STAT011 Statistical Methods I

Lecture 6 Regression and Two-Way Tables

Lu Chen
Swarthmore College
2/7/2019

Review

- Relationships between variables
- Relationship between two quantitative variables
- Correlation coefficient r
  - Definition and formula
  - Examples
- Least squares regression
  - How to find the best fitting line
    - *Minimize the sum of squares of vertical distances from the points to the line*
  - Least squares regression in R
    - `lm()`, `predict()`

Outline

- Assessing least squares regression line
  - Coefficient of determination r^2
  - Residual plot
  - Transformation
- Relationship between two categorical variables
  - Two-way tables
  - Bar plot
  - Interpreting two-way tables
    - Joint distribution
    - Marginal distribution
    - Conditional distribution

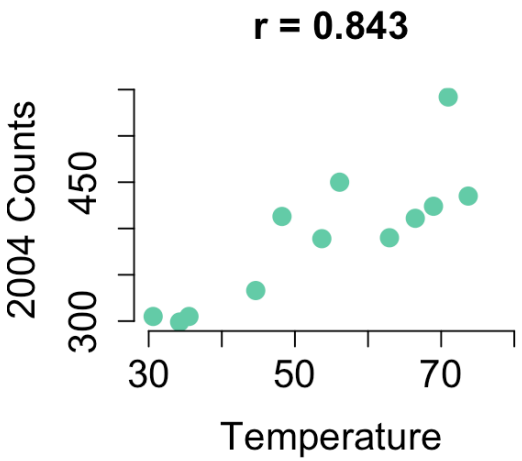
Correlation coefficient

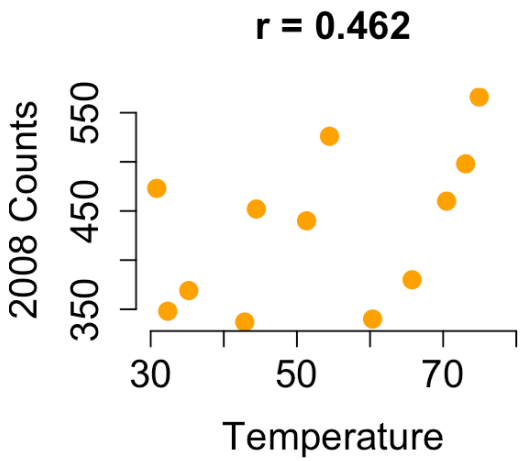
The **correlation** measures the *direction* and *strength* of the **linear relationship** between two quantitative variables. Correlation is usually written as ***r***.

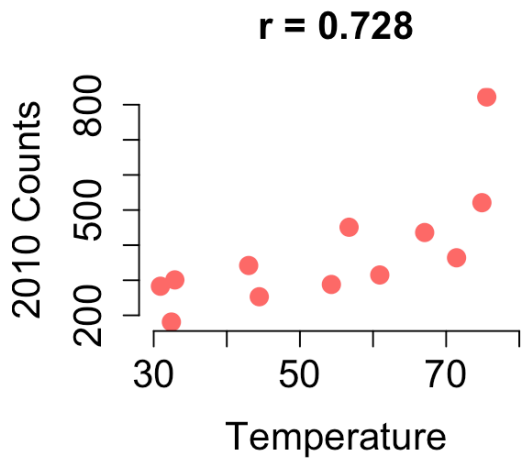
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

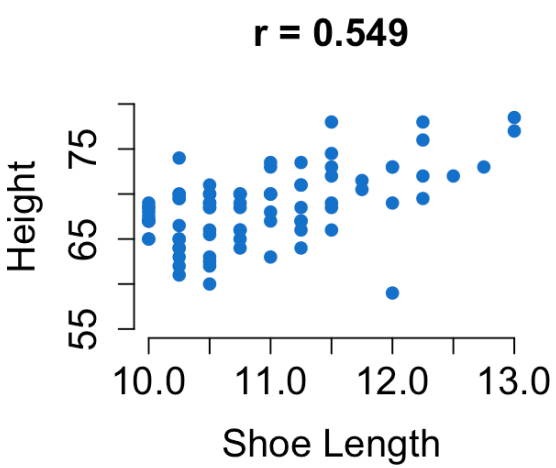
- $-1 \leq r \leq 1$
- $r > 0$: positive relationship
- $r < 0$: negative relationship
- $r = 0$: no relationship
- $r = \pm 1$: perfect relationship
  - For example: $y = 2x$

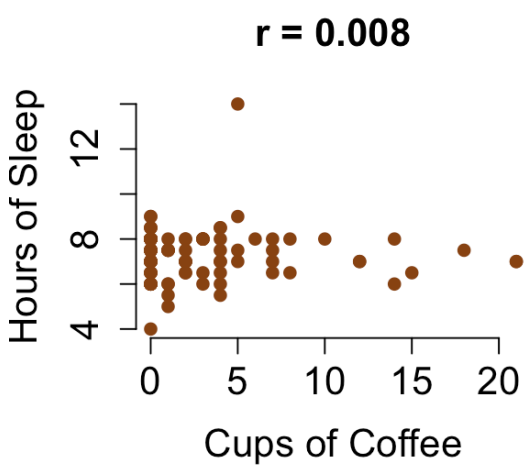
Correlation coefficient



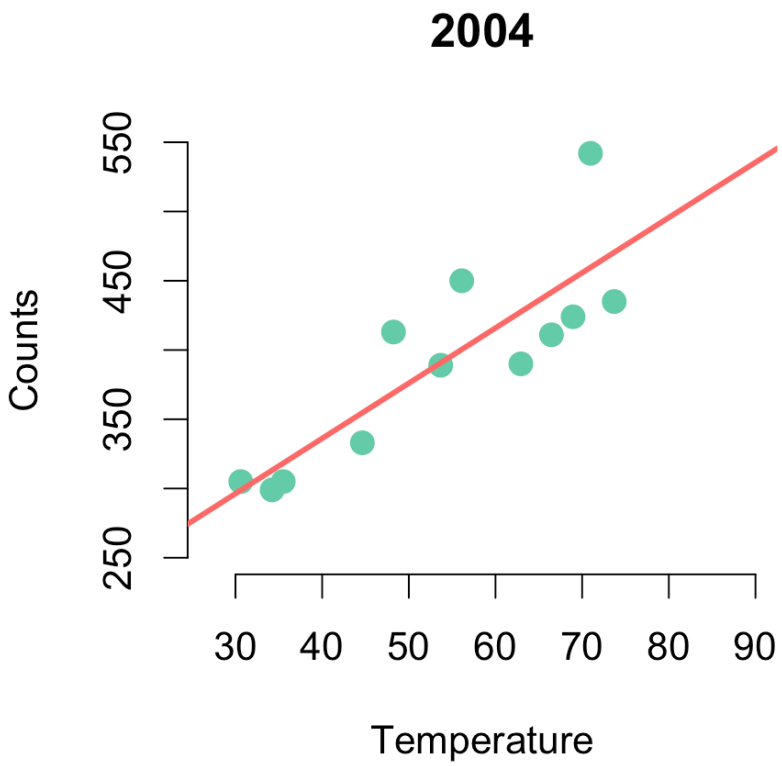








Least Square Regression (LSR)



$$\hat{y} = b_0 + b_1 x = 176.7 + 4.0x$$

Interpretation

- $b_0 = 176.7$: the predicted UFO *Count* is 176.7 when *Temperature* is 0.
  - b_0 is the **baseline** value.
  - Sometimes, value of b_0 does not have practical meaning.
- $b_1 = 4.0$: the predicted UFO *Count* increases 4 when *Temperature* increases 1 ° F.
  - b_1 measures the **rate of change**.

Least Square Regression (LSR)

> The **least-squares regression** line of y on x is the line that minimizes the sum of the squares of the vertical distances from the data points to the line.

$$\begin{aligned}
\text{Residual} &= \text{Observed } y - \text{Predicted } y \\
e &= y - \hat{y} \\
e_i &= y_i - \hat{y}_i \\
\text{To minimize } \sum(\text{residual})^2 &= \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2, \\
\text{we have slope } b_1 &= r\frac{s_y}{s_x} \text{ and intercept } b_0 = \bar{y} - b_1\bar{x}.
\end{aligned}$$

- How is correlation coefficient related to the least squares regression line?
- 1. r is related to the value of the slope b_1;
- 1. r^2 tells the success of the regression model.

Assessing LSR - Coefficient of determination

```r
# Square of correlation between Temperature and Count
cor(ufo2004$Temperature, ufo2004$Count)^2
```

```
## [1] 0.7104253
```

```r
# Variance of predicted y over Variance of observed y
m <- lm(Count ~ Temperature, data=ufo2004)
var(predict(m))/var(ufo2004$Count)
```

```
## [1] 0.7104253
```

- Correlation can be positive or negative. But ratio of SDs is always positive.
- **Coefficient of determination** r^2, the square of the correlation, is defined as

$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

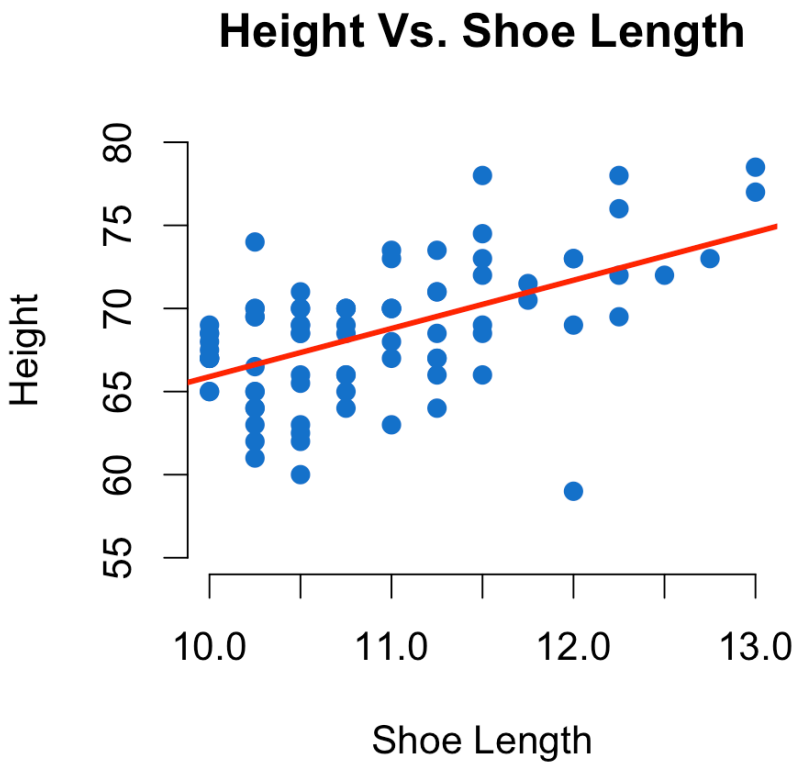
Assessing LSR - Coefficient of determination

Coefficient of determination (square of the correlation), r^2, is the **fraction of the variation** in the values of y that is explained by the least squares regression of y on x.

$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

- Residual $e = y - \hat{y}$
- Observed data $y = \hat{y} + e = b_0 + b_1 x + e$
- Variation in the observed data y comes from two parts:
  - The part of variation that *can* be explained by the regression line: $\hat{y} = b_0 + b_1 x$;
  - The part of variation that *cannot* be explained by the regression line: e.

Assessing LSR - Coefficient of determination



$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

- For a given *ShoeLength* value, there can be multiple values of *Height*. This means that the regression line cannot predict *Height* exactly. It only explains *Height* partially.
- r^2 : 30.2% of the variation in *Height* is explained by the least squares regression line that involves *ShoeLength*.
- r^2 is a measure of how successfully the regression line explains the reponse variable.

Assessing LSR - Coefficient of determination

$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

- In terms of value, r^2 is the square of r; they are essentially the same.
- In terms of meaning/interpreation, they are **different**.
  - The correlation r measures the direction and strength of a linear relationship. *No regression is involved.*
  - The coefficient of determination r^2 measures the fraction of variation in y explained by the least squares regression line $\hat{y}$. *It directly assesses a regression line.*

Assessing LSR - Residual plot

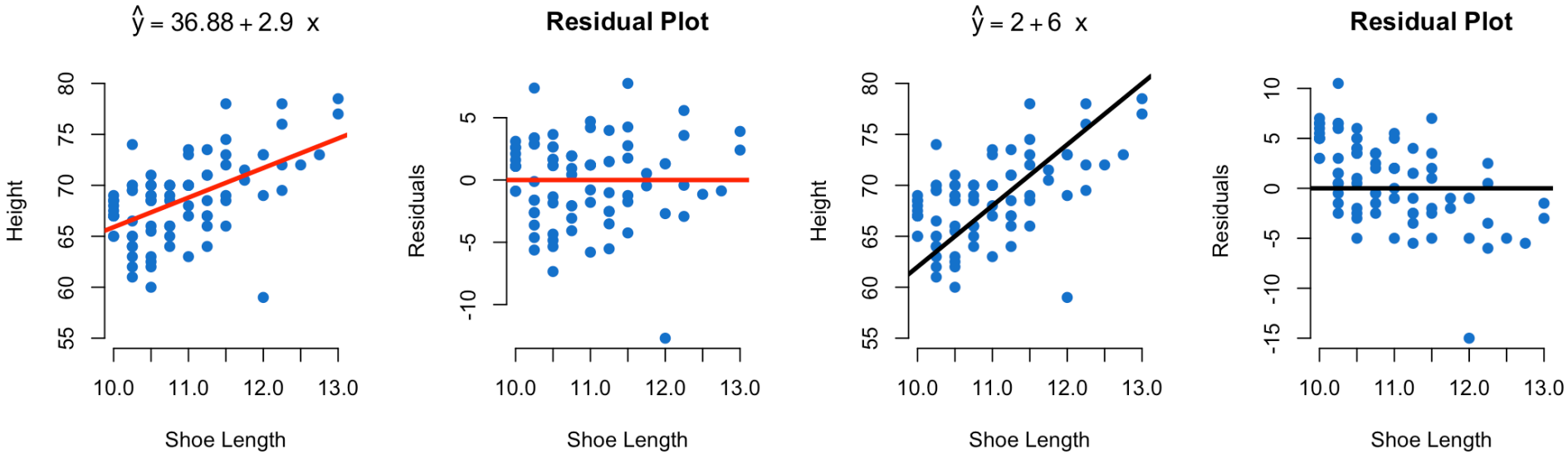
A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\text{Residual} = \text{Observed } y - \text{Predicted } y$$
$$e = y - \hat{y}$$

- Residuals can be either positive or negative.
- The sum and mean of the least-squares residuals is always zero.

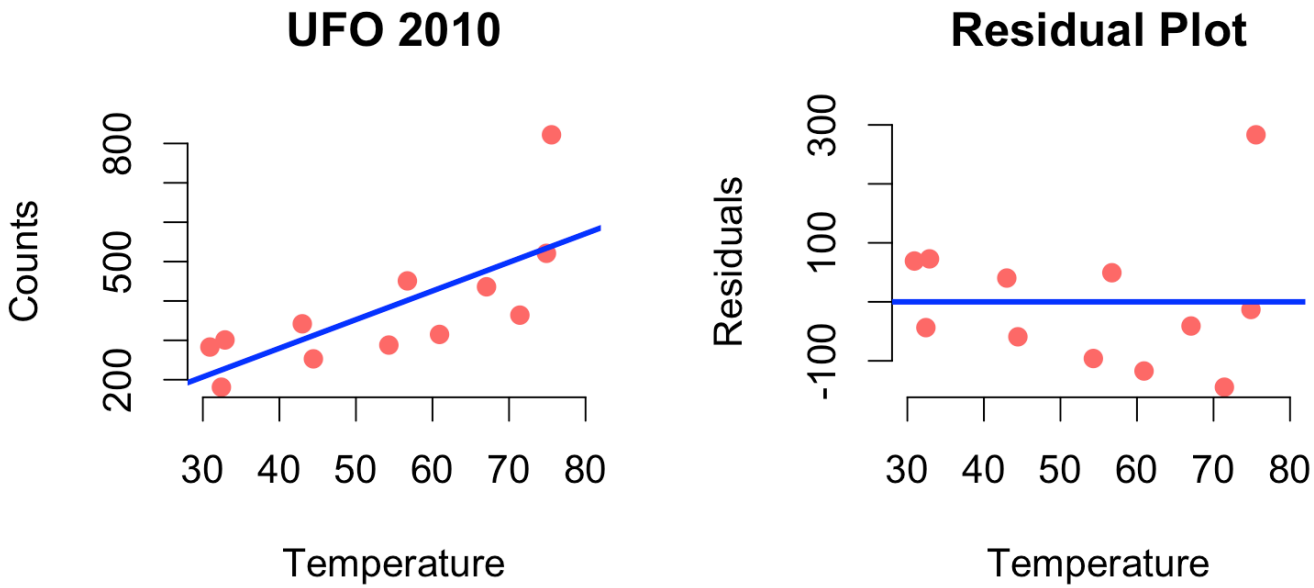
A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

Assessing LSR - Residual plot



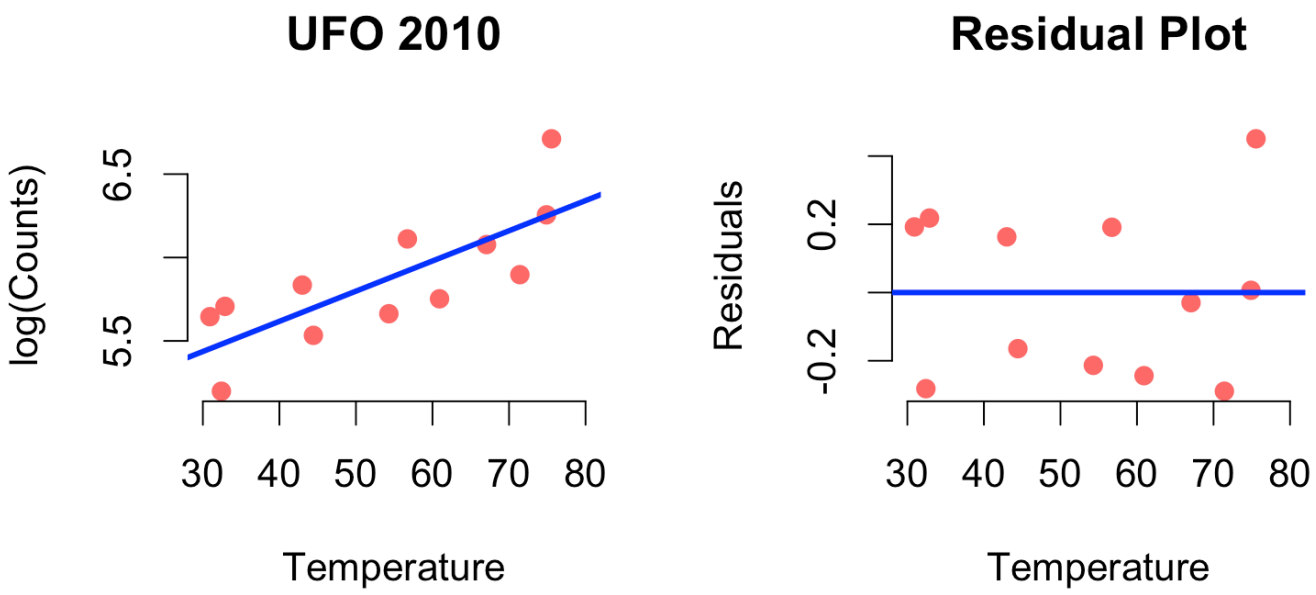
- If the regression line catches the overall pattern of the data, there should be *no pattern* in the residuals.
- If the residual plot shows any pattern, the regression line is NOT the best way to describing the data.

Assessing LSR - Residual plot



- This residual plot has some pattern (as *Temperature* increases, residual values descrease).
- When a residual plot displays any pattern, we may consider to **transform** the data or try some other regression methods.

Assessing LSR - Tranformation



- The response variable *Count* is tranformed by a natural logarithm function.
- In STAT 011, log() always denotes the natural logarithm function.

Assessing LSR in R

```r
# Coefficient of determination r^2
cor(X, Y)^2

# Residual plot
m <- lm(Y ~ X, data = ) # run a LSR model for Y on X
plot(X, m$residuals, xlab = , ylab = , main = , col = ) # plot residuals versus X
abline(h = 0, col = , lwd = ) # add a horizontal line y = 0

# log tranformation
log_Y <- log(Y)
m1 <- lm(log_Y ~ X, data = ) # Linear regression model
plot(X, log_Y, xlab = , ylab = , main = , col = ) # Scatterplot
abline(reg = m1, col = , lwd = ) # Add regression line
plot(X, m1$residuals, xlab = , ylab = , main = , col = ) # Residual plot
abline(h = 0, col = , lwd = ) # Add a horizontal line y = 0
```

Class Survey 2016

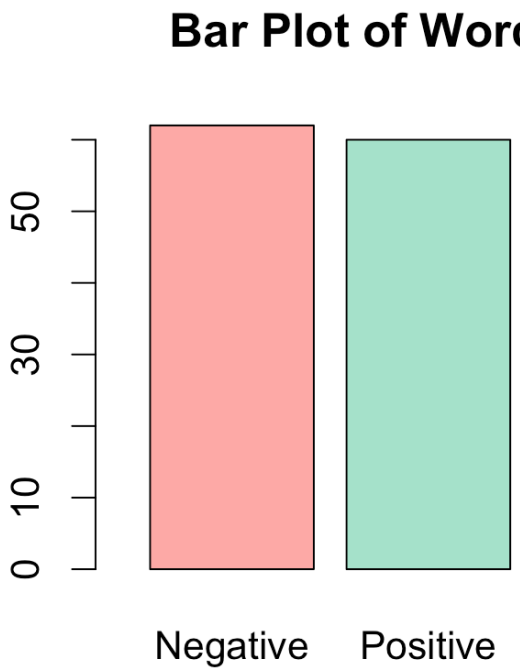
The last question has two versions. Half of the students received the first version, while the other half received the second.

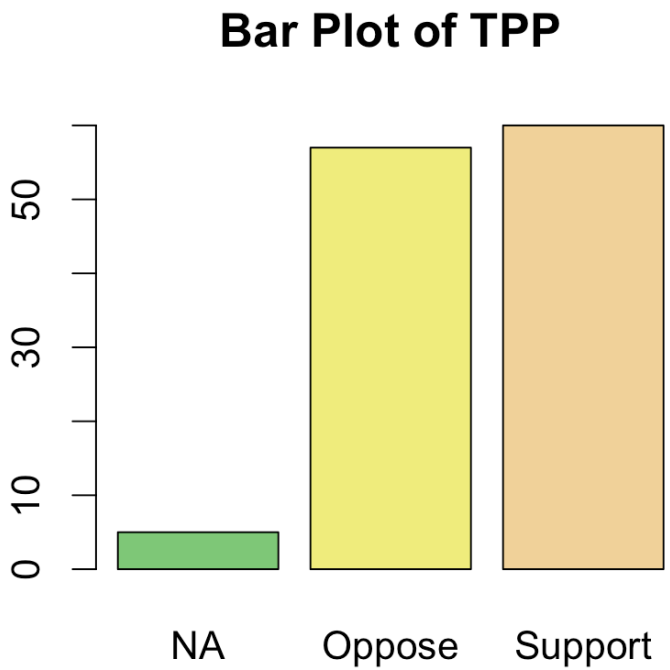
1. The Trans-Pacific Partnership (TPP) is a trade agreement among twelve Pacific Rim countries. Acording to the Obama administration, TPP will grow the American economy, support well-paying American jobs, and strengthen the American middle class. Do you SUPPORT or OPPOSE the TPP?
2. The Trans-Pacific Partnership (TPP) is a trade agreement among twelve Pacific Rim countries. Opponents of TPP say its regulations would undermine jobs in the US and work to the benefit of corporations rather than the 12 nations' workers. Do you SUPPORT or OPPOSE the TPP?

- To see how the wording of a survey question affects the respondents' answers, what are the two variables being studied in this problem?

Two categorical variables

2016 survey data, explore the two variables **separately**.





Two categorical variables

2016 survey data, explore the two variables **separately**.

Wording	Negative	Positive	Total
Count	62	60	122
Porportion	51%	49%	100%

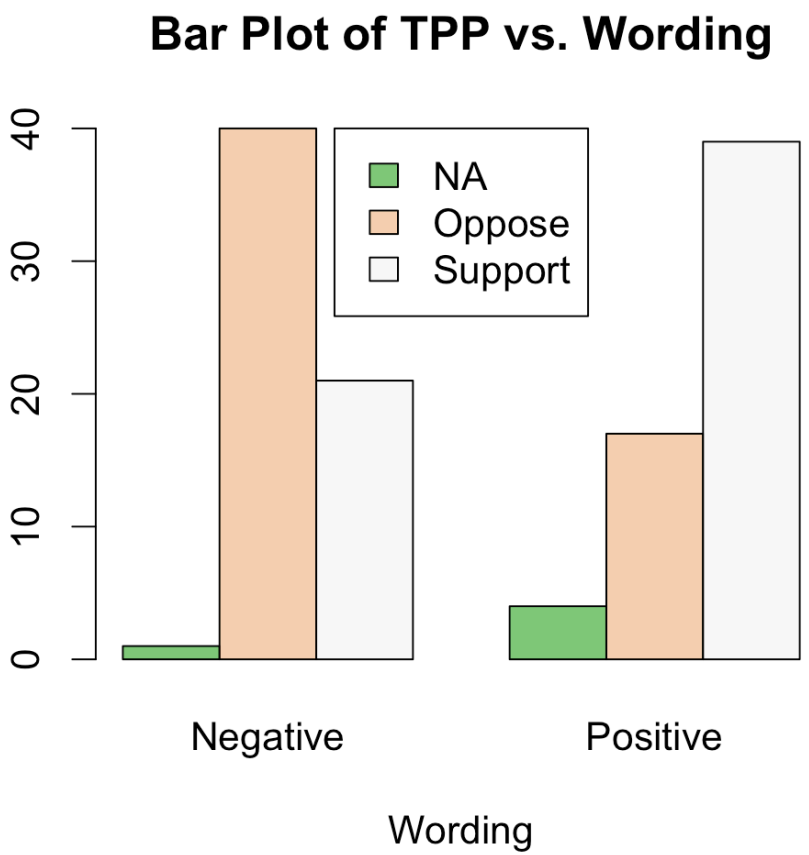
TPP	NA	Oppose	Support	Total
Count	5	57	60	122
Proportion	4%	47%	49%	100%

Relationship btw two categorical variables

Barplot

```
tab.tpp.wording <- table(Survey$TPP,
                         Survey$Wording)
tab.tpp.wording
barplot(tab.tpp.wording, beside=T,
        main="Bar Plot of TPP vs Wording",
        col= terrain.colors(3,alpha=0.6),
        xlab="Wording", legend.text = TRUE,
        args.legend = list(x = "top"))
```

```
##
##             Negative Positive
##    NA              1        4
##    Oppose         40       17
##    Support        21       39
```



Relationship btw two categorical variables

Two-way table

	Negative	Positive
NA	1	4
Oppose	40	17
Support	21	39

- Row variable: response variable (support or oppose TPP)
- Column variable: explanatory variable (wording of question)
- Cells: counts of observations taking the response variable value and the explanatory variable value
  - 40: there are 40 students who saw a negative wording of the question and decided to oppose TPP.
- R code for a two-way table: `table(Response, Explanatory)`
  - `table(Survey$TPP, Survey$Wording)`

Relationship btw two categorical variables

Usually for a **two-way table**, we include the **row totals**, **column totals** and **table total**.

	Negative	Positive	Total
NA	1	4	**5**
Oppose	40	17	**57**
Support	21	39	**60**
Total	**62**	**60**	**122**

- Row totals: distribution of the row (response) variable
- Column totals: distribution of the column (explanatory) variable
- Table total: total number of observations in the study

Joint distribution

$$\textbf{Joint distribution: cell proportion} = \frac{\text{cell count}}{\text{table total}}$$

```
prop.table(tab.tpp.wording)
```

	Negative	Positive
NA	1%	3%
Oppose	33%	14%
Support	17%	32%

Marginal distribution

Marginal distribution of row or column variable: $\frac{\text{row or column total}}{\text{table total}}$

```
prop.table(table(Survey$Wording))
prop.table(table(Survey$TPP))
```

Marginal distribution of Wording

	Negative	Positive	Total
Proportion	51%	49%	**100%**

Marginal distribution of TPP

	NA	Oppose	Support	Total
Proportion	4%	47%	49%	**100%**

Joint distribution and Marginal distribution

	Negative	Positive	Total
NA	1%	3%	**4%**
Oppose	33%	14%	**47%**
Support	17%	32%	**49%**
Total	**51%**	**49%**	**100%**

- 51% of the students saw negative wording of the question: 1% did not respond; 33% opposed TPP; 17% supported TPP. 49% of the students saw positive wording of the question: 3% did not respond; 14% opposed TPP; 32% supported TPP. In total, 4% did not respond; 47% opposed TPP; 49% supported TPP.

Conditional distribution

$$\textbf{Conditional distribution} = \frac{\text{cell count}}{\text{column total}}$$

	Negative	Positive	Total
NA	1	4	**5**
Oppose	40	17	**57**
Support	21	39	**60**
Total	**62**	**60**	**122**

- For example. Among the students who saw negative wording, $\frac{40}{62} = 65\%$ opposed TPP. Among the students who saw positive wording, $\frac{17}{60} = 28\%$ opposed TPP.

Conditional distribution

```
prop.table(tab.tpp.wording, margin = 2)
```

	Negative	Positive
NA	2%	7%
Oppose	65%	28%
Support	34%	65%
Total	100%	100%

- Negative wording results in more opposition of TPP than positive wording:
  - Condtional on negative wording, the proportion of opposition is 65%.
  - Condtional on positive wording, the proportion of opposition is 28%.
- Positive wording results in more support of TPP than negative wording:
  - Given negative wording, the proportion of support is 34%.
  - Given positive wording, the proportion of support is 65%.

Class Survey 2019

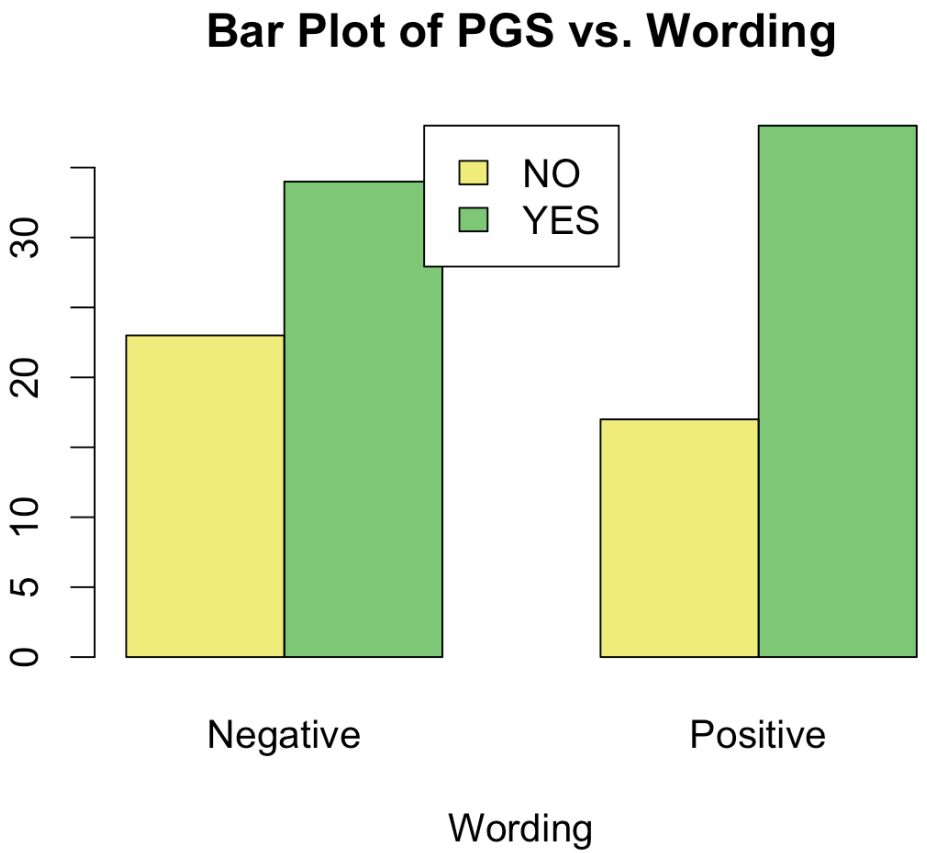
In this year's survey, you were also asked a question that had two versions of wording.

1. Personal Genome Services like *23andMe* provide ancestry information beyond what you can learn from relatives or from historical documentation. DNA variations can provide clues about where a person's ancestors might have come from. Would you consider using a Personal Genome Service?
2. Personal Genome Services like *23andMe* provide ancestry information to users who provide a DNA sample. There is currently little oversight or regulation of testing companies and there is concern people's genetic privacy is being compromised. Would you consider using a Personal Genome Service?

What are the two variables?

Personal Genome Services (*PGS*) and *Wording*

PGS	Negative	Positive	Total
NO	23	17	40
YES	34	38	72
Total	57	55	112



Personal Genome Services (*PGS*) and *Wording*

Joint and marginal distribution

PGS	Negative	Positive	Total
NO	21%	15%	36%
YES	30%	34%	64%
Total	51%	49%	100%

Conditional distribution

PGS	Negative	Positive
NO	40%	31%
YES	60%	69%
Total	100%	100%

Summary

- Assessing least squares regression line
  - Coefficient of determination $r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$
  - Residual plot
  - Transformation
- Relationship between two categorical variables
  - Two-way tables `table(Response, Explanatory)`
  - Bar plot `barplot()`
  - Interpreting two-way tables
    - Joint distribution `prop.table(two-way table)`
    - Marginal distribution `prop.table(table of each variable)`
    - Conditional distribution `prop.table(two-way table, margin = 2)`