214A: Lab 6

TA: Melissa Gordon Wolf

Fall 2019

Goals for today

1. Review: Data Screening
2. New: Compare means
3. New: Regression with one categorical predictor
4. New: Interpret p-values

Today, we're going to use two variables from the dataset:

- Predictor/Independent variable: Whether or not a math teacher took a college level applied math course (M1APPLIEDMTH)
- Outcome/Dependent variable: Time 1 math score (MTSCOR)

1. Review: Data Screening

We know that the indendent variable is categorical and the dependent variable is continuous. What are some methods that we could use to investigate each of these variables given their scale type?

In SPSS

Analyze > Descriptive Statistics > Frequencies

Frequencies

Statistics

M1 A14B Math teacher took college-level applied mathematics course(s)

N	Valid	17029
	Missing	6474

M1 A14B Math teacher took college-level applied mathematics course(s)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	10424	44.4	61.2	61.2
	Yes	6605	28.1	38.8	100.0
	Total	17029	72.5	100.0	
Missing	System	6474	27.5		
Total		23503	100.0		

Analyze > Descriptive Statistics > Explore

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
X1 Mathematics standardized score (time 1 math score)	21444	91.2%	2059	8.8%	23503	100.0%

Descriptives

			Statistic	Std. Error
X1 Mathematics standardized score (time 1 math score)	Mean		51.1096	.06882
	95% Confidence Interval for Mean	Lower Bound	50.9747	
		Upper Bound	51.2445	
	5% Trimmed Mean		51.1504	
	Median		50.9716	
	Variance		101.559	
	Std. Deviation		10.07767	
	Minimum		24.02	
	Maximum		82.19	
	Range		58.17	
	Interquartile Range		13.19	
	Skewness		-.061	.017
	Kurtosis		-.134	.033

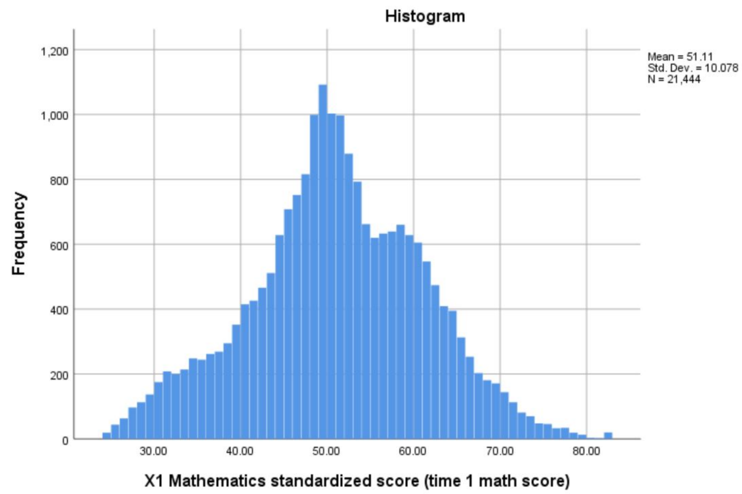
Percentiles

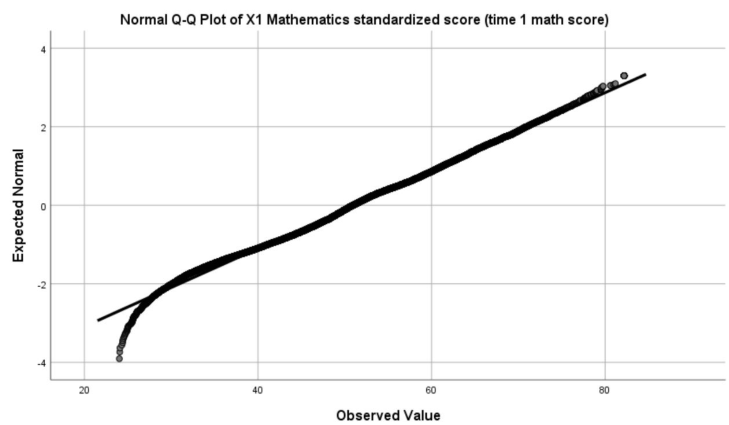
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	X1 Mathematics standardized score (time 1 math score)	33.1045	37.4729	44.9702	50.9716	58.1621	63.9912	67.5332
Tukey's Hinges	X1 Mathematics standardized score (time 1 math score)			44.9703	50.9716	58.1620		

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
X1 Mathematics standardized score (time 1 math score)	.024	21444	.000

a. Lilliefors Significance Correction





In R

I like to use the frq command from the sjmisc package because it gives me the value labels, counts, and percentages, both with and without NA's. It isn't necessarily the prettiest, but I really appreciate the value labels. What are some other commands and packages you can use to get this information? (summarytools::freq, descr::freq)

```
#sjmisc package
frq(df$M1APPLIEDMTH)
```

```
##
## M1 A14B Math teacher took college-level applied mathematics course(s) (x) <numeric>
## # total N=23503  valid N=17029  mean=0.39  sd=0.49
##
##  val                   label   frq raw.prc valid.prc cum.prc
##   -9                 Missing     0    0.00      0.00    0.00
##   -8       Unit non-response     0    0.00      0.00    0.00
##   -7 Item legitimate skip/NA     0    0.00      0.00    0.00
##    0                      No 10424   44.35     61.21   61.21
##    1                     Yes  6605   28.10     38.79  100.00
##   NA                    <NA>  6474   27.55        NA      NA
```

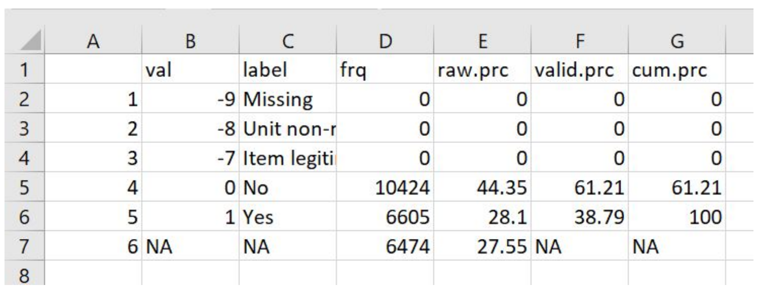
We can make this output look **MUCH** better with the kable package (we have to make the object a data frame, first!).

```
a<-frq(df$M1APPLIEDMTH)
a<-as.data.frame(a)
kable(a, booktabs=T)%>%
  kable_styling()
```

val	label	frq	raw.prc	valid.prc	cum.prc
-9	Missing	0	0.00	0.00	0.00
-8	Unit non-response	0	0.00	0.00	0.00
-7	Item legitimate skip/NA	0	0.00	0.00	0.00
0	No	10424	44.35	61.21	61.21
1	Yes	6605	28.10	38.79	100.00
NA	NA	6474	27.55	NA	NA

Alternatively, you could export it as a csv to excel and manipulate it there.

```
write.csv(a, file="frequencies.csv")
```



	A	B	C	D	E	F	G
1		val	label	frq	raw.prc	valid.prc	cum.prc
2	1	-9	Missing	0	0	0	0
3	2	-8	Unit non-r	0	0	0	0
4	3	-7	Item legiti	0	0	0	0
5	4	0	No	10424	44.35	61.21	61.21
6	5	1	Yes	6605	28.1	38.79	100
7	6	NA	NA	6474	27.55	NA	NA
8							

To summarize a continuous variable, I like to use the describe command from the psych package.

```
b<-describe(df$MTSCOR)
kable(b,digits=2,booktabs=T)%>%
  kable_styling()
```

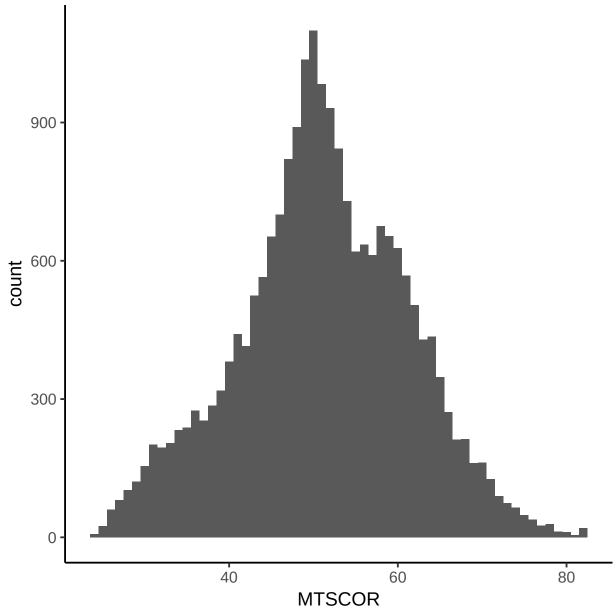
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	21444	51.11	10.08	50.97	51.24	9.83	24.02	82.19	58.17	-0.06	-0.13	0.07

```
#Check the calculation for the standard error if you'd like!
round(10.08/sqrt(21444),3)
```

```
## [1] 0.069
```

To plot a histogram, I like to use ggplot from tidyverse.

```
df%>%
  ggplot(aes(x=MTSCOR))+
  geom_histogram(binwidth = 1)+
  theme_classic()
```



To test if the data are normally distributed, SPSS defaults to using the KS-test in the explore command. However, the KS-test has been shown to have low power and therefore be unreliable. Instead, it is recommended to use the Shapiro-Wilks test (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/). You can still get this test in SPSS, but it doesn't pop up by default. However, when you run the SW test in R, you'll get an error message. What does that mean? How is R helping us?

```
ks.test(df$MTSCOR, "pnorm")
```

```
## Warning in ks.test(df$MTSCOR, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```

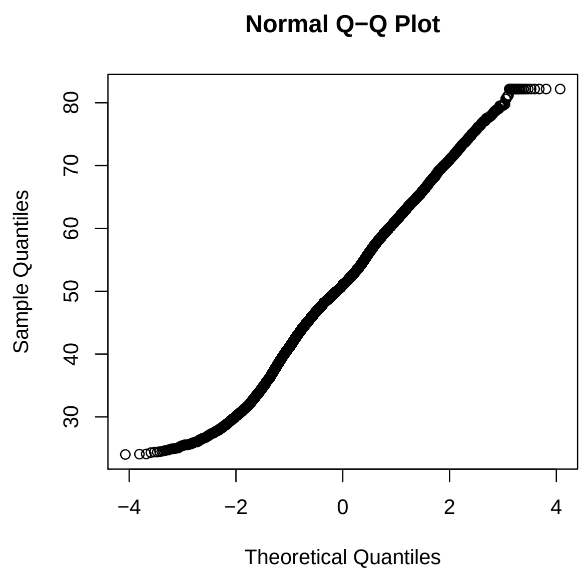
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df$MTSCOR
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(df$MTSCOR)
```

```
## Error in shapiro.test(df$MTSCOR): sample size must be between 3 and 5000
```

I don't normally plan to put qq-plot results in a report or journal article, so I can just use base R (since I don't care if it looks nice)

```
qqnorm(df$MTSCOR)
```



2. New: Compare means

There are two groups of teachers: those that took a college level applied math course, and those that did not. Let's see if their students score differently on their standardized math assessments by comparing the mean math score for each group.

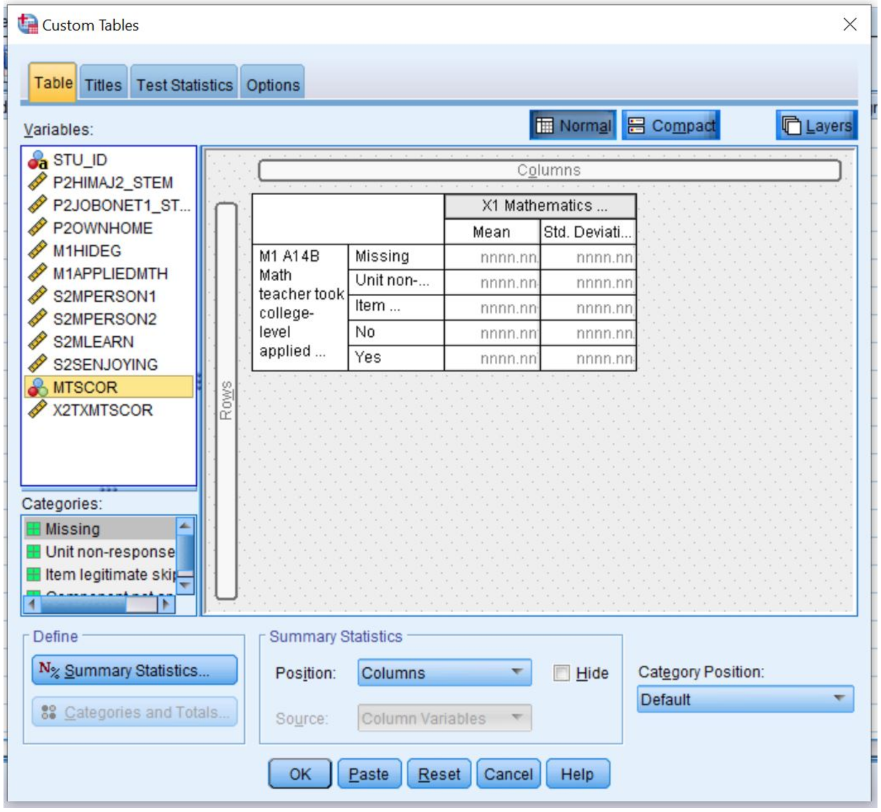
In SPSS

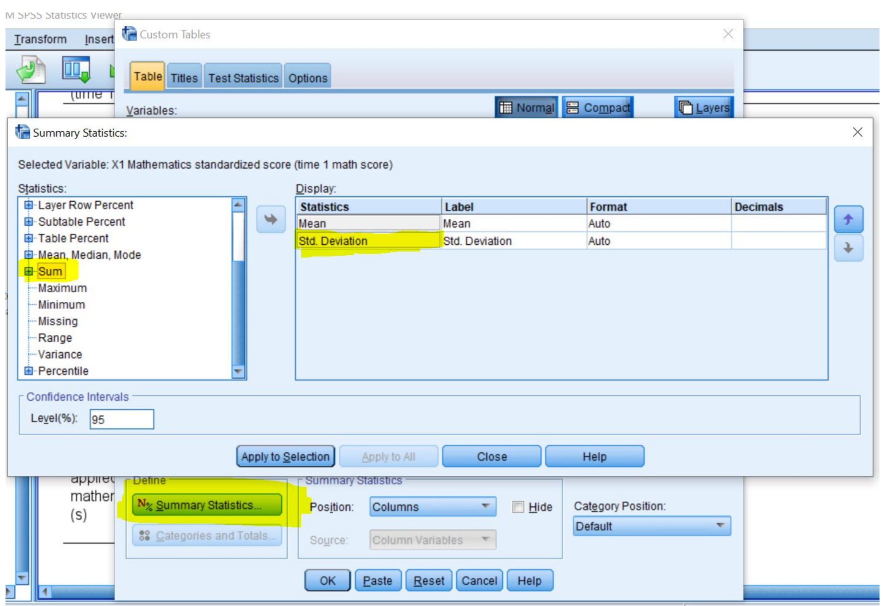
Analyze > Tables > Custom Tables

Make sure to change the variable type (measure) of M1APPLIEDMTH from 'Scale' to 'Nominal' since this is a nominal variable. You can do this by right clicking on the variable and selecting 'Nominal'.

Drag M1APPLIEDMTH onto the rows and MTSCOR onto the columns (you can reverse these - it doesn't matter - try it out and see what it looks like!).

Next, click on "Summary Statistics" on the bottom left. Under "Statistics" > "Sum", select "Standard Deviation" and drag it to the "Display" box under "Mean". Press "Apply to Selection" and "Close". Then, press "OK".


Custom Tables
Table
Titles
Test Statistics
Options
Variables:
Normal
Compact
Layers
STU_ID
P2HIMAJ2_STEM
P2JOBONET1_ST...
P2OWNHOME
M1HIDEG
M1APPLIEDMTH
S2MPERSON1
S2MPERSON2
S2MLEARN
S2SENJOYING
MTSCOR
X2TXMTSCOR
Columns
Rows
X1 Mathematics ...
Mean
Std. Deviati...
M1 A14B Math teacher took college-level applied ...
Missing
Unit non-...
Item ...
No
Yes
nnnn.nn
Categories:
Missing
Unit non-response
Item legitimate skip
Define
Summary Statistics...
Categories and Totals...
Summary Statistics
Position:
Columns
Hide
Source:
Column Variables
Category Position:
Default
OK
Paste
Reset
Cancel
Help


Custom Tables
Table
Titles
Test Statistics
Options
Normal
Compact
Layers
Variables:
Summary Statistics:
Selected Variable: X1 Mathematics standardized score (time 1 math score)
Statistics:
Layer Row Percent
Subtable Percent
Table Percent
Mean, Median, Mode
Sum
Maximum
Minimum
Missing
Range
Variance
Percentile
Display:
Statistics
Label
Format
Decimals
Mean
Mean
Auto
Std. Deviation
Std. Deviation
Auto
Confidence Intervals
Level(%):
95
Apply to Selection
Apply to All
Close
Help
Define
Summary Statistics...
Categories and Totals...
Summary Statistics
Position:
Columns
Hide
Source:
Column Variables
Category Position:
Default
OK
Paste
Reset
Cancel
Help

Custom Tables

		X1 Mathematics standardized score (time 1 math score)	
		Mean	Standard Deviation
M1 A14B Math teacher took college-level applied mathematics course (s)	Missing	.	.
	Unit non-response	.	.
	Item legitimate skip/NA	.	.
	No	51.23	10.03
	Yes	51.39	9.93

In R

I like to use dplyr from the tidyverse to compare means. Notice how I combined the kable command with the commands that I used to create the mean comparisons. What other commands could you use?

```
df%>%
  group_by(M1APPLIEDMTH)%>%
  summarise(mean = mean(MTSCOR, na.rm=TRUE),sd=sd(MTSCOR,na.rm=TRUE))%>%
  kable(digits=2,booktabs=T)%>%
  kable_styling()
```

M1APPLIEDMTH	mean	sd
0	51.23	10.03
1	51.39	9.93
NA	50.56	10.30

3. New: Regression with one categorical predictor

Of the methods we've learned, which ones could we use to evaluate if the means of each group are statistically significantly different from each other?

Let's revisit Andy's slides:

The four basic questions of statistics

- What do the data tend to be like? (*central tendency*)
  - Mean, median, mode
- How much do they tend to be like that? (*variation*)
  - Range, standard deviation
- How are two or more variables associated with one another? (*association*)
  - Correlation, regression, mean comparisons
- With how much confidence can we generalize from a sample to a population? (*inference*)
  - Statistical significance, p-values

UCSB

Let's run a regression model to determine if the mean differences of each group are statistically significantly different.

In SPSS

Analyze > Regression > Linear

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.008[a]	.000	.000	9.99517

a. Predictors: (Constant), M1 A14B Math teacher took college-level applied mathematics course(s)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	97.503	1	97.503	.976	.323[b]
	Residual	1598254.4	15998	99.903		
	Total	1598351.9	15999			

a. Dependent Variable: X1 Mathematics standardized score (time 1 math score)

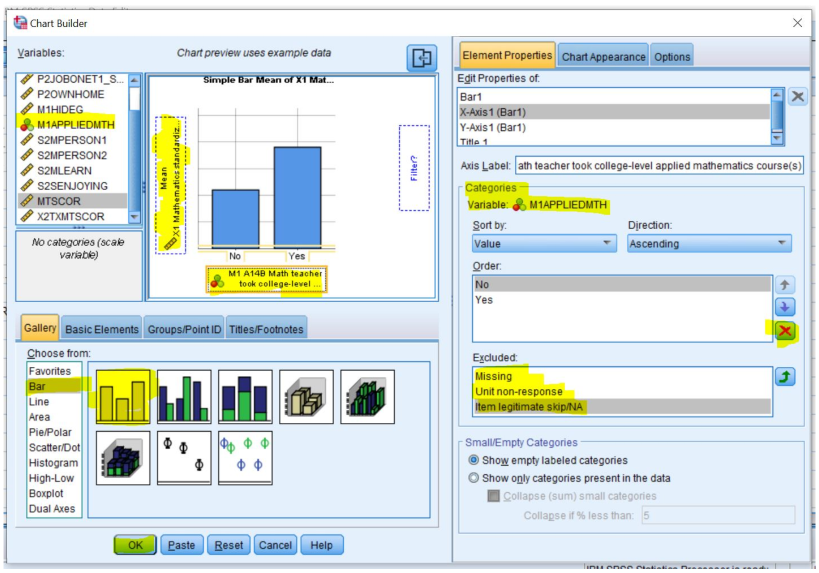
b. Predictors: (Constant), M1 A14B Math teacher took college-level applied mathematics course(s)

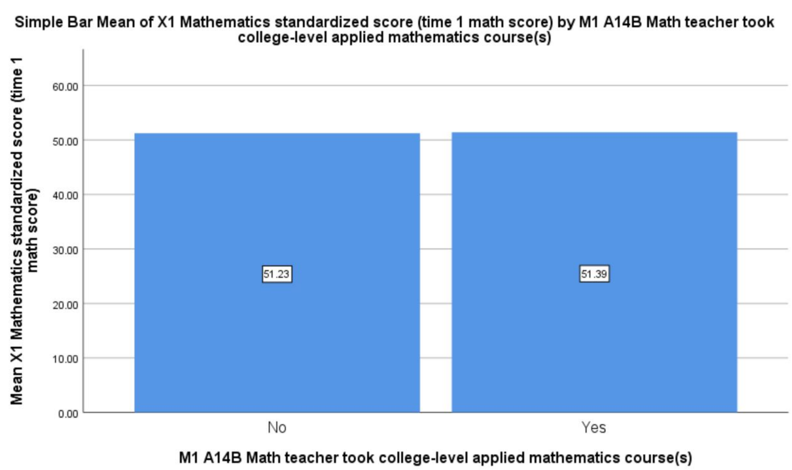
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51.235	.101		507.079	.000
	M1 A14B Math teacher took college-level applied mathematics course(s)	.160	.162	.008	.988	.323

a. Dependent Variable: X1 Mathematics standardized score (time 1 math score)

Now, let's plot a mean comparison. Select Graphs > Chart Builder. Drag the "bar plot" onto the "gallery chart preview". Change M1APPLIEDMTH to nominal, and drag it onto the x-axis. Move MTSCOR onto the y-axis. Now, click on the x-axis to trigger the axis options on the right. Under "categories", remove "missing", "unit non-response" and "item legitimate skip/NA" by selecting the red x. Press OK.





In R

lm and summary are base functions in R (no packages needed).

```r
summary(lm(MTSCOR~M1APPLIEDMTH,data=df))
```

```
##
## Call:
## lm(formula = MTSCOR ~ M1APPLIEDMTH, data = df)
##
## Residuals:
## <Labelled double>: X1 Mathematics standardized score (time 1 math score)
##      Min       1Q   Median       3Q      Max
## -27.2948  -6.1182  -0.1945   6.9404  30.9530
##
## Labels:
##  value                    label
##     -9                  Missing
##     -8         Unit non-response
##     -7  Item legitimate skip/NA
##     -6 Component not applicable
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   51.2346     0.1010 507.079   <2e-16 ***
## M1APPLIEDMTH   0.1602     0.1621   0.988    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.995 on 15998 degrees of freedom
##   (7503 observations deleted due to missingness)
## Multiple R-squared:  6.1e-05,    Adjusted R-squared:  -1.502e-06
## F-statistic: 0.976 on 1 and 15998 DF,  p-value: 0.3232
```

To make this output presentable, you'd have to export this to a csv file and manipulate it in Excel. But, first, you'd have to manipulate the results using the broom package.

```r
d<-summary(lm(MTSCOR~M1APPLIEDMTH,data=df))
write.csv(glance(d),"regression results1.csv")
write.csv(tidy(d),"regression results2.csv")
```

	A	B	C	D	E	F	G
1		r.squared	adj.r.squa	sigma	statistic	p.value	df
2	1	6.10E-05	-1.50E-06	9.995168	0.975977	0.323209	2

	A	B	C	D	E	F
1		term	estimate	std.error	statistic	p.value
2	1	(Intercept	51.23455	0.101039	507.0788	0
3	2	M1APPLIE	0.16017	0.162129	0.987915	0.323209
4						

Alternatively, you can use R Markdown and use the stargazer package. If you see yourself as a quant methods person, this is something to begin familiarizing yourself with! You'll be glad you did. Plus, look how pretty!

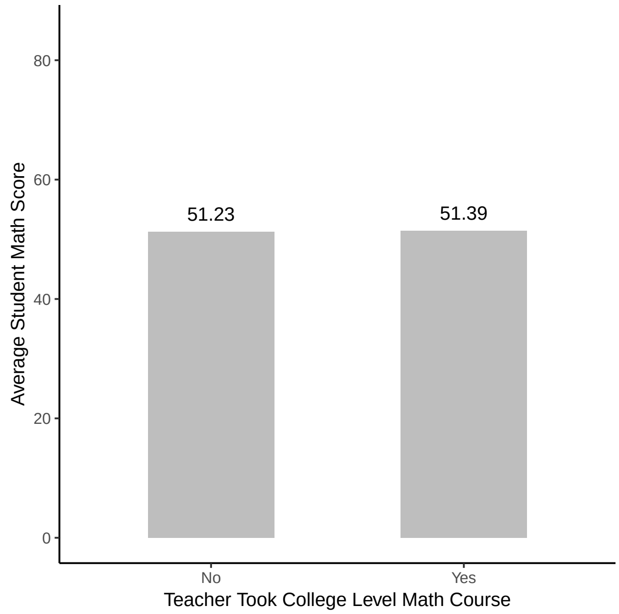
```r
e<-lm(MTSCOR~M1APPLIEDMTH,data=df)
stargazer(e,header=FALSE)
```

Table 1:

	Dependent variable:
	MTSCOR
M1APPLIEDMTH	0.160
	(0.162)
Constant	51.235***
	(0.101)
Observations	16,000
R^2	0.0001
Adjusted R^2	−0.00000
Residual Std. Error	9.995 (df = 15998)
F Statistic	0.976 (df = 1; 15998)
Note:	*p<0.1; **p<0.05; ***p<0.01

Now, let's plot this relationship. I like to use ggplot. Can you find another way to plot this? If so, tell me!

```r
df%>%
  group_by(M1APPLIEDMTH)%>%                            #compare these groups
  summarise(mean=mean(MTSCOR,na.rm=TRUE))%>%           #compute the means
  filter(is.na(M1APPLIEDMTH)==FALSE)%>%                #remove the NA values
  ggplot(aes(x=factor(M1APPLIEDMTH),y=mean))+          #plot. specify x as a factor
  geom_bar(stat="identity",fill="grey",width=.5)+      #aestheticsc
  geom_text(aes(label=round(mean,2)),nudge_y = 3)+     #add labels to bars
  xlab("Teacher Took College Level Math Course")+      #y-axis label
  ylab("Average Student Math Score")+                  #x-axis label
  scale_x_discrete(labels=c("No","Yes"))+              #label factor levels on x-axis
  ylim(0,85)+                                          #rescale y-axis
  theme_classic()                                      #aesthetics
```



4. New: Interpret p-values

Answer these questions on the quiz on Gauchospace

1. What is the null hypothesis?
2. What is the alternative hypothesis?
3. What kind of test statistic do we get for the regression coefficient?
4. What is the test statistic for the regression model?
5. What is the mean difference between the two groups?

   (hint: look at the regression coefficient for the parameter of interest)
6. Which group has a higher mean?
7. What is the p-value?
8. Is the test statistic statistically significant?
9. Are the means statistically significantly different from each other?
10. How much of the variation in average math test score is explained by the teacher's college coursework?
11. Would you reject or fail to reject the null hypothesis?