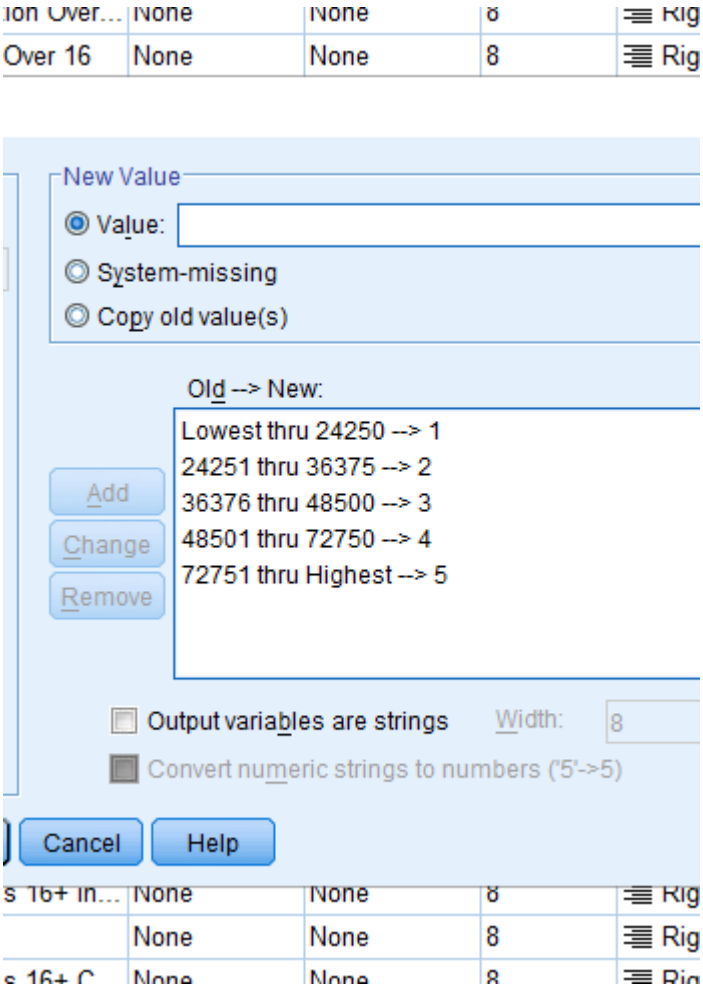


BASIC ANALYSIS IN SPSS

Exploring data, ANOVA, and T-Tests

ABSTRACT

In this tutorial, you will learn how to append data, decipher the characteristics of your dataset that help you best determine the type of test you would like to use, and finally run tests.

Ayanda Masilela
CSSCR Workshops (2018)

TO USE THIS DOCUMENT

The document is arranged in sections, which can be overviewed in the Table of Contents. It is subsequently broken down as an outline.

Descriptions and purposes are listed in black

Instructions for execution are highlighted in purple

Questions are highlighted in burnt orange

System dialogue will often be highlighted in blue

YOU WILL NEED

- SPSS 19 – this tutorial has not been tested on later generations of SPSS
- A basic understanding and vocabulary in statistics. I'll be walking through some conceptual basics, but the genuine mathematics behind each method and interpretation is beyond the scope of this tutorials

TABLE OF CONTENTS

GETTING STARTED

A. Download files
    a. <tinyurl.com/uwayatut>
        i. Navigate to "SPSS Tutorial Files"
        ii. Download the *SPSSAnaly.zip* file archive

B. Open the *EPAVehicles.sav* file
    a. Scroll to the right until you see the column "feScore"
        i. This is the *fuel economy score*. There are an awful lot of "-1" values in it
        ii. Run a Frequency analysis
            1. How many "-1" values were counted? _____
        iii. Open the *metadata.pdf* file (many variables have been removed to simplify this dataset)
        iv. What does a values of -1 mean in this dataset? _____
        v. We can correct this problem in the Variable window
            1. Scroll until you see "feScore" in the list, and then look to the right for the column that says "Missing"
            2. Click the […] button
            3. Click the radio button for Discrete Missing Values and type in "-1"
                a. Click **[OK]**
        vi. Run a Frequency analysis again
            1. Observe how now the "-1" values are now counted as "missing" in the output tables
C. Let's make sure to save your outputs, as well
    a. Click to your **Output** window
        i. Click **File → Save As…**
        ii. Navigate to the appropriate folder, and save the files as "VehiclesOutput.spv"
        iii. Don't forget to save this file periodically

APPENDING DATA

The dataset is fairly complete, but including more data on carbon dioxide emissions, greenhouse gas scores, and some other data may be interesting. To add variables, it will be necessary to "Append" one dataset to another. Fortunately, our data is coming from the same source, so it will be easy. We need to find a "Cardinal" or "Key" variable of some kind to link them together.

A. Open the Variable view in your SPSS window
B. Open the *vehiclesappend.csv* file
    a. Which variable do you suspect will make a good cardinal? _____
C. Click **Data → Merge Files → Add Variables**
    a. This new window will prompt you to select a dataset
        i. SPSS can only merge other SPSS data files, so it will be necessary to import the *vehiclesappend.csv* file
        ii. Close the window for now
        iii. Refer back to the *Intro to SPSS* packet for guidance on how to import files. If you are having trouble, feel free to use the *vehiclesappend_1.sav* file from the archive
        iv. When you have finished the import process, save the file as *vehiclesappend.csv*
D. According to the metadata, -1 is a common indicator for missing data. Make the "Missing" adjustment for each of the relevant variables
E. In both of your datasets, right-click the column header "id" and select Sort Ascending – **This is necessary for the merge to work**
F. In your *EPAVehicles.sav* window, **Click Data → Merge Files → Add Variables**
    a. We chose "add variables" because we are adding columns to the dataset that correspond to already existing cases
    b. The alternative, "add cases", would add case events from a similarly structured dataset, increasing our overall number of samples
        i. Adding 400 new cases to our already existing 40,000, totaling in 40,400 cases overall
    c. If you left the "vehiclesappend.sav" window open, it should appear in the list below for the An open dataset option
        i. Click it and then press **[Continue]**
        ii. Fill the checkbox for Match cases on key variables in sorted files
        iii. For Key Variables, add "id"
        iv. Click the radio button for Both datasets provide cases
            1. Click **[OK]**
        v. Data from *EPAVehicles.sav* should now be appended to the right of the data in *vehiclesappend.sav*

1. Save this new combined dataset as *FuelEcon.sav*

BASIC DATA ANALYSIS - One-Way ANOVA, One-Sample T-Test, Independent Samples T-Test, Paired Samples T-Test

This tutorial is not intended to be a substitute for your statistics class. I will be reviewing some basics that will satisfy the needs of working within software, but there are many other mechanics taking place beneath the software interface that with proper instruction can give you a better understanding of what the results are conveying.

These four analytical techniques are all seeking to decipher whether or not there is a statistically significant amount of variability between the means of the groups being compared. Each is suitable for specific situations…

- ANOVA – able to compare the means of multiple groups simultaneously
  - "Is the mean highway mileage of cars from 1979, 1989, 1999, and 2009 similar? Is there at least one difference within the group means?"

- One-Sample T-Test – compares a sample group to a known population mean
  - "The mean MPG of electric cars in 2012 was 85.44. I posit that the mean CityMPG of electric cars from 2018 has improved."

- Independent Samples T-Test – compares the means of two separate groups
  - "I would like to compare the combined gas mileage for all Manual 5-Speed cars vs. Automatic 4-Speed cars"

- Paired Samples T-Test– dependent variables being compared share a similar origin
  - "I would like to compare city gas mileage vs. highway gas mileage for each car manufactured in 2002"

Don't panic if this doesn't make sense yet. As we walk through each methodology, the logic of how these questions are structured and solved should become more clear.

ONE-WAY ANOVA – Exploring Data and Managing Outliers

In order to use One-Way ANOVA as an analysis technique, certain assumptions must be satisfied.

1. Independence of cases: each sample is an independent event, and cases do not directly influence each other
2. Normality: sample means are normally distributed. Alternatively each subdivision has a minimum of 30 samples (see Central Limit Theorem)
3. Homoscedasticity: variance among the testing groups is equal

We start with a question: "Do drive mechanisms (front-wheel drive, rear-wheel drive, etc.) have an influence on city fuel economy?"

Now we need to define a *null hypothesis*, or a starting declaration…

H_0 = drive mechanisms have no influence on city fuel economy

Next, we propose a counter-argument, a *test hypothesis*…

H_1 = drive mechanisms do have an influence on fuel economy

You may be ask yourself… "why did we start this hypothesis sequence with a declaration that seems to contradict the purpose of my investigation?"

How we structure the question can dictate our approach to answering it. As you continue through your in-person stats class, you will have opportunities to answer differently structured questions with additional information available for this particular question, the null hypothesis helps us start with an assumption…

H_0 makes the claim that "all drive systems have the same city fuel economy". It's okay if the claim seems unrealistic to you, but it is necessary to define it clearly.

H_1 makes the counter-claim, that there are differences in fuel economy between drive systems. At the conclusion of this test, we will be looking for evidence that there is, in fact, differences in fuel economy between drive systems

Our goal is to REJECT the null hypothesis.

A. First, we'll need to recode the data into a format that SPSS can interpret – it doesn't like to tally up string values, so we'll have to rename the drive characteristics as numbers
    a. **Analyze → Descriptive Statistics → Frequencies**

i. Run a Frequency analysis on the "drive" variable
ii. Your output table should yield 7 distinct categories
    1. That's a lot of typing for recoding, but there is a work-around
    2. In your output window, right-click the table and click **Export**
        a. Make sure that the radio button for Selected under Objects to Export has been chosen
        b. Leave the Type as "*.doc"
        c. Use the **Browse** button to direct the file to a desirable location, and name it *DriveTable.doc*
        d. Click **[OK]**
        e. Find and open the file, and you should now see a selectable copy of your output table

b. In your "FuelEcon.sav" window, click **Transform→Recode into Different Variables**
    i. You already know how to recode from the previous tutorial, but this may take some time. Feel free to tell your instructor if you need more time
    ii. Copy and paste from the table in the DriveTable.doc word document into the recoding window. When you are satisfied, complete the recode.
    iii. Feel free to adjust the **Values** entry in the Variable View. This will make your results window easier to interpret

B. Next, we'll need to see how well the data fits the assumptions of ANOVA
    a. Click **Analyze → Descriptive Statistics → Frequencies**
        i. Add "CityMPG" and "HwyMPG" to the Variables list
            1. Click the **Statistics** buttion
                a. Make sure that **Skewness** and **Kurtosis** are selected
                b. Click **[Continue]**
            2. Click the **Charts** button
                a. Select the radio button for **Histogram**
                    i. Click the checkbox for Show normal curve on histogram
                    ii. Click [**Continue**], then **[OK]** to run the analysis
            3. For skewness, value of between -1 and 1 indicates that the dataset is not skewed in either direction
                a. What are the skewness values for CityMPG and HwyMPG?

                _____
            4. Kurtosis measures whether the data is "flat" or "peaked". A good Kurtosis score is less than 3 times the standard error.
                a. What is the standard error of CityMPG and HwyMPG?

                _____

b. What is the Kurtosis value, and what does it indicate?

_____

5. Scroll down to the histograms
    a. Both datasets seem to be concentrated between 0 and 50, but there are values that are well in excess of this general range
    b. From here, it is worthwhile to explore whether or not some of these values are outliers and, if so, remove the and reevaluate our dataset for skewness

b. Click **Analyze → Descriptive Statistics → Descriptives**
    i. Add "CityMPG" and "HwyMPG" to the Variables box
    ii. Fill the checkbox for the save standardized values as variables dialog
    iii. Click **[OK]**
c. Click **Variable** view
    i. Click the header for the corresponding new "CityMPG" variable and call it "ZCityMPG"
    ii. Click the header for the corresponding new "HwyMPG" variable and call it call it "ZHwyMPG
d. Right-click the header for "CityMPG" and select Sort Ascending
    i. This function has calculated a "Z-Score". In a nutshell, Z-Scores tell us how distant a value is from the mean. The scale of this measurement is the standard deviation.
    ii. The conventional estimation for where 99.9% of the data lies is within 3.29 standard deviations – anything beyond these bounds is an outlier

Let's take a look at the outputs...

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CityMPG	40508	6	150	18.26	7.501
Valid N (listwise)	40508				

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HwyMPG	40508	9	123	24.40	7.456
Valid N (listwise)	40508				

The mean for CityMPG is 18.26 with a standard deviation of 7.501. The outlier cutoff is a Z-Score of 3.29. We can translate the Z-Score into actual miles per gallon values simply by multiplying the standard deviation by some easy, whole intervals[1].

Standard Deviations **Below the Mean**				**Mean (City MPG)**	Standard Deviations **Above the Mean**			
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st		Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
-11.74	-4.243	3.26	10.76	18.26	25.76	33.26	40.76	42.94
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st	**Mean (Hwy MPG)**	Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
-4.166	2.032	9.488	16.94	24.40	31.86	39.31	46.77	48.93

The table above is telling us that 99.9% of all automobiles in our City Miles per Gallon sample have a fuel economy of between -11.74 MPG and 42.94 MPG. Obviously, we can't fuel a car to go negative distances, so in real-world terms, 99.9% of cars get between 0 MPG and 42.94 MPG. Any car that gets gas mileage greater than 42.94 MPG is an outlier.

[1] In a real stats class, you will discuss standard deviation cutoffs as Z-Scores with an absolute value of 1.96 being inclusive of 95% of data, 2.58 inclusive of 99% of data, and 3.29 as 99.9% of data. At this stage for illustrative purposes, using the outlier cutoff and simple intervals within is the easiest option. This will be addressed in a later section.

e. Click **Variable** view, and observe the Z-Score for the CityMPG results sorted ascending
   i. What was the lowest Z-Score? What is the year, make, and model of that car? _____
   ii. Right-click the column header for CityMPG and Sort Descending. What is the highest Z-Score? What is the year, make, and model of that car?
   _____

"WTF does the Z-Score mean in less abstract terms?"

Bear with me. We're going to break down the Z score formula based on the examples we've been working with above.

Z-Score = [(sample mean) – (population mean)] / (standard deviation)

Your statistics textbook probably says…

$$z = \frac{x - \mu}{\sigma}$$

Let's loosely define some terms in the context of this question…

- **Z-Score (*z*):** The number of standard deviations from the mean
- **Sample Mean (*x*):** The value of a specific case (Such as the BMW M6 entry)
- **Population Mean (μ):** The mean of the CityMPG across the entire dataset
- **Standard Deviation (σ):** An interval representing variation within the dataset as it relates to the *Population Mean*

Next, we'll examine an entry from the list…

f. Right-click the column header for **Model** and click Sort Descending
   i. In line 1, select the row header to highlight the **2006 Lincoln Zephyr**
   ii. What information do we have about this car?
      1. Z-score for City MPG = -0.1679
      2. Population Mean (CityMPG) = 18.26
      3. Standard Deviation = 7.501
      4. But we're missing the **Sample Mean**… let's solve for it!

Plug it into the formula!

$$-0.1679 = (x - 18.26)/7.501$$

…and solve…

$$-0.1679(7.501) = x - 18.26$$

$$-1.259 = x - 18.26$$

$$-1.259 + 18.26 = x$$

$$17.00 = x$$

Great, so the Sample Mean (x) is equal to 17. Now what?

g. Click **Data View**, and scroll so that you can see the "CityMPG" column
   i. What is the CityMPG listed for the 2006 Lincoln Zephyr?f

_____

Congratulations, you are now a mathmagician! Let's take a moment to break down how this all fits together.

Z-Score = [(sample mean) – (population mean)] / (standard deviation)

$$-0.1679 = (x - 18.26)/7.501$$

- The negative Z-Score tells us that the **Sample Mean** will be lower than the **Population Mean**
- Since the Z-Score has an absolute value of less than 1, we know that the **Sample Mean** is less than one **Standard Deviation** in distance from the **Population Mean**

Again, think of **Standard Deviation** as an interval. The distance of one interval unit is 7.501.

As we continue to solve…

$$-0.1679(7.501) = x - 18.26$$

The value of -1.259 is the proportion of the Standard Deviation interval unit. Since 1 Standard Deviation is equal to 7.501, -0.1679 (or 16.79%) of the Standard Deviation is equal to -1.259.

$$-1.259 = x - 18.26$$

$$17.00 = x$$

The value of -1.259 represents "to what extent less than the mean" is the actual value of the sample mean. Thus 18.26 but 1.259 fewer MPG is the fuel economy of the 2006 Lincoln Zephyr.

Therefore, the **Sample Mean** = 17.

For a visual, let's return to the Z-Score table from Page 8. We know the CityMPG (x) of the 2006 Lincoln Zephyr is 17. I've marked an approximate location of where it falls on the scale below.

Standard Deviations **Below the Mean**				**Mean (City MPG)**	Standard Deviations **Above the Mean**			
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st		Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
-11.74	-4.243	3.26	10.76	18.26	25.76	33.26	40.76	42.94
				x 17.0				
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st	**Mean (Hwy MPG)**	Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
-4.166	2.032	9.488	16.94	24.40	31.86	39.31	46.77	48.93

The Zephyr's city fuel economy is definitely lower than the mean, but is still less than one standard deviation away from the overall mean.

h. Look at your results for the questions on Page 10, Section e., Parts i and ii.
   i. Place an "x" on the table above for each answer based on the Z-Score
   ii. Calculate the sample mean based on the Z-Score
      1. When you are done, check the sample value listed in the dataset. Do they match?

i. Walk through the same steps as above to calculate the sample mean of the 2006 Lincoln Zephyr's Highway Fuel Economy (HwyMPG). Do your results match what appears in the data table? Use the space provided below.

j. Go back to the **Data** view and find the entry with the highest Z-Score.
   i. What is that value? What is its CityMPG? _____

FILTERING OUTLIERS

There are a few ways to accomplish this task. Some tutorials will advise you to generate a new variable with missing values and use the new variable for calculations. Others, such as this one, will recommend "Selecting" out cases and applying "Filters". Generating filters does indeed generate a new variable, but the variable is considered temporary until the filter is removed.

A. Click **Data → Select From Cases**
    a. Click the radio button for If condition is satisfied
    b. Click the [If] button
        i. Add "ZCityMPG" to the formula box
        ii. Click the "<=" button
        iii. Type 3.29
            1. This will select all cases where ZCityMPG is less than or equal to 3.29
            2. Since there were no lower outliers, we can be flexible with range
            3. Click [**Continue**], then **[OK]**
    c. Sort the ZCityMPG column Descending
        i. Look to the left at the row headers. How many cases have been excluded?
B. Click **Analyze → Descriptive Statistics → Descriptives**
    a. Add "CityMPG" to the variables box
    b. Make sure that "HwyMPG" **is *not*** in the variables box
        i. Click the **Statistics** button
            1. Make sure that **Skewness** and **Kurtosis** are selected
            2. Click **[Continue]**
        ii. Click the **Charts** button
            1. Select the radio button for **Histogram**
                a. Click the checkbox for Show normal curve on histogram
                b. Click **[Continue]**, then **[OK]** to run the analysis
    c. What is the new valid count for the total number of cases? _____
    d. What is the new mean after applying the filter? _____
    e. What is the new Skewness value? _____
    f. What is the new Kurtosis value? _____

Overall, much improved after removing the outliers. Skewness and Kurtosis have been greatly reduced. While not perfect-perfect, this is about as close as we'll get to meeting the assumptions of ANOVA, especially considering that real-world data is rarely ever as cooperative as we would like. Fortunately, ANOVA testing is somewhat robust (ask your stats teacher about robustness) when dealing with non-normal or steep data. We take these steps to ensure the most correct usage of the ANOVA method, and solve problems to the best of our ability.

ONE-WAY ANOVA (for real, this time)

A. Click **Analyze → Compare Means → One-Way ANOVA**
    a. Add "CityMPG" to the Dependent List
    b. Add "Drive_Codes" to the Factor
    c. Click the **[Options]** button
        i. Fil the check boxes for Descriptive, Brown-Forsythe, Welch, Homogeneity of Variences Test, and Means Plot
        ii. Click **[Continue]**, then **[OK]**

B. The results are in!

ANOVA

CityMPG

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	283181.818	6	47196.970	3333.795	.000
Within Groups	552496.120	39026	14.157		
Total	835677.938	39032			

    a. Looking at the descriptives table, there does appear to be some difference in means between drive types, but are they statistically significant?
        i. Scroll to the first ANOVA table. The significance is 0.000… that indicates that **we can reject the null hypothesis** – there is indeed a difference in fuel economy based on the drive system of the car

C. *"What were those other options that we selected?"* – The theoretical specifics of these tests far exceed the scope of this tutorial, but they are still important. The robustness tests are a second and third set of eyes for assessing your dataset. Ask your professor or someone at CSSS for the details.
    a. **Test of Homogeneity of Variances** – This screening seeks to discover whether differences in values between subset groups of the data are similar. Basically, does the Z-Score of Front-Wheel Drive cars scale equally the Z-Score of Rear-Wheel Drive Cars , and do these Z-Scores also scale equally the Z-Score of Four-Wheel Drive cars… etc.
        i. The significance score of 0.000 indicates that **NO**, the Z-Scores of each subset are not equal to each other – there is a statistically significant difference in variance
        ii. This is often the case in real-world examples, especially if our subsets have differing sample sizes. It would be nice if things were so neat, and

that our datasets satisfied all assumptions exactly, but this time we don't have the luxury. It's okay though!

Test of Homogeneity of Variances

CityMPG

Levene Statistic	df1	df2	Sig.
265.527	6	39026	.000

b. **Welch's Robustness Test** – Welch's accounts for heterogeneous variances, as well as heterogeneous sample sizes.
    i. **The verdict:** 0.000, Welch finds that there is a statistically significant amount of variability among means in the dataset
c. **Brown-Forsythe's Robustness Test** – Offers additional screening for datasets that are not normally distributed
    i. **The Verdict:** 0.000, Brown-Forsythe finds that there is a statistically significant amount of variability among means in the dataset
d. **Means Plot:** This is a diagram depicting the distributions of the means of each category. Feel free to browse and observe contrasts

Robust Tests of Equality of Means

CityMPG

	Statistic[a]	df1	df2	Sig.
Welch	3036.126	6	2182.450	.000
Brown-Forsythe	3729.025	6	5920.014	.000

a. Asymptotically F distributed.

We were lucky this time that ANOVA, Welch's, and Brown-Forsythe all indicated statistically significant differences in the means. That doesn't always happen. Again, what would cause a hypothesis to succeed on a basic One-Way ANOVA and fail a robustness test is beyond the scope of this tutorial. Nonetheless, when you produce your own outputs, be sure to include these tests and consult with an experienced researcher as you interpret these results.

At this stage, I'll leave you with some questions to experiment with…

After evaluating the collection of 293 automobiles that were excluded as outliers, I noticed that (unsurprisingly) many of them were purely electric cars. At the beginning of the tutorial, we recoded fuel classifications base on the "fuelType1" column. Follow these steps and rerun the analysis that we did above and see if anything changes.

1. Filter out all electric vehicles
2. Explore descriptive statistics using "fuelType_Gen1"
3. Note that all gasoline classifications, whether regular, midgrade, or premium, were simplified to all gasoline - Feel free to recode again to separate the grades and run further analyses
4. Calculate Z-Scores
5. What is the CityMPG that constitutes outlier status?
    a. Hint: this corresponds to any case entry with a Z-Score absolute value of 3.29
6. Create a new filter that excludes electric vehicles AND excludes outliers
    a. Use the syntax: *fuelTypeGen1 ~= 4 AND ZSco01 <= 3.29*
7. Run a One-Way ANOVA test that evaluates whether drive type has an impact on fuel economy
    a. Use CityMPG and Drive_Codes just like last time
    b. Include all of the same robustness tests

T-TESTS and THEIR ASSUMPTIONS (PARAMETERS)

Just like ANOVA, T-Tests also have several assumptions that must be met.

1. That the data is either scalar or interval in nature
2. That the sample data has been randomly collected
3. That the data is normally distributed
4. That the sample size is large enough – ideally no fewer than 30
5. Homogeneity of Variance

Sounds simple enough. With a few minor adjustments, such as filtering away outliers, we should meet these criteria without any problems.

BASIC DATA ANALYSIS - One-Sample T-Test

"The mean MPG of electric cars in from 1984-2019 was 100.98. I posit that the mean CityMPG of Honda electric cars is greater than that of the overall population in the same period."

With this question, we have a comparative mean to another population with similar qualities to the sample we are interested in. Knowing that, we can test our One Sample, Honda electric cars, to the population mean of electric cars from all years[2].

H_0 = Honda electric cars **do not** have a different fuel economy than electric cars from the broader electric car population from 1984-2019

H_1 = Honda electric cars **do** have a different fuel economy than electric cars from 1984-2019

A. Recode your data so that Honda is equal to 1
    a. Utilize the "make" column to complete this task
    b. Name the new variable "Honda"
B. Click Data → Select Cases
    a. Click the [Reset] button
    b. Click the If condition is satisfied radio button
    c. Click the [If] button
        i. Generate the syntax that will select fuelTypeGen1 = 4 AND Honda = 1
        ii. Click **[Continue]**, then **[OK]**
C. Click **Analyze → Compare Means → One-Sample T Test**
    a. Set the Test Value to "100.98"

[2] Honda was selected due to its long history of participating in the electric car market.

b. Add "CityMPG" to the **Test Variables** box
c. In the **[Options]** window, leave the confidence interval at 95%
d. Click **[OK]**

Let's look at the results…

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
CityMPG	6	102.33	41.399	16.901

One-Sample Test

	Test Value = 100.98					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
CityMPG	.080	5	.939	1.353	-42.09	44.80

Honda electric cars appear to have a higher mean CityMPG than the population mean (102.33 versus 100.98), but the Standard Deviation is quite broad.

Look at the "Sig. 2-tailed", that's our P-Value, and it is well in excess of 0.05. That high Significance Values (P-Value) indicates that there is no statistically significant difference between the population mean of all electric cars and Honda electric cars.

Thus, we fail to reject the null hypothesis.

BASIC DATA ANALYSIS – Independent Samples T-Test (aka Two-Sample T-Test)

"I would like to compare the combined gas mileage for all Manual 5-Speed cars vs. Automatic 4-Speed cars"

This test is quite similar to the One Sample T-Test. In the One-Sample T-Test, we tested the mean unknown mean of one sample (Hondas) versus the known mean of all electric cars from the years of 1984-2019.

In a Two-Sample T-Test, we are testing two means that we don't know. We'll try to answer the question above, do Manual 5-Speed cars get better gas mileage than Automatic 4-speed cars?

H_0: Manual 5-Speeds and Automatic 4-Speeds have equal gas mileage

$H_{1:}$ Manual 5-Speeds have a different gas mileage from Automatic 4-Speeds

A. Data → Select Cases
    a. Click [Reset] to clear any previous filters
B. First we'll need to recode transmission types
    a. Analyze → Descriptive Statistics → Frequencies
        i. Add "transmission" to the **Variables** box
        ii. Make sure that display frequency tables is checked
        iii. Click **[OK]**
    b. Looks like we have a lot of results. Luckily we only need to look at two of them – feel free to copy the gold text below when recoding
        i. Automatic 4-spd
        ii. Manual 5-spd
    c. In the recode window, assign the following values and click [Change]:
        i. Name: "Trans_Code"
        ii. Label: "Auto or Manual"
    d. In the [Old and New Values] window
        i. "1" to Automatic 4-spd
        ii. "2" to Manual 5-spd
        iii. "0" to All other values
    e. In the **Variable View**, change the **Values** accordingly
        i. 1 = Automatic
        ii. 2 = Manual
        iii. 0 = Other
    f. Here we will try a different variation in removing data we don't want to assess

i. In the neighboring **Missing** column, click the corresponding cell to your new variable and press the […] button
ii. Click the Discrete missing values radio button
iii. Add "0" to the first box
iv. Click **[OK]**

C. **Analyze → Compare Means → Independent Samples T-Test**
   a. Add "CombMPG" to the **Test Variables** box
   b. Add "Trans_Code" to the **Grouping Variable** box
   c. In the **[Options]** window, leave the confidence interval at 95%
   d. Click **[Define Groups]**
      i. Assign "1" to Group 1
      ii. Assign "2" to Group 2
      iii. Click **[OK]**

D. Let's compare means…
   a. Automatic Mean? _____ Standard Deviation? _____
   b. Manual Mean? _____ Standard Deviation? _____
   c. **The Verdict:** the **Significance Value** of 0.000 indicates that we **REJECT** the null hypothesis

BASIC DATA ANALYSIS – Paired-Samples T-Test

"Do cars get better gas mileage in cities or on highways?"

This final assessment compares the means of two characteristics, but these characteristics are tied together in some way. In the case of fuel economy, we are looking at each car individually, and then comparing the characteristic of highway MPG versus city MPG.

H_0: HwyMPG = CityMPG

H_1: HwyMPG ≠ CityMPG

We'll play with another method for cutting up our datasets…

A. **Data → Split Dataset**
    a. Click the radio button for Compare Groups
        i. Add "Year" to the Groups Based on: box
        ii. Leave all the other options as they are, and click **[OK]**
B. Now filter out the outliers (this will be helpful for clarifying the final section)
    a. **Data → Select Cases**
        i. Add the following syntax to Cases If: ZCityMPG <= 3.29 AND Z_HwyMPT <= 3.29

This assessment is going to provide comparative results for every single year, so processing my take a while.

C. **Analyze → Compare Means → Paired-Samples T-Test**
    a. Add "City MPG" to Variable 1
    b. Add "HwyMPG" to Variable 2
    c. Leave the confidence interval at 95%
    d. Click **[OK]**
D. Scroll through the results, compare means – in all cases, highway mileage is higher than city mileage, thus with Significance of 0.000 for each year, we can REJECT the Null Hypothesis.

UNDERSTANDING CONFIDENCE INTERVALS and SIGNIFICANCE VALUES (aka P-Values)

CONFIDENCE INTERVALS

I left this section for the end since I dumped so much other conceptual information at the front of this tutorial. I'll dig a little deeper into this concept in an additional tutorial that discusses Z-Scores in relation to significance vales (P-Values in your stats textbook).

A **confidence interval** is an indicator of where we can infer that our sample means will lie. Conventionally 95% confidence interval us used, and has been the measure for this tutorial.

Think back to when we were trying to decipher outliers in the ANOVA section. From our assessment, we determined that 99.9% of automobiles in our dataset fell within 3.29 standard deviations. The remaining 0.01% of our cars, those with fuel economies above 43 MPG, did not represent 99.9% of our data.

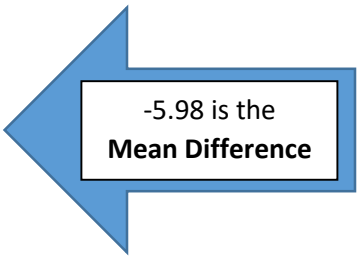
In other words… *"We can say with 99.9% confidence that the cars in this sample have a fuel economy of less than 43 MPG. Anything beyond that probably represents and anomaly in our data."*

A **confidence interval** applies the same standards for inclusion or exclusion from the norm, but instead looks at differences in means between the two groups that we are comparing. With the 95% confidence interval, we declare a boundary ahead of completing the whole test by declaring…

"95% of our mean difference values fall between this lower boundary and that upper boundary."

- The interval we calculated for determining outliers in our dataset applied to a **single** sampling group
- The **confidence interval** we calculate for our T-Tests is applied to a situation where we are evaluating the means of **multiple groups**
    - Furthermore, this can be best understood by comparing the plausible range of mean differences

	CityMPG	HwyMPG	Difference
	19	25	-6
	9	14	-5
	23	33	-10
	10	12	-2
	17	23	-6
	21	24	-3
	22	29	-7
	23	26	-3
	23	31	-8
	23	30	-7
	23	30	-7
	18	26	-8
	21	29	-8
	18	26	-8
Mean	19.29	25.57	-5.98



Every T-Test follows this basic structure of calculating the mean difference between groups. Though the sample in the above figure is tiny in comparison to our dataset of 42,000, the basic structure is the same. From the Independent-Samples T-Test, we calculated the mean difference of (Manual 5-Speed) – (Automatic 4-speed).

Just like the groups we sampled, the **Mean Difference** also has a **Standard Deviation**.

You may have breezed over a footnote on Page 10 discussing this topic. reiterate, in a two-tailed test like ours, a Z-Score of 1.96 is a marker for where 95% of our sampled data will lie within our model. This will help us calculate the appropriate interval.

Let's take a look at the output tables from the Paired-Samples T-Test.

Group Statistics

	trans_code	N	Mean	Std. Deviation	Std. Error Mean
CombMPG	Automatic	11045	18.07	3.738	.036
	Manual	8355	21.74	5.315	.058

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
CombMPG	Equal variances assumed	683.043	.000	-56.510	19398	.000	-3.675	.065	-3.803	-3.548
	Equal variances not assumed			-53.920	14265.374	.000	-3.675	.068	-3.809	-3.542

The output gives us a lot of information. Fortunately, it also provides us the value of our lower and upper confidence intervals. We can quickly decipher it based on the upper or lower confidence interval in relation to the mean difference.

We'll take the absolute value of the upper confidence interval and subtract it from the absolute value of the mean difference.

3.675 - 3.548 = Standard Deviation

0.127 = Standard Deviation

In a nutshell, this tells us that…

"In 95% of comparisons, Automatic vehicles will get between 3.548 and 3.803 fewer miles per gallon than Manual vehicles."

Based on this range, the mean of Automatic vehicles will always be lower than that of Manual vehicles – there is no possibility of overlap.

"What does an outcome look like if we don't find a statistically significant difference?"

Let's look at an analysis comparing CityMPG from 1997 and 1998 that found no statistical significance.

Group Statistics

	year	N	Mean	Std. Deviation	Std. Error Mean
CityMPG	1997	762	17.14	4.021	.146
	1998	810	17.03	3.942	.139

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
CityMPG	Equal variances assumed	.658	.417	.519	1570	.604	.104	.201	-.290	.498
	Equal variances not assumed			.519	1559.798	.604	.104	.201	-.290	.499

The lower confidence interval indicates that there is a possibility of a 1997 sample mean being lower than that of 1998. The upper confidence interval indicates that there is **also** a possibility of a 1997 sample mean being higher than that of 1998.

"In 95% of comparisons, 1997 vehicles will get between 0.290 fewer and .498 greater miles per gallon than Manual vehicles."

Based on this this range, the mean of 1997 vehicles **could be** lower than that of 1998, and it also **could be** higher. There is a possibility of overlap. Thus, there is no statistically significant difference in the mean CityMPG of 1997 and 1998 vehicles.

SIGNIFICANCE VALUES (P-VALUES)

A **Significance Value** is the standard with which we measure the fitness of our **Test Hypothesis (H_0)**. It is the converse of the confidence interval. In any analysis, there are two **Significance Values**

- **Given Significance Value** – the standard that the analysis sets for confirmation
- **Derived Significance Value** – the result that we calculate in the process of verifying our claim

For example…

IF Confidence Interval = 95% (0.95), THEN Given Significance Value = 5% (0.5)

IF Confidence Interval = 99% (0.99), THEN Given Significance Value = 1% (0.1)

For our analysis to confirm the **Test Hypothesis** we want our calculated **Significance Value** to be **LESS THAN** the **Significance Value *given*** to us as the standard in the analysis.

In the analysis of Manual vs. Automatic cars, the **GIVEN Significance Value** was 0.05, based on our decision to keep the **Confidence Interval** at 0.95. The **DERIVED Significance Value** was 0.000. Therefore, we can confirm a statistically significant difference.

In the analysis of 1997 vs. 1998 cars, the **GIVEN Significance Value** was 0.05. The **DERIVED Significance Vale** was .417. Therefore, we can confirm there **is not** a statistically significant difference.

TL, DR: All you need to know for now is that your **Derived Significance Value** must be lower than the **Given Significance Value** in order to confirm your **Test Hypothesis**.

VISUALIZING OUTCOMES

If you've made it this far, great job! Keep practicing. Much of this section tries to explain statistics without actually doing any math. I'll be rendering a more math-oriented update in the near future.

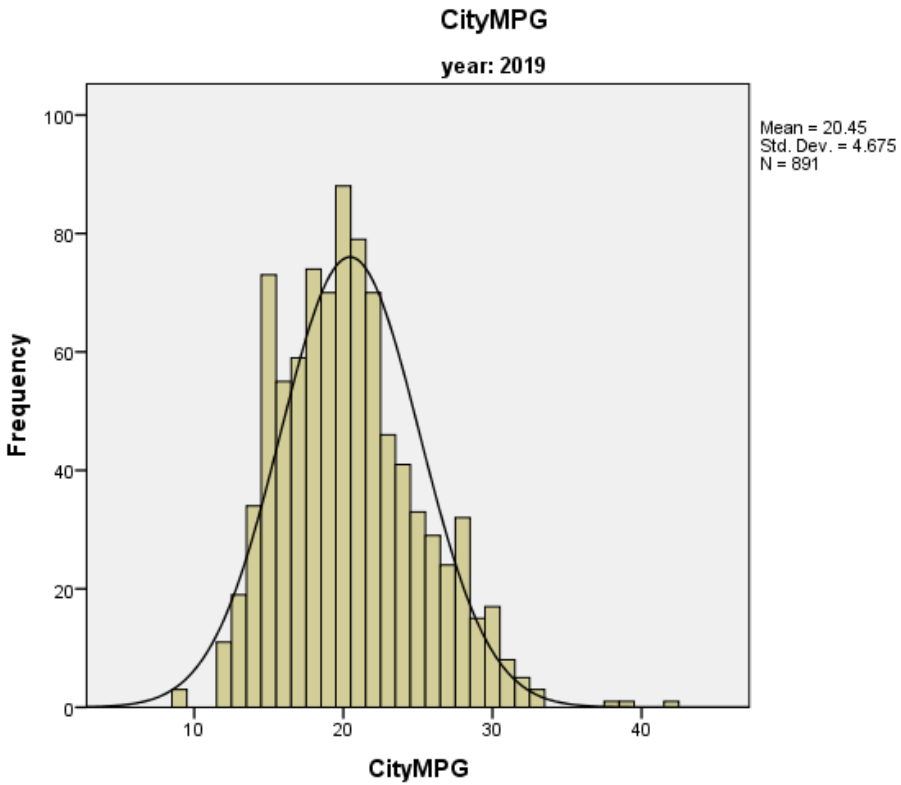
Let's look at 2019 Specifically. We'll use the MINI Cooper S Hardtop 4 door…

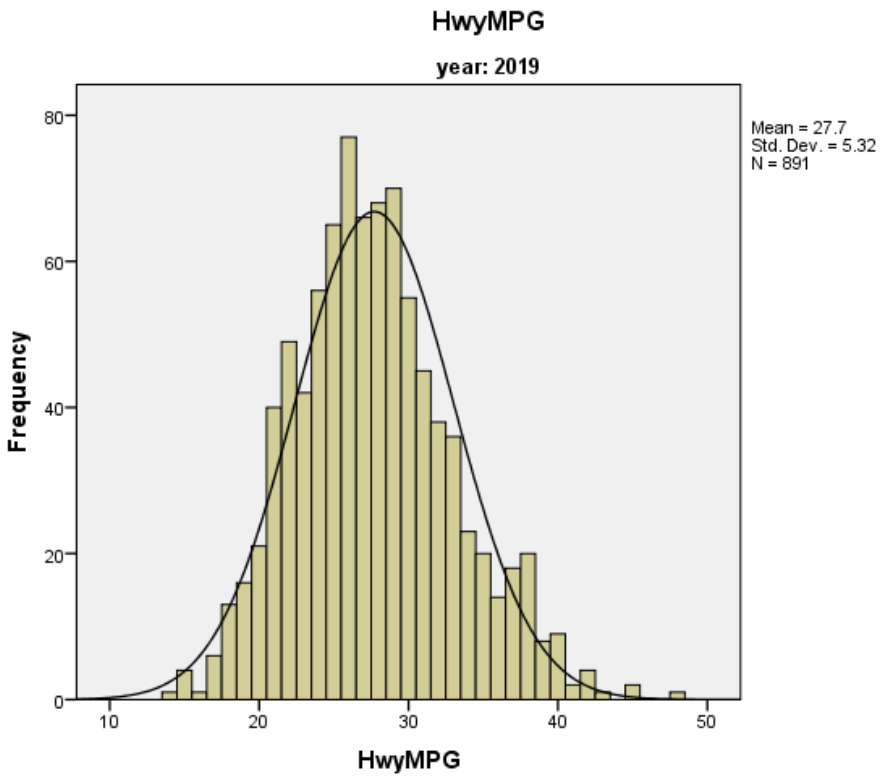
Standard Deviations **Below the Mean**				**Mean (City MPG)**	Standard Deviations **Above the Mean**			
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st		Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
5.069	6.425	11.1	15.78	20.45	25.125	29.8	34.473	35.83
				X 23				
Z = -3.29 Lower Outlier	Z = -3 3rd	Z = -2 2nd	Z = -1 1st	**Mean (Hwy MPG)**	Z = 1 1st	Z = 2 2nd	Z = 3 3rd	Z = 3.29 Upper Outlier
10.2	11.74	17.06	22.38	27.7	33.02	38.34	43.66	45.20
					X 32			

The MINI Cooper seems to be performing slightly above the population mean in both cases. It also follows the general trend that Highway MPG tends to be higher than City MPG.

From our Significance Value, we can also infer that 95% of the cars in our analysis will fall within these comparative norms, with the mean Highway MPG being higher than the Mean City MPG.

Take a look at these actual values first and how they fall within the calculated curve.





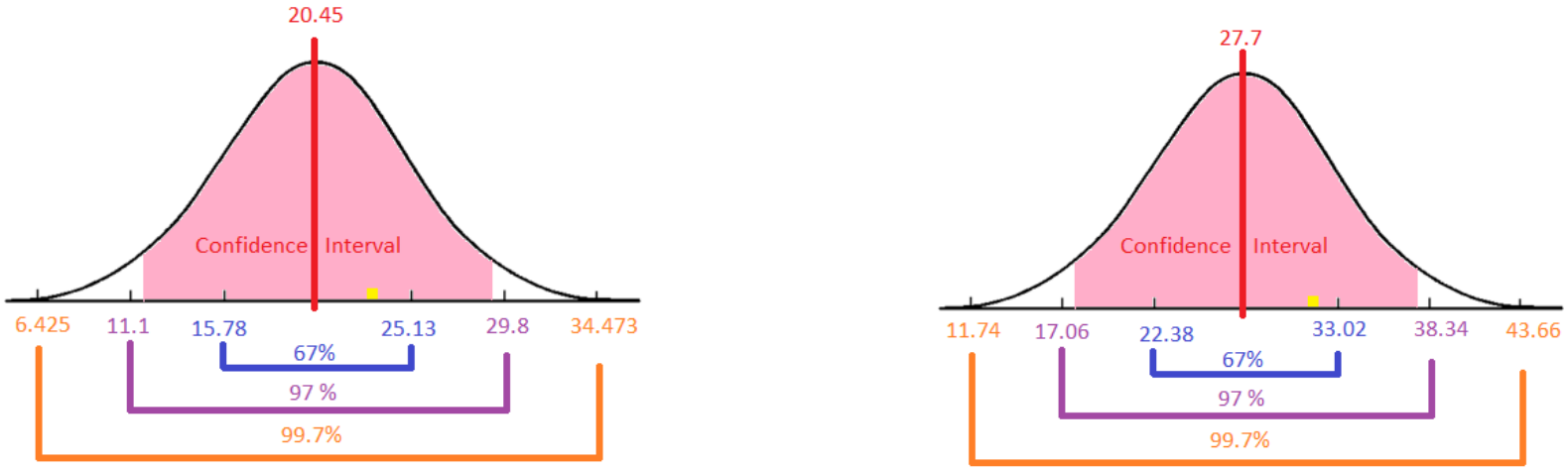
As you can see, the results fall more or less under a normal curve, with a small selection falling outside of the estimations.

This is where things get a little abstract. These curves represent the most likely range of values within each compared group (City MPG vs Highway MPG). Note how these intervals match the tables we have been working with throughout the lesson.

City MPG

The yellow notch represents the MINI Cooper

Highway MPG



Look at the pink regions – these denote the 95% confidence interval – where 95% of our data points are anticipated to be. In the *Confidence Intervals* section above, we noted that 95% confidence falls at 1.96 standard deviations from the mean. It is very close to 2 standard deviations, but is indeed a narrower range, as the depicted curves show.

You've probably heard your professor talk about **Predictive Modeling or Inferential Statistics**. At its most basic, predictive modeling is calculating a theoretical range of likely outcomes. The pink curves directly above are a smoothed representation of the somewhat jagged histograms that display real-world data. Looking at the jagged histograms, we can eyeball where the bulk of the samples land. An analysis of distribution ascribes a numerical value to these distributions, and defines how they relate to one-another.

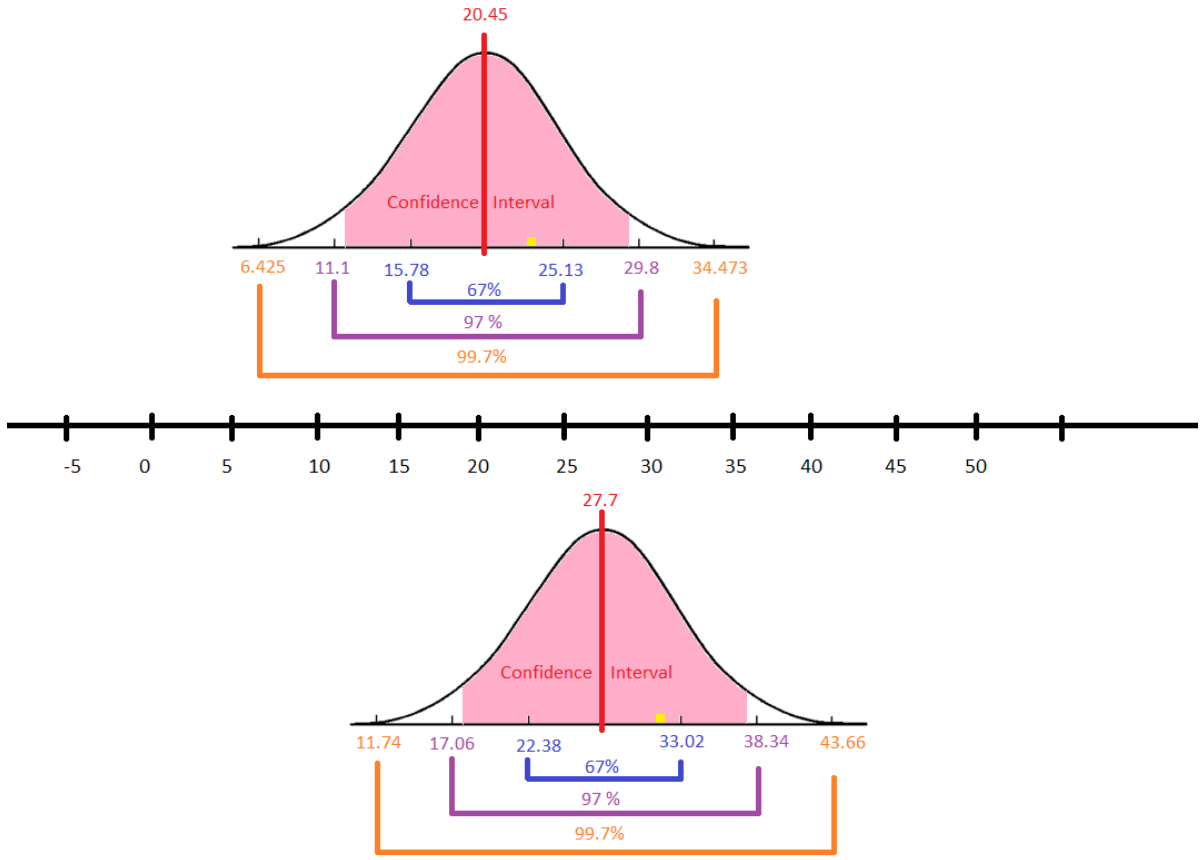
A T-Test simply takes two separate analyses of distributions and compares them. It's an analysis of two analyses. The outcome is a few numerical details that can tell us how two different datasets compare to each other.

The first half of the analysis, deciphering distributions, investigates **variation within** the two datasets. The second half of the analysis, comparing the distributions of the two datasets, investigates **variation** between the datasets. How similar or different are these distributions?

The T-Test achieves this by putting them on a scale. With this scale, we can compare the means. And distributions.

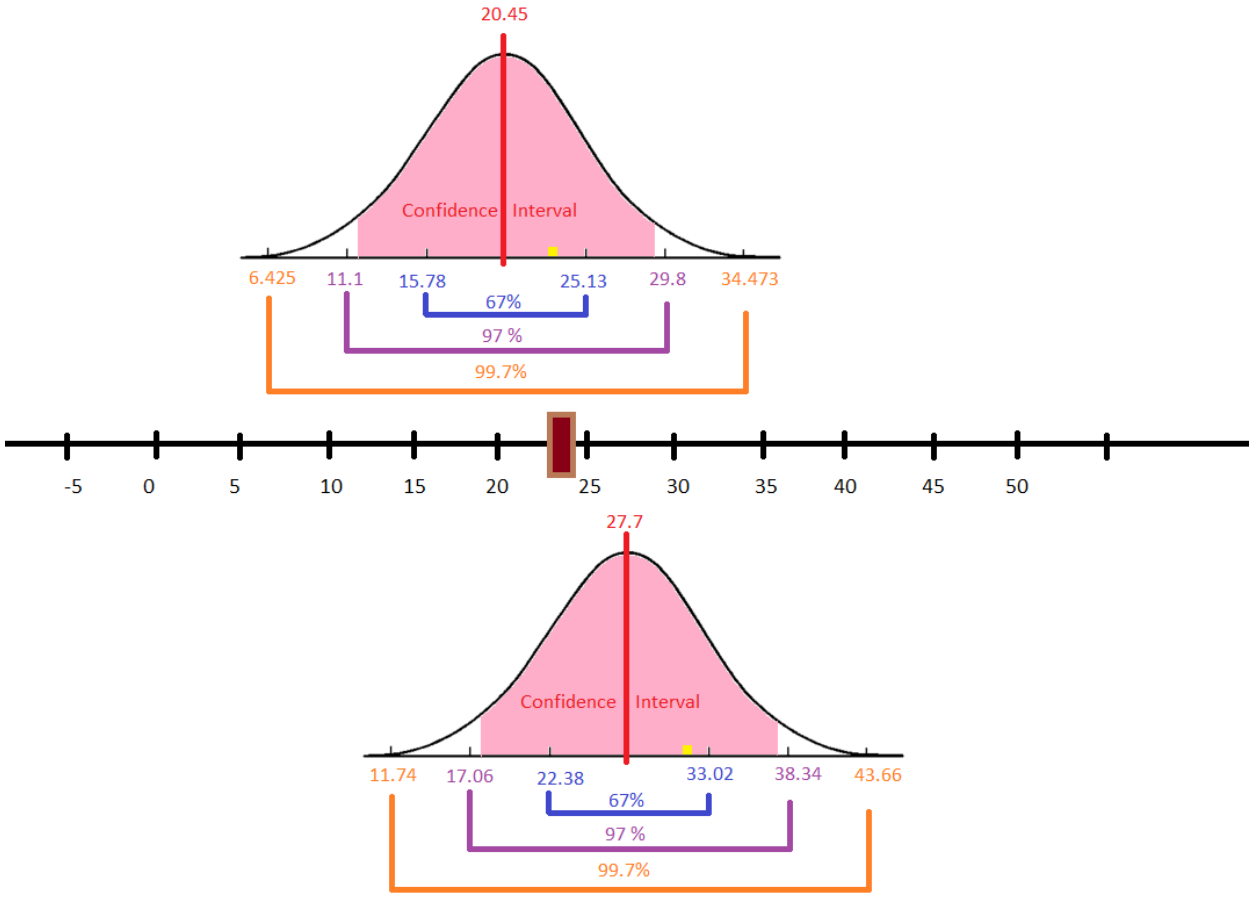
Looking at the means alone, there appears to be a difference. There is also some difference in the standard deviation of each subset.

We can infer with 95% confidence that the mean highway fuel economy will be greater than that of city fuel economy.



The next step evaluates the range of means.

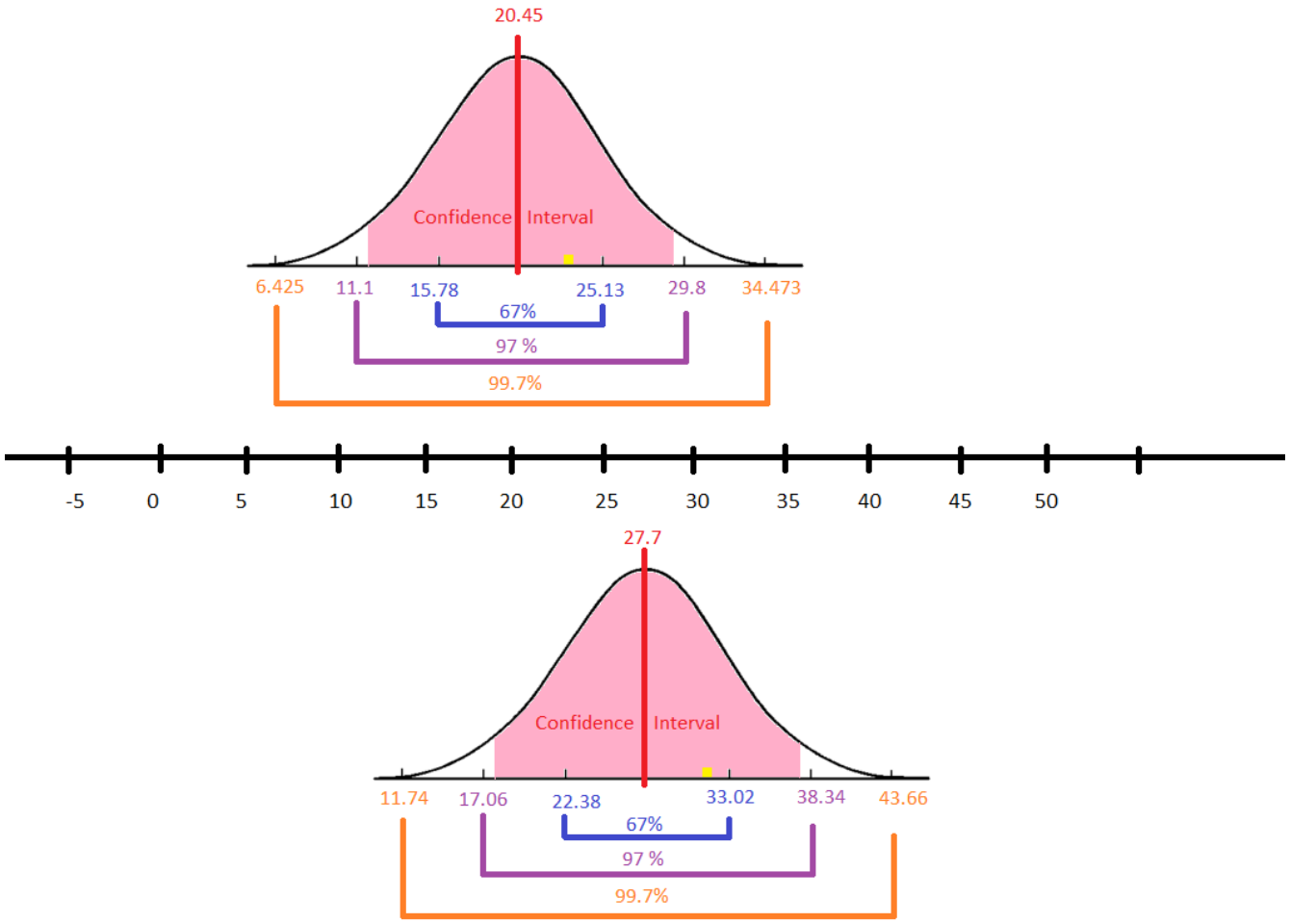
In the analysis, we concluded that the difference of means had a 95% confidence interval of being within 3.548 and 3.803 miles per gallon below the mean of Manual Cars. This is represented in the brown box. There is no possibility of overlap given such a contrast in the means of the two subgroups.



What if mean values are very close? Can there still be statistical significance?

Yes. The answer lies in the standard deviation.

Let's refer back to the 2019 dataset and how it relates to the mean scale. This dataset doesn't lend itself to demonstrating a close mean with disparate standard deviation scenario, so we'll rely on a strong contrast for explanation.



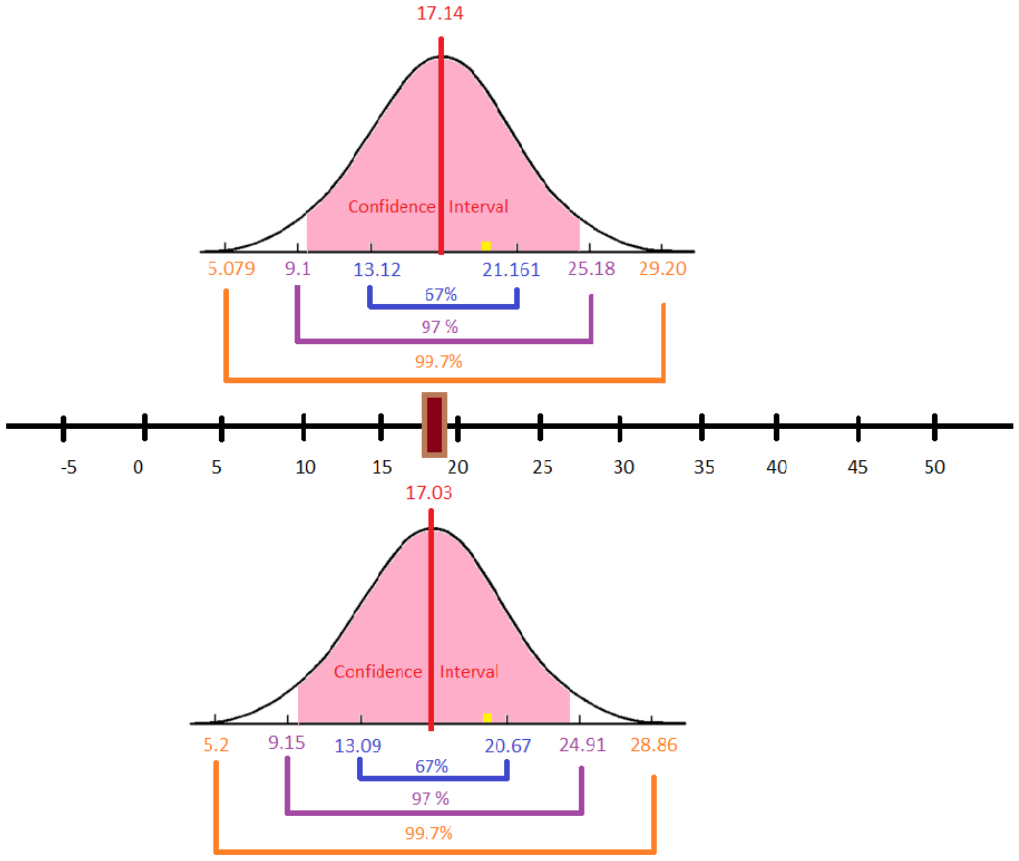
Below the pink curves, you'll see a blue, purple, and orange bar. Those are the standard deviations.

Those standard deviations operate as their own unit of intervals. The intervals derived are unique to each data subset. They are their own independent scale calculated on-the-fly based on the real-world values of the dataset. Spread and/or tightness of standard deviations contributes to statistically significant difference.

In this next case comparing City MPG in 1997 and 1998 (Independent Samples T-Test), no statistical significance was found.

1997: Mean = 17.14; Standard Deviation 4.021 1998: Mean = 17.03; Standard Deviation 3.942

Significance Value = 0.417



In this case, based on the means as well as the standard deviations being so similar, no statistically significant difference was found between these two datasets.

We concluded in the section that 1997 cars got between 0.290 fewer and 0.498 greater gas mileage than 1998 cars. While this measure is narrow, there is a possibility of overlap, therefore there is no statistical significance in means.