Week 8 - Describing data

Dr Micah

Research report update

Instructions on scoring data after completing the 20-item depression scale (https://github.com/jjcurtin/arc_measures/raw/main/CESD/CESD.pdf) for your research report.

Completing the Depression inventory

To score the CES-D scale, the following steps should be taken:

- Add up the scores for each of the 20 items.
- Reverse score the highlighted items **4, 8, 12,** and **16**. Note that these items are worded in a *positive* way (e.g. I felt happy), meaning *higher* scores here correspond with *lower* depression.
- Reversing the highlighted items ensures that high (low) scores on these items indicate low (high) levels of depression.
- Add up the scores for all 20 items, including the reversed scored items.

The total score can range from 0 to 60, with higher scores indicating greater levels of depression symptoms.

Center for Epidemiologic Studies Depression Scale (CES-D), NIMH

Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past week.

	During the Past Week				
	Rarely or none of the time (less than 1 day)	Some or a little of the time (1-2 days)	Occasionally or a moderate amount of time (3-4 days)	Most or all of the time (5-7 days)	
1. I was bothered by things that usually don't bother me.	☐	☐	☑	☐	3
2. I did not feel like eating; my appetite was poor.	☐	☑	☐	☐	2
3. I felt that I could not shake off the blues even with help from my family or friends.	☐	☑	☐	☐	2
4. I felt I was just as good as other people.	☐	☐	☑	☐	2 ← Reverse scored
5. I had trouble keeping my mind on what I was doing.	☑	☐	☐	☐	1
6. I felt depressed.	☐	☑	☐	☐	2
7. I felt that everything I did was an effort.	☐	☐	☑	☐	2
8. I felt hopeful about the future.	☐	☑	☐	☐	4 ← Reverse scored
9. I thought my life had been a failure.	☐	☐	☑	☐	3
10. I felt fearful.	☐	☑	☐	☐	2
11. My sleep was restless.	☐	☑	☐	☐	2
12. I was happy.	☐	☐	☑	☐	3
13. I talked less than usual.	☐	☑	☐	☐	2
14. I felt lonely.	☐	☐	☑	☐	3
15. People were unfriendly.	☐	☐	☐	☑	4
16. I enjoyed life.	☐	☐	☑	☐	2 ← Reverse scored
17. I had crying spells.	☐	☑	☐	☐	2
18. I felt sad.	☐	☐	☑	☐	3
19. I felt that people dislike me.	☐	☑	☐	☐	2
20. I could not get "going."	☑	☐	☐	☐	1

SCORING: zero for answers in the first column, 1 for answers in the second column, 2 for answers in the third column, 3 for answers in the fourth column. The scoring of positive items is reversed. Possible range of scores is zero to 60, with the higher scores indicating the presence of more symptomatology.

Sample Depression Inventory

Highlighted values are *reverse scored*. This is because, relative to the other items on the scale, higher scores for items $4, 8, 12$ and 16 are negatively related to depression.

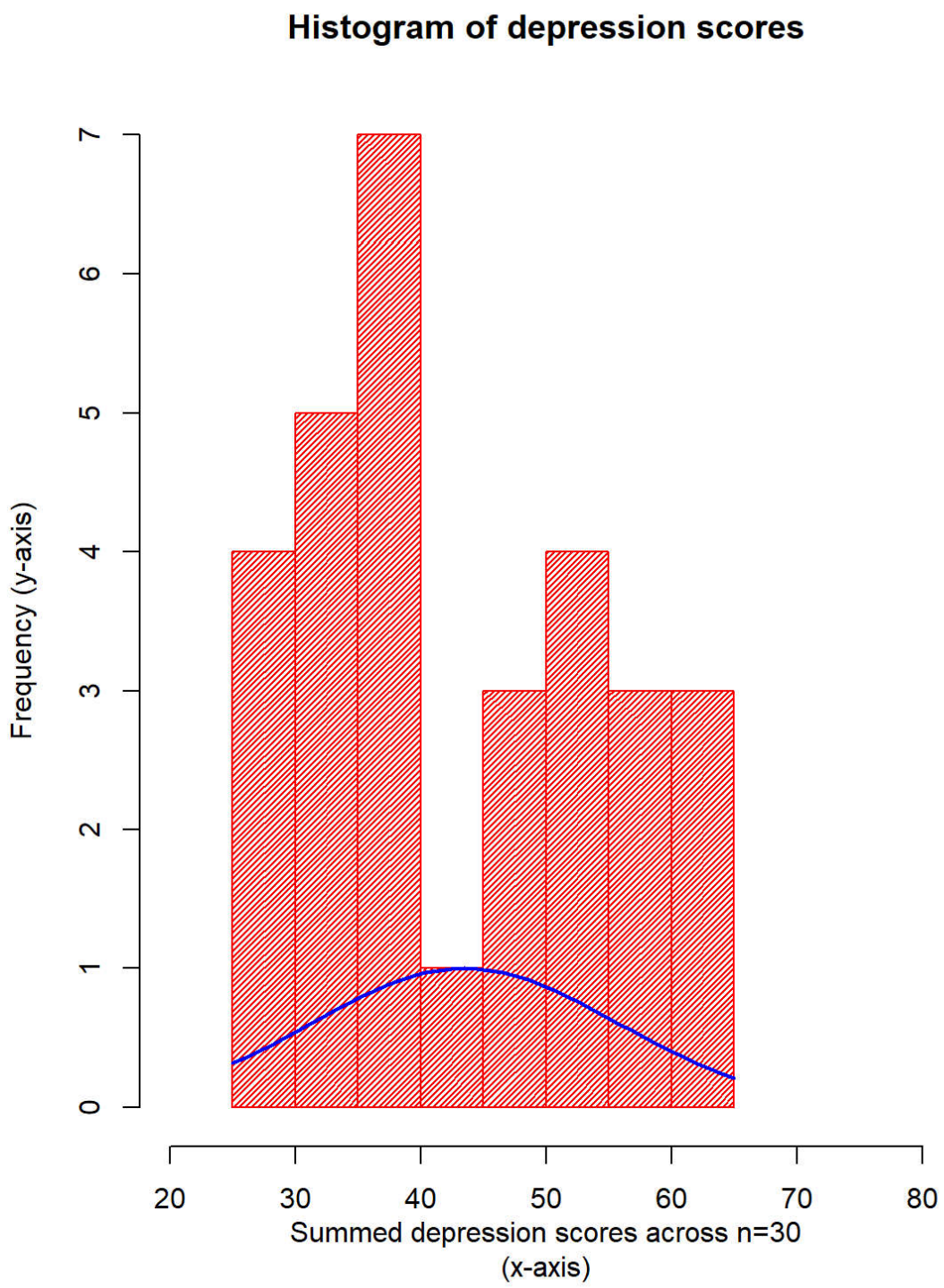
The raw values recorded our hypothetical participant are 3, 2, 2, 2, 1, 2, 2, 4, 3, 2, 2, 3, 2, 3, 4, 2, 2, 3, 2.

The total ($\sum$) score for this participants is 46. This is the *depression score* for one participant.

Exploring how scores are distributed

Suppose we have collected depression inventory data from $n = 30$ persons. The individual depression scores for 30 hypothetical participants are 52, 55, 32, 65, 31, 46, 60, 36, 30, 43, 50, 35, 40, 30, 37, 47, 27, 32, 33, 37, 25, 53, 58, 59, 65, 63, 36, 36, 52, 40.

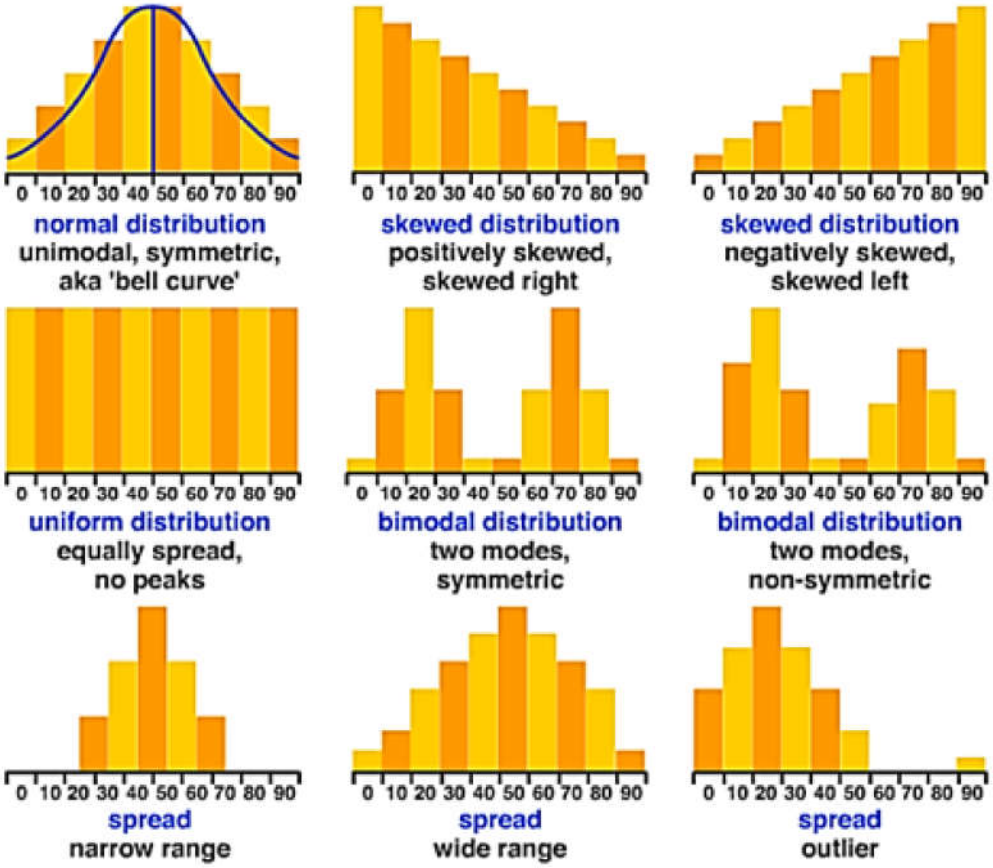
We can visually explore how the data is distributed using a *histogram*



In the histogram above, the *y*-axis represents the relative frequencies with which individual scores (*x*-axis) were observed. We can observe that our sample of $n = 30$ participants produced depression scores ranging between 25 and 65. Within this range, we can see the scores are **bimodally** distributed (there are two frequency 'peaks'). Many statistical tests require data to be **normally-distributed** (note the *blue* curve). It is good practice to explore the shape of your data before running any tests!

A histogram graphically represents the distribution of a dataset. It shows the frequency of different values in the dataset by dividing the range of values into *bins*, then plotting the number of data points that fall into each bin as a *bar*. The x-axis of a histogram represents the bins, and the y-axis represents the frequency of the data points in each bin. The height of each bar represents the number of data points that fall into that bin. The shape of the histogram can reveal information about the distribution of the data, such as whether it is symmetric or skewed, and whether it has one or multiple peaks. In general, if the histogram is symmetric and bell shaped, this indicates that the data is normally distributed, while a skewed histogram indicate a non-normal data distribution.

Some examples of distributions



Varieties of distributions

- A normal distribution is a probability distribution that is *symmetric about the mean*, with the majority of the data points concentrated around the mean and the number of data points decreasing as the distance from the mean increases. This type of distribution is also known as a Gaussian distribution or a bell-shaped distribution.
- A skewed distribution shows data points which *are not* evenly distributed around the mean. One tail of the distribution is longer or fatter than the other. If the tail is on the left side of the graph, it is said to be negatively skewed and if the tail is on the right side of the graph, it is said to be positively skewed.
- A bimodal distribution is a distribution with *two peaks*, indicating that there are two distinct groups of data points with different modes (the most common value). This type of distribution is often seen when there are two distinct groups of observations or when there is a mix of two different types of observations.

> Histograms are a useful tool for visualizing the distribution of a dataset and understanding the frequency of different values in the data. However, they do not always provide a complete picture of the data or convey certain information. For example, histograms do not provide a single summary value that describes the typical or center of the data, which can be important when trying to make comparisons or draw conclusions about the data. In addition, histograms can be misleading when the data is skewed, showing a false representation of the data. Central tendency statistics can give a better idea of the true center of the data, even when it is skewed.

Central tendency statistics

Central tendency statistics, such as the mean, median, and mode, provide a single value that represents the center of a dataset. They are useful for summarizing the data and providing a general idea of where the data is concentrated.

Central tendency statistics can include the mean, median, and mode as single point estimates representing the center of a dataset.

> The mean is the sum of all values divided by the number of values and is sensitive to outliers.

> The median is the middle value of the dataset that is *not* affected by outliers.

> The mode is the value that appears most frequently in the dataset.

Depending on the distribution of the data and the purpose of the analysis, different central tendency statistics may be more appropriate.

Recall our hypothetical depression scores from earlier, which are sorted below from smallest to largest.

25, 27, 30, 30, 31, 32, 32, 33, 35, 36, 36, 36, 37, 37, 40, 40, 43, 46, 47, 50, 52, 52, 53, 55, 58, 59, 60, 63, 65, 65

Let's look at the mean, median and mode of these values:

1. **Mean/average**: Sum up all individual data ($\sum$) and divide by sample size, $Mean = \frac{\sum 52+55+32+65+31+46+60+36+30+43+50+35+40+30+37+47+27+32+33+37+25+53+58+59+65+63+36+36+52+40}{N=30} = 43.5$
2. **Median**: Data point which occurs most frequently in the middle of the distribution, $Median = 40$
3. **Mode**: Most frequently occurring data point (not used extensively), $Mode = 36$

Median and mode estimates can be generally derived from 'eyeing' the data (recall our earlier histogram), whereas mean estimates typically require computation. Instead of summing up all the values manually and dividing the total by the sample, R has built-in functions for estimating means, medians (and many other useful estimates), which we cover in a moment.

Assigning numeric values to variables in R

In R, you can assign a number to a variable using the assignment operator '<-'. For example, you can assign the number 42 to a variable x like this

```
x <- 42
```

You can also assign multiple numbers to multiple variables at once. For example:

```
x <- 1
y <- 2
z <- 3
```

You can also use the ‘c()’ function to create *a vector of numbers* and assign it to a variable, For example:

```
numbers <- c(1, 2, 3, 4, 5)

# Now if you type in the variable in the console and press Enter, you will see the following:

numbers
```

[1] 1 2 3 4 5

Using R to estimate means

We can implement then use `R` ’s built-in functions for computing the `mean` and `median` estimates.

Let’s first assign the raw values to a variable we will be calling `raw.data` .

```
# Input raw numbers
raw.data <- c(52, 55, 32, 65, 31, 46, 60, 36, 30, 43, 50, 35, 40, 30, 37, 47, 27, 32, 33, 37, 25, 53, 58, 5
9, 65, 63, 36, 36, 52, 40)

# Estimate mean
mean(raw.data)
```

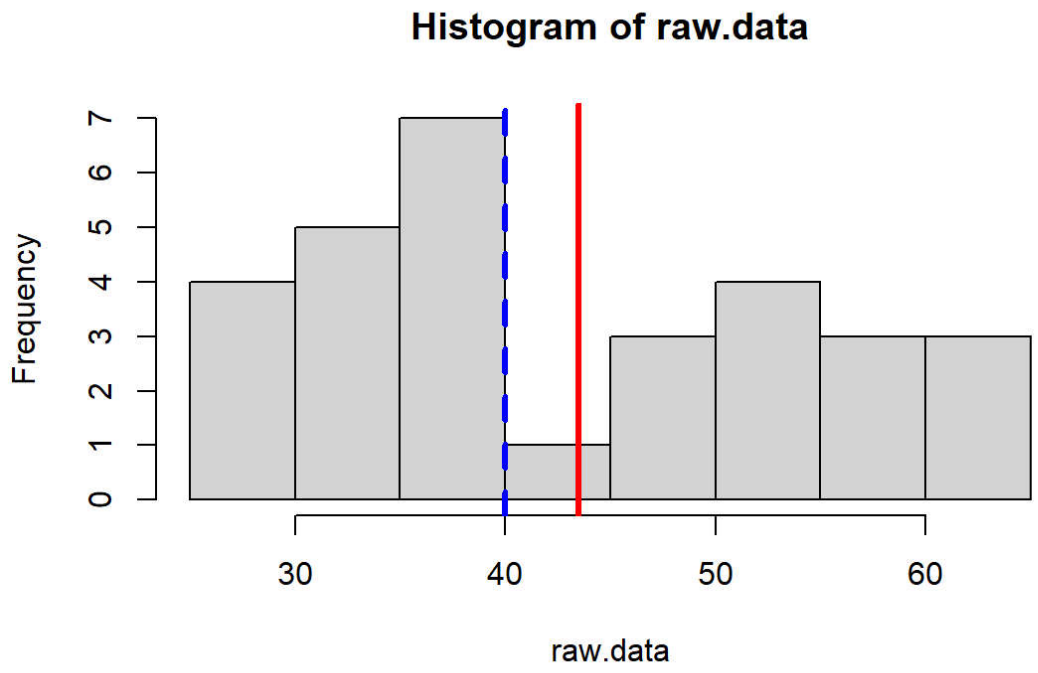
[1] 43.5

```
# Estimate median
median(raw.data)
```

[1] 40

`mean(raw.data) =` 43.5
`median(raw.data) =` 40

We can overlay the *mean* (red line) and *median* (blue line) values across a histogram to see how these central tendency estimates may represent the data.



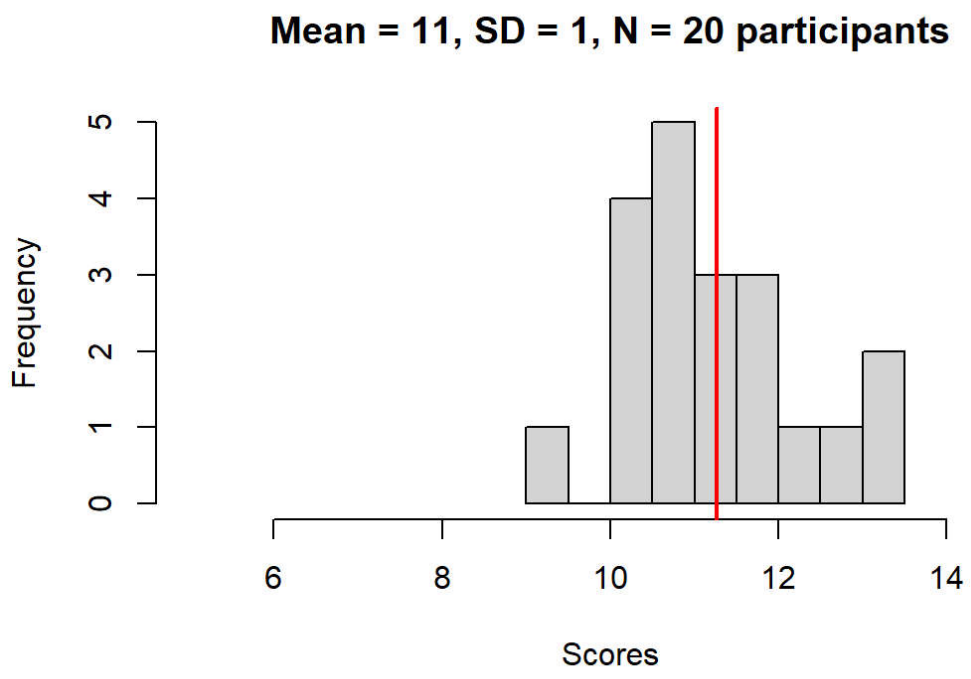
A good rule of thumb when looking at the mean and median estimates is noting their difference. A distribution is considered to be more normal if the mean and median are close together, indicating that the distribution is symmetric and not skewed to one side or the other. On the other hand, if a distribution is skewed to the left, the mean will be less than the median. If a distribution is skewed to the right, the mean will be greater than the median.

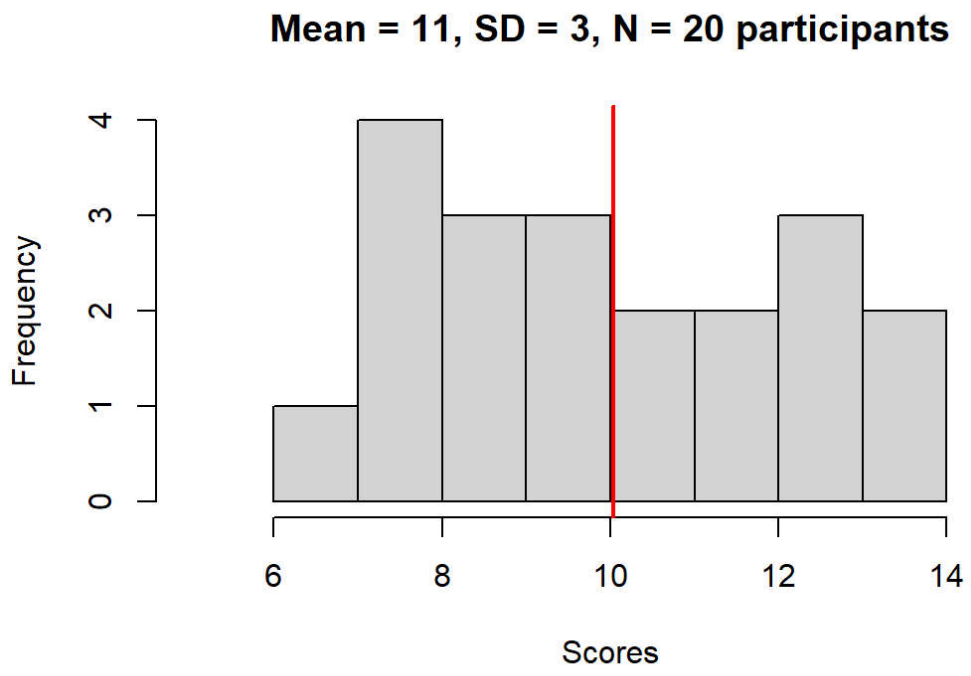
Note that this is only one measure of normality and is meant as a guiding tool.

Estimating variance

Measures of central tendency, such as the mean, median, and mode, provide a single value that represents the center or typical value of a dataset. However, these measures alone do not provide a complete picture of the data and its distribution.

For example, note how both histograms below have the same mean value (red line) from 20 observations, but have different ‘widths’.





The 'width' of the histogram is the *standard deviation*, or *SD*, which is a measure of variance that provides additional information about the spread or dispersion of the data. Measures of variance, such as the standard deviation and variance, indicate how much the data deviates from the central tendency measure, and how much the data values vary from one another.

Measures of variance are necessary to calculate other important statistics such as the standard error and confidence intervals, which provide information about the level of accuracy of a sample statistic and how it compares to the population parameter.

In summary, measures of central tendency provide a summary of the center of a dataset, while measures of variance provide information about the spread and variability of the data. Together, they give a complete picture of the data distribution and a better understanding of the dataset.

Defining Standard Deviation

The standard deviation is a measure of how much the numbers in a group are spread out from the average, or mean, of the group.

For example, imagine you have a group of five numbers: 1, 2, 3, 4, and 5. The mean of this group is (1+2+3+4+5)/5 = 3.

The standard deviation tells you how much these numbers deviate from the mean. If all the numbers are very close to the mean of 3, then the standard deviation will be low. But if some numbers are much higher or lower than the mean, the standard deviation (SD) will be high.

The SD is a common way to quantify the spread of a dataset. If the standard deviation is low, it means the data points are clustered around the mean, whereas if the SD is high, it means the data points are spread out over a wide range of values.

Estimating Standard Deviation manually

To calculate the standard deviation of a set of numbers, you can use the following 5 steps:

1. Find the mean of the numbers by adding them all up and dividing by the total number of numbers. For example, if the numbers are x1, x2, x3, …, xn, the mean is: $\bar{x} = \frac{x_1+x_2+x_3+...+x_n}{n}$
2. Subtract the mean from each number to find the deviations. For example, the deviation of x1 is: $x_1 - \bar{x}$
3. Square each deviation to eliminate negative values.
4. Sum all the squared deviations. $\sum_{i=1}^{n}(x_i - \bar{x})^2$
5. Divide the sum of squared deviations by the total number of numbers (n) and then take the square root of that quotient. $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$

This final value is the standard deviation of the set of numbers.

Fortunately, we can use the `sd()` function built within `R` to calculate this.

```
sd(raw.data)
```

```
## [1] 12.19878
```

We can round the number of decimal places to two digits

```
round(sd(raw.data),2)
```

```
## [1] 12.2
```

We can combine multiple values together (e.g., the *mean* and *sd*)

```
c(mean(raw.data),sd(raw.data))
```

```
## [1] 43.50000 12.19878
```

Let's output the *mean, median and standard deviations*, rounded to 2 decimal places, using the code we have learned so far

```
round(c(mean(raw.data),median(raw.data),sd(raw.data)),2)
```

```
## [1] 43.5 40.0 12.2
```

Recall that our `raw.data` variable represented raw depression scores from 20 participants.We can now describe the central tendency and variability of that distribution using our computed values in APA format.

> We implemented a 20-item depression survey across $n = 20$ participants. A histogram of depression scores indicated a bi-modal (non-symmetric) distribution. The Mean and SD depression scores recorded for our sample were $M = 43.5; SD = 12.2$.

Effect sizes and practically important differences

Imagine our initial sample of $n = 30$ participants were all USP students. We already know the mean and SDs for our intial sample of 20 participants were $M_{USP} = 43.5$ and $SD_{USP} = 12.2$ respectively.

Suppose we are provided data from a sample of 20 FNU students, with means and SDs of $M_{FNU} = 42.1$ and $SD_{FNU} = 8.65$ respectively. We can now test whether the *difference* (d) between the two groups are **practically meaningful**.

> One means to answer this is to explore for the **effect size** between the two variables. Briefly, effect size measures the strength of the relationship between two variables in an experiment. Effect sizes help researchers understand whether a difference or association they have found is practically meaningful or not. A large effect size indicates that there is a strong relationship or a large treatment effect, while a small effect size indicates that the relationship or effect is weak.

Cohen's d

A common effect size measure is known as Cohen's d, which is used to compare the means of two groups and to express the difference between them in standard deviation units.

Cohen's d can be estimated as $d = \frac{M_1 - M_2}{SD_{Pooled}}$ where $SD_{Pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$, where M_1 and M_2 are the two group means, and SD_1 and SD_2 are the two group standard deviations.

You will note a new term here, called SD_{Pooled}, which describes the 'pooled' standard deviation, which is simply the common standard deviation of the two groups being compared. The pooled standard deviation is calculated by combining the standard deviations of the two groups and then taking the square root of the average of the squared standard deviations.

The pooled standard deviation can be estimated as follows:

$$SD_{Pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}} = \sqrt{\frac{(11.98^2 + 8.65^2)}{2}} = \sqrt{\frac{143.52 + 74.82}{2}} = \sqrt{\frac{218.34}{2}} = \sqrt{}109.17 = 10.45$$

Once the pooled standard deviation is calculated, it is then used to calculate Cohen's d by dividing the difference between the means of the two groups by the pooled standard deviation. We can enter our pooled SD into the earlier formula for a difference score, $d = \frac{M_1 - M_2}{SD_{Pooled}} = \frac{45.7 - 42.1}{10.45} = 0.34$.

> A positive value of Cohen's d indicates that the first group mean is greater than the second group mean and vice versa. Cohen's d values of around 0.2, 0.5, and 0.8 are commonly used to indicate small, medium, and large effects respectively. Out difference score of $d = .34$ implies a *small to medium* effect. The *size* of the effect corresponds with the practical significance of the difference.

Cohen's d is a useful measure of effect size because it allows researchers to compare the size of the effect across studies, regardless of the specific units of measurement used.

Running R

Please go through the following resources to help you setup R if you haven't already done so:

- Set up an RStudio cloud account (https://www.youtube.com/watch?v=uK1Va_UWQFc)
  - Familiarize yourself with the *Console* and *Global Environment*
  - All executable code are typed into the *Console* (bottom left)
  - Stored variables appear in the *Global environment* (top right)
- Introduction to R (https://www.youtube.com/watch?v=SWxoJqTqo08)
  - A free and powerful data analytics platform
- Once you've familiarized yourself with the RStudio console, try some basic operations (https://www.youtube.com/watch?v=9v3fy04pDho)
  - The code in the video is run on an older console, but the operations are the same. Try and replicate them in the console of your RStudio workspace.
- Once you have setup `R`, you can complete this week's Teaching Evaluation (TE).

Teaching Evaluation: Basic R operations

1. Install R *or* open a RStudio account
2. Assign the following values into a variable called `data1`
   18, 20, 14, 15, 7, 5, 7, 14, 9, 16
3. Assign the following values into a variable called `data2`
   20, 18, 23, 18, 16, 22, 18, 10, 11, 15
4. Using the `R` console, estimate the Mean and Standard Deviation for `data1` and `data2`
5. Calculate the Mean and Standard Deviation for both variables manually (use the formulas in slides 6 and 9) and **show each step**.
6. Report your outcomes using the format illustrated in slide 11. Remember to provide means and SDs for *both* `data1` and `data2`.

Submit your finished report into the TE5 dropbox.