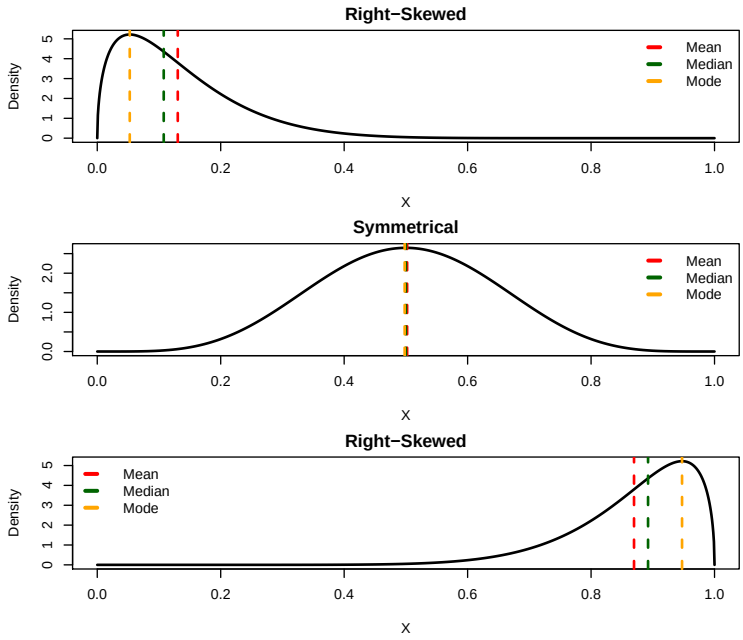
- Similarly, if a continuous variable Z is *left-skewed*, then it is usually true that

$$\bar{Z} \leq \check{Z} \leq \text{mode}(Z) \tag{10}$$

Figure 4 illustrates these general rules:

[4]Because if it was, of course, then all the values of D_i would have to be identically zero or one, in which case it's no longer a variable, but a constant.

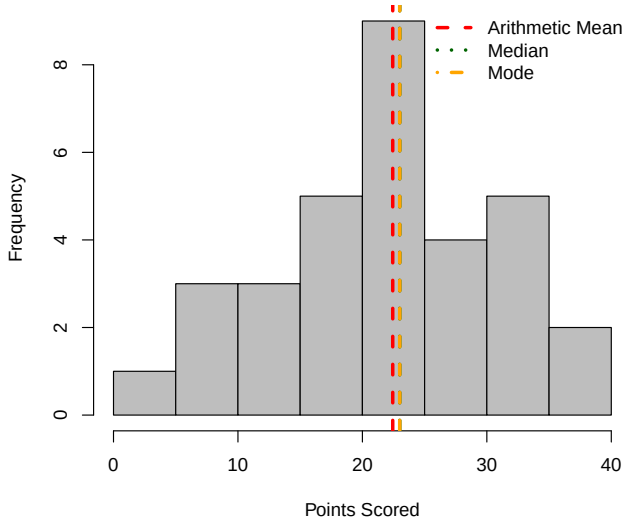
Figure 4: Locations of Means, Medians, and Modes



These are good rules of thumb, though as this blog post suggests, they are not mathematical certainties. In both instances, the mean is the most affected by "outliers."

We can see this in Figure 5. The data are effectively symmetrical, which means that the mean (at 22.44) is almost identical to the median and the mode (both of which are 23).

Figure 5: Points Scored in Week One of the 2016 NFL Season Redux



Practical Tips

- Never, never, never use anything but the mode for nominal-level data.
- Means can be used for ordinal-level variables, but should be applied with care.
- Harmonic and geometric means are *very* rarely used in political science, even when they should be.
- Be careful doing things like "averaging" means (across data sets and the like). The fact that means incorporate frequency information makes that a dangerous proposition.
- Finally, remember that a figure will almost always tell you more about your data than a number...