Math 107

Exploratory Data Analysis: Two Variables
(Sections 2.1, 2.4, 2.5)

Two Categorical Variables

Two-Way Tables

A **two-way table** shows how the values are distributed along each variable, contingent on the value of the other variable.

Example:

910 randomly sampled voters in Tampa, FL were asked if they supported the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children.

DREAM Act

		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

DREAM Act

		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

What proportion of these voters supported the DREAM Act?

DREAM Act

		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

What proportion of conservatives supported the dream act?

DREAM Act

		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

What proportion of moderates supported the dream act?

DREAM Act

		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

What proportion of liberals supported the dream act?

DREAM Act

What is the difference in the proportion of conservatives and liberals that support the DREAM Act?

DREAM Act

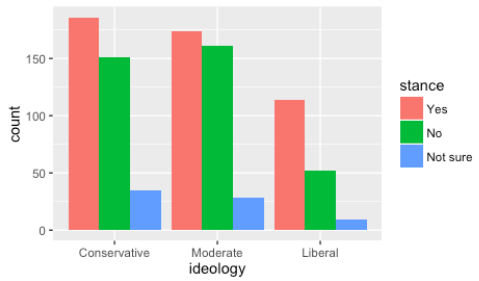
		Ideology			
		Conservative	Moderate	Liberal	Total
Response	Yes	186	174	114	474
	No	151	161	52	364
	Not sure	35	28	9	72
	Total	372	363	175	910

What proportion the voters that do not support the DREAM Act in this sample are conservatives?

```
# Load the data set
dream <- read.csv("dream.csv")


# Obtaining counts by stance & ideology pair
library(dplyr)
ctbl <-
  dream %>%
  group_by(stance, ideology) %>%
  summarise(count = n())


# Making the two-way table (this is just formatting)
library(tidyr)
two_way_tbl <-
  ctbl %>%
  spread(key = ideology, value = count)
```

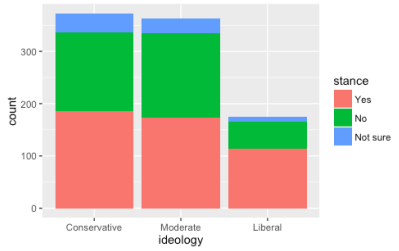
Side-by-Side Bar Chart

In a **side-by-side bar chart**, the height of each bar is equal to the corresponding cell in the two-way table.



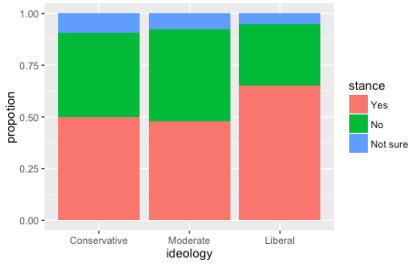
Segmented Bar Chart

A **segmented bar chart** (using frequencies) is like a side-by-side bar chart, but the bars are stacked.



Segmented Bar Chart

A **segmented bar chart** (using relative frequencies) divides the bar into segments corresponding to the proportion of the sample in each “level” of a variable.




```
# Load ggplot2
library(ggplot2)

# Side-by-side bar chart
ggplot(data = dream) +
  geom_bar(mapping = aes(x = ideology, fill = stance),
           position = "dodge")

# Stacked bar chart using counts
ggplot(data = dream) +
  geom_bar(mapping = aes(x = ideology, fill = stance))

# Stacked bar chart using proportions
ggplot(data = dream) +
  geom_bar(mapping = aes(x = ideology, fill = stance),
           position = "fill") +
  ylab("propotion")
```

Simpson's Paradox

Surgeries in Springfield

Springfield has two doctors: Dr. Hibbert and Dr. Nick.

Dr. Hibbert is a highly respected medical professional, while Dr. Nick appears on infomercials on late night TV claiming that he has the best overall surgical success rate in Springfield.

Let's explore the data a bit...

Surgeries in Springfield

Dr. Hibbert

	Heart	Bandaid
Success	70	10
Failure	20	0

Dr. Nick

	Heart	Bandaid
Success	2	81
Failure	8	9

Calculate the overall success rate for each doctor.

Surgeries in Springfield

Dr. Hibbert

	Heart	Bandaid
Success	70	10
Failure	20	0

Dr. Nick

	Heart	Bandaid
Success	2	81
Failure	8	9

Calculate the success rate for each surgery, for each doctor. How do these compare to the overall success rates?

Simpson's Paradox

Simpson's Paradox:

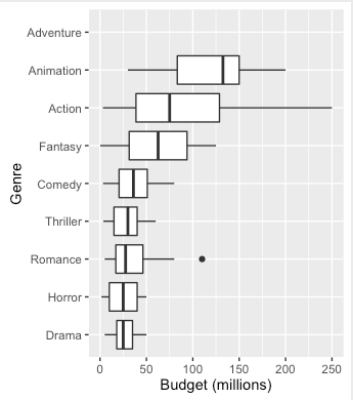
An observed relationship between two variables can change, or even reverse, when a third variable is considered.

Lesson: be careful averaging over the levels of a variable, the overall average may be misleading.

One Categorical and One Quantitative Variable

Audience Scores 2011

Budgets for all movies that came out in 2011 can be displayed using **side-by-side boxplots** (right). How do the budgets differ by genre?




```
# Basic side-by-side boxplots
ggplot(data = HollywoodMovies2011) +
  geom_boxplot(mapping = aes(x = Genre, y =
                             RottenTomatoes))


# You can also order them by medians and flip the
coordinates
ggplot(data = HollywoodMovies2011) +
  geom_boxplot(mapping = aes(x = reorder(Genre, Budget,
               median, na.rm = TRUE), y = Budget)) +
  coord_flip() +
  xlab("Genre") +
  ylab("Budget (millions)")
```

Statistics by Group

After looking at the side-by-side boxplots for the budget data, we may wish to see the summary statistics broken down by genre.

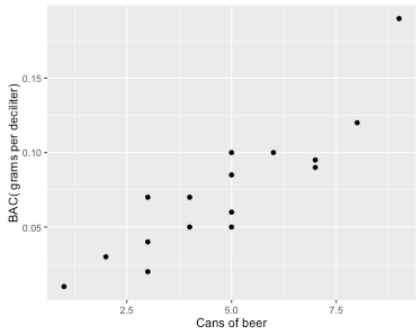
Genre	min	Q1	median	Q3	max	mean	sd	n	missing
Action	3.0	38.75	75.0	128.75	250	89.62500	61.53559	32	0
Adventure	*NA*	*NA*	*NA*	*NA*	*NA*	NaN	*NA*	1	1
Animation	30.0	83.25	132.5	150.00	200	114.91667	51.11922	12	0
Comedy	3.2	20.50	36.0	50.95	80	38.50370	23.28754	27	0
Drama	5.0	18.00	25.0	35.00	50	25.33333	12.89315	21	0
Fantasy	0.2	31.40	62.6	93.80	125	62.60000	88.24693	2	0
Horror	1.5	10.00	25.0	40.00	50	25.79412	16.65932	17	0
Romance	5.0	17.00	27.5	46.25	110	38.40000	33.22048	11	1
Thriller	3.5	15.00	30.0	40.00	60	30.78462	18.01975	13	0

```
HollywoodMovies2011 %>%
  group_by(Genre) %>%
  summarise(min = min(Budget, na.rm = TRUE),
            Q1 = quantile(Budget, .25, na.rm = TRUE),
            median = median(Budget, na.rm = TRUE),
            Q3 = quantile(Budget, .75, na.rm = TRUE),
            max = max(Budget, na.rm = TRUE),
            mean = mean(Budget, na.rm = TRUE),
            sd = sd(Budget, na.rm = TRUE),
            n = n(),
            missing = sum(is.na(Budget)))
```

Two Quantitative Variables

Scatterplot

A **scatterplot** is a graph of the relationship between two quantitative variables



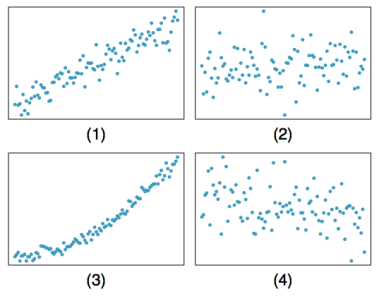
```
# Load the data
bac <- read.table("bac.txt", header = TRUE)
head(bac)

# Create a scatterplot
ggplot(data = bac) +
  geom_point(mapping = aes(x = Beers, y = BAC)) +
  xlab("Cans of beer") +
  ylab("BAC( grams per deciliter)")
```

Scatterplot

From a scatterplot we can learn about the following:

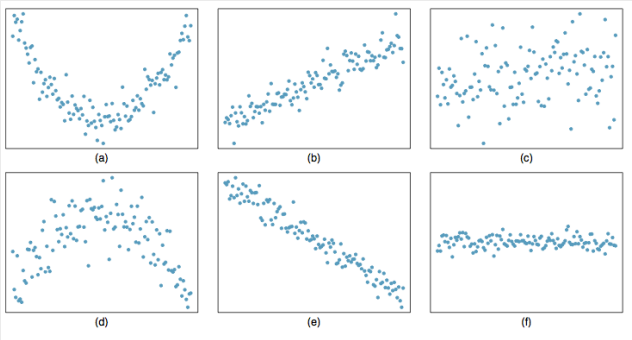
- **Direction** of association
- **Strength** of the association
- **Form/trend**
- **Outliers**

Your Turn



- Indicate whether each plot shows: (a) positive association; (b) negative association; (c) no association.
- If positive/negative, is the association linear?

Your Turn



For each of the plots, identify the strength of association (weak, moderate, or strong).

Correlation

Correlation is a numerical measure of the **strength** and **direction** of a the **linear** association between two *quantitative* variables.

Notation:

sample correlation:

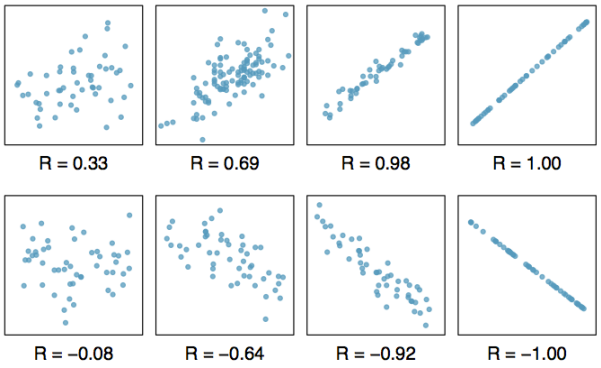
population correlation:

In R: `cor(bac$BAC, bac$Beers)`

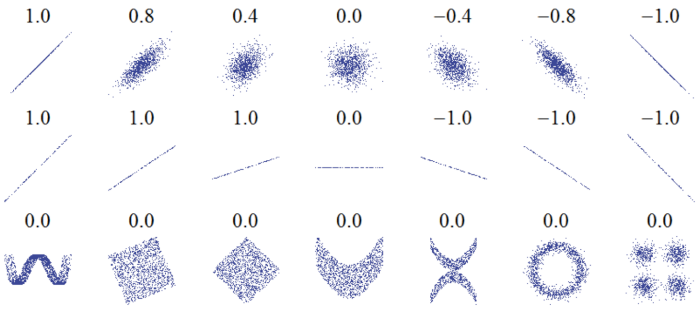
Properties

1. $-1 \leq r \leq 1$
2. The sign indicates the direction of the linear association.
   - $r > 0 \rightarrow$ positive association
   - $r < 0 \rightarrow$ negative association
   - $r = 0 \rightarrow$ no **linear** association
3. The closer r is to ±1, the stronger the association
4. r has no units
5. r does not depend on the units of measurement
6. The correlation between x and y is the same as the correlation between y and x

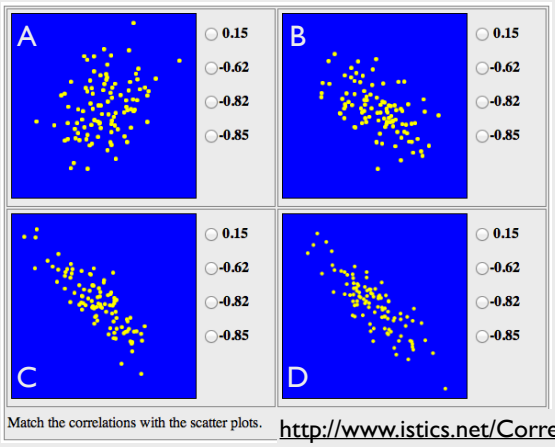
Correlation



Correlation

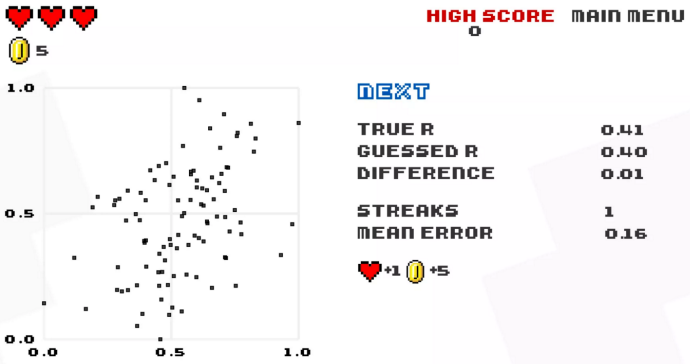


Your Turn



http://www.istics.net/Correlations/

Guess the Correlation



http://guessthecorrelation.com/

Cautions

1. Correlation can be heavily influenced by outliers. Don't just look at the correlation! **Always plot your data!**
2. r = 0 indicates that there is no *linear* association between the two variables, but the variables could still be associated! **Always plot your data!**
3. Correlation does not imply causation! Remember to **think!**

Recap

Variables	Statistical Graphic	Summary Statistics
Categorical	bar chart, pie chart	frequency table, relative frequency table, proportion
Quantitative	dotplot, histogram, boxplot	mean, median, max, min, standard deviation, range, IQR, five number summary
Categorical vs Categorical	side-by-side bar chart, segmented bar chart	two-way table, difference in proportions
Quantitative vs Categorical	side-by-side boxplots	statistics by group, difference in means
Quantitative vs Quantitative	scatterplot	correlation