UBC Library Research Commons

Workshop: R for Statistical Analysis

Data Analysis Team:

- Albina Gibadullina (GAA)
- Amir Michalovich (GAA)
- Jeremy Buhler (Data Librarian)
- Sarah Parker (Data Librarian)

Land Acknowledgement

UBC is located on the traditional, ancestral, and unceded territory of the xʷməθkʷəy̓əm (Musqueam) people.

- Traditional: recognizes lands traditionally used and/or occupied by the Musqueam people or other First Nations in other parts of the country.
- Ancestral: recognizes land that is handed down from generation to generation.
- Unceded: refers to land that was not turned over to the Crown (government) by a treaty or other agreement.

Pre-workshop setup

Download and install R

For Windows:

1. Visit [R Project (https://www.r-project.org/)](https://www.r-project.org/) to learn about R versions.
2. Download and install R from your preferred CRAN mirror [here (https://cran.r-project.org/mirrors.html)](https://cran.r-project.org/mirrors.html)
   A. Choose "0-Cloud" or a mirror site near you.

For Mac:

1. Check that your macOS system is up-to-date
2. Download and install R from [The Comprehensive R Archive Network (https://cran.r-project.org/)](https://cran.r-project.org/)

Download and install R studio

For Windows and Mac:

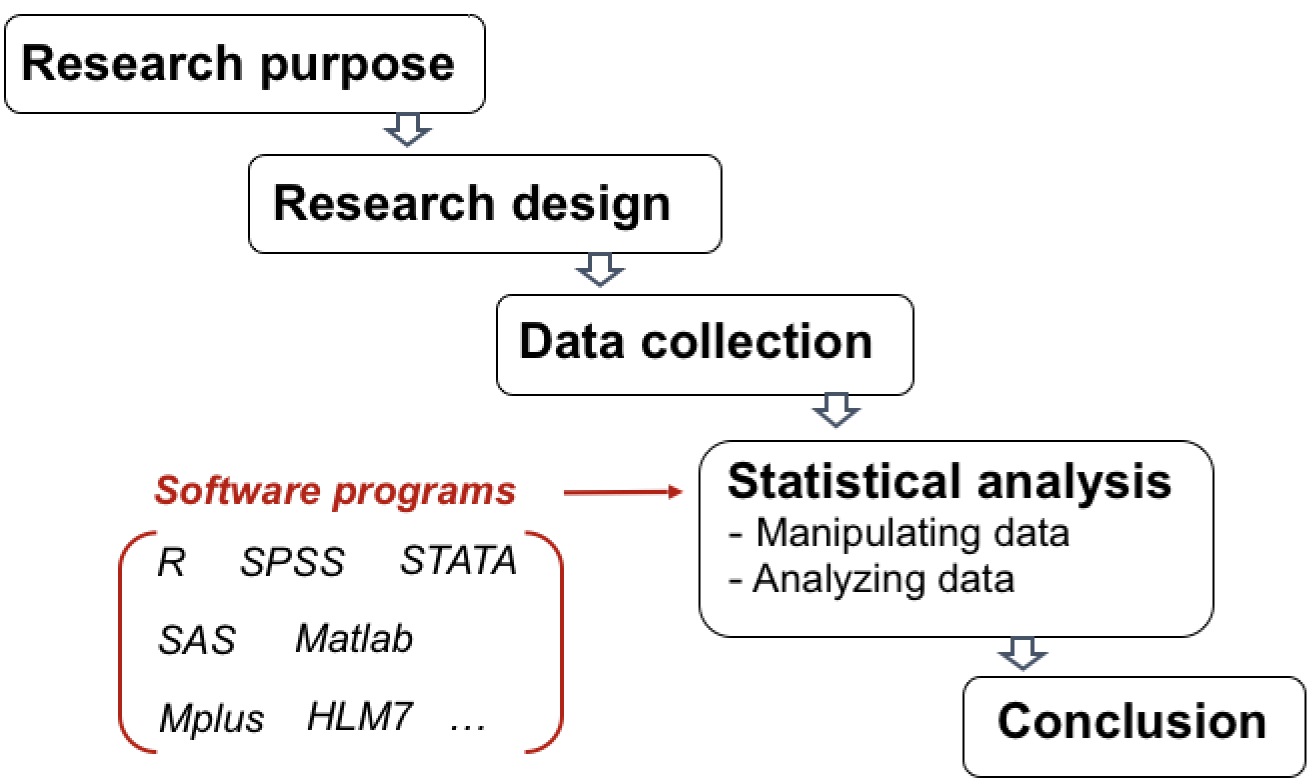
1. Download and install R Studio from [here (https://rstudio.com/products/rstudio/download/#download)](https://rstudio.com/products/rstudio/download/#download)

Learning Objectives

- Learn how to identify the types of variables in R
- Learn the basic commands for descriptive statistics
- Learn the basic commands for inference statistics

Overview of Quantative Research Process

A systamatic research process that involves collecting objective, measureable data, using statistics to analyze the data, and generalizing the results to a larger population to explain a phenomena. Usually, software programs assist on data analysis.



Data Analysis in Quantitative Research

Definitions

- Data refers to facts or pieces of information that can either be quantitative or qualitative.
- Variable refers to any property that can be observed or measured.

Types of Variables

It is important to understand the different types of variables because they will determine the statistical analysis method.

Type	Description	Example
Nominal	Labels or Descriptions that cannot be ordered	Gender
Ordinal	Labels or Descriptions that can be ordered	Education Level
Interval	Numeric values with equal magnitude, doesn't have absolute zero	SAT scores
Ratio	Numeric values with equal magnitude, does have absolute zero	Age

Categorize these variables in R

Nominal/Ordinal -> Character or Factor

Intervel/Ratio -> Numeric or Integer

Definitions

- Character: Text
- Factor: Integer associated with a specific category
- Numeric: Number with decimal point
- Integer: Number with no decimal point

Getting Started

Set working directory in R studio

You can set the working directory using **Session** > **Set Working Directory** > **Choose Directory**.

Loading a built-in R dataset

About the data

3 Measures Of Ability: SATV, SATQ, ACT: "Self reported scores on the SAT Verbal, SAT Quantitative and ACT were collected as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. Age, gender, and education are also reported. The data from 700 subjects are included here as a demonstration set for correlation and analysis" (Revelle et al., 2009). (https://www.rdocumentation.org/packages/psych/versions/1.9.12.31/topics/sat.act)

Format

A data frame with 700 observations on the following 6 variables.

`gender`

males = 1, females = 2

`education`

self reported education 1 = high school ... 5 = graduate work

`age`

age

`ACT`

ACT composite scores may range from 1 - 36. National norms have a mean of 20.

`SATV`

SAT Verbal scores may range from 200 - 800.

`SATQ`

SAT Quantitative scores may range from 200 - 800

Write and run the following commands to load the dataset

In [67]:
```
# Installing packages

install.packages("psych") # a statistical analysis package
install.packages("rstatix") # a statistical analysis package
install.packages("dplyr")  # a data manipulation package
install.packages("ggplot2") # a data visualization package
install.packages("GGally") # a data visualization package (extension of ggplot
2)
```

```
Installing package into ‘/home/jupyter/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)

Installing package into ‘/home/jupyter/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)

Installing package into ‘/home/jupyter/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)

Installing package into ‘/home/jupyter/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)

Installing package into ‘/home/jupyter/R/x86_64-pc-linux-gnu-library/4.0’
(as ‘lib’ is unspecified)
```

In [71]:
```
# Loading packages

library(psych)
library(rstatix)
library(dplyr)
library(ggplot2)
library(GGally)
```

In [72]:
```
# Creating a data frame `scores` using `sat.act` dataset from `psych` package

scores <- sat.act
```

In [73]:
```
# scores <-read.csv("sat.act.csv")
```

Describing and manipulating data structure

In [74]:
```
head(scores) #See the first six rows of the data frame
```

A data.frame: 6 × 6

	gender	education	age	ACT	SATV	SATQ
	<int>	<int>	<int>	<int>	<int>	<int>
29442	2	3	19	24	500	500
29457	2	3	23	35	600	500
29498	2	3	20	21	480	470
29503	1	4	27	26	550	520
29504	1	2	33	31	600	550
29518	1	5	26	28	640	640

In [75]:
```
# Count the number of observations

nrow(scores) #To check the number of rows
ncol(scores) #To check the number of columns
```

700

6

str(df): To check the structure of your data

In [76]:
```
str(scores)
```

```
'data.frame':	700 obs. of  6 variables:
 $ gender   : int  2 2 2 1 1 1 2 1 2 2 ...
 $ education: int  3 3 3 4 2 5 5 3 4 5 ...
 $ age      : int  19 23 20 27 33 26 30 19 23 40 ...
 $ ACT      : int  24 35 21 26 31 28 36 22 22 35 ...
 $ SATV     : int  500 600 480 550 600 640 610 520 400 730 ...
 $ SATQ     : int  500 500 470 520 550 640 500 560 600 800 ...
```

Question: What do you notice?

In [77]:
```
# Check current data format of the  `gender` variable

typeof(scores$gender)
```

'integer'

as.factor(df$columnname): To change a variable to factor

In [78]:
```
scores$gender <- as.factor(scores$gender)
```

is.factor(df$columnname): To check if a variable is defined as factor

In [79]:
```
is.factor(scores$gender)
```

TRUE

Extra information

is.integer(df$columnname): To check if a variable is defined as integer

is.numeric(df$columnname): To check if a variable is defined as numeric

is.character(df$columnname): To check if a varible is defined as character

is.logical(df$columnname): To check if a varible is defined as logical: TRUE/FALSE

as.integer(df$columnname): To change a variable to integer

as.numeric(df$columnname): To change a variable to numeric

as.character($columnndfame): To change a variable to character

as.logical(df$columnname): To change a variable to logical: TRUE/FALSE

Exercise #1

- Using typeof command, check the current format of `education`
- Using as.factor command, change `education` to factor.
- Using is.factor command, check if `education` is defined as factor.

Answer to Exercise #1

In [80]:
```
typeof(scores$education)
```

'integer'

In [81]:
```
scores$education <- as.factor(scores$education)
```

In [82]:
```
is.factor(scores$education)
```

TRUE

In [83]:
```
# Check updated data structure

str(scores)
```

```
'data.frame':	700 obs. of  6 variables:
 $ gender   : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 1 2 2 ...
 $ education: Factor w/ 6 levels "0","1","2","3",..: 4 4 4 5 3 6 6 4 5 6 ...
 $ age      : int  19 23 20 27 33 26 30 19 23 40 ...
 $ ACT      : int  24 35 21 26 31 28 36 22 22 35 ...
 $ SATV     : int  500 600 480 550 600 640 610 520 400 730 ...
 $ SATQ     : int  500 500 470 520 550 640 500 560 600 800 ...
```

Transforming values of categorical variables

`Gender` is now factor but it is still coded as "1" (men) and "2" (women) - it would be helpful for later analysis to change "1" to "men" and "2" to "women"

In [84]:
```
# Replacing values

# Step 1: Change the data format to charachter
scores$gender <- as.character(scores$gender)

# Step 2: Replace "1" with "men" and "2" with "women"
scores$gender[scores$gender=="1"] <- "Men"
scores$gender[scores$gender=="2"] <- "Women"

# Step 3: Change the data format back to factor
scores$gender <- as.factor(scores$gender)

# Step 4: Check levels for `gender`
levels(scores$gender)
```

'Men' · 'Women'

In [85]:
```
# Optional: If you wanted to re-arrange the order of levels: "Women" first, "M
en" second

# Step 5: Reorder levels for `gender`
scores <- scores %>%
  reorder_levels(gender, order = c("Women","Men"))

# Step 6: Check levels for `gender`
levels(scores$gender)
```

'Women' · 'Men'

Descriptive Statistics

Descriptive statistics summarize the data in a meaningful way. The purpose of using descriptive statistics is to explore the observed data and not to draw inferences.

Describing Categorical Data

In [86]:
```
# Get frequency for the `gender` variable
table(scores$gender)
```

```
Women   Men
  453   247
```

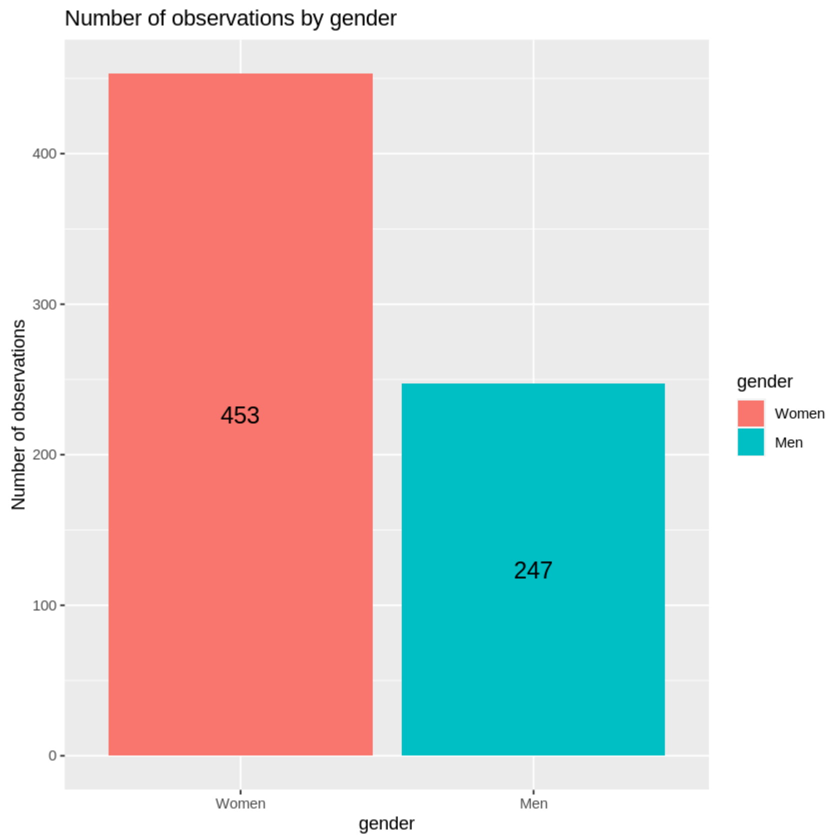
In [87]:
```
# Get cross-tabulation for `gender` and `education`
table(scores$gender, scores$education)
```

```
         0   1   2   3   4   5
  Women 30  25  21 195  87  95
  Men   27  20  23  80  51  46
```

Visualizing Categorical Data

In [88]:
```
# Make a bar-chart showing counts of observations by gender

scores %>%
  ggplot(aes(fill=gender, x=gender)) +
    geom_bar(aes(y=..count..))+
    geom_text(aes(label = ..count..,y= ..count..), stat="count", position = po
sition_stack(vjust = 0.5), size=5) +
    labs(title="Number of observations by gender",y="Number of observations")
```

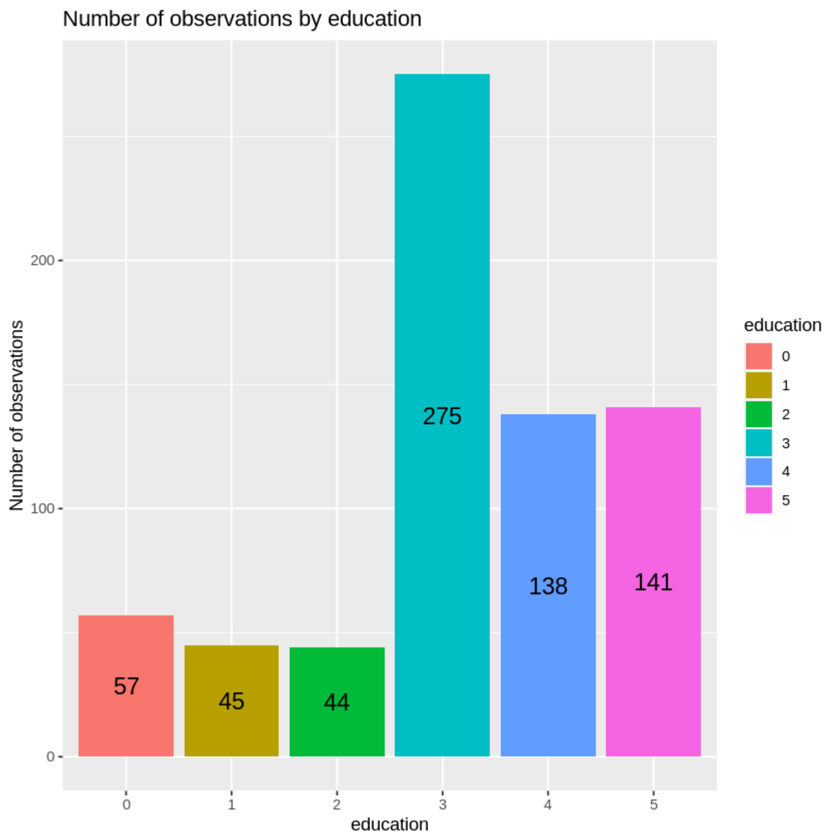


Exercise #2

Make a bar-chart showing counts of observations by levels of education

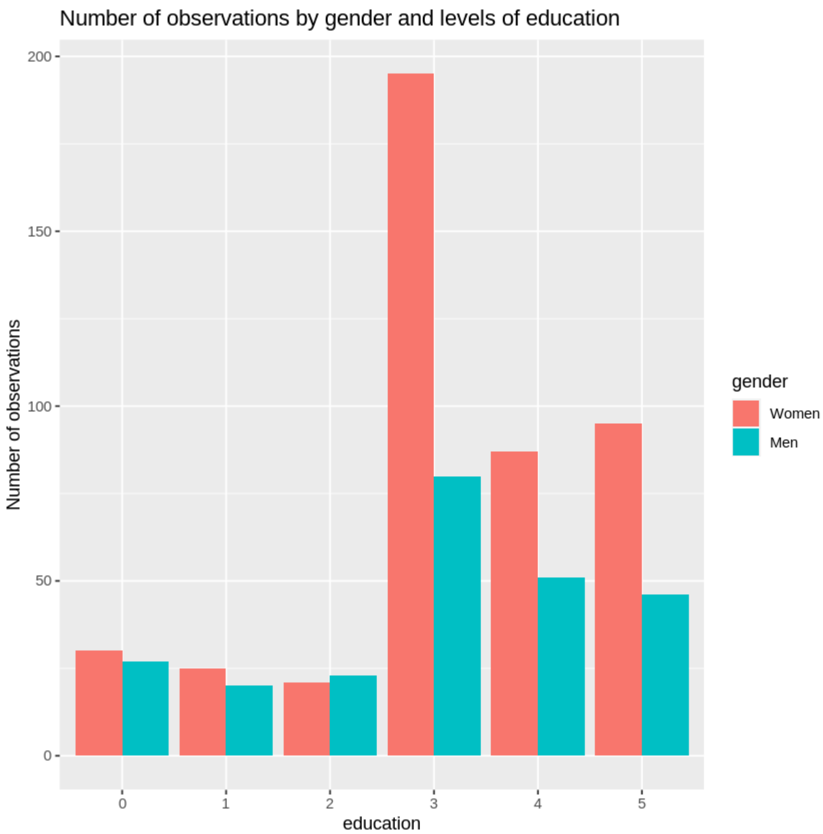
In [89]:
```
# Answer

scores %>%
  ggplot(aes(fill=education, x=education))+
    geom_bar(aes(y=..count..))+
    geom_text(aes(label = ..count..,y= ..count..), stat="count", position = po
sition_stack(vjust = 0.5), size=5) +
    labs(title="Number of observations by education",y="Number of observation
s")
```



In [90]:
```
# Make a grouped bar-chart showing counts of observations by gender and levels of education

scores %>%
  ggplot(aes(fill=gender,x=education, y=..count..)) +
    geom_bar(aes(y=..count..), position="dodge")+
    labs(title="Number of observations by gender and levels of education",y="Number of observations")
```



Describing Quantitative Data

In [91]:
```
# Find summary statistics for each quantitative variable

get_summary_stats(scores)
```

A tibble: 4 × 13

variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ACT	700	3	36	29	25	32	7	4.448	28.547	4.824	0.182
age	700	13	65	22	19	29	10	5.930	25.594	9.499	0.359
SATQ	687	200	800	620	530	700	170	118.608	610.217	115.639	4.412
SATV	700	200	800	620	550	700	150	118.608	612.234	112.903	4.267

In [92]:
```r
# Find summary statistics for `ACT` scores, grouped by `gender`

scores %>%
  group_by(gender) %>%
  get_summary_stats(ACT)
```

A tibble: 2 × 14

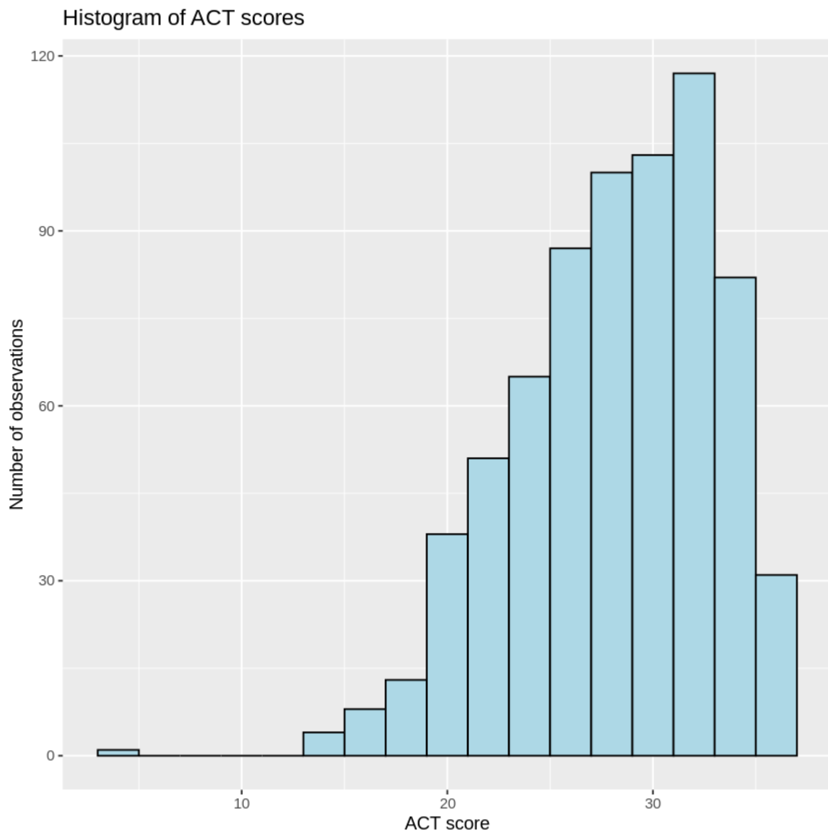
gender	variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	s
<fct>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<db
Women	ACT	453	15	36	29	25	32.0	7.0	4.448	28.417	4.688	0.22
Men	ACT	247	3	36	30	25	32.5	7.5	4.448	28.785	5.064	0.32

Visualizing quantitative data

Histograms and Density Plots

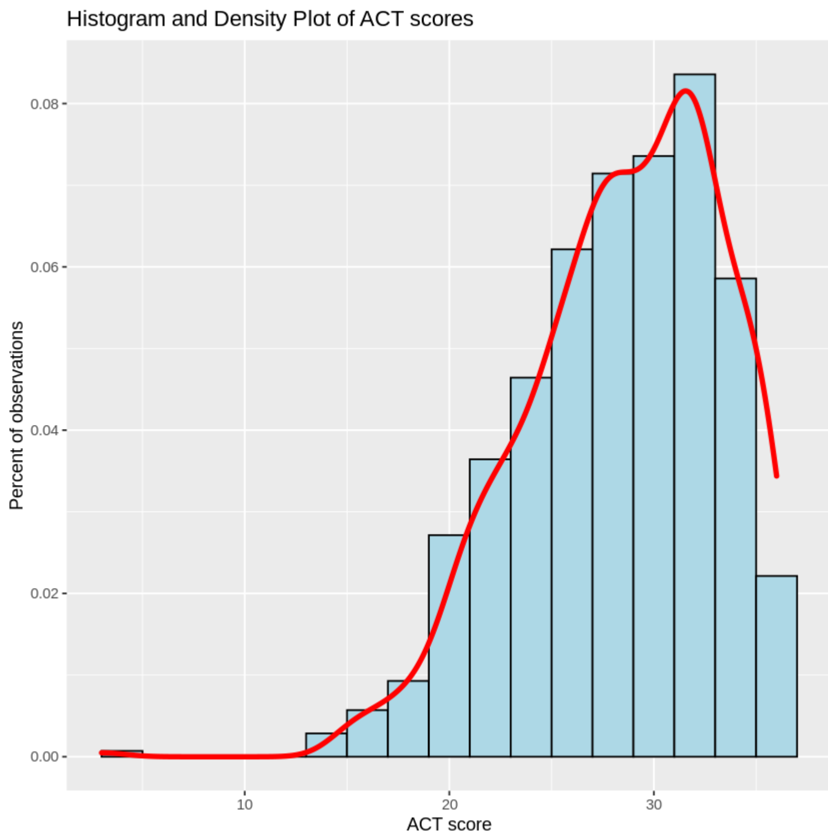
In [93]:
```
# Visualizes the distribution of ACT scores

scores %>%
  ggplot(aes(x=ACT)) +
    geom_histogram(binwidth=2,color="black", fill="lightblue") +
    labs(title="Histogram of ACT scores ", x="ACT score", y="Number of observa
tions")
```



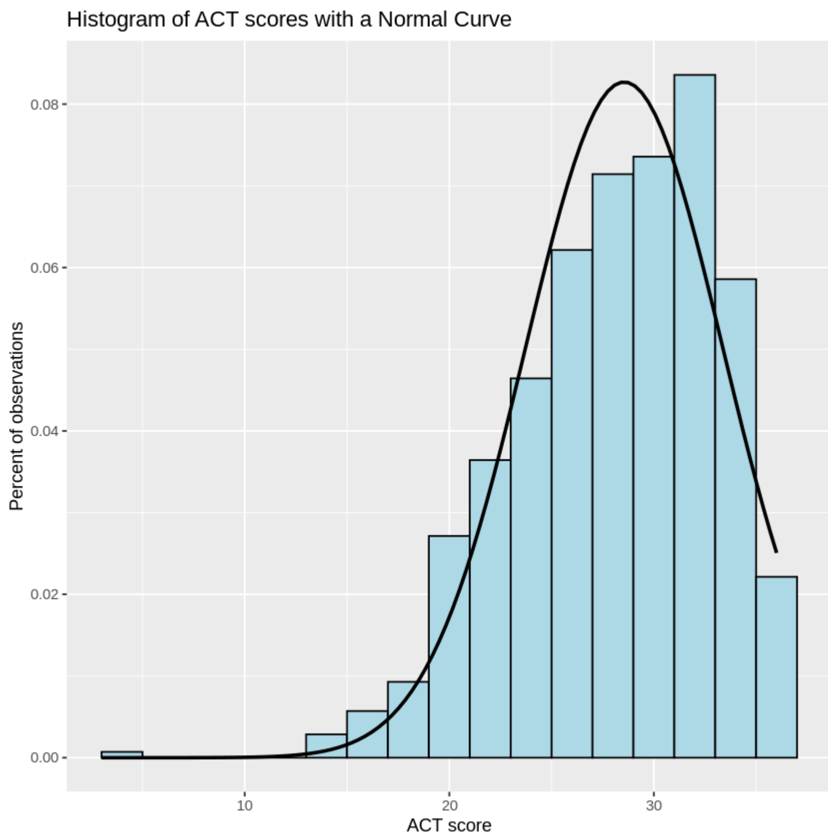
In [94]:
```
# Add a red density plot to the histogram of ACT scores

scores %>%
  ggplot(aes(x=ACT)) +
    geom_histogram(aes(y = stat(density)), binwidth=2, color="black", fill="li
ghtblue") +
    geom_density(size=1.5, color="red") +
    labs(title="Histogram and Density Plot of ACT scores ", x="ACT score", y=
"Percent of observations")
```



In [95]:
```
# What if you wanted to add a normal distribution curve instead of a density p
lot?

scores %>%
  ggplot(aes(x=ACT)) +
    geom_histogram(aes(y = stat(density)), binwidth=2, color="black", fill="li
ghtblue") +
    stat_function(fun = dnorm, args = list(mean = mean(scores$ACT), sd = sd(sc
ores$ACT)),size=1)+
    labs(title="Histogram of ACT scores with a Normal Curve", x="ACT score", y
="Percent of observations")
```

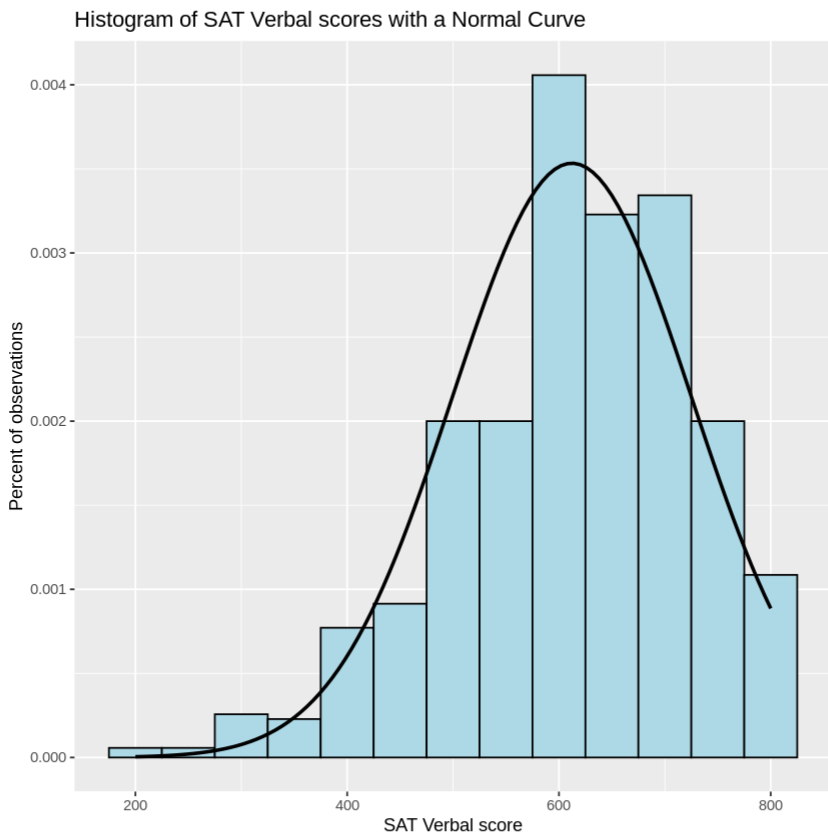


Exercise #3

Make a histogram of `SATV` scores with an added Normal Curve

In [96]:
```
# Answer

scores %>%
  ggplot(aes(x=SATV)) +
    geom_histogram(aes(y = stat(density)), binwidth=50, color="black", fill="l
ightblue") +
    stat_function(fun = dnorm, args = list(mean = mean(scores$SATV), sd = sd(s
cores$SATV)),size=1)+
    labs(title="Histogram of SAT Verbal scores with a Normal Curve", x="SAT Ve
rbal score", y="Percent of observations")
```

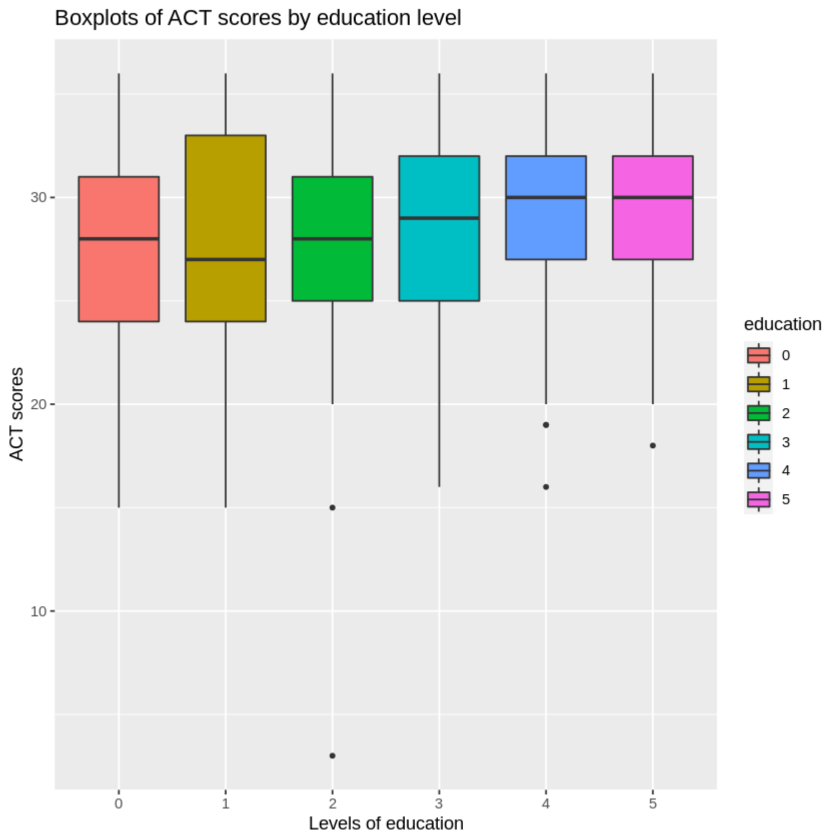


Boxplots

Boxplots are used to compare distributions of one quantitative variable across multiple categories.

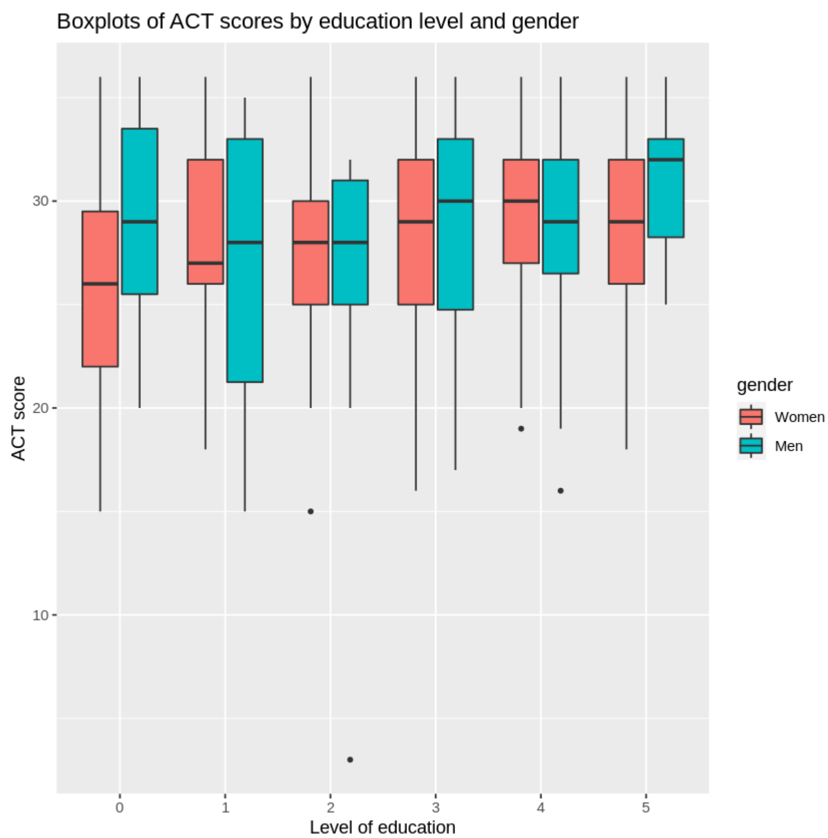
In [97]:
```
# Comparing distribution of `ACT` scores variable by education level:

scores %>%
ggplot(aes(x=education, y=ACT, fill=education))+
    geom_boxplot(outlier.size=1)+
    labs(title="Boxplots of ACT scores by education level",x="Levels of educat
ion", y="ACT scores")
```



In [98]:
```
# Comparing distribution of `ACT` scores variable by education level and gende
r (grouped by education level first, gender second)

scores %>%
ggplot(aes(x=education, y=ACT, fill=gender))+
    geom_boxplot(outlier.size=1)+
    labs(title="Boxplots of ACT scores by education level and gender",x="Level
of education", y="ACT score")
```

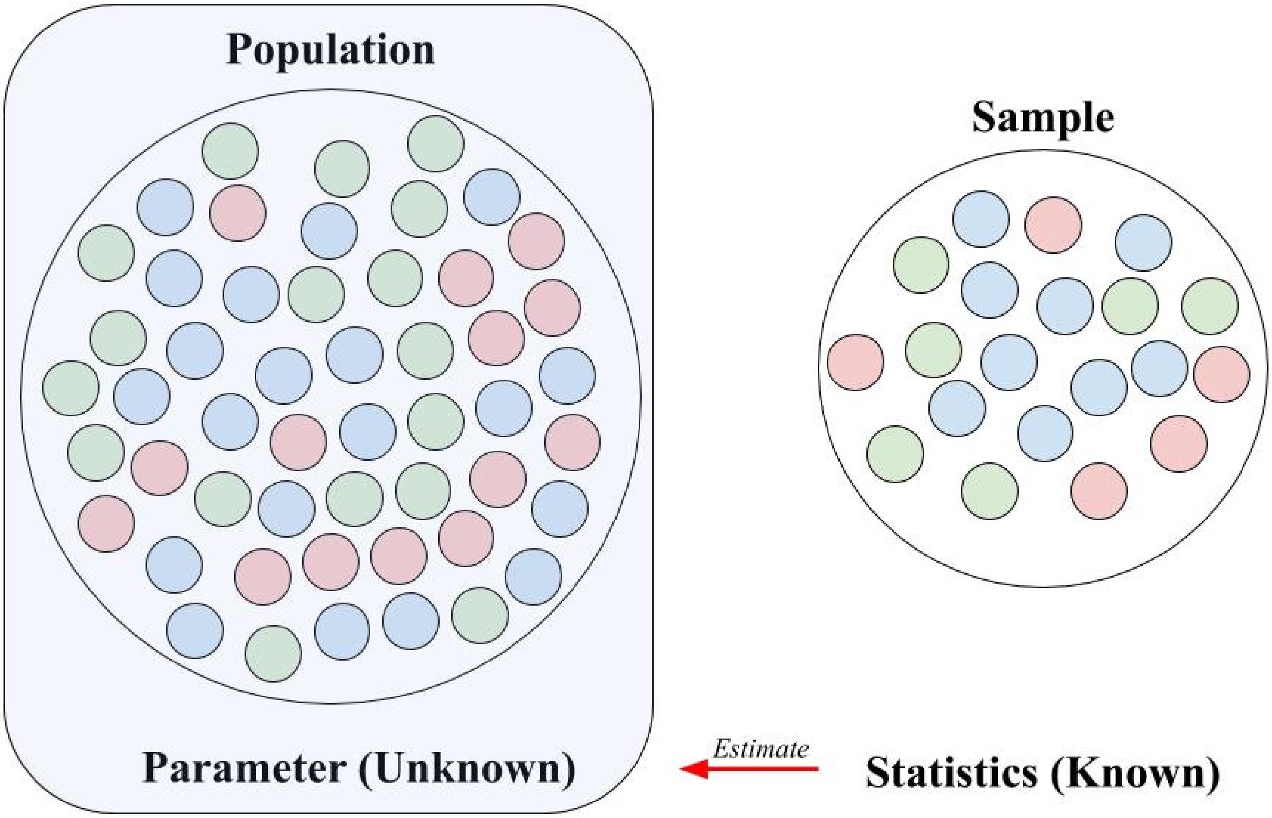


Inferential Statistics

Unlike descriptive statistics, inferential statistics use the observed data to make inferences about the population.

What is Inferential Statistics?

A series of statistical techniques that enable us to make generalizations about populations from sample data.



How to select a statistical test?

Selecting the appropriate statistical test should depend upon data type of your variables and your hypothesis:

Independent Variable	Dependent Variable	Statistical Test or Model	Parametric Test
None (use hypothesized value)	Quantitative (Interval or Ratio)	One sample t-test	No
None (use hypothesized value)	Categorical (2 or more categories)	Chi-square goodness of fit test	No
Categorical (Binary, 2 categories)	Quantitative (Interval or Ratio)	Two sample t-test	Yes
Categorical (2 or more categories)	Quantitative (Interval or Ratio)	One-way analysis of variance (ANOVA)	Yes
2x Categorical (2 or more categories)	Quantitative (Interval or Ratio)	Two-way analysis of variance (ANOVA)	Yes
Categorical (2 or more categories)	Categorical (2 or more categories)	Chi-square test of independence	No
Quantitative (Interval or Ratio)	Quantitative (Interval or Ratio)	Simple linear regression	Yes
Multiple (2 or more) Quantitative	Quantitative (Interval or Ratio)	Multiple linear regression	Yes
Quantitative (Interval or Ratio)	Categorical (Binary, 2 categories)	Simple logistic regression	No
Multiple (2 or more) Quantitative	Quantitative (Interval or Ratio)	Multiple logistic regression	No

See [R Companinion Handbook (https://rcompanion.org/handbook/D_03.html)](https://rcompanion.org/handbook/D_03.html) for a more extended list of statistical tests.

Differences between parametric and non-parametric tests

Common model assumptions found in parametric tests:

1. Independence
2. Normality
3. Equal variance

Simple linear regression have some additional assumptions. For more information: [Simple linear regression assumptions (https://www.statisticssolutions.com/assumptions-of-linear-regression/)](https://www.statisticssolutions.com/assumptions-of-linear-regression/)

Process of Hypothesis Testing

1. Identify H-null (your starting guess) and H-alt (what you are trying to prove)
2. Compute **test-statistics**

    - Test-statistics measures how many SDs away your sample value is from your hypothesized pop. parameter
3. Compute **p-value** associated with your test-statistics

    - P-value measures the likelihood of attaining your sample statistics, given that H-null is true
4. Make a conclusion:

    - P-value and **α** (significance level)
        - P-value < α -> Reject Ho
        - P-value > α -> Fail to reject Ho
5. State the conclusion

    - Reject Ho -> We have enough evidence to state that Ha is true at α significance level
    - Fail to reject Ho -> We don’t have enough evidence to state that Ha is true at α significance level

Interpreting the results

- t: test-statistics
- df: degrees of freedom
- p-value: Statistical significance (to reject H-null, p-value < α)

One sample t-test

It is used to see whether the hypothesized value of the population mean matches actual (true) value.

For example, is the average ACT score for all participants 27?

H-null: Mean ACT = 27

H-alt: Mean ACT != 27

In [99]:
```
t.test(scores$ACT, mu = 27)
```

```
	One Sample t-test

data:  scores$ACT
t = 8.4862, df = 699, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 27
95 percent confidence interval:
 28.18920 28.90509
sample estimates:
mean of x
 28.54714
```

Conclusion

t(df=699) = 8.4862, p-value = 0.00% < 5% (reject the null)

The average ACT score is not 27.

Chi-square goodness of fit

It is used to see whether the actual distribution (from a sample) of a categorical variable mathes the expected distribution.

For example, is gender distribution 50%-50% in this data set?

H-null: p(women) = 0.5, p(men) = 0.5

H-alt: p(women) != 0.5, p(men) != 0.5

In [100]:
```
chisq.test(table(scores$gender), p = c(0.5,0.5))
```

```
	Chi-squared test for given probabilities

data:  table(scores$gender)
X-squared = 60.623, df = 1, p-value = 6.913e-15
```

Conclusion

X-squared(df=1) = 60.623, p-value = 0.00% < 5% (reject the null)

The gender distribution is not 50%-50%.

Two sample t-test

It is used to see whether there are group differences in population means between two groups.

For example, do men and women have different average SAT verbal scores?

H-null: Mean SATV for males = Mean SATV for females

H-alt: Mean SATV for males != Mean SATV for females

In [101]:
```r
#Write and run this command:

t.test(scores$SATV ~ scores$gender, var.eq = TRUE)
```

```
        Two Sample t-test

data:  scores$SATV by scores$gender
t = -0.49792, df = 698, p-value = 0.6187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.99137  13.09357
sample estimates:
mean in group Women   mean in group Men
           610.6645            615.1134
```

Conclusion

t(df=698) = -0.49792, p-value = 61.87% > 5% (fail to reject the null)

There was no statistically significant difference in average SAT verbal scores between men and women.

One-way ANOVA

It is used to determine whether there are group differences in numeric data between two or more groups.

For example, do SAT verbal scores significantly differ by educational levels (1= HS, 2= some college degree, 3 = 2-year college degree, 4= 4-year college degree, 5= graduate work, 6=professional degree)?

H-null: Mean SATV of students who have HS degree = Mean SATV of students who have some college degree = ...

H-alt: Mean SATV of students who have HS degree != Mean SATV of students who have some college degree != ...

In [102]:
```
ANOVA_SATV_ed <- aov(scores$SATV ~ scores$education)

summary(ANOVA_SATV_ed)
```

```
                  Df  Sum Sq Mean Sq F value Pr(>F)
scores$education   5   80754   16151   1.269  0.275
Residuals        694 8829392   12722
```

Interpreting the results

- df: degree of freedom
- sum sq: sum of squares
- mean sq: mean squares
- F value: computing statistics
- Pr(>F): statistical significance

Conclusion

F value(df Model=5, df Residuals=694) = 1.269, p-value = 27.5% > 5% (fail to reject the null)

There were no significant group differences in SAT verbal scores according to students' educational levels.

We do not have to run the post hoc tests because the group differences are not significant.

Extra information

There are different types of post hoc tests (https://www.rdocumentation.org/packages/DescTools/versions/0.99.36/topics/PostHocTest), but the Tukey's HSD is the most popular post hoc test for comparing multiple pairings.

In [103]:
```
# R command for Tukey's HSD:

TukeyHSD(aov(scores$SATV ~ scores$education, data = scores), conf.level=.95)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = scores$SATV ~ scores$education, data = scores)

$`scores$education`
           diff        lwr       upr     p adj
1-0 -16.8421053  -81.11952  47.43531 0.9756332
2-0 -40.4860447 -105.17035  24.19826 0.4737629
3-0  -4.3742265  -51.28442  42.53597 0.9998182
4-0   0.4405034  -50.31023  51.19124 1.0000000
5-0   4.8883912  -45.70428  55.48106 0.9997835
2-1 -23.6439394  -91.98228  44.69440 0.9215425
3-1  12.4678788  -39.36489  64.30064 0.9833397
4-1  17.2826087  -38.05008  72.61529 0.9482786
5-1  21.7304965  -33.45725  76.91824 0.8708907
3-2  36.1118182  -16.22468  88.44832 0.3596722
4-2  40.9265481  -14.87828  96.73138 0.2906084
5-2  45.3744358  -10.28669 101.03556 0.1835826
4-3   4.8147299  -28.81095  38.44041 0.9985288
5-3   9.2626177  -24.12403  42.64926 0.9687311
5-4   4.4478878  -34.14924  43.04502 0.9994867
```

Two-way ANOVA

It is used to determine whether there are group differences in numeric data between groups charachterized by two different categorical variables.

For example, do SAT verbal scores significantly differ by educational levels and gender?

H-null: Mean SATV of female students who have HS degree = Mean SATV of male students who have some college degree = ...

H-alt: Mean SATV of female students who have HS degree != Mean SATV of male students who have some college degree != ...

In [104]:
```
ANOVA_SATV_ed_g <- aov(scores$SATV ~ scores$education+scores$gender)

summary(ANOVA_SATV_ed_g)
```

```
                  Df  Sum Sq Mean Sq F value Pr(>F)
scores$education   5   80754   16151   1.269  0.276
scores$gender      1    6740    6740   0.529  0.467
Residuals        693 8822652   12731
```

Conclusion

- Education: F value(df Model=5, df Residuals=693) = 1.269, p-value = 27.6% > 5% (fail to reject the null)
- Gender: F value(df Model=1, df Residuals=693) = 0.529, p-value = 46.7% > 5% (fail to reject the null)

There were no significant group differences in SAT verbal scores according to students' educational levels or gender.

Chi-square test of independence

It is used to determine whether two categorical variables are dependent or independent.

For example, is gender independent of education levels?

H-null: Gender and education levels are independent

H-alt: Gender and education levels are dependent

In [105]:
```
chisq.test(table(scores$gender, scores$education))
```

```
	Pearson's Chi-squared test

data:  table(scores$gender, scores$education)
X-squared = 16.085, df = 5, p-value = 0.006605
```

Conclusion

X-squared(df=5) = 16.085, p-value = 0.006% < 5% (reject the null)

Gender and education levels are dependent.

Simple Linear Regression

It is used to identify a presense of a linear relationship between two quantitative variables.

For example, is there a linear relationship between SAT Verbal and SAT Quantitative scores?

H-null: there is no linear relationship between SAT Verbal and SAT Quantitative scores (Beta1 = 0)

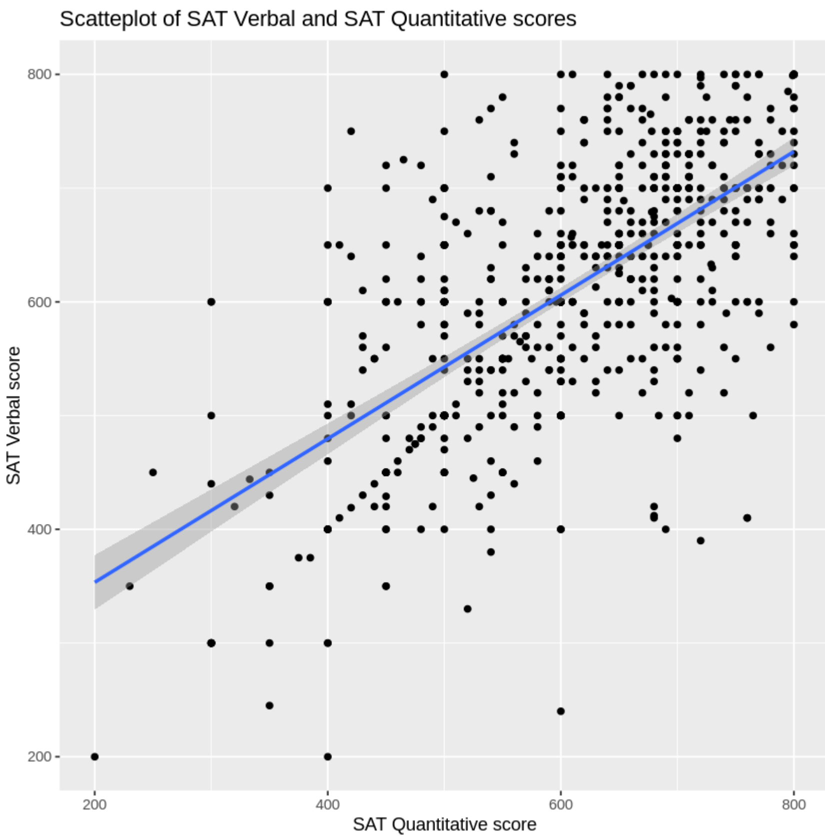
H-alt: there is a linear relationship between SAT Verbal and SAT Quantitative scores (Beta1 != 0)

In [106]:
```
# Plotting a scatterplot with a best-fit line

ggplot(scores, aes(x=SATQ, y=SATV)) +
  geom_point()+
  geom_smooth(method=lm)+
  labs(title="Scatteplot of SAT Verbal and SAT Quantitative scores",x="SAT Qua
ntitative score", y="SAT Verbal score")
```

```
`geom_smooth()` using formula 'y ~ x'

Warning message:
“Removed 13 rows containing non-finite values (stat_smooth).”
Warning message:
“Removed 13 rows containing missing values (geom_point).”
```



In [107]:
```
SATV_SATQ <- lm(scores$SATV ~ scores$SATQ)
summary(SATV_SATQ)
```

```
Call:
lm(formula = scores$SATV ~ scores$SATQ)

Residuals:
    Min      1Q  Median      3Q     Max 
-365.89  -50.57   -3.23   54.68  257.74 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 227.14322   17.77978   12.78   <2e-16 ***
scores$SATQ   0.63124    0.02863   22.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.71 on 685 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.4151,	Adjusted R-squared:  0.4143 
F-statistic: 486.2 on 1 and 685 DF,  p-value: < 2.2e-16
```

Conclusion

T-test for Beta1:

- t-stat(SATQ) = 22.05, p-value = 0.00% < 5% (reject the null)
- There is a linear relationship between SAT Quantitative scores and SAT Verbal scores.

ANOVA for Regression:

- F-stat(1,685) = 486.2, p-value = 0.00% < 5% (reject the null)
- The overall model is worthwhile.

Interpreting the results

- The estimated regression line equation: SATV = 227.14 + 0.63(SATQ). We would expect 0.63 points increase in SAT Verbal scores for every one point increase in SAT Quantitative score, assuming all the other variables are held constant.
- 41.51% of the variability in the SAT verbal scores was explained by the SAT quantitative scores.

Multiple Linear Regression

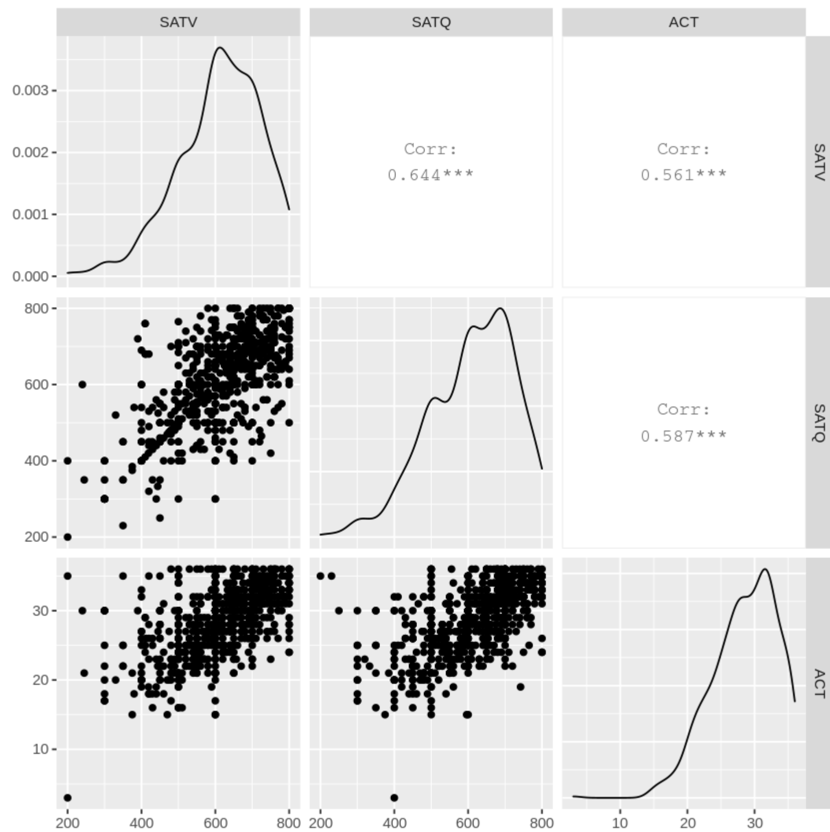
It is used to explain/predict one quantitative variable using multiple explanatory variables (one of which has to be quantitative).

For example, can you explain/predict SAT Verbal using SAT Quantitative scores and ACT scores?

In [108]:
```
# Making a correlation matrix for SATV,SATQ,ACT

scores %>%
  select(SATV,SATQ,ACT) %>%
  ggpairs(ggplot2::aes())
```

```
Warning message in ggally_statistic(data = data, mapping = mapping, na.rm = n
a.rm, :
“Removed 13 rows containing missing values”
Warning message:
“Removed 13 rows containing missing values (geom_point).”
Warning message:
“Removed 13 rows containing non-finite values (stat_density).”
Warning message in ggally_statistic(data = data, mapping = mapping, na.rm = n
a.rm, :
“Removed 13 rows containing missing values”
Warning message:
“Removed 13 rows containing missing values (geom_point).”
```



In [109]:
```
summary(lm(scores$SATV ~ scores$SATQ +  scores$ACT))
```

```
Call:
lm(formula = scores$SATV ~ scores$SATQ + scores$ACT)

Residuals:
    Min      1Q  Median      3Q     Max 
-376.97  -46.64    1.89   50.25  243.20 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 138.57027   20.25101   6.843 1.73e-11 ***
scores$SATQ   0.47130    0.03382  13.934  < 2e-16 ***
scores$ACT    6.52075    0.80963   8.054 3.56e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.93 on 684 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.4658,	Adjusted R-squared:  0.4642 
F-statistic: 298.2 on 2 and 684 DF,  p-value: < 2.2e-16
```

Conclusion

T-test for Beta1 (SATQ):

- t-stat(SATQ) = 13.93, p-value = 0.00% < 5% (reject the null) -> There is a linear relationship between SAT Quantitative scores and SAT Verbal scores.

T-test for Beta1 (ACT):

- t-stat(ACT) = 8.05, p-value = 0.00% < 5% (reject the null) -> There is a linear relationship between ACT scores and SAT Verbal scores.

ANOVA for Regression:

- F-stat(2,684) = 298.2, p-value = 0.00% < 5% (reject the null)
- The overall model is worthwhile.

Interpreting the results

- The estimated regression line equation: SATV = 138.57 + 0.47(SATQ)+6.52(ACT).
- 46.58% of the variability in the SAT verbal scores was explained by the SAT Quantitative and ACT scores. Adding ACT as explanatory variable increased R-square by 5.07%.

What if we wanted to include all of the variables in our dataset?

In [110]:
```
summary(lm(scores$SATV ~ scores$SATQ + scores$ACT + scores$age + scores$educat
ion + scores$gender))
```

```
Call:
lm(formula = scores$SATV ~ scores$SATQ + scores$ACT + scores$age + 
    scores$education + scores$gender)

Residuals:
    Min      1Q  Median      3Q     Max 
-385.57  -44.03    2.22   50.87  238.47 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       162.60918   24.05344   6.760 2.98e-11 ***
scores$SATQ         0.47611    0.03464  13.745  < 2e-16 ***
scores$ACT          6.60273    0.82472   8.006 5.17e-15 ***
scores$age         -0.80005    0.44468  -1.799   0.0724 .  
scores$education1  -8.55593   16.80347  -0.509   0.6108    
scores$education2  -9.22061   17.35465  -0.531   0.5954    
scores$education3  -4.78783   12.36804  -0.387   0.6988    
scores$education4  -0.42819   14.44457  -0.030   0.9764    
scores$education5   0.85816   15.42466   0.056   0.9556    
scores$genderMen  -16.55463    6.78574  -2.440   0.0150 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.66 on 677 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.4747,	Adjusted R-squared:  0.4677 
F-statistic: 67.98 on 9 and 677 DF,  p-value: < 2.2e-16
```

Conclusion

T-tests for Beta1:

- SATQ: p-value = 0.00% < 5% (reject the null) -> There is a linear relationship between SATQ and SATV scores.
- ACT: p-value = 0.00% < 5% (reject the null) -> There is a linear relationship between ACT and SATV scores.
- Age: p-value = 7.24% > 5% (fail to reject the null) -> There is no linear relationship between age and SATV scores.
- Education: p-value >50% (fail to reject the null) -> There is no linear relationship between education and SATV scores.
- Gender: p-value = 1.50% < 5 % (reject the null) -> There is a linear relationship between gender and SATV scores.

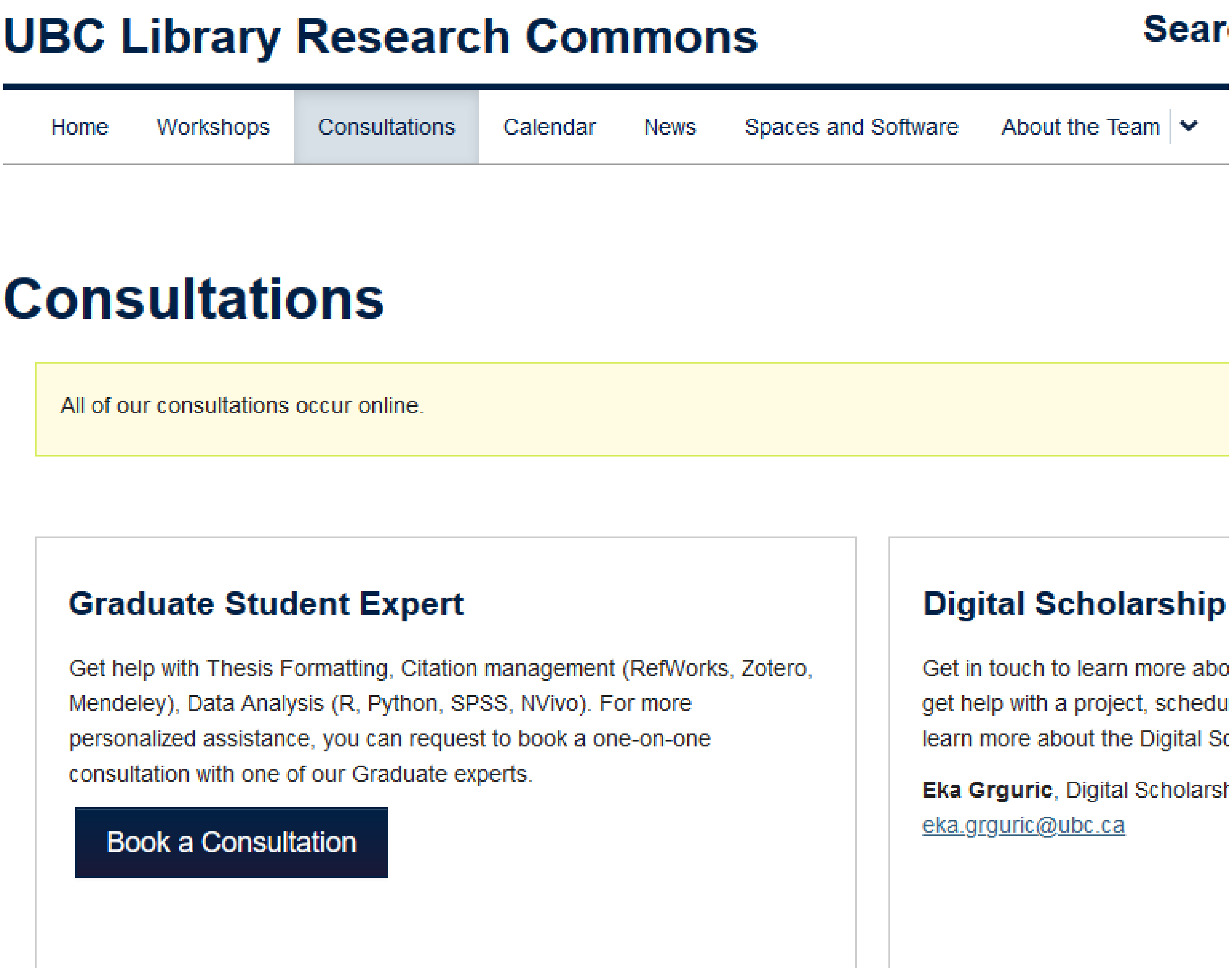
ANOVA for Regression:

- F-stat(9,677) = 67.98, p-value = 0.00% < 5% (reject the null)
- The overall model is worthwhile.

Interpreting the results

- The estimated regression line equation: SATV = 162.61 + 0.48(SATQ)+6.60(ACT)-16.55(Men).
- 46.77% of the variability in the SAT verbal scores was explained by the model as a whole.

Questions?



Reference(s);

Revelle, William, Wilt, Joshua, and Rosenthal, Allen. (2009). Personality and Cognition: The Personality-Cognition Link. In Gruszka, Alexandra and Matthews, Gerald and Szymura, Blazej (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer.