214A: Lab 7

TA: Melissa Gordon Wolf

Fall 2019

Goals for today

1. Descriptives & recoding variables
2. T-tests & confidence intervals
3. Compute effect sizes

- **Our research question**: Is income related to academic achievement?
- **Testable hypothesis**: Do students in poverty score lower on math tests?
- **Null hypothesis**: Students in poverty do not score differently on math tests than students who are not poverty.
- **Alternative hypothesis**: Students in poverty do score differently on math tests than students who are not poverty.

- **Independent/Grouping variable**: X1Poverty
- **Dependent/Outcome variable**: X1TXMSCR

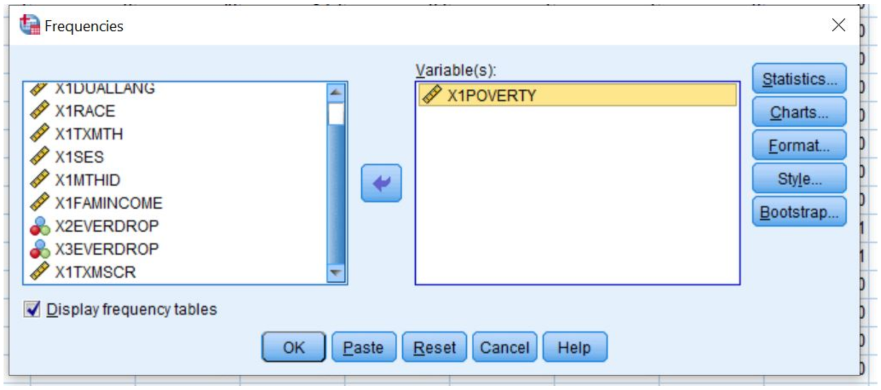
1. Descriptives & recoding variables

We always begin by investigating the variables we want to use in our analysis.

In SPSS

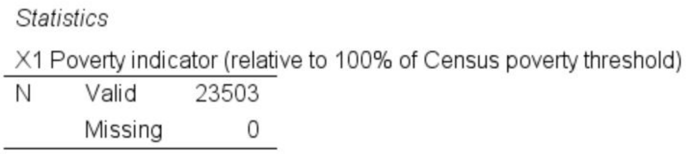
For categorical variables:

Analyze > Descriptive Statistics > Frequencies



Variable(s): X1POVERTY

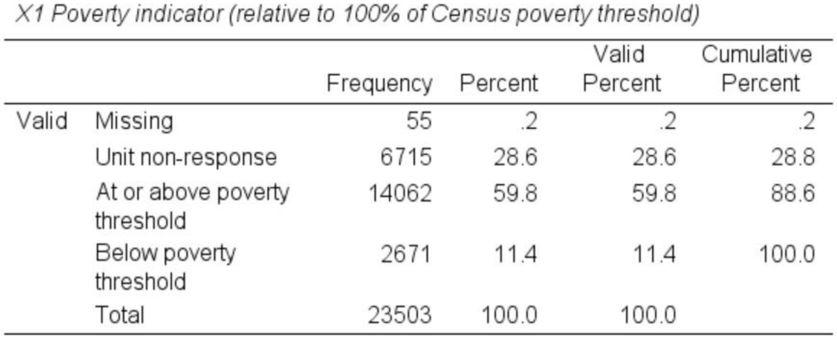
☑ Display frequency tables



Statistics

X1 Poverty indicator (relative to 100% of Census poverty threshold)

N	Valid	23503
	Missing	0

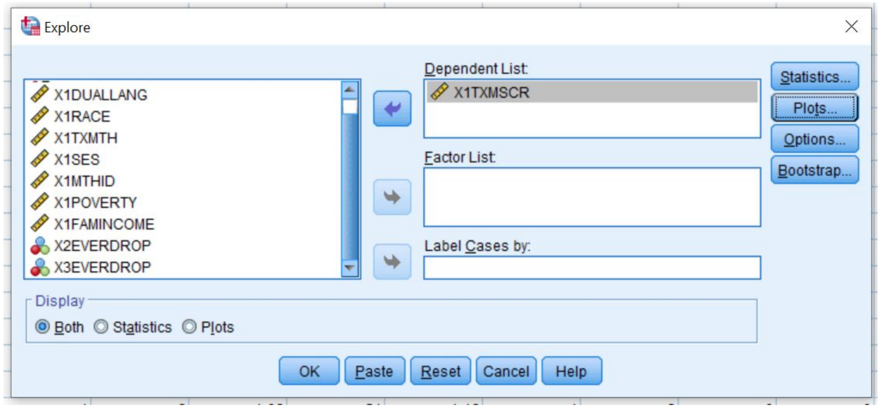


X1 Poverty indicator (relative to 100% of Census poverty threshold)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Missing	55	.2	.2	.2
	Unit non-response	6715	28.6	28.6	28.8
	At or above poverty threshold	14062	59.8	59.8	88.6
	Below poverty threshold	2671	11.4	11.4	100.0
	Total	23503	100.0	100.0	

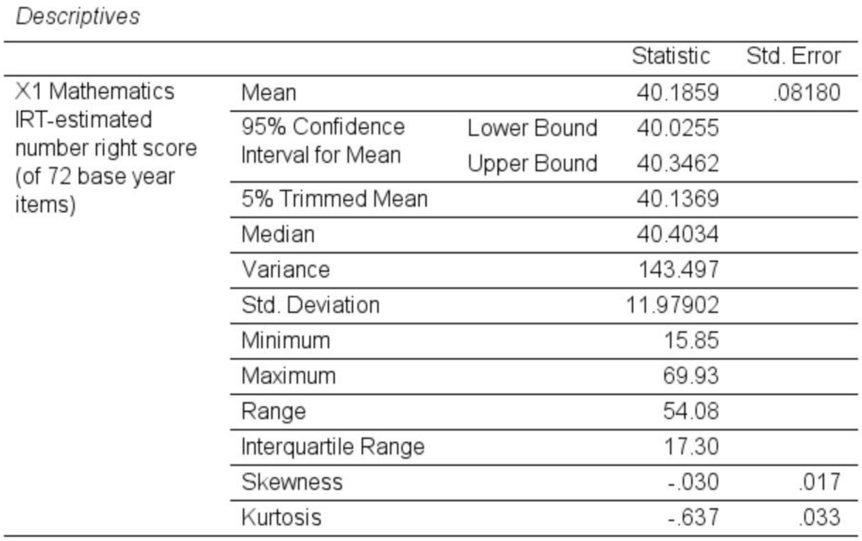
For continuous variables:

Analyze > Descriptive Statistics > Explore



Dependent List: X1TXMSCR

Display: ● Both ○ Statistics ○ Plots

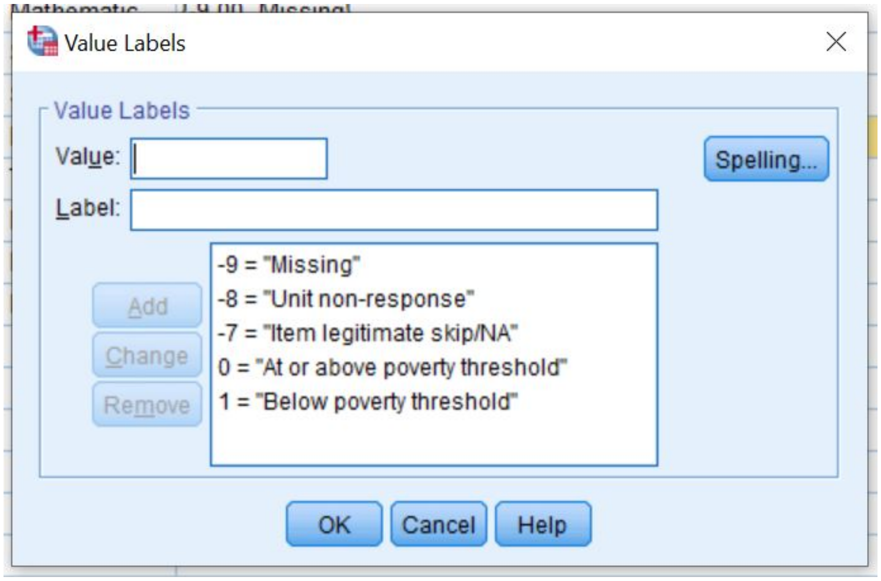


Descriptives

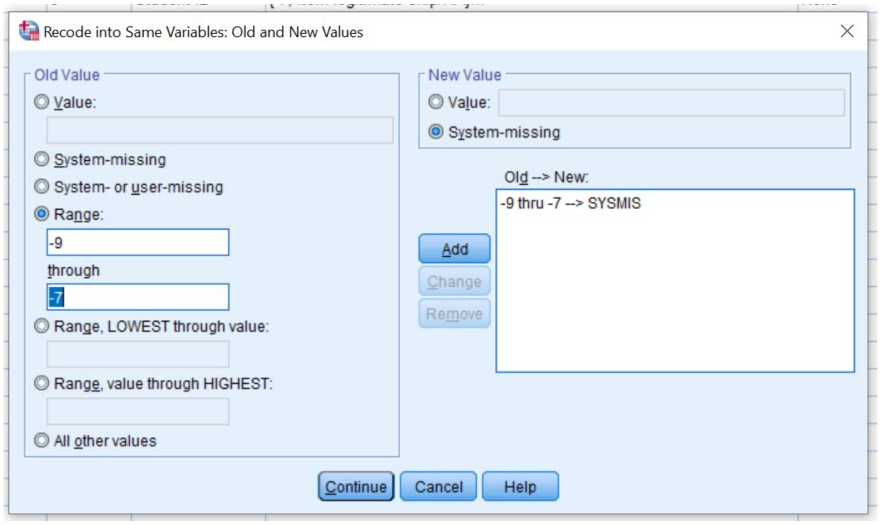
			Statistic	Std. Error
X1 Mathematics IRT-estimated number right score (of 72 base year items)	Mean		40.1859	.08180
	95% Confidence Interval for Mean	Lower Bound	40.0255	
		Upper Bound	40.3462	
	5% Trimmed Mean		40.1369	
	Median		40.4034	
	Variance		143.497	
	Std. Deviation		11.97902	
	Minimum		15.85	
	Maximum		69.93	
	Range		54.08	
	Interquartile Range		17.30	
	Skewness		-.030	.017
	Kurtosis		-.637	.033

We can see that we need to recode our categorical variable because we have a bunch of missing values that aren't correctly coded as missing.

Transform > Recode into Same Variables



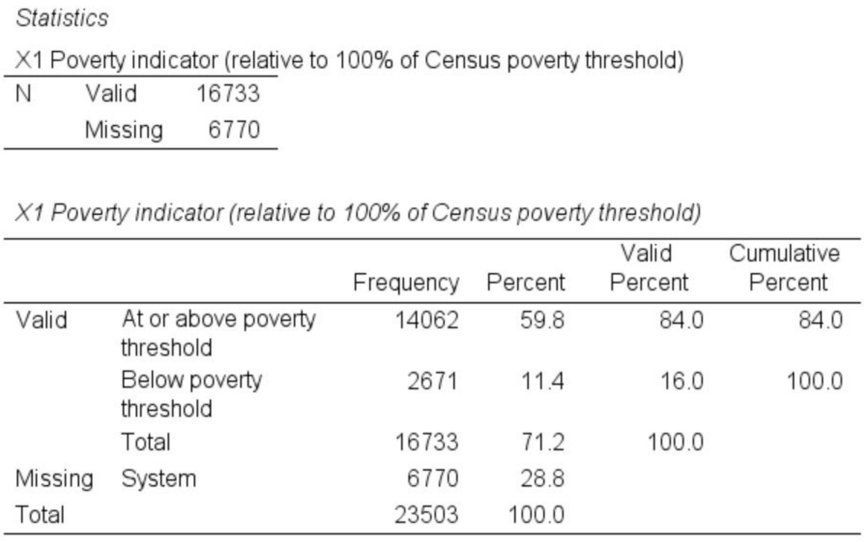
-9 = "Missing"
-8 = "Unit non-response"
-7 = "Item legitimate skip/NA"
0 = "At or above poverty threshold"
1 = "Below poverty threshold"


Recode into Same Variables: Old and New Values
Old Value
○ Value:
○ System-missing
○ System- or user-missing
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values
New Value
○ Value:
Add
Change
Remove
Continue
Cancel
Help

● Range: -9 through -7

● System-missing

Old --> New:
-9 thru -7 --> SYSMIS



Statistics

X1 Poverty indicator (relative to 100% of Census poverty threshold)

N	Valid	16733
	Missing	6770

X1 Poverty indicator (relative to 100% of Census poverty threshold)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	At or above poverty threshold	14062	59.8	84.0	84.0
	Below poverty threshold	2671	11.4	16.0	100.0
	Total	16733	71.2	100.0	
Missing	System	6770	28.8		
Total		23503	100.0		

In R

```
#Read in the data
library(haven)
lab8data <- read_sav("C:/Users/Melissa/Documents/UCSB/214/Lab 7/lab7data.sav")
df<-lab8data

#For categorical variables (sjmisc package)
frq(df$X1POVERTY)
```

```
##
## X1 Poverty indicator (relative to 100% of Census poverty threshold) (x) <numeric>
## # total N=23503  valid N=23503  mean=-2.19  sd=3.71
##
##  val                         label   frq raw.prc valid.prc cum.prc
##   -9                       Missing    55    0.23      0.23    0.23
##   -8             Unit non-response  6715   28.57     28.57   28.80
##   -7        Item legitimate skip/NA     0    0.00      0.00   28.80
##    0 At or above poverty threshold 14062   59.83     59.83   88.64
##    1       Below poverty threshold  2671   11.36     11.36  100.00
##   NA                          <NA>     0    0.00        NA      NA
```

```
#For continuous variables (psych package)
describe(df$X1TXMSCR)
```

```
##    vars     n  mean    sd median trimmed   mad   min   max range  skew kurtosis
## X1    1 21444 40.19 11.98   40.4   40.25 12.85 15.85 69.93 54.08 -0.03    -0.64
##       se
## X1 0.08
```

```
summary(df$X1TXMSCR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.85   31.79   40.40   40.19   49.09   69.93    2059
```

We want to make sure that we only use the 0 and 1 values from the poverty indicator in our analysis. To do this, let's recode -9 and -8 as NA, so that we can omit NA values in our dataset.

```
#sjmisc package
frq(df$poverty<-rec(df$X1POVERTY, rec="-9=NA;-8=NA;else=copy"))
```

```
##
## X1 Poverty indicator (relative to 100% of Census poverty threshold) (x) <numeric>
## # total N=23503  valid N=16733  mean=0.16  sd=0.37
##
##  val   frq raw.prc valid.prc cum.prc
##    0 14062   59.83     84.04   84.04
##    1  2671   11.36     15.96  100.00
##   NA  6770   28.80        NA      NA
```

```
#Check the dataset to see that we added the variable correctly
View(df)

#Add value labels to the new variable (Base R)
df$poverty<-factor(df$poverty,
                   levels=c(0,1),
                   labels=c("At or above poverty threshold",
                           "Below poverty threshold"))

#Check to see if the variable labels were added properly
frq(df$poverty)
```

```
##
## x <categorical>
## # total N=23503  valid N=16733  mean=1.16  sd=0.37
##
##                              val   frq raw.prc valid.prc cum.prc
##  At or above poverty threshold 14062   59.83     84.04   84.04
##        Below poverty threshold  2671   11.36     15.96  100.00
##                             <NA>  6770   28.80        NA      NA
```

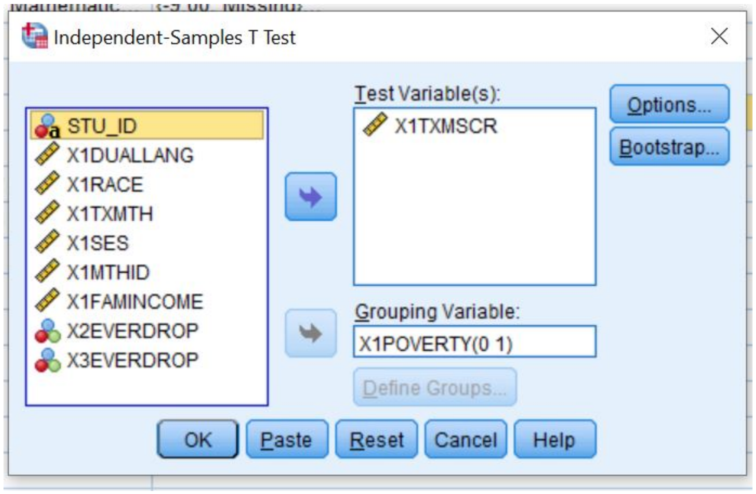
```
#Pro-tip: If you wanted to delete the variable you created, you could use "df$poverty<-NULL"
```

2. T-tests & confidence intervals

Let's run a t-test to see if these group means are statistically significantly different. In other words, is the average math test score statistically significantly different for students that are in poverty and students that are not in poverty?

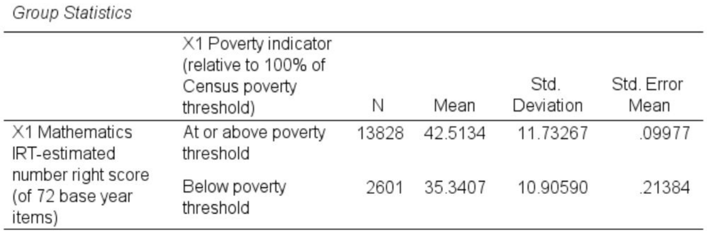
In SPSS

Analyze > Compare Means > Independent Samples T-test



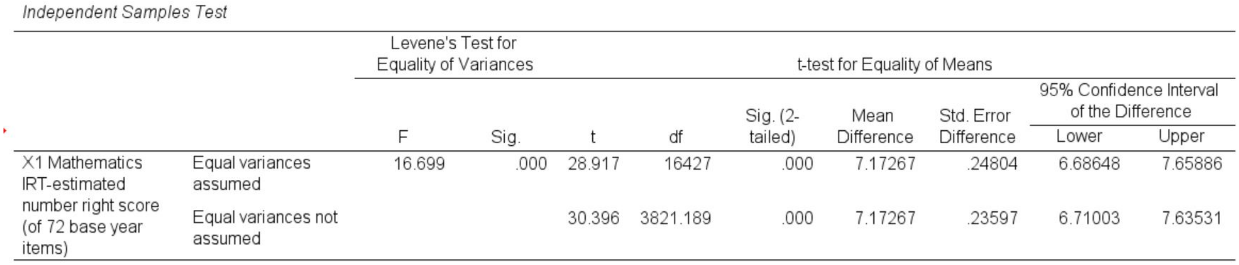
Test Variable(s): X1TXMSCR

Grouping Variable: X1POVERTY(0 1)



Group Statistics

	X1 Poverty indicator (relative to 100% of Census poverty threshold)	N	Mean	Std. Deviation	Std. Error Mean
X1 Mathematics IRT-estimated number right score (of 72 base year items)	At or above poverty threshold	13828	42.5134	11.73267	.09977
	Below poverty threshold	2601	35.3407	10.90590	.21384



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
X1 Mathematics IRT-estimated number right score (of 72 base year items)	Equal variances assumed	16.699	.000	28.917	16427	.000	7.17267	.24804	6.68648	7.65886
	Equal variances not assumed			30.396	3821.189	.000	7.17267	.23597	6.71003	7.63531

Quiz questions

(*Answer on Gauchospace*)

In R

```
#We can run a t-test in Base R
t.test(df$X1TXMSCR~df$poverty)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$X1TXMSCR by df$poverty
## t = 30.396, df = 3821.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.710027 7.635314
## sample estimates:
## mean in group At or above poverty threshold
##                                  42.51341
##        mean in group Below poverty threshold
##                                  35.34074
```

3. Compute effect sizes

We can see that the difference between groups is statistically significant, but let's see how meaningfully different it is by computing an effect size measure like Cohen's D.

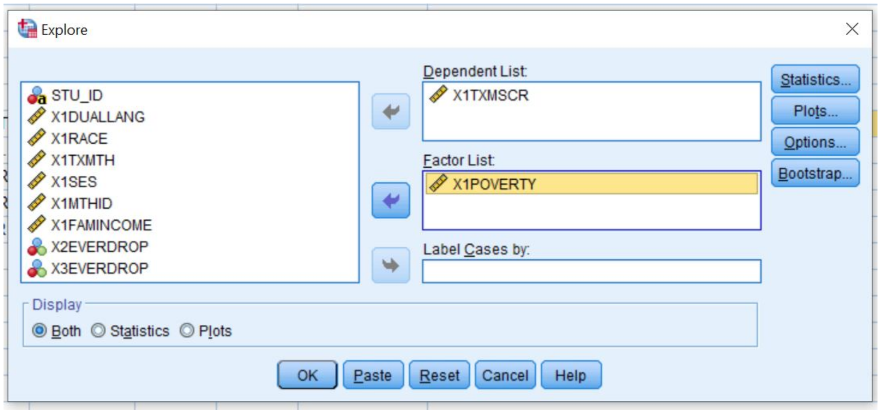
In SPSS

SPSS actually cannot give us an effect size measure. Thus, we have to do this in Excel.

First, we need the equation:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)*s_1^2+(n_2-1)*s_2^2}{n_1+n_2-2}}}$$

We see that we need the mean, standard deviation, and sample size from each group. Let's get this from Analyze > Descriptives > Explore:

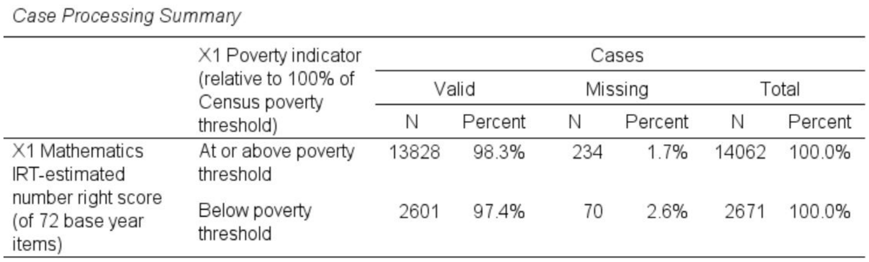


Dependent List: X1TXMSCR

Factor List: X1POVERTY

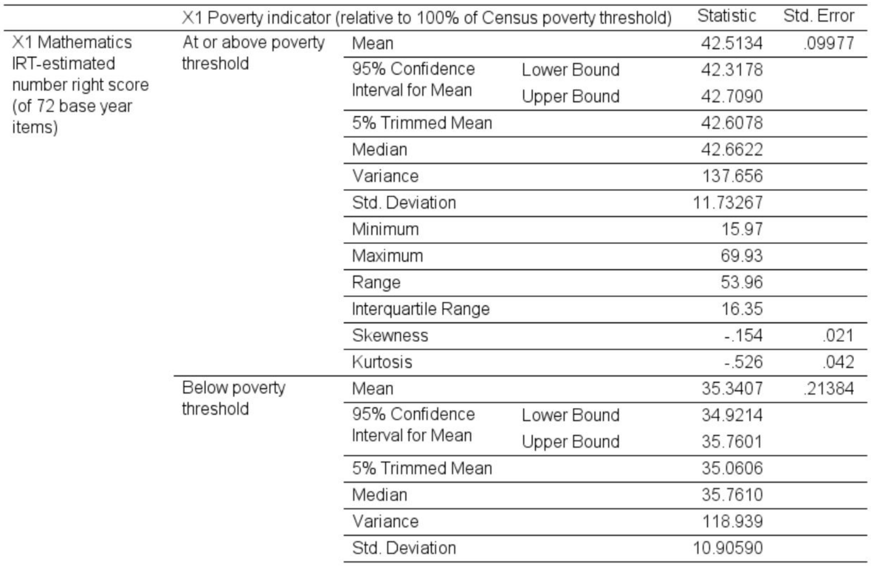
Label Cases by:

Display
● Both ○ Statistics ○ Plots



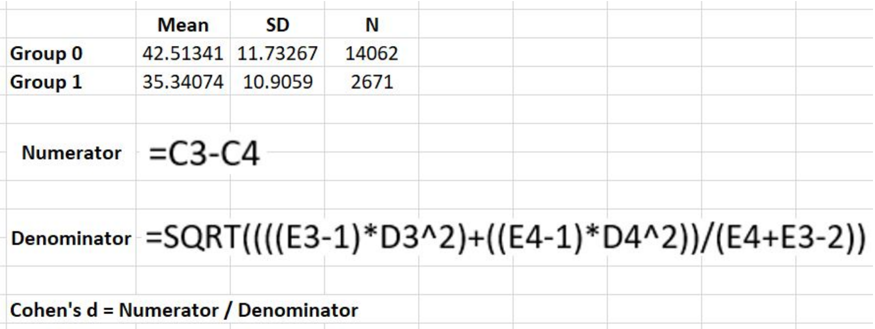
Case Processing Summary

	X1 Poverty indicator (relative to 100% of Census poverty threshold)	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
X1 Mathematics IRT-estimated number right score (of 72 base year items)	At or above poverty threshold	13828	98.3%	234	1.7%	14062	100.0%
	Below poverty threshold	2601	97.4%	70	2.6%	2671	100.0%



	X1 Poverty indicator (relative to 100% of Census poverty threshold)			Statistic	Std. Error
X1 Mathematics IRT-estimated number right score (of 72 base year items)	At or above poverty threshold	Mean		42.5134	.09977
		95% Confidence Interval for Mean	Lower Bound	42.3178	
			Upper Bound	42.7090	
		5% Trimmed Mean		42.6078	
		Median		42.6622	
		Variance		137.656	
		Std. Deviation		11.73267	
		Minimum		15.97	
		Maximum		69.93	
		Range		53.96	
		Interquartile Range		16.35	
		Skewness		-.154	.021
		Kurtosis		-.526	.042
	Below poverty threshold	Mean		35.3407	.21384
		95% Confidence Interval for Mean	Lower Bound	34.9214	
			Upper Bound	35.7601	
		5% Trimmed Mean		35.0606	
		Median		35.7610	
		Variance		118.939	
		Std. Deviation		10.90590	

Now, let's compute Cohen's D using the above equation. You can do this on a calculator, you can use Excel, or you can google it.



	Mean	SD	N
Group 0	42.51341	11.73267	14062
Group 1	35.34074	10.9059	2671
Numerator	=C3-C4		
Denominator	=SQRT((((E3-1)*D3^2)+((E4-1)*D4^2))/(E4+E3-2))		
Cohen's d = Numerator / Denominator			

Quiz question: What is the effect size?

(*Answer on Gauchospace*)

In R

To practice, let's calculate this statistic computationally and then ask R to replicate the results using a package.

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)*s_1^2+(n_2-1)*s_2^2}{n_1+n_2-2}}}$$

Begin by getting the mean, standard deviation, and sample size for each group.

```
#dplyr package

df%>%
  group_by(poverty)%>%
  summarise(mean=mean(X1TXMSCR,na.rm=TRUE),sd=sd(X1TXMSCR,na.rm=TRUE),n=n())
```

```
## Warning: Factor `poverty` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 3 x 4
##   poverty                          mean    sd     n
##   <fct>                           <dbl> <dbl> <int>
## 1 At or above poverty threshold    42.5  11.7 14062
## 2 Below poverty threshold          35.3  10.9  2671
## 3 <NA>                             36.3  11.4  6770
```

Next, let's save these values as objects and then use the objects to write out the equation.

```
x1=42.51341
x2=35.34074
sd1=11.73267
sd2=10.90590
n1=14062
n2=2671

numer=x1-x2
denom=(sqrt((((n1-1)*sd1^2)+((n2-1)*sd2^2))/(n1+n2-2)))
numer/denom
```

```
## [1] 0.6180842
```

That was a lot of work! Let's see if we can get a package to replicate that for us.

```
#There are quite a few packages that will give us Cohen's D

#lsr package
cohensD(df$X1TXMSCR~df$poverty)
```

```
## [1] 0.618028
```

```
#effsize package
#df$X1TXMSCR is numeric, but effsize doesn't recognize that because it isn't
#compatible with the haven package
class(df$X1TXMSCR)
```

```
## [1] "haven_labelled"
```

```
#Relabel it as numeric and use that variable
df$num <- as.numeric(df$X1TXMSCR)

#There are two packages that use the function cohen.d: psych and effsize. To
#tell R that we want it to use the effsize package, start with effsize:: and
#then type cohen.d.  You can think of this as typing "library::function".
effsize::cohen.d(df$num~df$poverty)
```

```
##
## Cohen's d
##
## d estimate: 0.618028 (medium)
## 95 percent confidence interval:
##     lower     upper
## 0.5756058 0.6604501
```

```
#We get the effect size, the confidence interval, and the magnitude.
```