**Dependent T-test**

This file is designed to act as a walkthrough in how to manage, analyze, and visualize a dataset for comparison between two groups. The dataframe is a publicly available dataset from AddHealth's longitudinal study in Adolescent Development. For this exercise we will be using Wave 1 & Wave 4's data to test the hypothesis that weight changes over time.

Extraction, Transformation, Loading (ETL):

* The two dataframes used for this project have already been reduced to Weight, ID, and Sex
* Our first step is joining the dataframes using an all join function (as opposed to left join or right join) this will create NAs for areas where there is no data
* As a result of the merge function columns with the same name were given a .x and .y column name. So, we’re going to rename them to Wave 1 and 4
* Using the codebook values, we’ll change all non-answer values to NA and convert the weight fields to integers
* We will be removing the SEX column because it isn’t part of our hypothesis
* We’ll also be making a new column named “diff” which represents the difference between wave 1 and wave 4 weights
* As an additional note, for visualization purposes we will also be transforming the dataframe from having multiple columns for weight to having a single column for weight and a column for time period

Data Cleaning:

* NAs were removed
* While not a statistical assumption of dependent t-test, we are going to remove the univariate outliers of our weights columns because the codebook shows there are some extreme outliers which may be recording errors.

Statistical Assumptions:

1. The Dependent Variable is Continuous and the Independent Variable is Categorical
2. There is no relationship or overlap between the Independent Variable categories
3. Difference variable between conditions is normally distributed among participants

In our dataset, the first statistical assumption is met because we are looking at how changes in time period lead to changes in weight. If your Dependent Variable is not truly intervallic (such as in Likert data or ranking systems) you can use the non-parametric Wilcoxon signed rank test.

Each wave is a different period in time and therefore does not have the possibility of multiple classification. Thusly, we can also say that the second statistical assumption has been met.

The third requirement can be checked using the descriptives function with the diff column we made. Liberal cut-offs for Skew and kurtosis are -3.00 to 3.00 and -10.00 to 10.00 respectively. If you'd like something a little more conservative, Curran West and Finch (1996) recommend -2.00 to 2.00 and -7.00 to 7.00. You can also visually verify this by looking at the histograms and seeing if they look like they are roughly centered at the highest point with symmetrical decreases from there out (like a bell). If you find that skew or kurtosis exceed these values, you'll need to remove univariate outliers to make sure they aren't pulling your data.

Outcome and interpretation:

With all of our assumptions met, we are able to conduct the paired t-test. As expected, there is a significant difference in weight between Wave1 and Wave4, t(4805) = -93.6, *p* <.001, Cohen’s d = -1.35. The difference in weight averages between the waves is 41lbs.

Visualization:

Like the independent t-test, code for visualizing this as a bar graph has been included. Additional code has been included for creating line graphs for participants by ID.

**Important Note**: While it can be tempting to conduct multiple t-tests on data (for example comparing variables across one categorical IV and then by another) the likelihood of falsely seeing a change when there is none increase with each test. This is known as inflated alpha error or familywise error.