

Lecture 05: Relationships between 2 variables

Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Lecture 05: Relationships between 2 variables

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Recap of chapters 1 and 2

Mostly looking at a single variable:

- ▶ Graphs to explore the distribution of single variables (histograms, bar charts)
- ▶ Summary numbers to describe our distributions:
 - ▶ Measures of central tendency (mean, median)
 - ▶ Measures of spread (standard deviation, IQR)

One example of two variables:

- ▶ Time plots to examine what happens to a variable over time

Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Learning objectives for today

Lecture 05: Relationships between 2 variables

- ▶ Explore the relationship between two quantitative variables
 - ▶ Directionality
 - ▶ Association vs causation
- ▶ Make scatter plots to look at relationships visually
 - ▶ using `geom_point()`
- ▶ Use the correlation coefficient to quantify the strength of linear relationships
 - ▶ calculate correlations using `cor()`

Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Relationships between two quantitative variables

**Relationships between two
quantitative variables**

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Explanatory (X) and response (Y) variables

Bi-directional:

- ▶ “X predicts Y”, or “Y predicts X”
- ▶ “X is associated with Y”, or “Y is associated with X”

Unidirectional:

- ▶ “X causes Y”

Which variable is x and which is y?

In prediction we generally use X to denote the variable we are using to predict the variable of interest (Y)

In causation we generally use X to denote the explanatory (independent) and Y to denote the response (dependent)

Graphically the X variable is on the X (horizontal) axis and the Y variable is the Y (vertical) axis

Which variable is x and which is y?

1. Each hospital's rate of hospital-acquired infections, and whether the hospital has implemented a hand-washing intervention as part of a cluster randomized trial.
2. The weight in kilograms and height in centimeters of a person
3. Inches of rain in the growing season and the yield of corn in bushels per day
4. A person's leg length and arm length, in centimeters

How to investigate causation?

- ▶ Randomized controlled trials (RCTs) to randomize individuals to different levels
- ▶ Observational study that is *designed* to investigate causation and reduce the risk of bias

Relationships between two
quantitative variables

**Looking at relationships
visually: Scatterplots**

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Looking at relationships visually: Scatterplots

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

- ▶ Scatterplots are a good way to visualize a relationship between two variables
- ▶ When we look at a scatterplot we want to evaluate:
 - ▶ The overall Pattern of the dots
 - ▶ Any notable exceptions to the pattern
 - ▶ Direction (positive or negative)
 - ▶ Form (straight line or curved)
 - ▶ Strength (how closely the points follow a line)
 - ▶ Are there any obvious outliers

Scatterplot Syntax in R

Relationships between two
quantitative variables

**Looking at relationships
visually: Scatterplots**

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
name of plot <- ggplot(data = dataset, aes(x = xvariable, y = yvariable)) +  
geom_point(na.rm=TRUE) + theme_minimal(base_size = 15)+  
labs(x = "xlabel", y = "ylabel", title = "Title")
```

Bi-directional relationships ex: systolic and diastolic BP

Read in NHANES dataset

```
nhanes_dataNA <- read_csv("nhanes.csv")  
nhanes_data <- nhanes_dataNA[rowSums(is.na(nhanes_dataNA[, 15:18])) == 0, ]  
names(nhanes_data)
```

```
## [1] "ridageyr" "agegroup" "gender" "military" "born" "citizen"  
## [7] "drinks" "drinkscat" "bmxt" "bmxt" "bmxbmi" "bmicat"  
## [13] "bpxpls" "bpxsy1" "bpxsy2" "sys1d" "sys2d" "bpxdi1"  
## [19] "bpxdi2" "dias1d" "dias2d" "bpcat" "chest" "fs1"  
## [25] "fs2" "fs3" "lbdhdd" "hdlcat" "highhdl" "hi"  
## [31] "asthma" "vwa" "vra" "va" "aspirin" "sleep"  
## [37] "is" "hs" "lbdldl" "highldl"
```

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Bi-directional relationships ex: systolic and diastolic BP

Relationships between two
quantitative variables

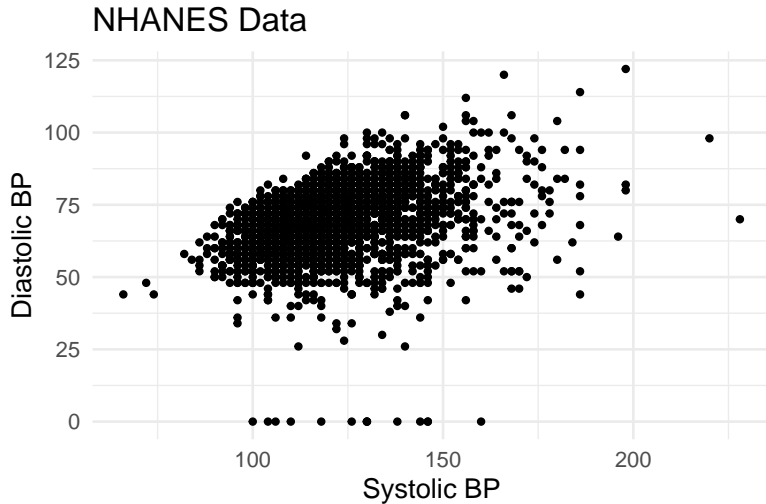
Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
nhanes_scatter <- ggplot(data = nhanes_data, aes(x = bpxsy1, y = bpxdi1)) +  
  geom_point(na.rm=TRUE) + theme_minimal(base_size = 15) +  
  labs(x = "Systolic BP",  
       y = "Diastolic BP",  
       title = "NHANES Data")
```

Bi-directional relationships ex: systolic and diastolic BP



Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Bi-directional relationships ex: systolic and diastolic BP

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

What do we notice from the plot?

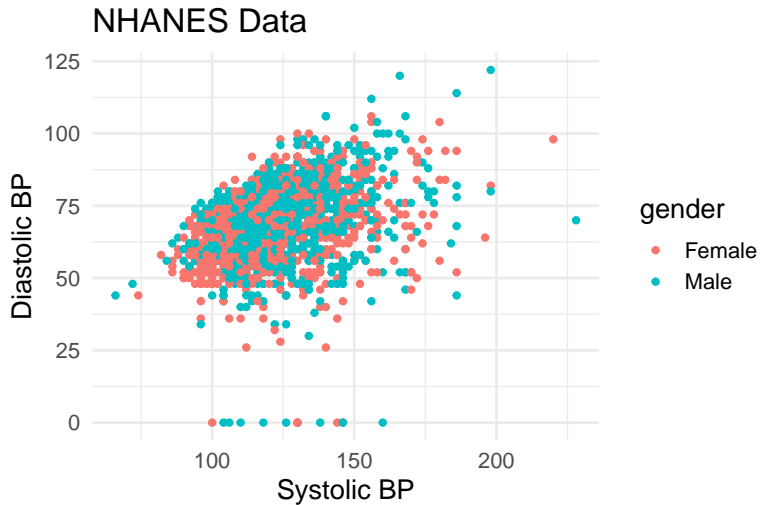
- ▶ Is there a visible association?
- ▶ Any notable exceptions to the pattern
- ▶ Direction (positive or negative)
- ▶ Form (straight line or curved)
- ▶ Strength (how closely the points follow a line)
- ▶ Are there any obvious outliers

Bi-directional relationships ex: systolic and diastolic BP

We can add a third variable to our graph by coloring the dots

```
nhanes_scatter <- ggplot(data = nhanes_data, aes(x = bpxsy1, y = bpxd11)) +  
  geom_point(aes(col=gender), na.rm=TRUE) + theme_minimal(base_size = 15)+  
  labs(x = "Systolic BP",  
       y = "Diastolic BP",  
       title = "NHANES Data")
```

Bi-directional relationships ex: systolic and diastolic BP



Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Association with a plausible direction

Manatee data set from your textbook:

```
mana_data <- read_csv("Ch03_Manatee-deaths.csv")  
head(mana_data)
```

```
## # A tibble: 6 x 3  
##   year powerboats deaths  
##   <dbl>     <dbl>   <dbl>  
## 1  1977         447     13  
## 2  1987         645     39  
## 3  1997         755     54  
## 4  2007        1027     73  
## 5  1978         460     21  
## 6  1988         675     43
```

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Power boats and Manatees

Lecture 05: Relationships between 2 variables

Relationships between two
quantitative variables

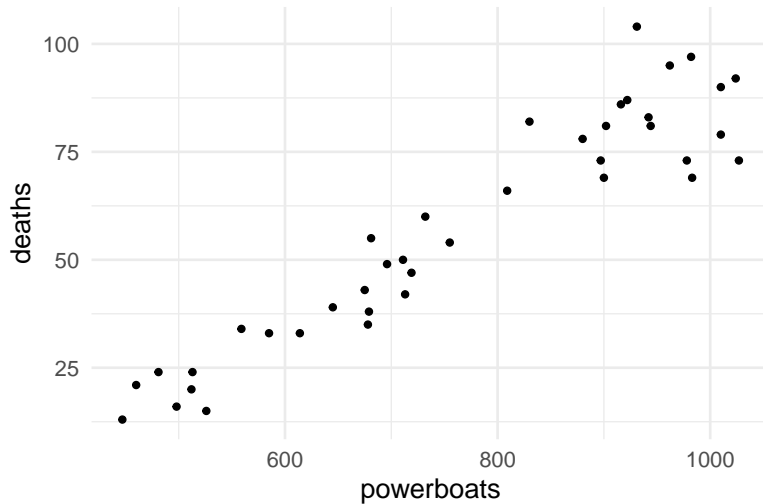
Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
mana_scatter <- ggplot(data = mana_data, aes(x = powerboats, y = deaths)) +  
  geom_point() + theme_minimal(base_size = 15)
```

Power boats and Manatees



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

What do we notice from the plot?

- ▶ Is there a visible association?
- ▶ Any notable exceptions to the pattern
- ▶ Direction (positive or negative)
- ▶ Form (straight line or curved)
- ▶ Strength (how closely the points follow a line)
- ▶ Are there any obvious outliers

Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Power boats and Manatees

Lecture 05: Relationships between 2 variables

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

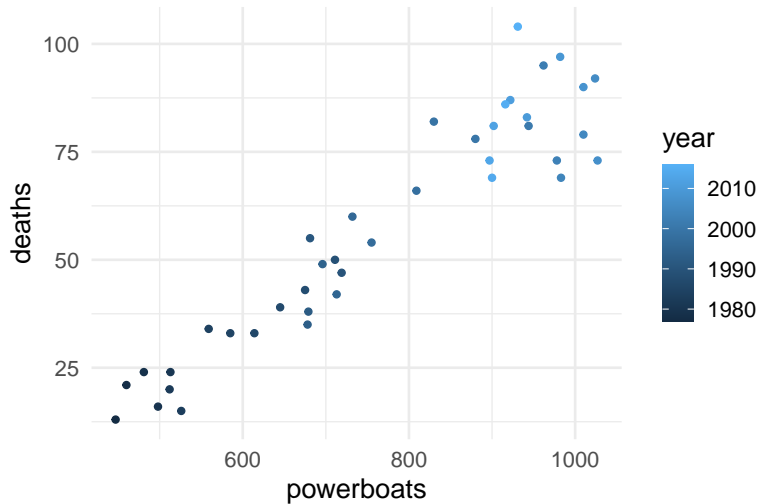
Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

What if we layer in a continuous third variable?

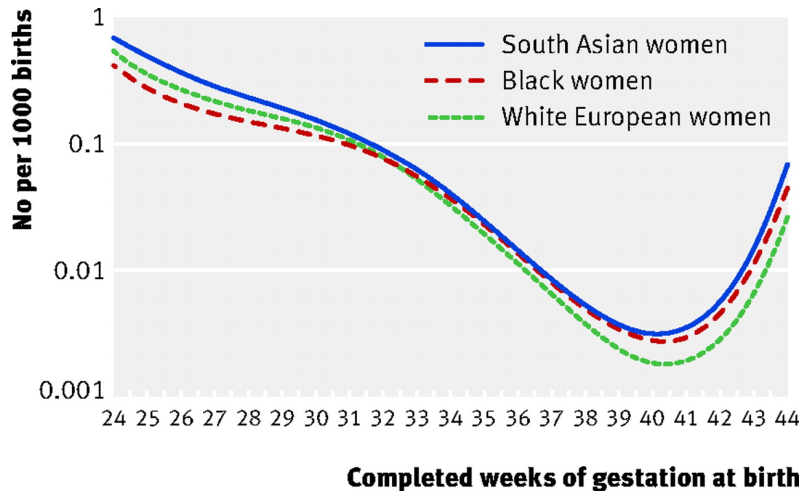
```
mana_scatter <- ggplot(data = mana_data, aes(x = powerboats, y = deaths)) +  
  geom_point(aes(col=year)) + theme_minimal(base_size = 15)
```

Power boats and Manatees



A non-linear example

Gestational age and perinatal mortality



Source: Balchin et al. BMJ. 2007.

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Exploratory analysis using scatterplots

Lean body mass and metabolic rate: Problem and Plan

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Problem: Is lean body mass (person's weight after removing the fat) associated with metabolic rate (kilocalories burned in 24 hours)?

Plan: A diet study was conducted on 12 women and 7 men that measured lean body weight and metabolic rate for each individual.

Lean body mass and metabolic rate: DATA

Data: In the textbook

Subject	Sex	Mass (kg)	Rate (Cal)	Subject	Sex	Mass (kg)	Rate (Cal)
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

What would the corresponding data frame look like? How many variables would it have? How many rows?

Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Lean body mass and metabolic rate: DATA

Note: you won't be tested on writing code using tibble::tribble()

Do be able to look at the code and recognize that it is creating a data set

```
weight_data <- tibble::tribble(  
  ~subject, ~gender, ~mass, ~rate,  
  1, "M", 62.0, 1792,  
  2, "M", 62.9, 1666,  
  3, "F", 36.1, 995,  
  4, "F", 54.6, 1425,  
  5, "F", 48.5, 1396,  
  6, "F", 42.0, 1418,  
  7, "M", 47.4, 1362,  
  8, "F", 50.6, 1502,  
  9, "F", 42.0, 1256,  
  10, "M", 48.7, 1614,  
  11, "F", 40.3, 1189,  
  12, "F", 33.1, 913
```

Lean body mass and metabolic rate: Analysis

Lecture 05: Relationships between 2 variables

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

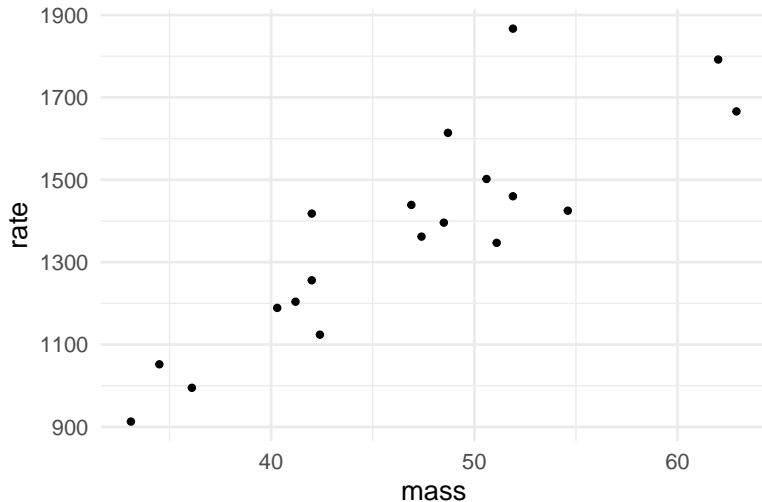
**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Exploratory data analysis using scatter plots

```
weight_scatter <- ggplot(weight_data, aes(x = mass, y = rate)) +  
  geom_point() +  
  theme_minimal(base_size = 15)
```

Lean body mass and metabolic rate: Analysis



Relationships between two quantitative variables

Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

Analysis: Colour the points by gender

Relationships between two
quantitative variables

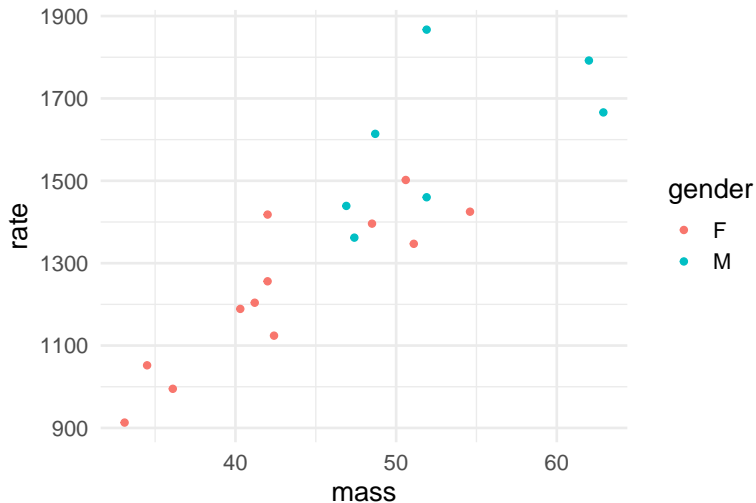
Looking at relationships
visually: Scatterplots

**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
weight_scatter <- ggplot(weight_data, aes(x = mass, y = rate)) +  
  geom_point(aes(col=gender)) +  
  theme_minimal(base_size = 15)
```


Analysis: Colour the points by gender



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Analysis: Create separate plots for men and women

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

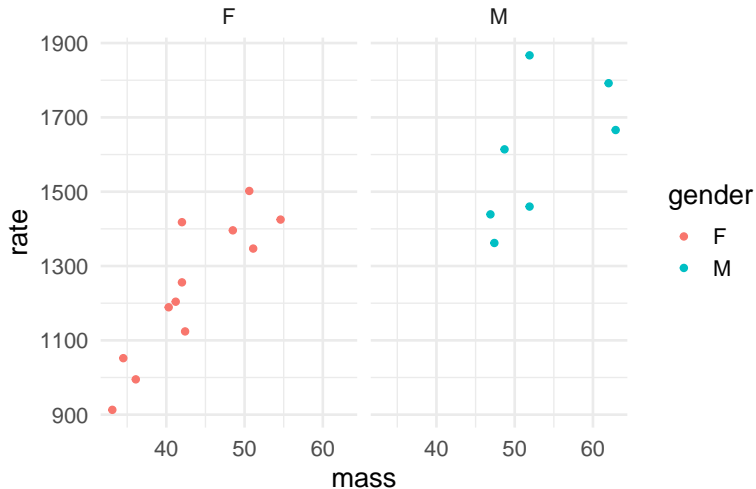
**Exploratory analysis using
scatterplots**

Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
weight_scatter <- ggplot(weight_data, aes(x = mass, y = rate)) +  
  geom_point(aes(col=gender)) +  
  theme_minimal(base_size = 15)+  
  facet_wrap(~ gender)
```

Analysis: Create separate plots for men and women

weight_scatter



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Assessing a relationship between two variables with a number: Pearson's correlation

Pearson's correlation

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Using just our eyes, we can often say something about whether an association between two variables is weak or strong.

But we can also use a numeric value to describe the direction and strength of an association

Pearson's correlation

- ▶ For **linear** associations, we can use **Pearson's correlation coefficient** (denoted by r) to **quantify the strength** of a linear relationship between two variables.
- ▶ The correlation between x and y is:

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

Notice that because we are dividing by the standard deviation the values become unitless

Relationships between two quantitative variables

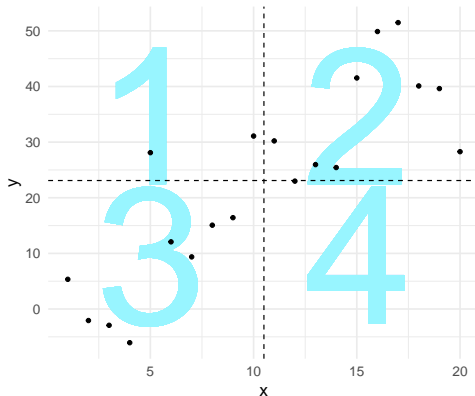
Looking at relationships visually: Scatterplots

Exploratory analysis using scatterplots

Assessing a relationship between two variables with a number: Pearson's correlation

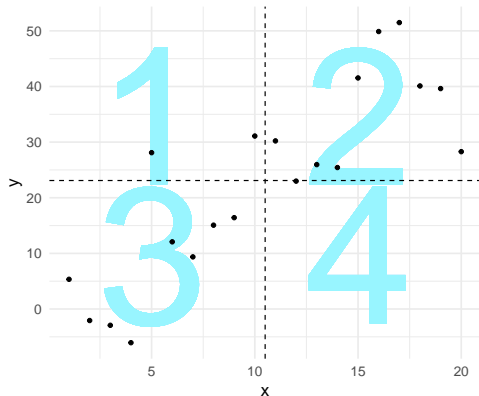
Intuition about Pearson's correlation

To understand this formula, first only consider the numerators of the fractions (i.e., $x_i - \bar{x}$ and $y_i - \bar{y}$). If you imagine a scatter plot of x and y , we can also add a dashed line at the mean x value of \bar{x} and a dashed line at the mean y value (\bar{y}):



Intuition about Pearson's correlation

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



- Points in Q2 and Q3 contribute positive products to r
- Points in Q1 and Q4 contribute negative products to r

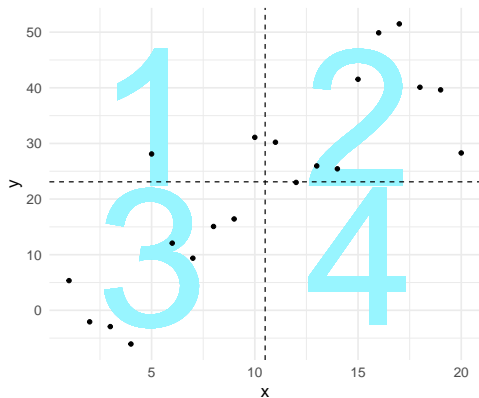
Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Intuition about Pearson's correlation



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

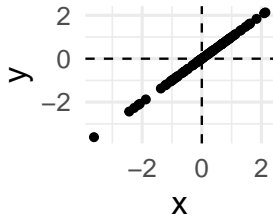
- ▶ The more there are points in Q2 and Q3 vs. Q1 and Q4, the more the value of the correlation coefficient will be higher and positive
- ▶ If you want even more of an explanation see the response to this [stack overflow post](#)

Properties of the correlation coefficient

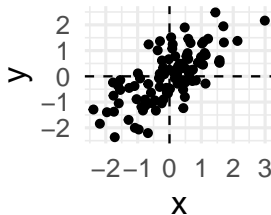
- ▶ Always a number between -1 and 1.
 - ▶ -1: A perfect, negative linear association
 - ▶ 1: A perfect, positive linear association
 - ▶ 0: No linear association
- ▶ Is used to measure the association between two *quantitative* variables.
- ▶ Only useful for *linear* associations!

Correlation and direction

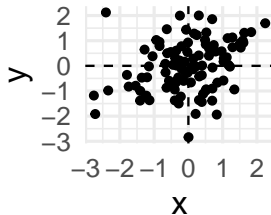
Correlation = 1



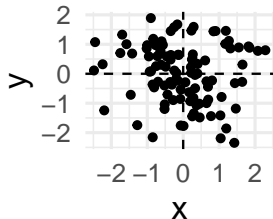
Correlation = 0.7



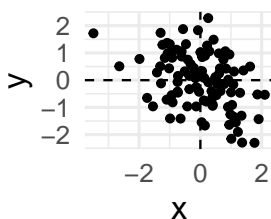
Correlation =



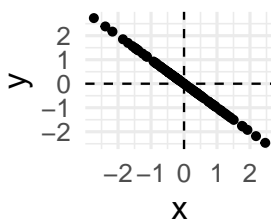
Correlation = -0.1



Correlation = -0.4



Correlation =



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Syntax: Pearson's correlation using `cor()`

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

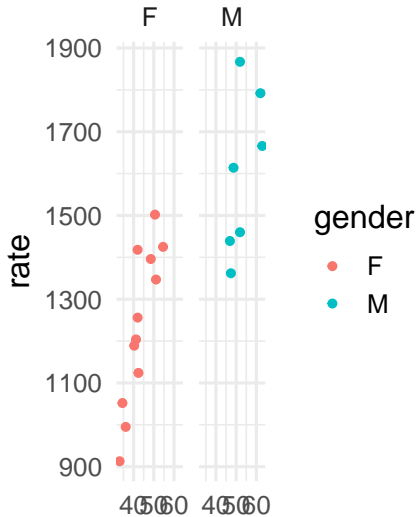
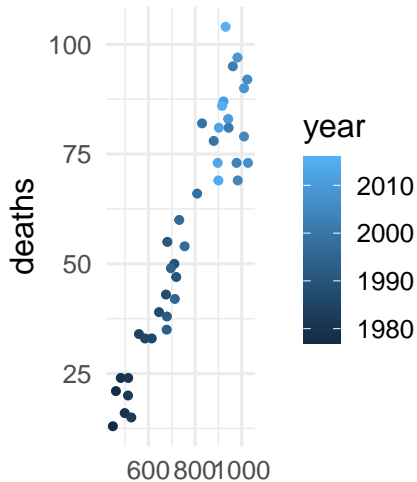
Assessing a relationship
between two variables with
a number: Pearson's
correlation

```
correlation coefficient <- dataset %>%
```

```
summarize(newvar = cor(xvar, yvar))
```

Syntax: Pearson's correlation using `cor()`

Remember the manatee plot and the weight plot:



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Syntax: Pearson's correlation using cor()

Now, calculate the correlations between X and Y for manatees:

```
mana_cor <- mana_data %>%  
  summarize(corr_mana = cor(powerboats, deaths))  
mana_cor
```

```
## # A tibble: 1 x 1  
##   corr_mana  
##   <dbl>  
## 1      0.945
```

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Syntax: Pearson's correlation using cor()

And for the weight data:

```
weight_cor <- weight_data %>%  
  summarize(corr_weight = cor(mass, rate))  
weight_cor
```

```
## # A tibble: 1 x 1  
##   corr_weight  
##   <dbl>  
## 1      0.865
```

Relationships between two
quantitative variables

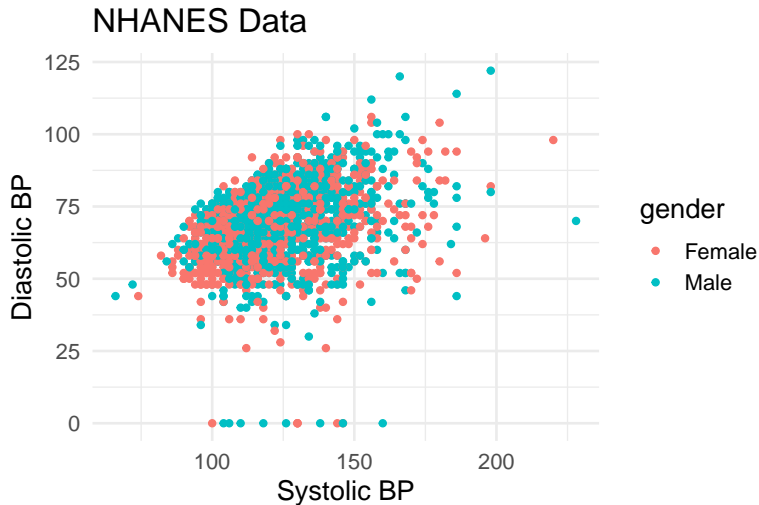
Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Syntax: Pearson's correlation using `cor()`

What about our blood pressure data from NHANES?



Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

Syntax: Pearson's correlation using cor()

```
bp_cor <- nhanes_data %>%  
  summarize(corrbp = cor(bpxsy1, bpxdi1))  
bp_cor
```

```
## # A tibble: 1 x 1  
##   corrbp  
##   <dbl>  
## 1  0.322
```

Relationships between two
quantitative variables

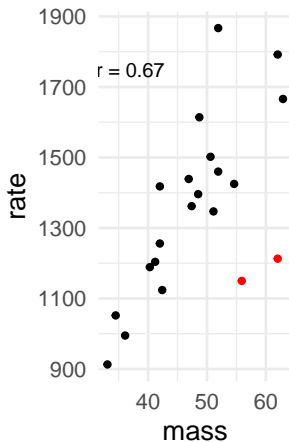
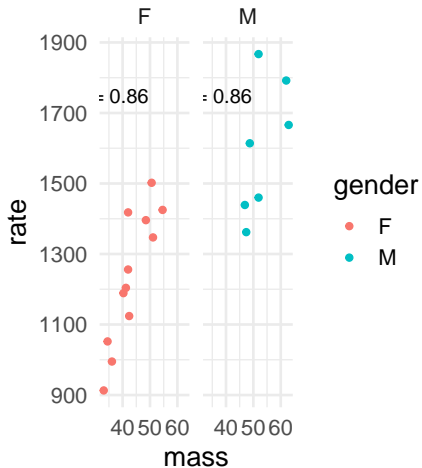
Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

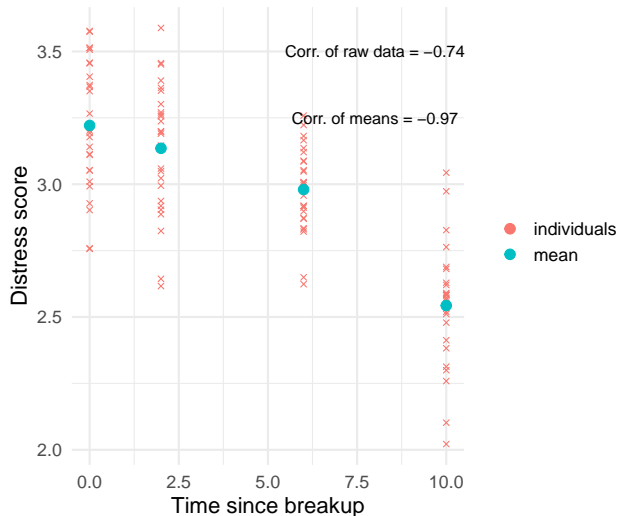
Properties of the correlation coefficient

The correlation coefficient is not resistant to outliers, notice what happens when we add two outliers (in red) to the `weight_data` and recalculate correlation



Properties of the correlation coefficient

- ▶ Correlations for average measures is typically stronger than correlations for individual data



Important concepts

Lecture 05: Relationships between 2 variables

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

- ▶ Determine which variable is explanatory and which is response, or when it doesn't matter
- ▶ Visually describe the relationship between two variables (form, direction, strength, and outliers)
- ▶ Numerically describe the relationship with the correlation coefficient r

R Recap: What functions did we use?

Relationships between two
quantitative variables

Looking at relationships
visually: Scatterplots

Exploratory analysis using
scatterplots

Assessing a relationship
between two variables with
a number: Pearson's
correlation

- ▶ `geom_point()`,
- ▶ `aes(col = gender)` to color points by levels of gender
- ▶ `summarize()` to calculate correlation using `cor(var1, var2)`

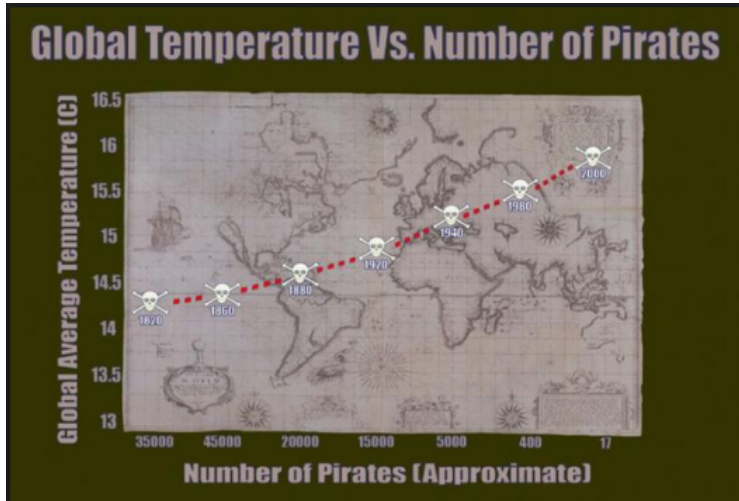
Reminder: Association does not equal causation

Remember that just because two variables are associated, does not mean there is a causal relationship

The correlation coefficient measures association *not* causation.

Even a very strong association doesn't mean that one variable causes the other.

Reminder: Association does not equal causation



This image is one from a Forbes.com article but this example pops up in lots of places