

Relationships between two
quantitative variables

Overview

Questions about the worksheet

Review of a few concepts

Scatterplots

Correlation

Midterm self-evaluations due on Wednesday (Oct 3)

All first year students need to complete midterm self-evaluations by tomorrow (Wednesday Oct 3rd)

Types of questions to answer:

- What is going well in this course?
 - Is there anything you need to improve?
- What is your plan for making that improvement?
- Any suggestions for ways to improve the class

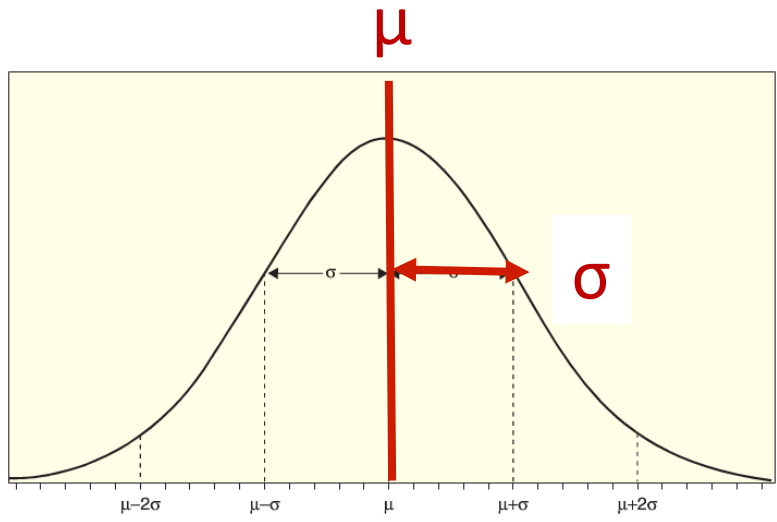
Questions about worksheet 3?

How did worksheet 3 go?

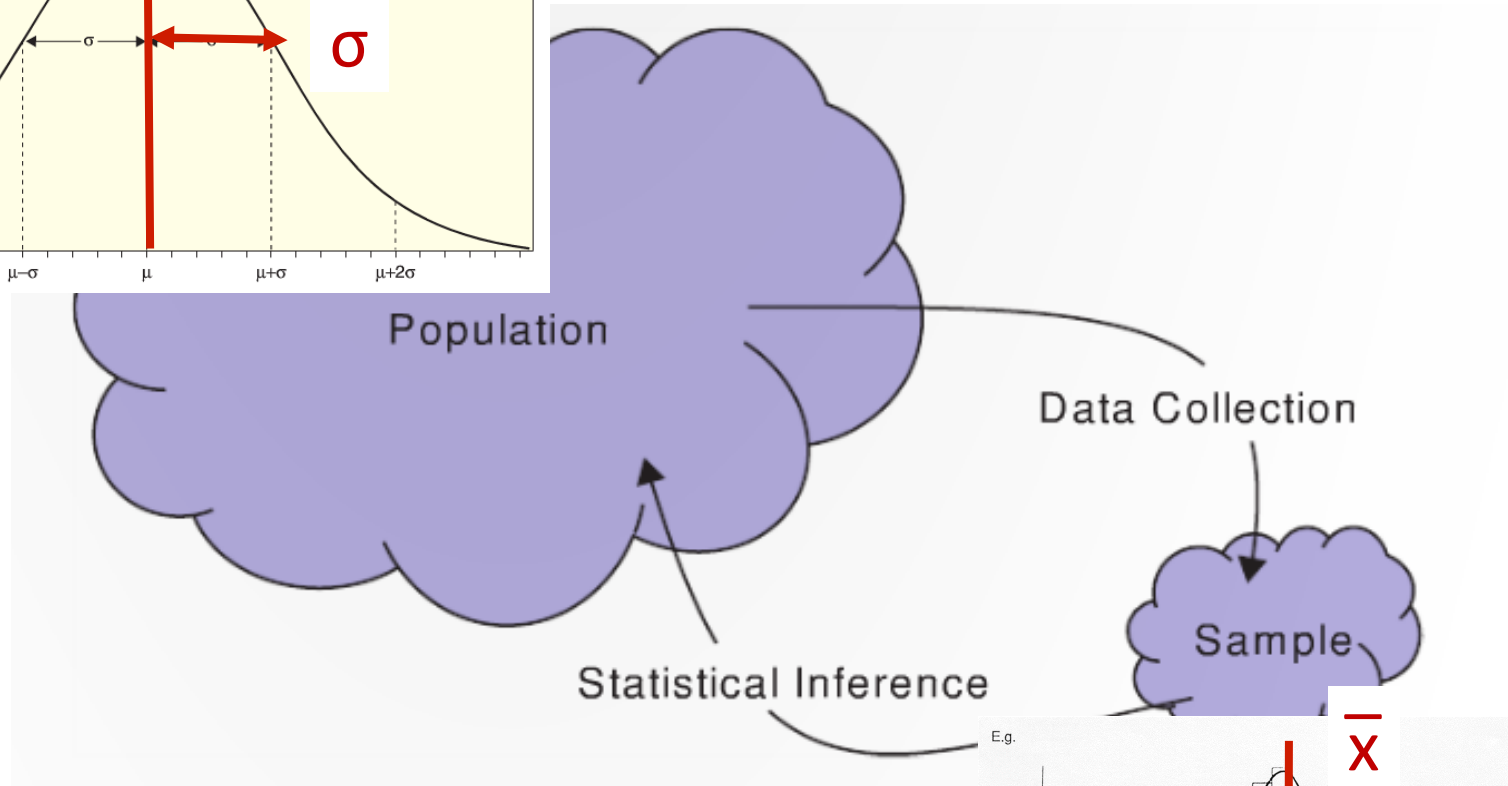
- Too easy, too hard, just right?

Any questions?

Please try to get your own a copy of the Lock 5 book!

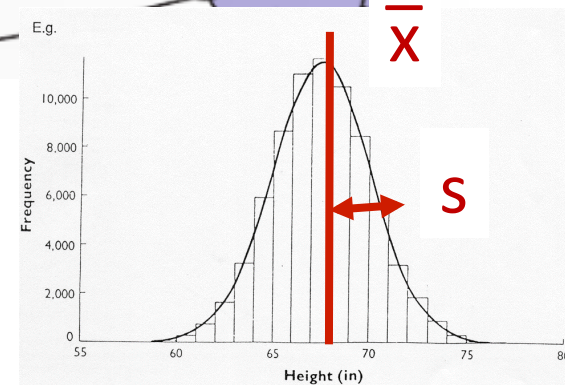


Parameters



$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Statistics



Review: z-scores

The z-scores tells how many standard deviations a value is from the mean

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

Which statistic is most impressive?

Z-score FGPct = 0.868

Z- score Points = 2.698

Z-score Assists = 1.965

Z-score Steals = 1.771



Normal pillow



Question: What percent of the pillow's mass is ± 2 standard deviations from the mean?

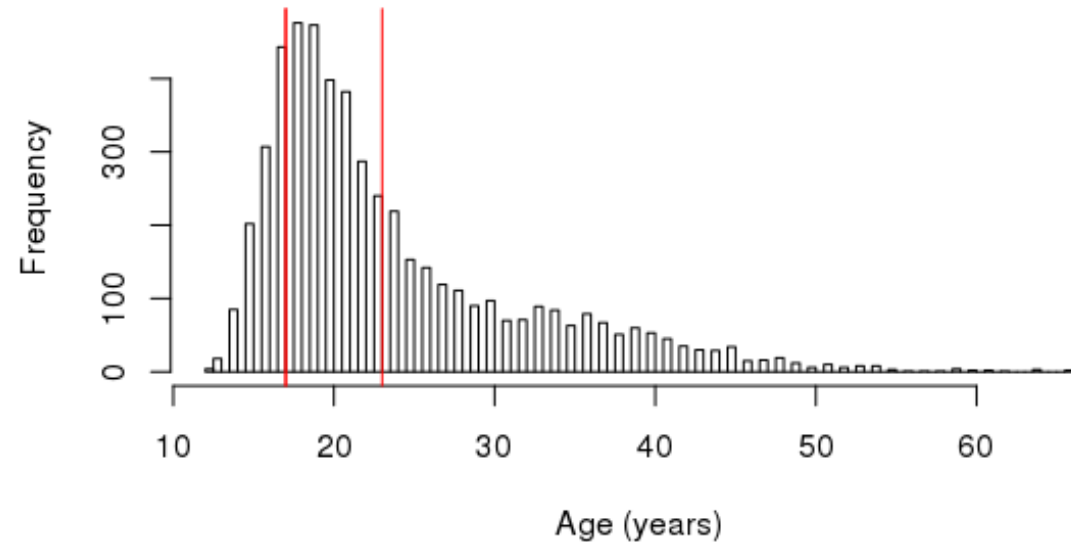
Answer: 95%

Review: quantiles (percentiles)

The **P^{th} percentile** is the value of a quantitative variable which is greater than P percent of the data



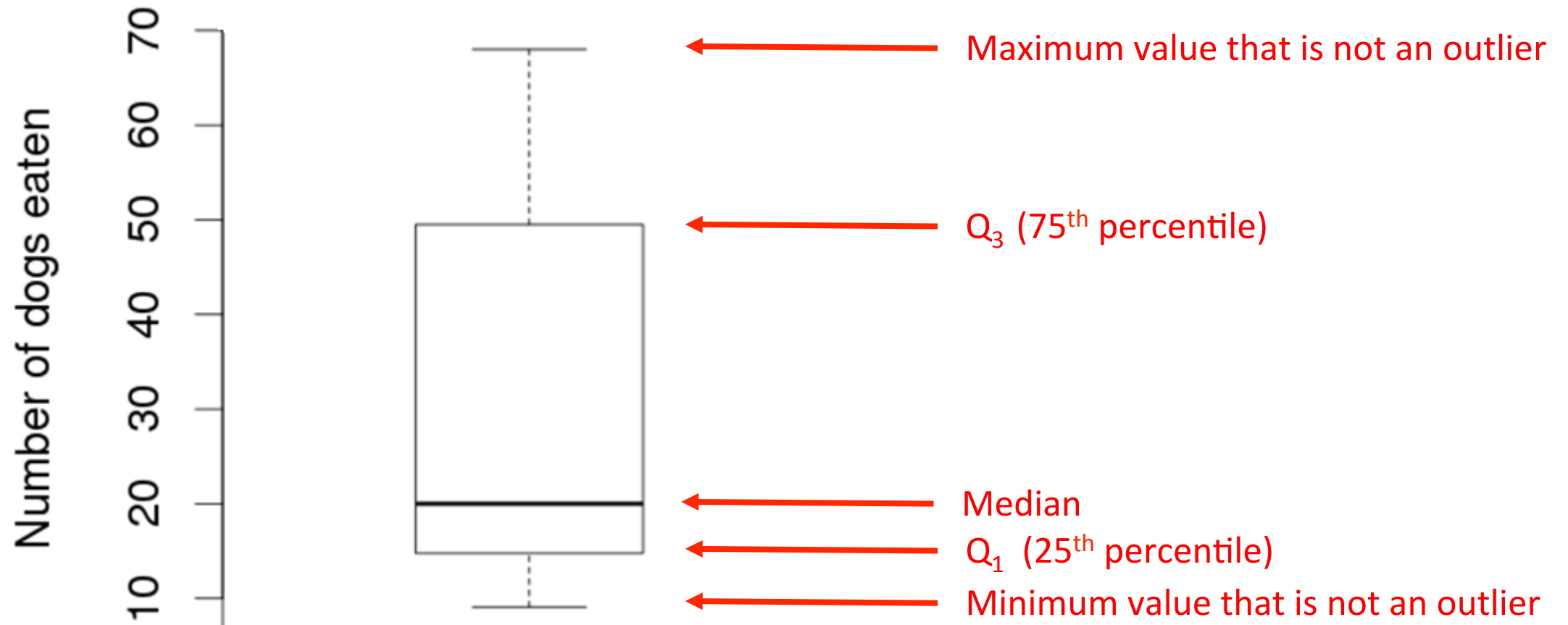
Histogram of Ages of people arrested for marijuana use



60th percentile value is 23

i.e., 60% of the arrests were of ages 23 or less

Review: boxplot (5 number summary)



Outlier quiz

1. What is an **outlier**?

- A: An observed value that is notably distinct from the other values in a dataset

2. Why are they problematic?

- A: can potentially have a large influence on the statistics you calculate

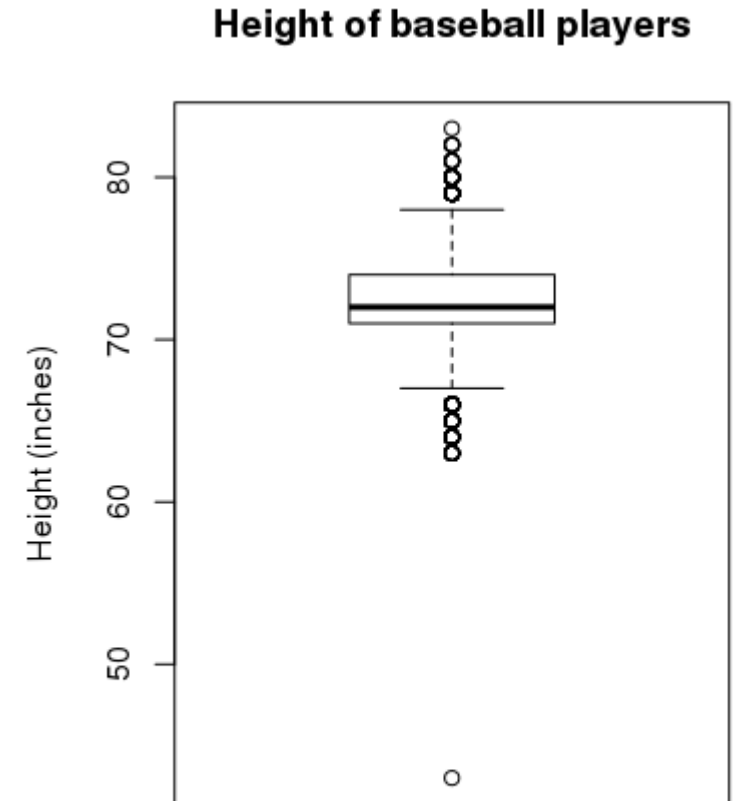
3. What should you do if you have an outlier in your data?

A: See if you can understand what is causing it!

- If it's an error, delete the point
- If it's a real value, make sure it is not having a bit effect on your conclusions, and/or use resistant statistics

4. Is the mean and/or median is resistant?

- A: The median is resistant when the mean is not



Outlier quiz

1. Is there an **outlier**?

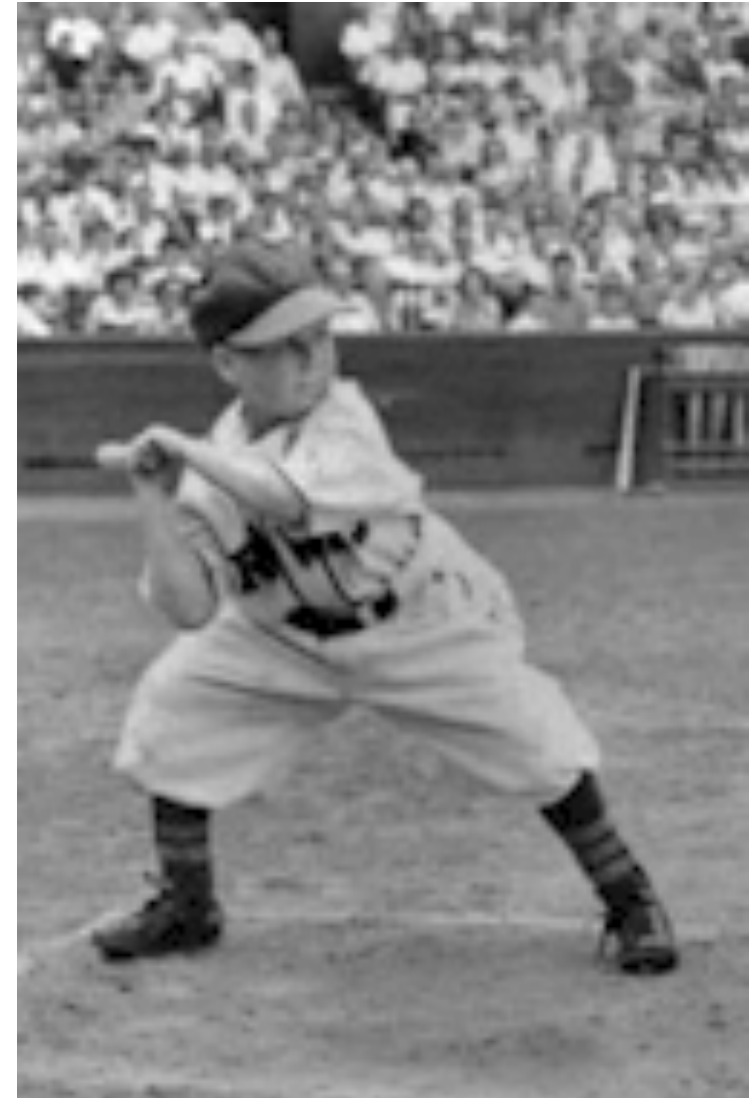
- A: Yes!

2/4. Is the outlier problematic?

- A: it will probably not have a big effect since there are so many other data points
 - Also it will have minimal effect on the boxplot because the median (and quantiles) are resistant

3. What does this outlier correspond to?

A: [Eddie Gaedel](#)



Relationships between two
quantitative variables

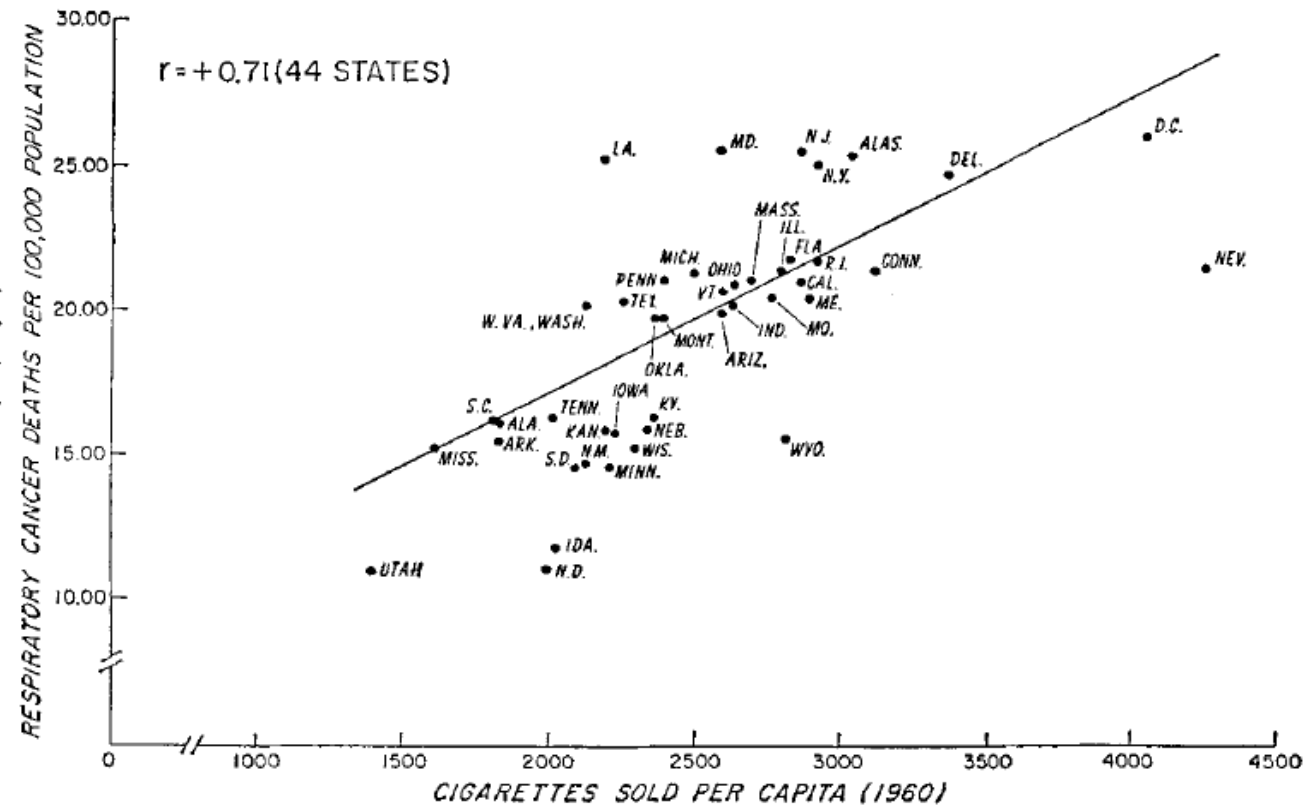
Two quantitative variables

In 1968, Joseph Fraumeni published a paper published in the Journal of the National Cancer Institute that examined the relationship between smoking and different types of cancer

State	Cig per capita	Bladder	Lung	Kidney	Leukemia
AL	18.2	2.9	17.05	1.59	6.15
AZ	25.82	3.52	19.8	2.75	6.61
AR	18.24	2.99	15.98	2.02	6.94
CA	28.6	4.46	22.07	2.66	7.06
CT	31.1	5.11	22.83	3.35	7.2
DE	33.6	4.78	24.55	3.36	6.45
DC	40.46	5.6	27.27	3.13	7.08

Relationship between smoking and lung cancer

TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956-61) and *per capita* cigarette sales (1960) in 44 States.



Scatterplot

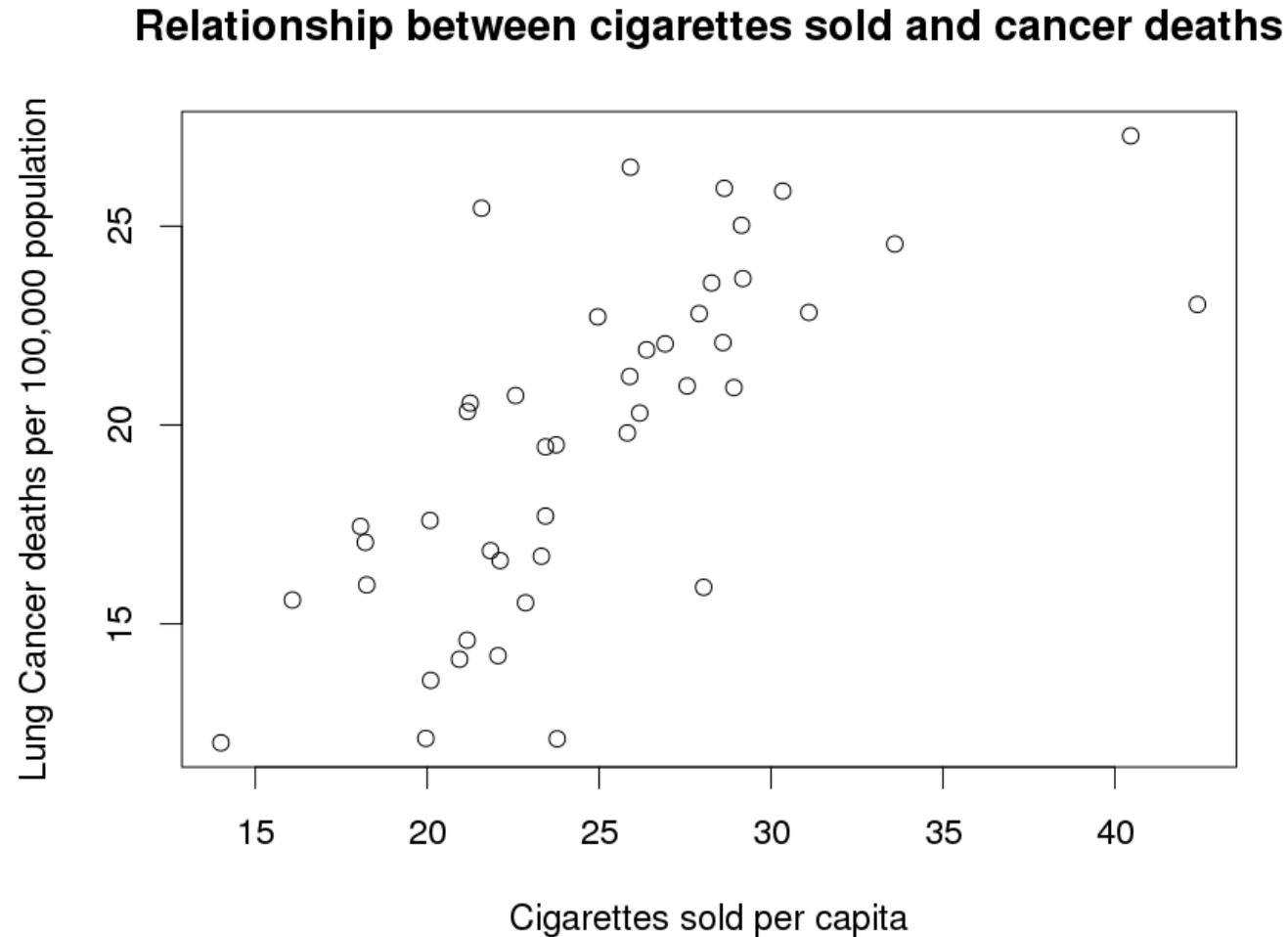
A **scatterplot** graphs the relationship between two variables

- Each axis represents the value of one variables

- Each point the plot shows the value for the two variables for a single data case

If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis

Relationship between smoking and lung cancer



R: `plot(x, y)`

Questions when looking at scatterplots

Do the points show a clear trend?

Does it go upward or downward?

How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Questions when looking at scatterplots

Do the points show a clear trend?

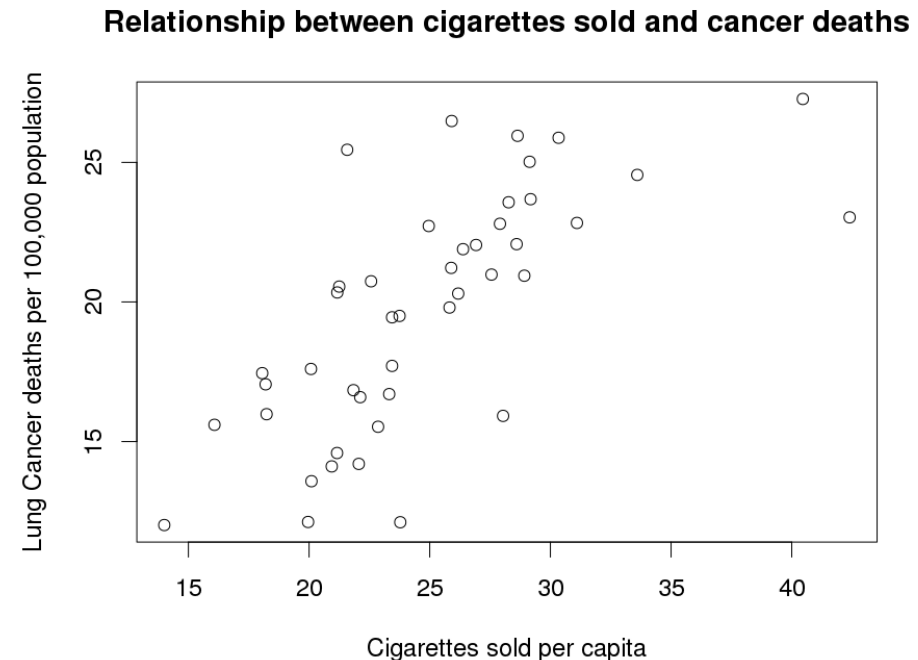
Does it go upward or downward?

How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Smoking and cancer



Questions when looking at scatterplots

Do the points show a clear trend?

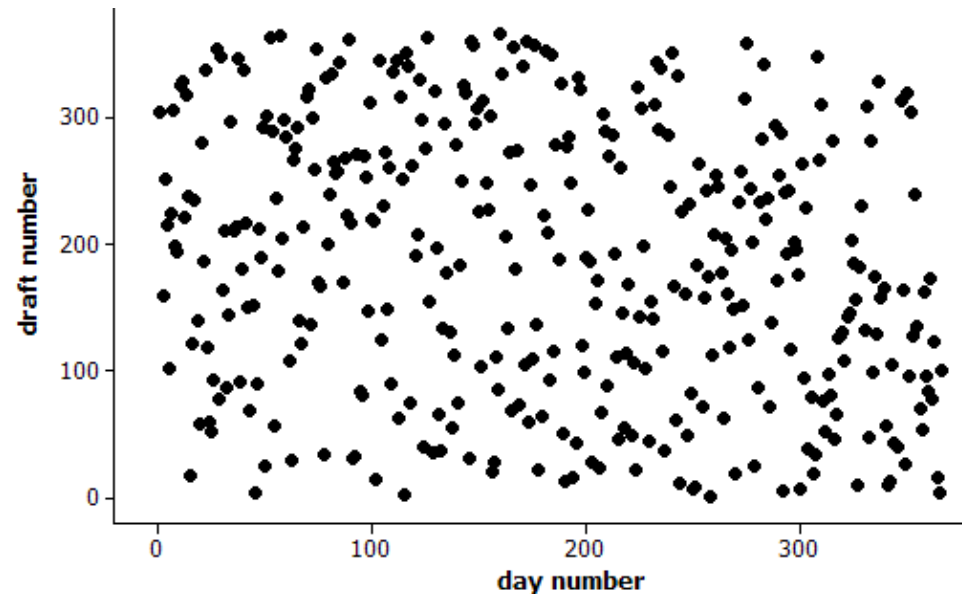
Does it go upward or downward?

How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

1969 Draft Lottery for
the Vietnam War



Questions when looking at scatterplots

Do the points show a clear trend?

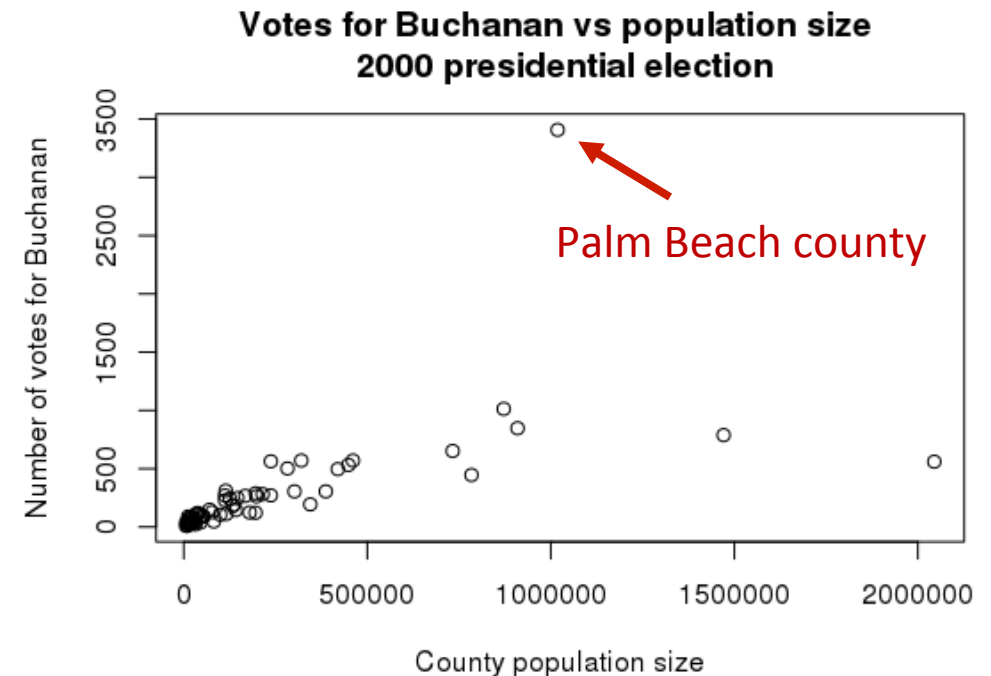
Does it go upward or downward?

How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

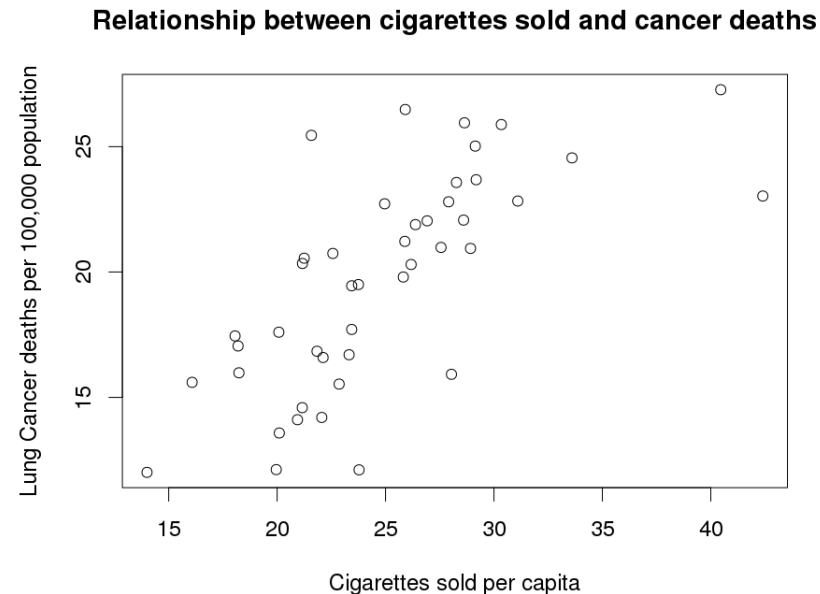
Votes for Buchanan vs county population size
2000 presidential election



statistics

We can plot relationships visually but it would be nice to have a number that summarizes the strength of a relationship

- A number that summarizes a sample of data is called a...



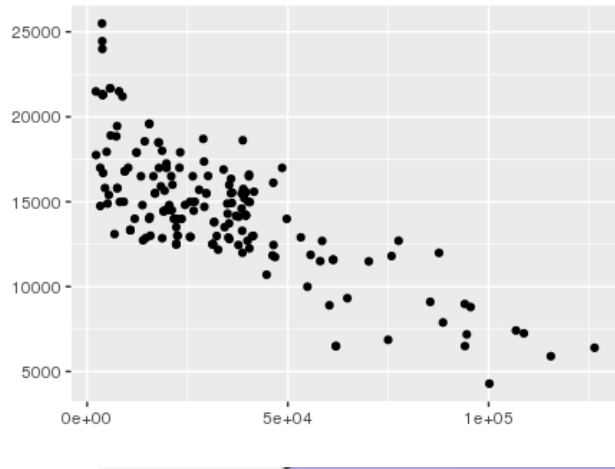
The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables.

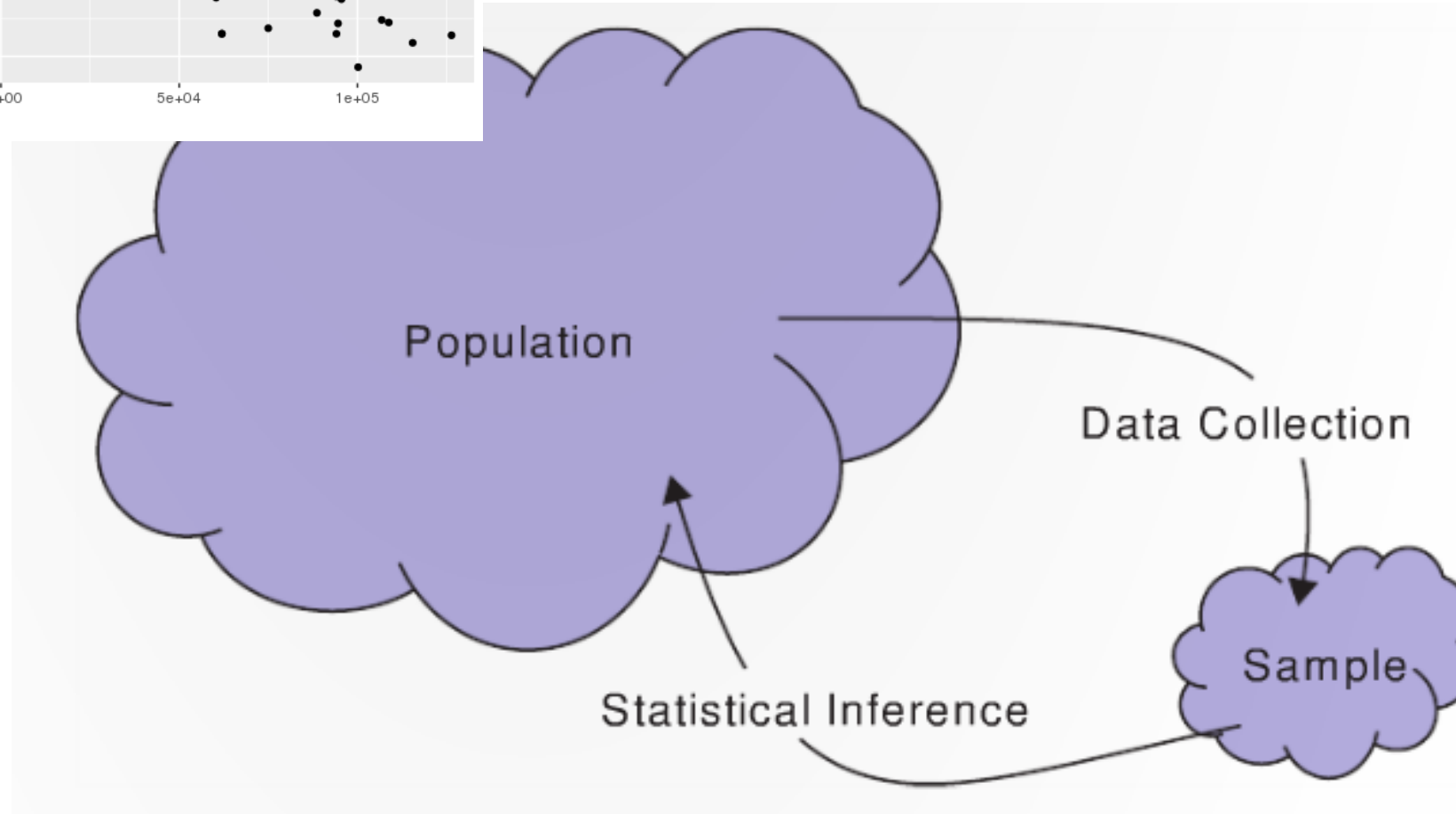
$$r = \frac{1}{(n - 1)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- The correlation for a sample is denoted with **r**
- The correlation in the population is denoted with **ρ**
(the Greek letter rho)

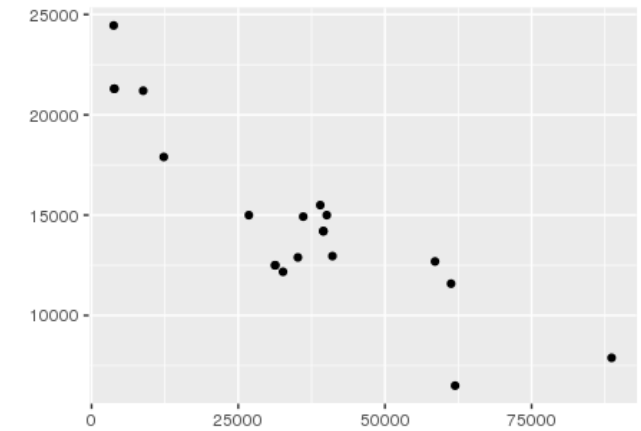
R: `cor(x, y)`



ρ parameter



r statistic

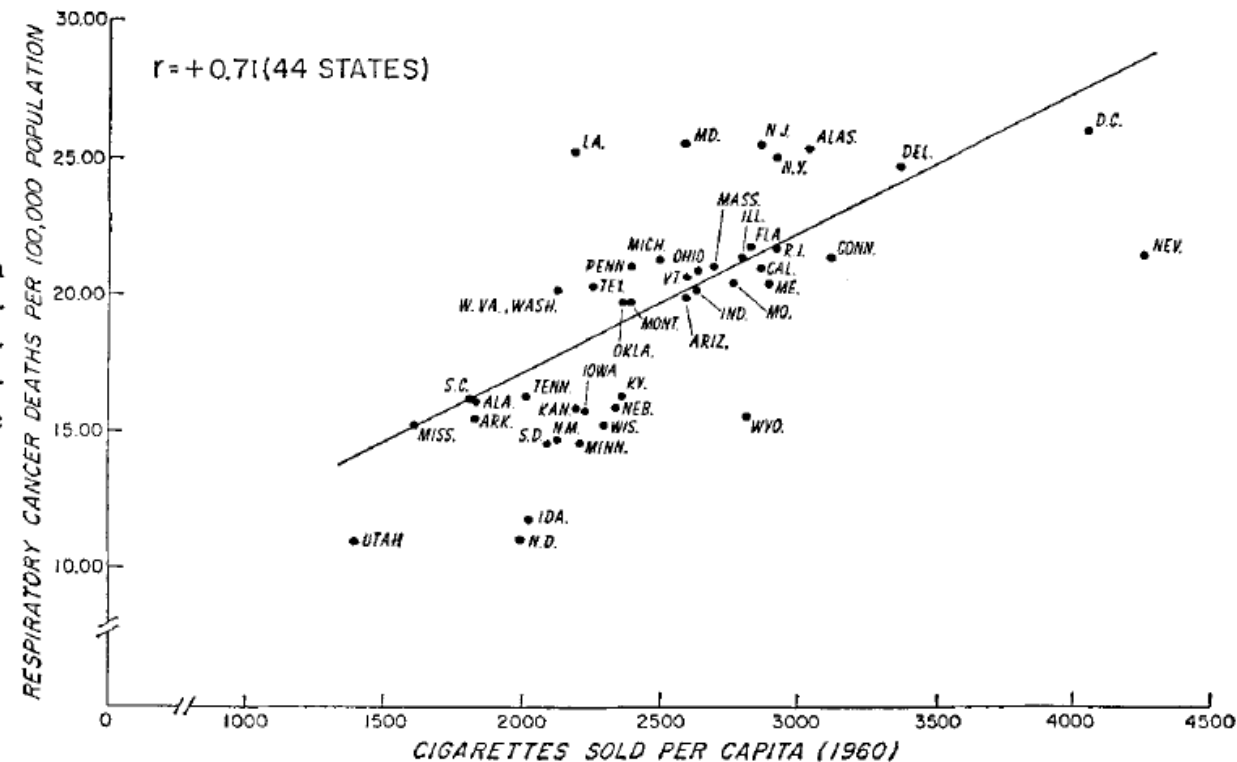


Smoking and lung cancer correlation?

The **correlation** is measure of the strength and direction of a linear association between two variables.

TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956-61) and *per capita* cigarette sales (1960) in 44 States.

$r = 0.71$



Properties of the correlation

Correlation is always between -1 and 1: $-1 \leq r \leq 1$

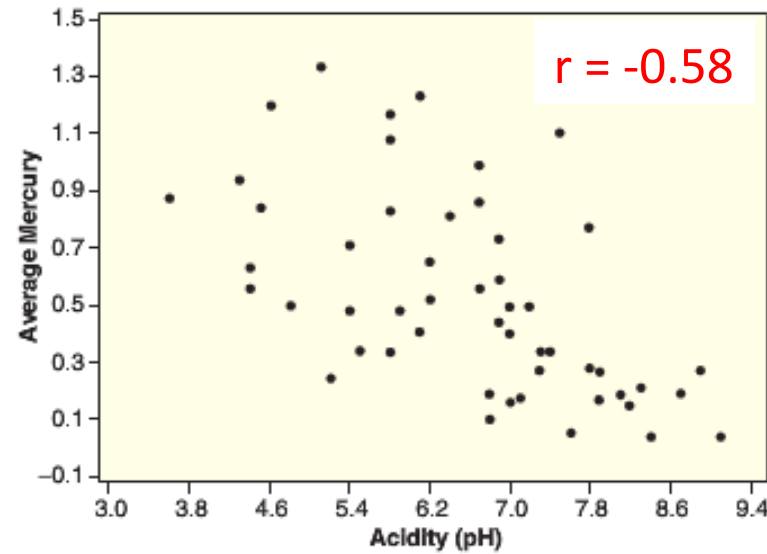
The sign of r indicates the direction of the association

Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

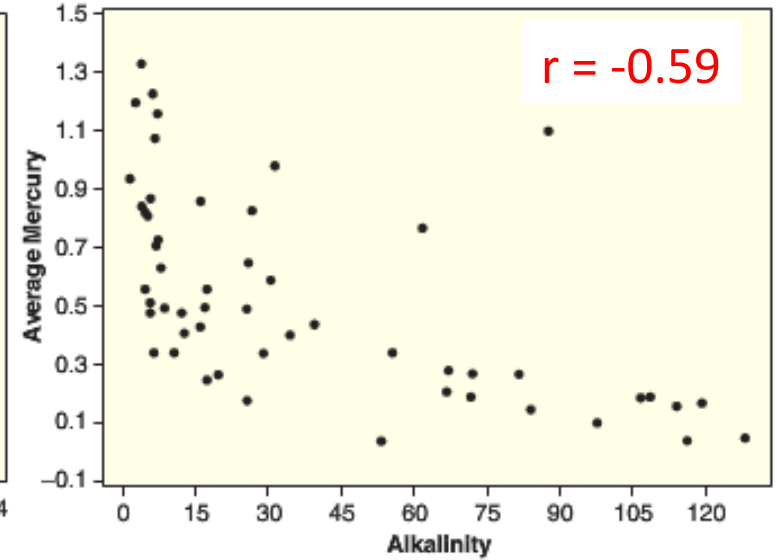
Correlation is symmetric: $r = \text{cor}(x, y) = \text{cor}(y, x)$

Florida lakes

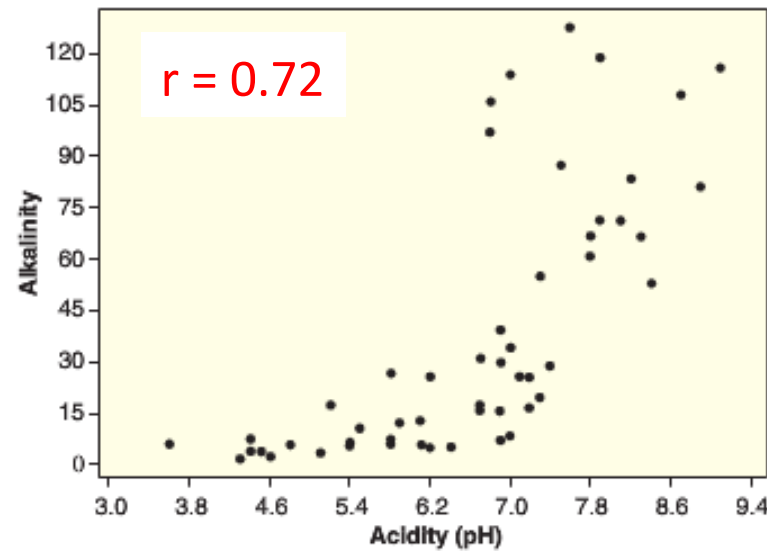
Correlation game



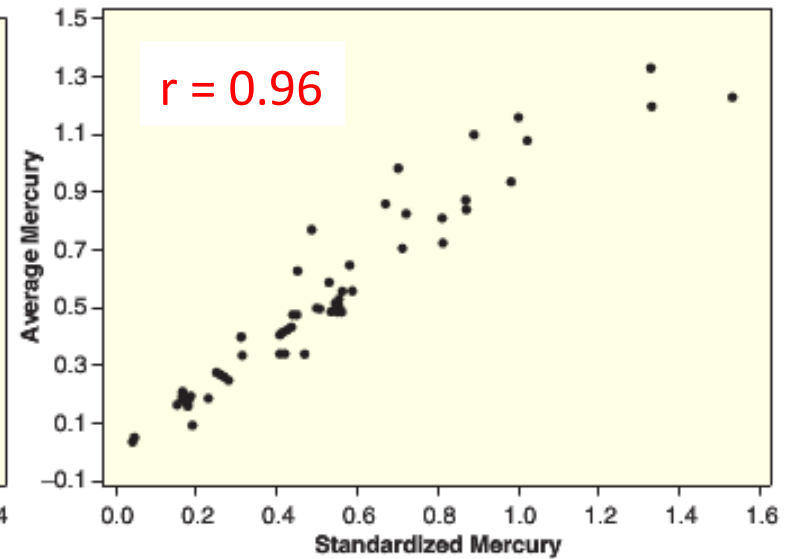
(a) Average mercury level vs acidity



(b) Average mercury level vs alkalinity



(c) Alkalinity vs acidity



(d) Average vs standardized mercury levels

Let's calculate some correlations

Is there an associate between cigarettes sold per capita and other types of cancer?

- Bladder cancer (BLAD)
- Kidney cancer (KID)
- Leukemia (LEUK)

load the data

```
> load("/home/shared/intro_stats/cs206_data/smoking_cancer.rda")
```

create a scatter plot and calculate the correlation

```
> plot(smoking$CIG, smoking$LUNG)
```

```
> cor(smoking$CIG, smoking$LUNG)
```

For next class – practice problems

Lock5 exercises first edition: 2.153, 2.155, 2.159, 2.177

Lock5 exercises second edition: 2.165, 2.167, 2.170, 2.191