Summation & Central Tendency (based on Kirk, Ch. 3)

Summation

Summation is just a short-hand notation for adding up a bunch of numbers.

So, if our goal is to add up all the scores for variable X, we could write

\[ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 \]

But this takes up a lot of space and assumes that we know how many cases there are. We could also write:

\[ X_1 + X_2 + ... + X_7 \]

This takes up less space, but it’s a bit clumsy and also assumes we know how many cases there are. A more general form takes care of the number of cases problem:

\[ X_1 + X_2 + ... + X_n \]

Here, \( n \) stands for the number of cases there are for that variable, so we just keep adding until we get to the end. But we still have the “clumsiness” problem. The standard notation that takes care of this is:

\[ \sum_{i=1}^{n} X_i \]

This means, start with case number 1 (\( i = 1 \)), and add up (\( \Sigma \)) every case until you get to the last (\( n^{th} \)) case (in the example above, this was 7).

If we wanted to start with the 2\(^{nd} \) case and end with the next to last case, our notation would be:

\[ \sum_{i=2}^{n-1} X_i \]

So, for the example above, \( \sum_{i=2}^{n-1} X_i = X_2 + X_3 + X_4 + X_5 + X_6 \)

Here are a few more examples:

\[ \sum_{i=2}^{3} X_i = X_2 + X_3 \]
\[ \sum_{k=2}^{3} X_k = X_2 + X_3 \] (don’t have to use \( i \) for the index)

\[ \sum_{i=5}^{1} X_i = X_7 + X_6 + X_5 + X_4 + X_3 + X_2 + X_1 \] (can sum backward)

\[ \Sigma X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 \] (the lack of subscripts/superscripts means to sum all values)

There are often times when we first want to perform some operation on scores and then add up the results. The summation notation works here, too:

\[ \sum_{i=1}^{n} (X_i - \bar{X}) = (X_1 - \bar{X}) + (X_2 - \bar{X}) + ... + (X_7 - \bar{X}) \]

Once we start involving more than one operation, however, we need to remember to use the order of operations. Summing/adding (along with subtraction) goes last, so we first want to perform the specified operations and then add.

**Order of Operations Review**

First, we compute what’s in parentheses, then any exponents (i.e., powers; remember that this includes roots), then multiplication and division (these are at the same level because division is just a special case of multiplication), then adding and subtracting (again, these are at the same level because subtraction is just a special case of addition...adding a negative number).

Some people use an acronym to remember this order: **PEMDAS** (for Parentheses, Exponentiation, Multiplication/Division, Addition/Subtraction), sometimes expressed as the mnemonic "Please Excuse My Dear Aunt Sally"

Practice using: \( X = \{7, 8, 10, 12, 13\} \) \( Y = \{10, 8, 11, 17, 19\} \)

1) \[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \]
   first, calculate what’s in the parentheses for each \( i \) (case), then square each of these results, then add them up  ...remember to use a table to keep calculations neat!
2) \[ \sum_{i=1}^{n} (X_i - \bar{X})^2 - 5 \]

3) \[ \sum_{i=1}^{n} X_i^2 \]

4) \[ \left( \sum_{i=1}^{n} X_i \right)^2 \]

5) \[ \sum_{i=1}^{n} X_i^2 - \frac{\left( \sum_{i=1}^{n} X_i \right)^2}{n} \]

6) \[ \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y}) \]

Some Summation Theorems (short cuts you can use if you want...handy for proofs...see Kirk, p. 88-91)

\[ \Sigma(X + Y) = \Sigma X + \Sigma Y \] (adding pairs of Xs and Ys is the same as adding Xs and Ys separately and then adding together)

\[ \Sigma kX = k\Sigma X \] (k is a constant...this is really just factoring...5(8) + 5(4) + 5(3) = 5(8+4+3))
\[ \Sigma (X+k) = \Sigma X + nk \] (\(k\) is a constant...since \(k\) remains constant, we are adding it up \(n\) times)

Will use summation throughout the quarter!

**Central Tendency**

**Mode** \((Mo)\): score or category that occurs most frequently (distribution can be bimodal or multimodal)

**Mean** \((\bar{X})\): average; most frequently use “arithmetic mean”

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

**Calculating from Ungrouped Frequency Distribution:**

- Sometimes scores are reported as an ungrouped frequency distribution (see Ch. 2 handout)
- Instead of listing the same score multiple times (e.g., 7, 7, 7, 7, 7), the frequency distribution tells us how many times \((f = 5)\) this score \((X = 7)\) was observed
- We could use the raw score formula (above) to calculate the mean, but a shortcut is just to multiply each score by its associated frequency
- This becomes particularly useful when there are large numbers of the same score (and we don’t have a computer around to help us out!)
- Formally, this formula is written:

\[
\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n}, \text{ where } j \text{ indexes the number of class intervals (scores); } k \text{ is the number of class intervals, and } f \text{ is the number of cases with a given score}
\]

**Median** \((Mdn)\): point in the distribution that divides the data into two groups having equal frequency

determining the median is straightforward if there is only one case for the middle score(s):
- if \(n\) is odd, the median is simply the middle score
- if \(n\) is even, the median is halfway between the two middle scores
if there are multiple cases where the median lies, we must “interpolate”

**logical interpolation:**

1) mentally work your way in to the middle of the distribution (e.g., start from the outsides of the distribution and *evenly* take away cases from each side) until you reach the middle score with multiple observations (if there is only one observation, that score is your median, and there is no need to proceed)

2) using the real limits as the boundaries for the score (e.g., a score of 7 is really 6.5 to 7.5), divide that score up into pieces...one piece for each case (if there are 4 cases, the score is divided into quarters; if there are 3 cases, the score is divided into thirds, etc.)

3) now continue working your way into the middle of the distribution; where you end up is the median

**interpolation by formula**

for this method, it is necessary to complete step (1) above

then compute $Mdn = X_{ll} + i \left( \frac{n/2 - \sum f_b}{f_i} \right)$, where:

- $X_{ll}$ is the *real* lower limit of the interval containing the median (the logic is that you are starting at the bottom of the interval and counting up into it)
- $i$ is the size of the class interval (1 if you are working with ungrouped distributions)
- $n$ is the number of scores in the entire distribution (thus, $n/2$ is the location of the midpoint)
- $\sum f_b$ is the number of scores below $X_{ll}$ (this tells you how many scores up from the bottom that you’ve already come...so, taking this away from $n/2$ tells you how many more scores you have to go)
- $f_i$ is the number of scores there are in the interval containing the median (this tells you how many pieces to divide that interval into)

Example:

$X = \{3, 3, 3, 4, 4, 4, 5, 6\}$

$Mdn (X) =$


caution: do not use the interpolation procedure if the median falls in a class interval that does not contain any cases (use the “straightforward” method instead); e.g., \( X = \{2, 2, 3, 8, 9, 11\} \)

Comparison of Measures of Central Tendency

<table>
<thead>
<tr>
<th>properties/stability</th>
<th>applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td><strong>applicability</strong></td>
</tr>
<tr>
<td>responds to change in every score (i.e., ((X_1 + X_2 + \ldots + X_n)/n)) acts as the balancing point of the distribution (i.e., (\sum(X - \bar{X}) = 0)) is the most sensitive to extreme scores (think Bill Gates) <strong>stability</strong>: most resistant to chance sampling variation when the population distribution is bell-shaped</td>
<td><strong>measurement scales</strong>: near-interval ((?)), interval, ratio; not not nominal or pure ordinal; not open-ended distributions (e.g., with “50 or above”) <strong>distribution shape</strong>: best for fairly symmetrical distributions (can be problematic for highly skewed) <strong>usefulness</strong>: used in many mathematical procedures/statistical computations (because it is highly “tractable”)</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td><strong>applicability</strong></td>
</tr>
<tr>
<td>does not respond to how far above or below each score is therefore it is much less sensitive to extreme scores <strong>stability</strong>: second most resistant to chance sampling variation when the population distribution is bell-shaped; more stable (than mean) when the population distribution is highly skewed or has extreme outliers</td>
<td><strong>measurement scales</strong>: ordinal, near-interval, interval, ratio; can be used for open-ended distributions; not for nominal <strong>distribution shape</strong>: most commonly used for skewed distributions or distributions with extreme scores; also used descriptively for symmetrical distributions <strong>usefulness</strong>: little use in inferential statistics (less tractable)</td>
</tr>
<tr>
<td><strong>Mode</strong></td>
<td><strong>applicability</strong></td>
</tr>
<tr>
<td><strong>stability</strong>: affected greatly by sampling variation (i.e., the value changes a lot); it is also heavily tied to the width and location of class intervals in grouped distributions</td>
<td><strong>measurement scales</strong>: can be used for all measurement scales (only one for nominal data); some say mode is best for quantitative, discrete variables (e.g., number of kids) <strong>usefulness</strong>: rarely used other than descriptively</td>
</tr>
</tbody>
</table>

Relative Locations:

In most skewed distributions, the mean is “pulled down” by the extreme distributions in the tails, the mode is at the “hump,” and the median is in between (note that this is in alphabetical order)
But, as von Hippel (2005) points out, this is not always the case; the most common exception occurs when discrete variables are used from von Hippel (2005). Mean, Median, and Skew: Correcting a Textbook Rule. (see http://www.amstat.org/publications/jse/v13n2/vonhippel.html):

Under these definitions, discrete distributions can easily break the rule. For example, in the General Social Survey, respondents are asked how many people older than 18 live in their household. Figure 2 gives the responses for 2002 (1996 was similar). The skew is clearly to the right, yet the mean is left of the median and mode.