Measures of spread continued and relationships between two quantitative variables

Overview

Quick review:

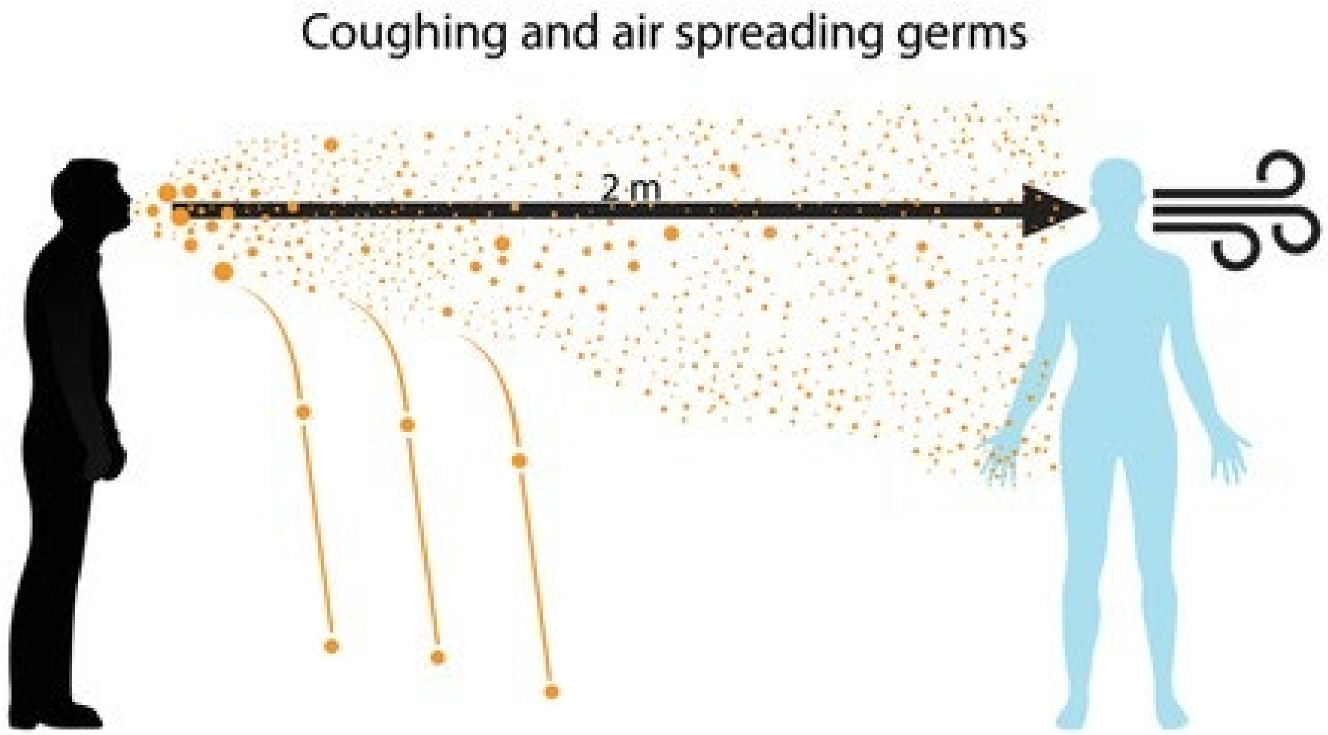
- Standard deviations, z-scores, percentiles

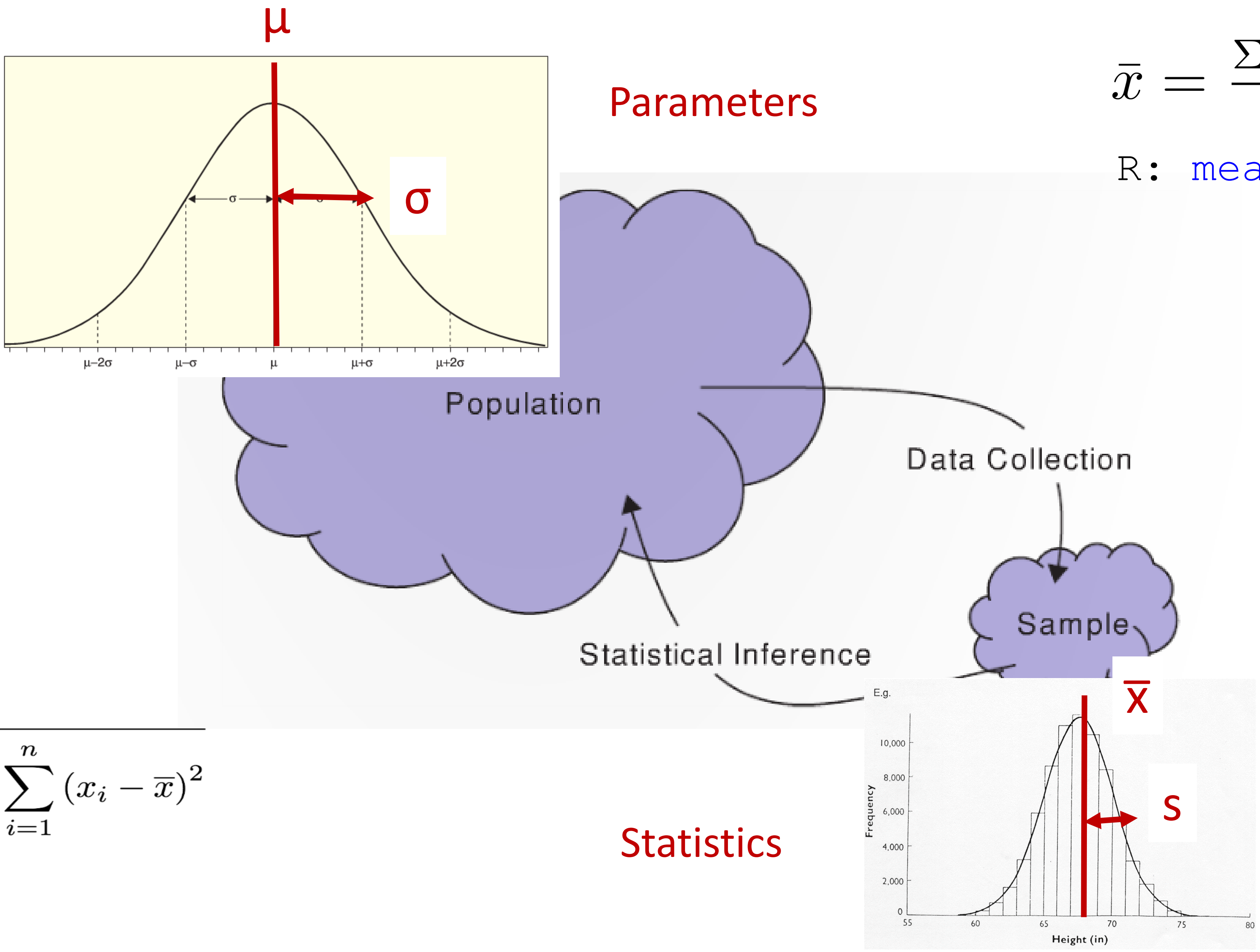
Boxplots

Scatter plots

Correlation

Review and continuation of measures of spread…





Parameters

$$\bar{x} = \frac{\Sigma_i^n x_i}{n}$$

```
R: mean(x)
```

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

```
R: sd(x)
```

Statistics

Review: z-scores

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

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

<u>Which statistic is most impressive?</u>

Z-score FGPct	=	0.868
Z- score Points	=	2.698
Z-score Assists	=	1.965
Z-score Steals	=	1.771

The normal pillow

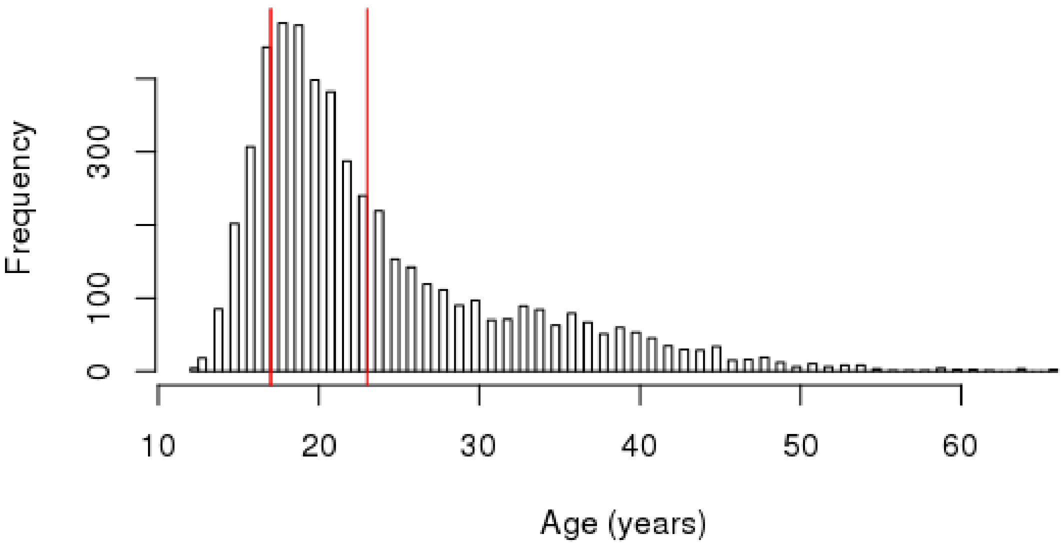
Question: What percent of the pillow’s mass is ± 2 standard deviations from the mean?

Answer: 95%

Review: quantiles (percentiles)

The **pth percentile** is a quantitative value ***x*** 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

In R: quantile(Arrests$age, .6)

The quantile universe

Five-Number Summary = (minimum, Q_1, median, Q_3, maximum)

Q_1 = 25th percentile, Q_3 = 75th percentile

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

As a rule of thumb, we call a data value an **outlier** if it is:

Smaller than: Q_1 - 1.5 * IQR

Larger than: Q_3 + 1.5 * IQR

In R: `fivenum(v)`

Box plots

A **box plot** is a graphical display of the five-number summary and consists of:

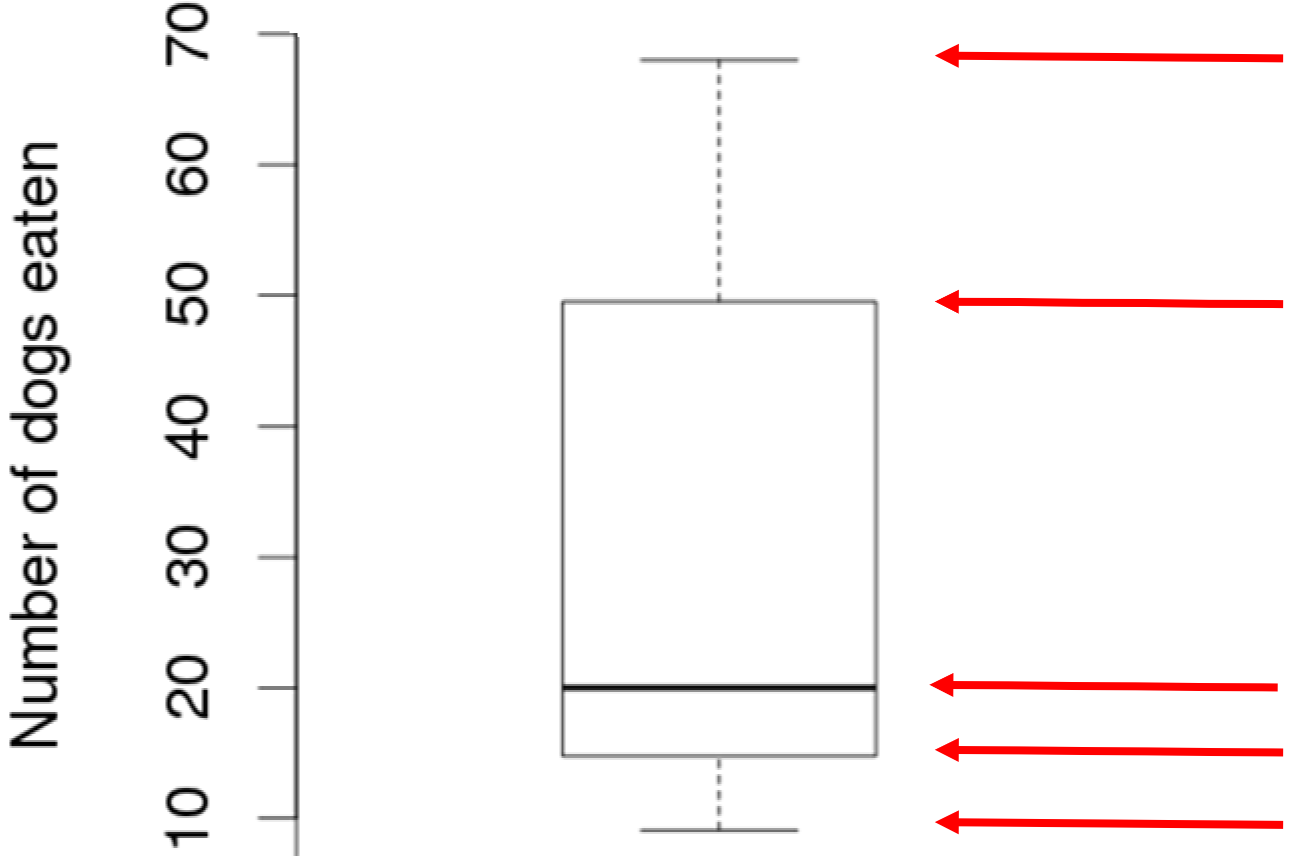
1. Drawing a box from Q_1 to Q_3

2. Dividing the box with a line (or dot) drawn at the median

3. Draw a line from each quartile to the most extreme data value that is not and outlier

4. Draw a dot/asterisk for each outlier data point.

Box plot of the number of hot dogs eaten by the men’s contest winners 1980 to 2010



Maximum value that is not an outlier

Q_3

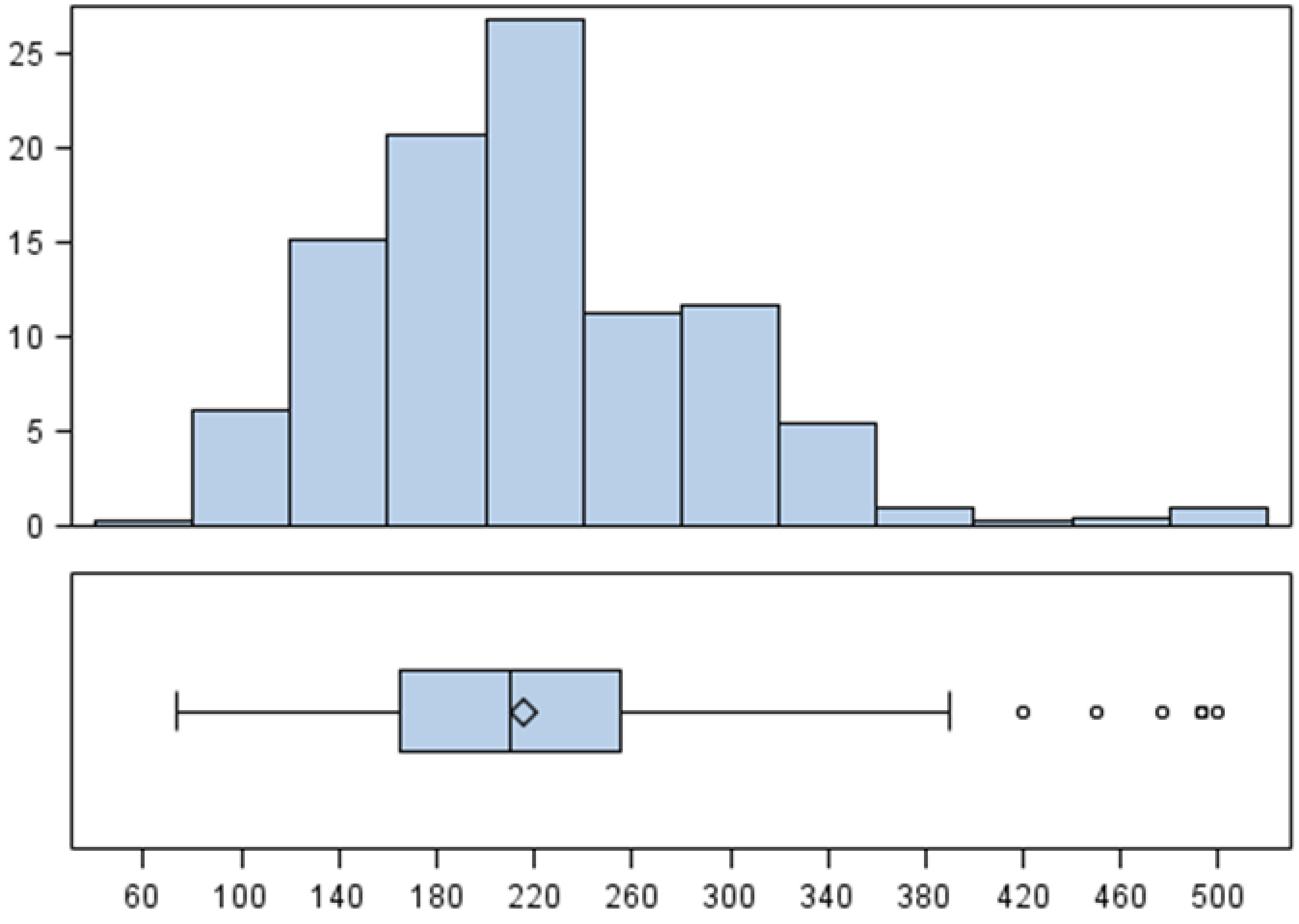
Median

Q_1

Minimum value that is not an outlier

```
R: boxplot(v)
```

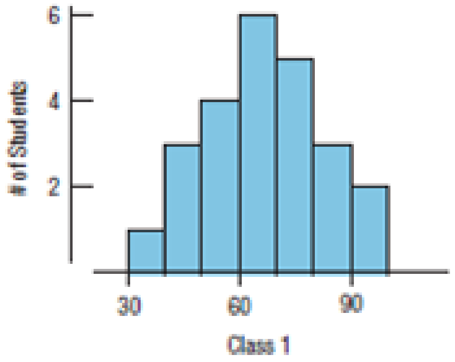
Box plots extract key statistics from histograms



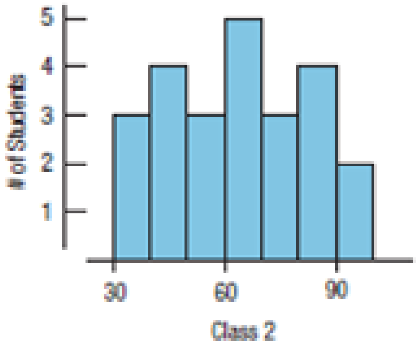
Box plots extract key statistics from histograms

Question: which Box plot goes with which histogram?

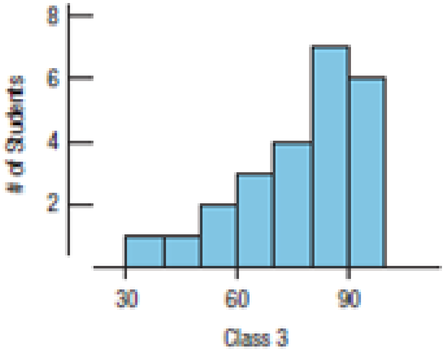
Histogram 1

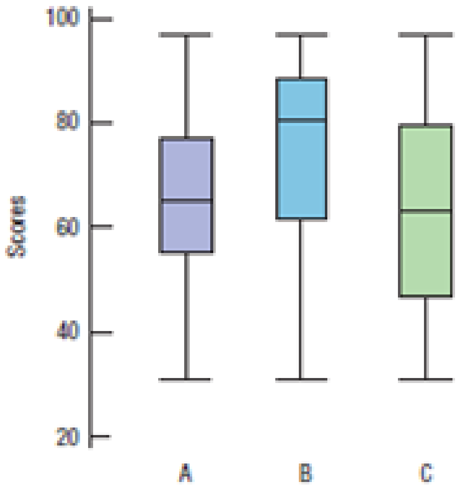


Histogram 2



Histogram 3

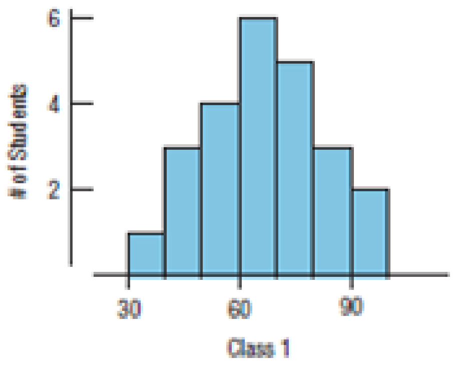




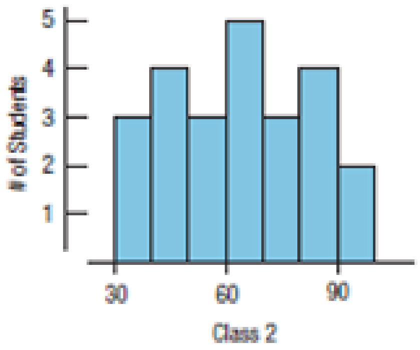
Box plots extract key statistics from histograms

Question: which Box plot goes with which histogram?

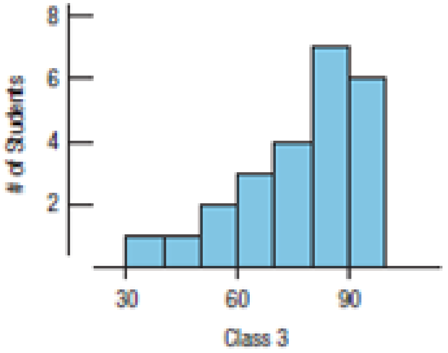
A

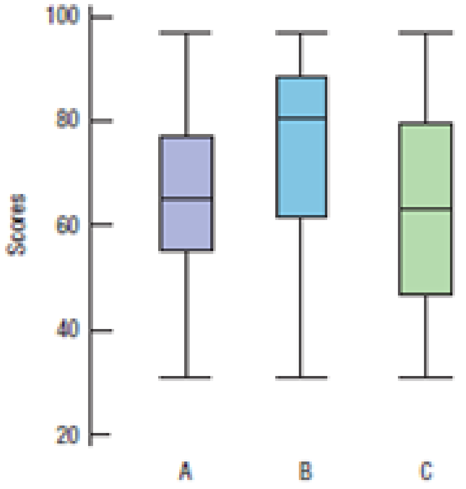


C



B



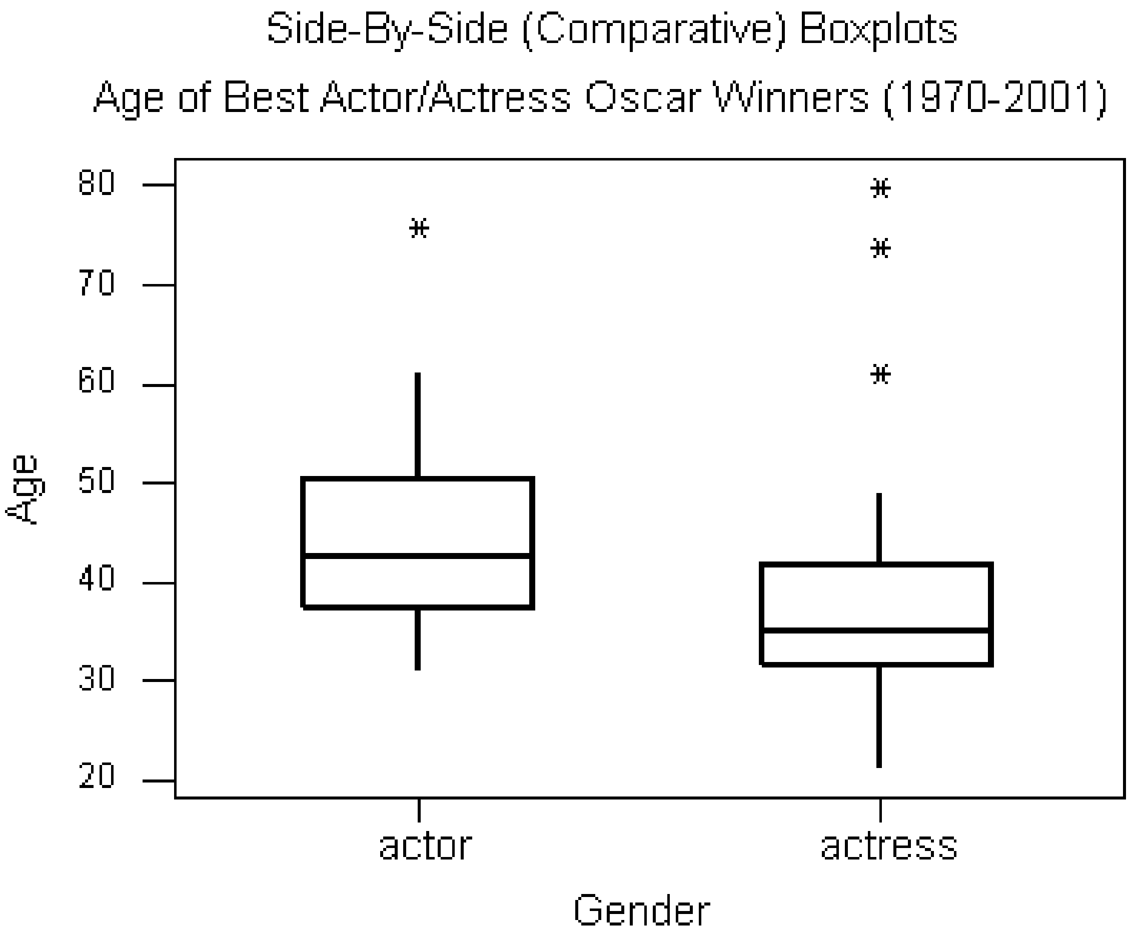


Comparing quantitative variables across categories

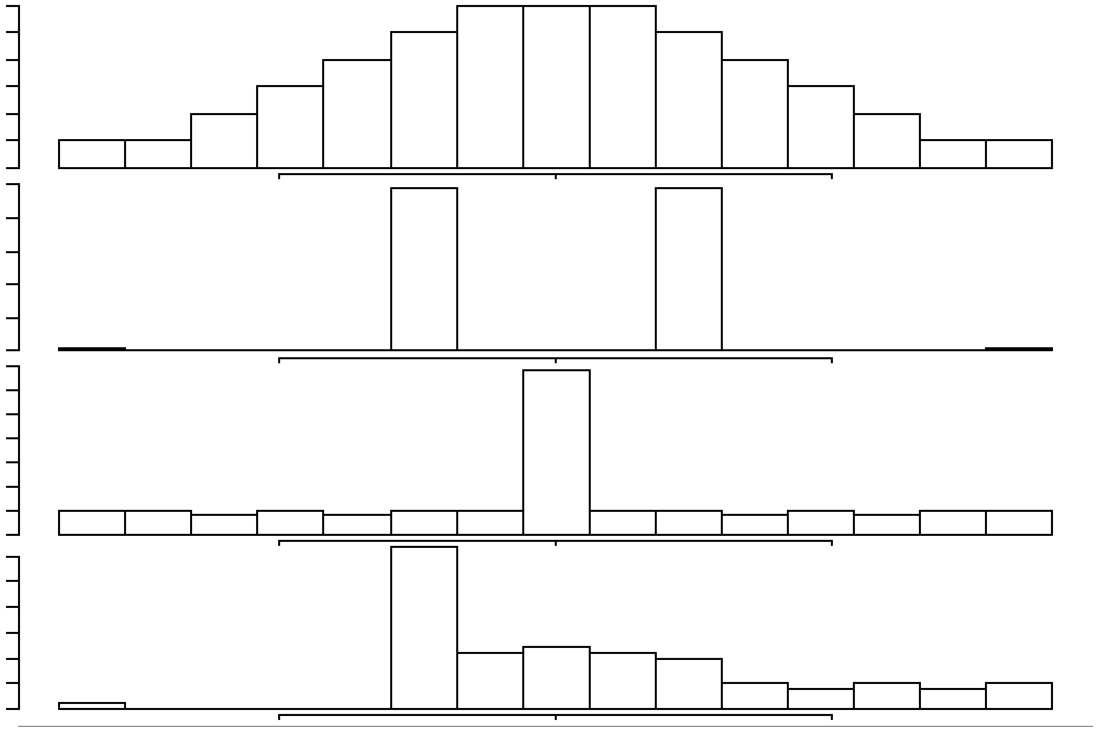
Often one wants to compare quantitative variables across categories

Side-by-Side graphs are a way to visually compare quantitative variables across different categories

Side-by-side box plots

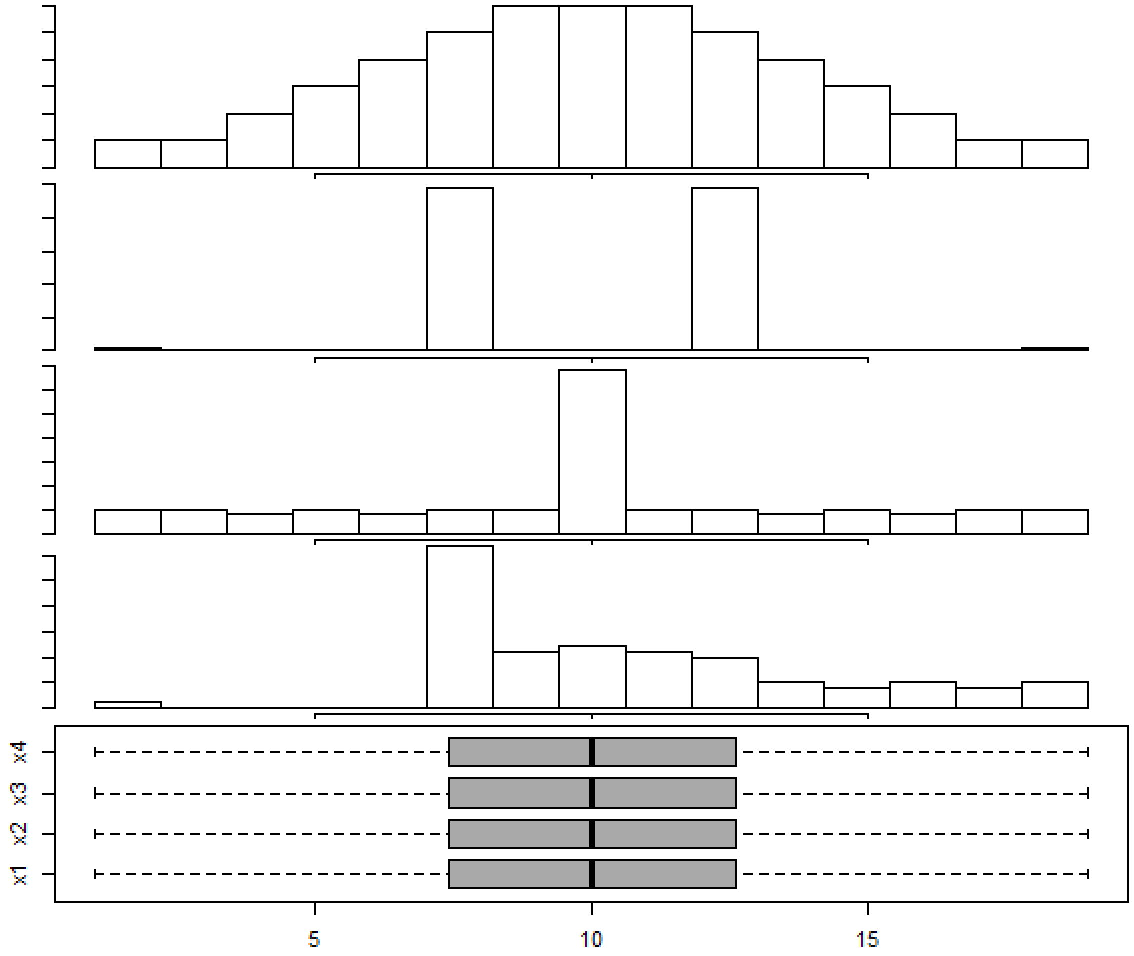


Box plots don’t capture everything



Do you think the box plots for these distributions look similar?

Box plots don’t capture everything



Side-by-size boxplots in R

```
> boxplot(v1, v2,                              # compare two vectors v1 and v2
     names = c("name 1", "name 2"),            # labels below each box plot
      ylab =  "y-axis name"                    # y-axis label name
  )
```

Try it yourself, create histograms and boxplots for this data:

```
> download_data("distribution_vs_boxplot.Rda")
> load("distribution_vs_boxplot.Rda")
> boxplot(x1, x2, x3, x4)
```

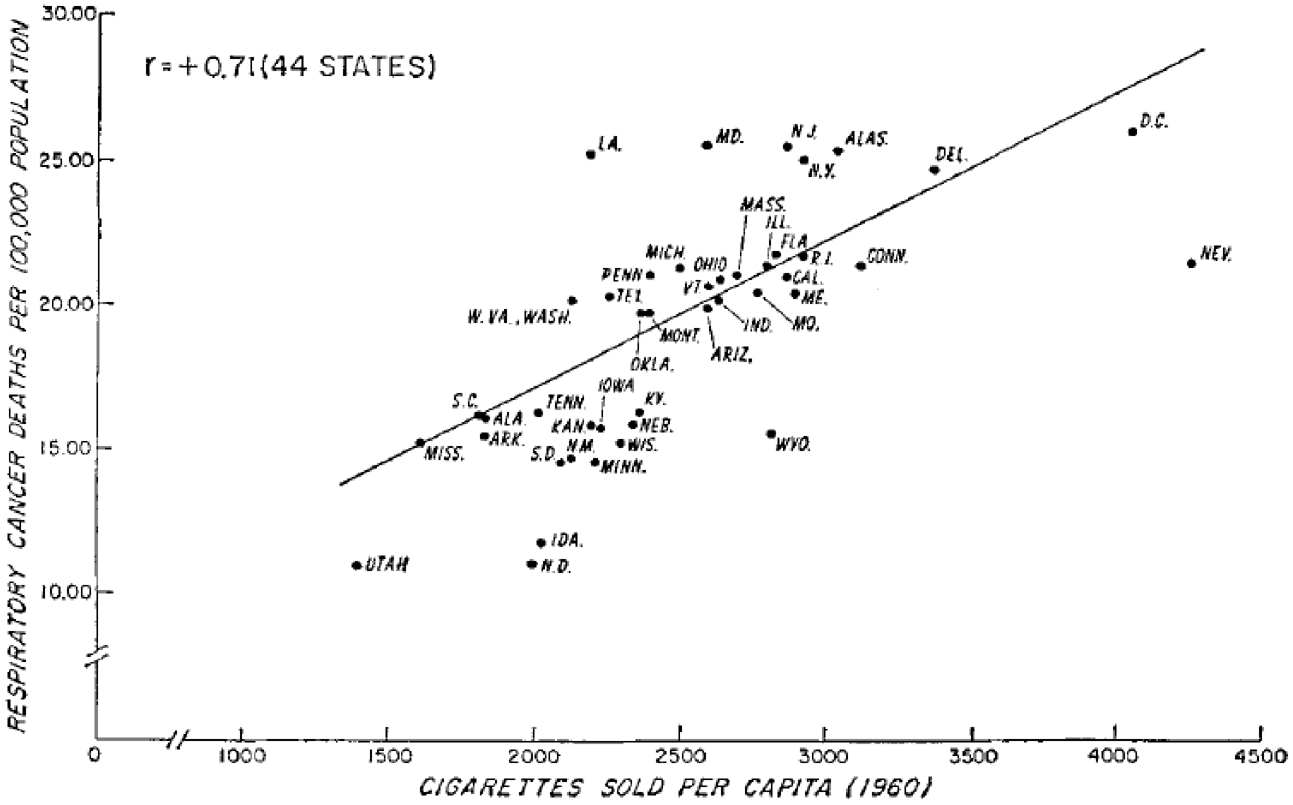
Relationships between two quantitative variables

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.

JOURNAL OF THE NATIONAL CANCER INSTITUTE

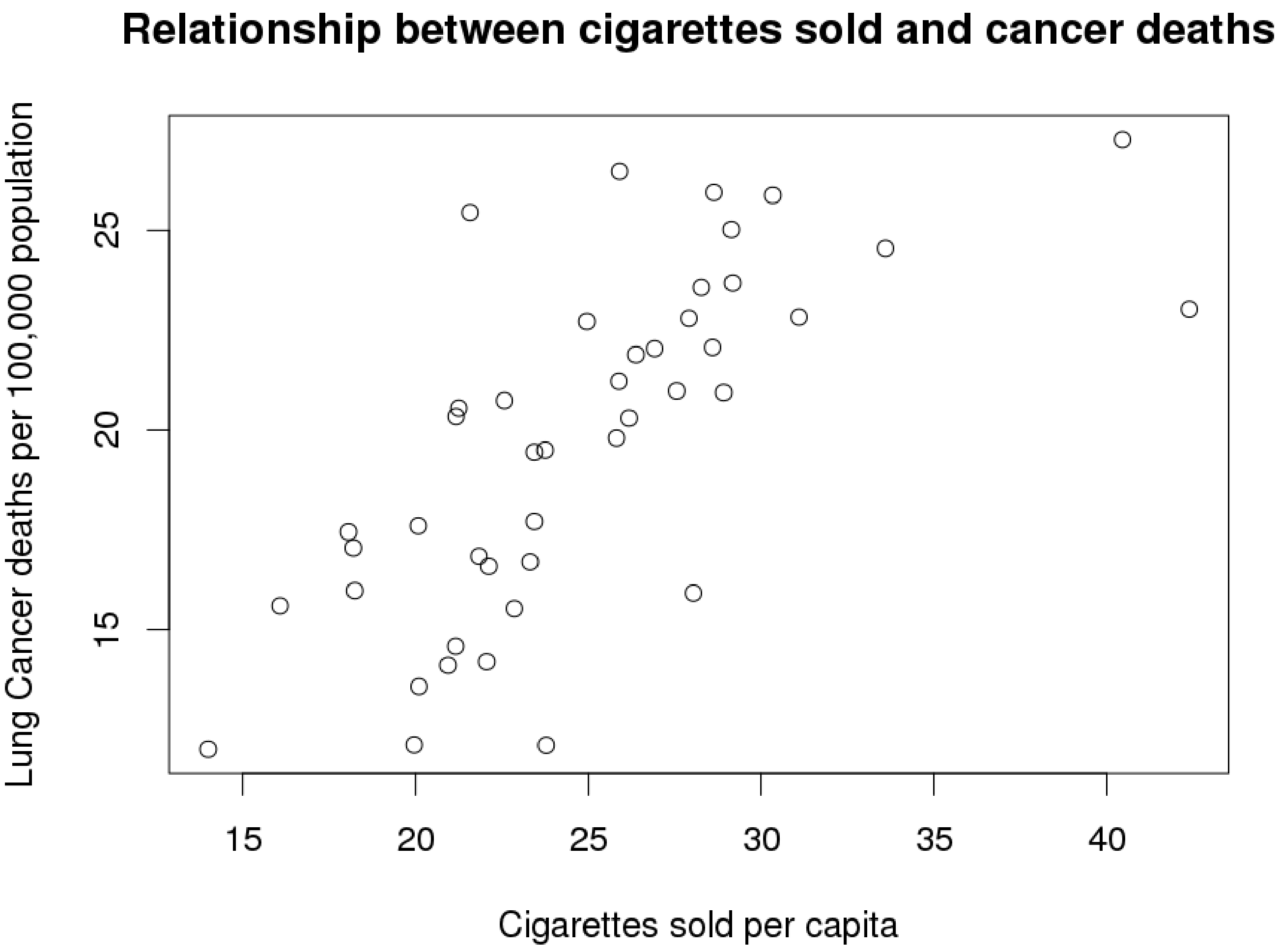
Scatterplot

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?

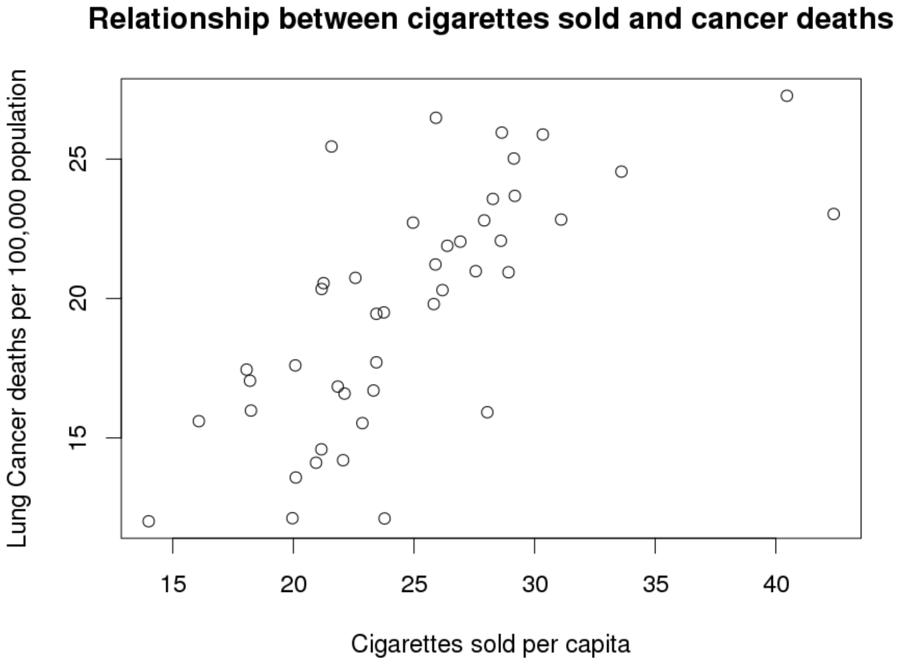
  Does it go upward or downward?

  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



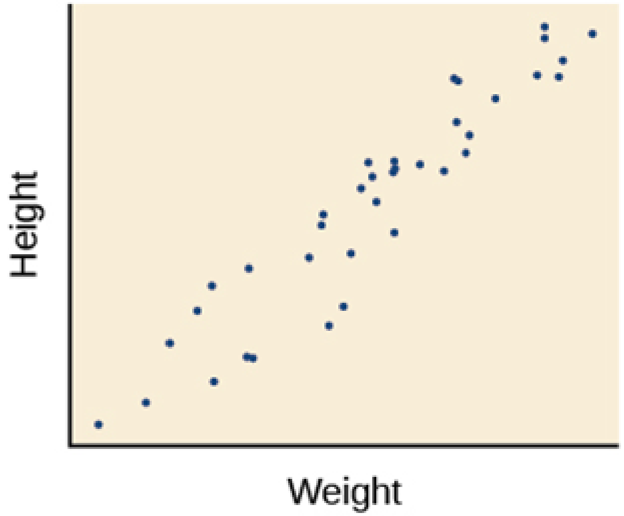
Positive, negative, no correlation

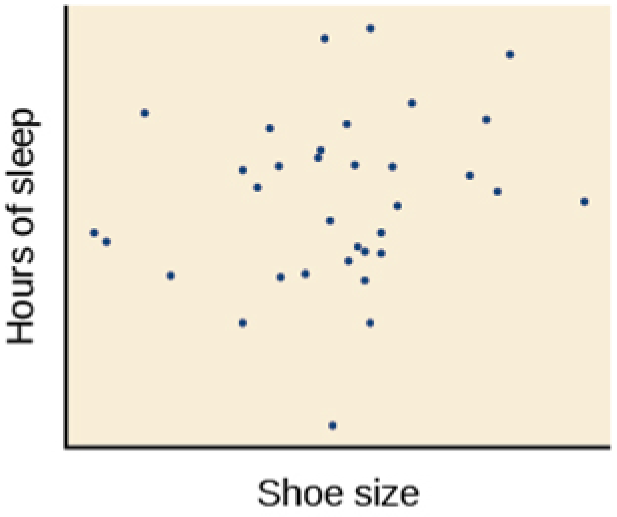
Do the points show a clear trend?

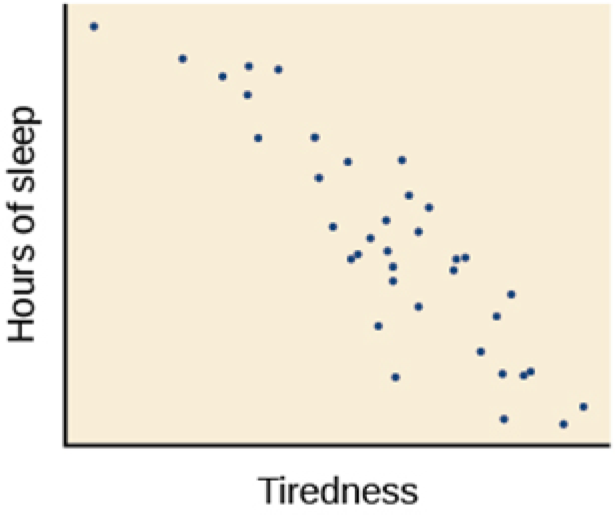
- Does it go upward or downward?
- How much scatter around the trend?

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

Are there any outlier points?







The correlation coefficient

The **correlation** is measure of the strength and direction of a <u>linear association</u> between two variables

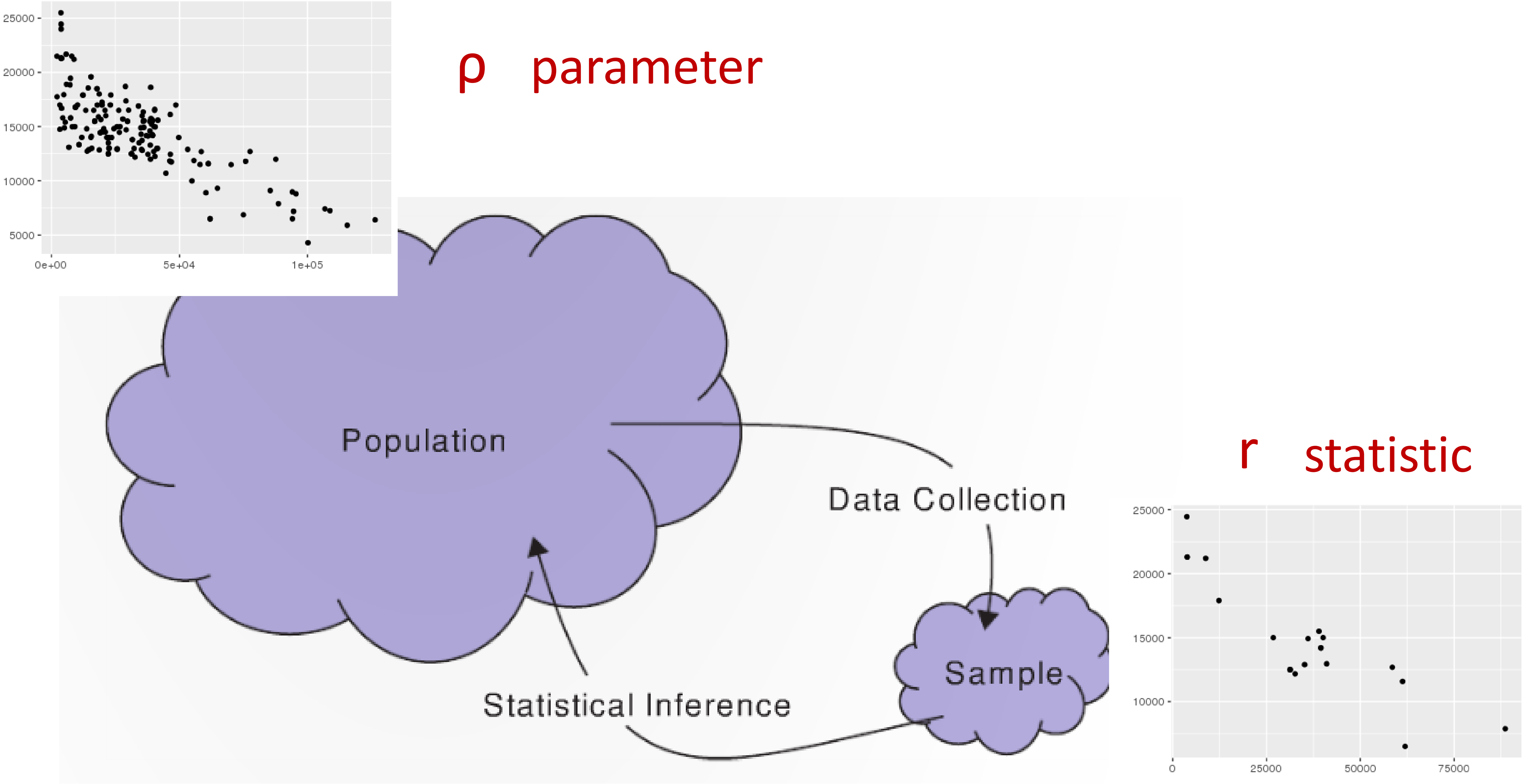
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

- 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

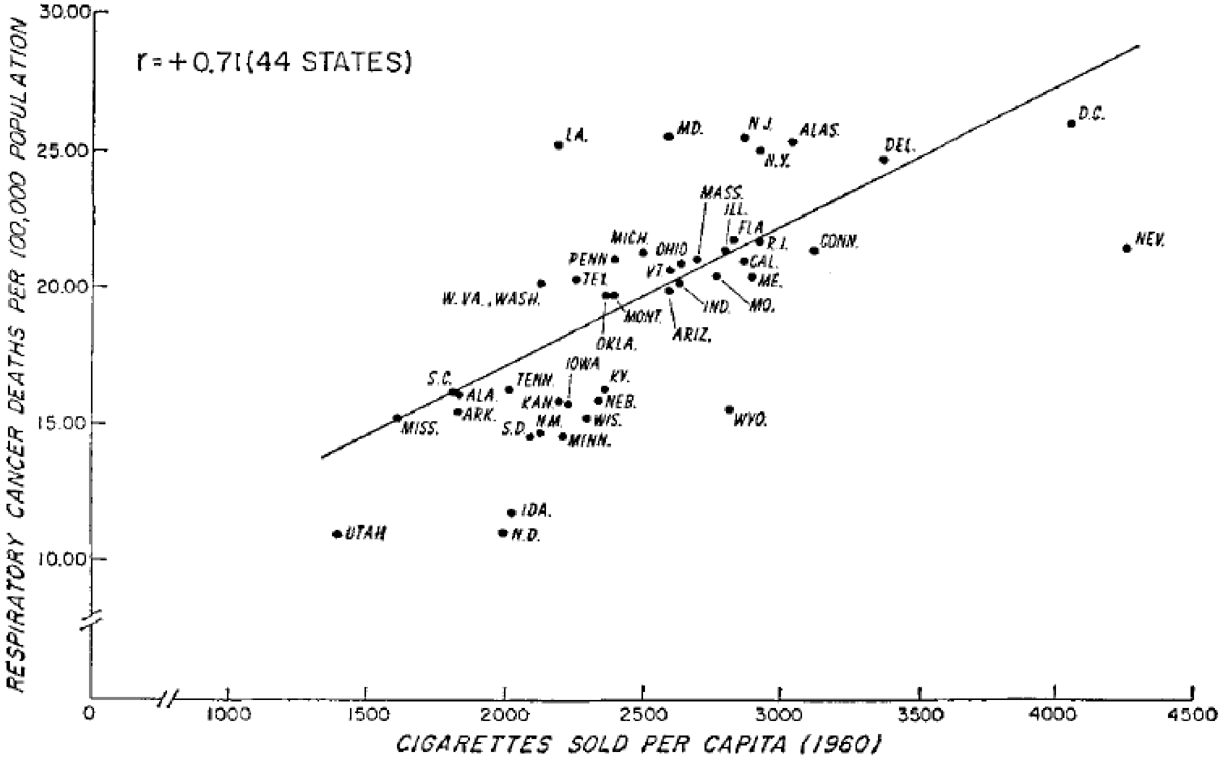
r statistic



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

JOURNAL OF THE NATIONAL CANCER INSTITUTE

Properties of the correlation

Correlation as always between -1 and 1: -1 ≤ r ≤ 1

The sign of r indicates the direction of the association

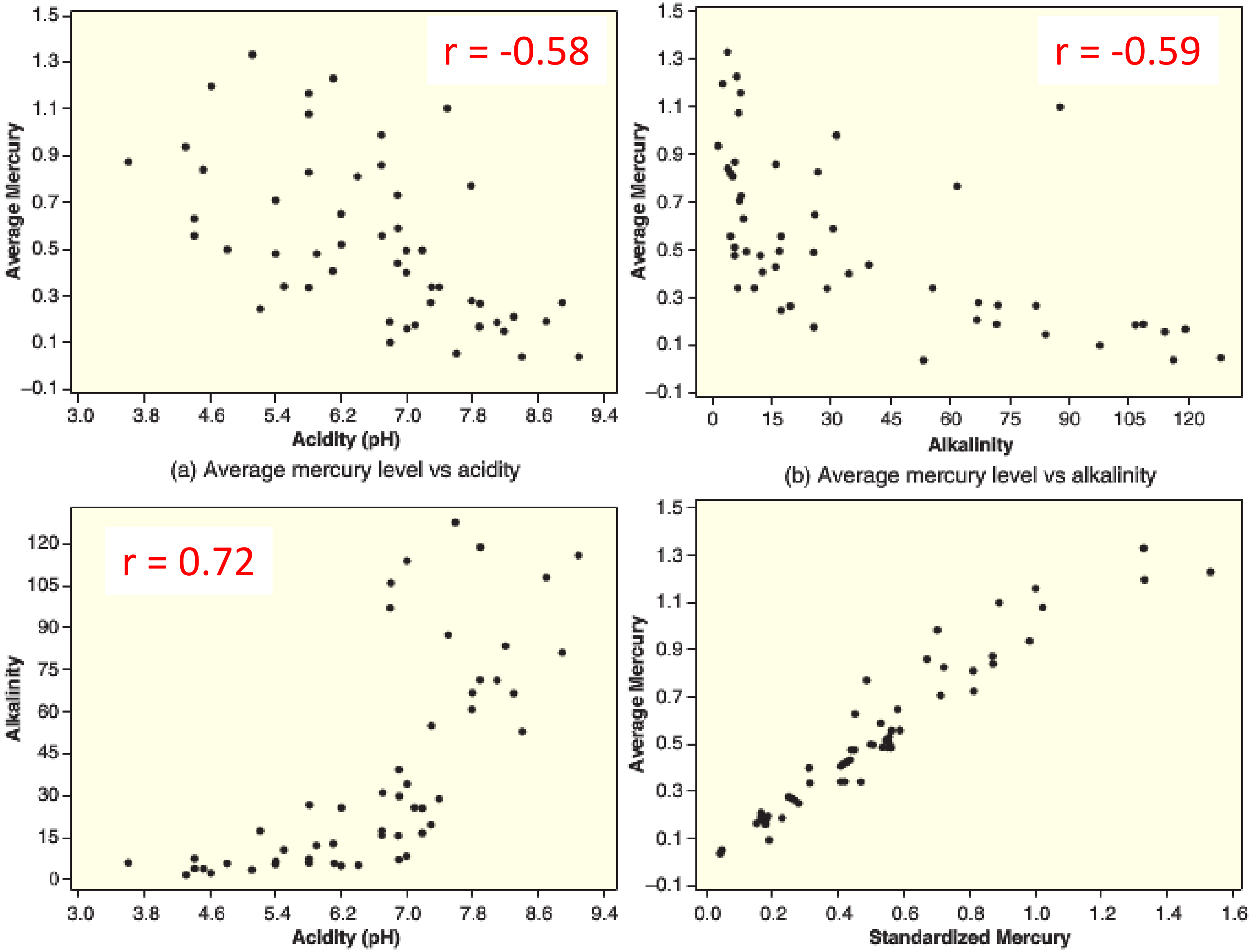
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: r = cor(x, y) = cor(y, x)

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

Florida lakes

Correlation game



r = -0.58

(a) Average mercury level vs acidity

r = -0.59

(b) Average mercury level vs alkalinity

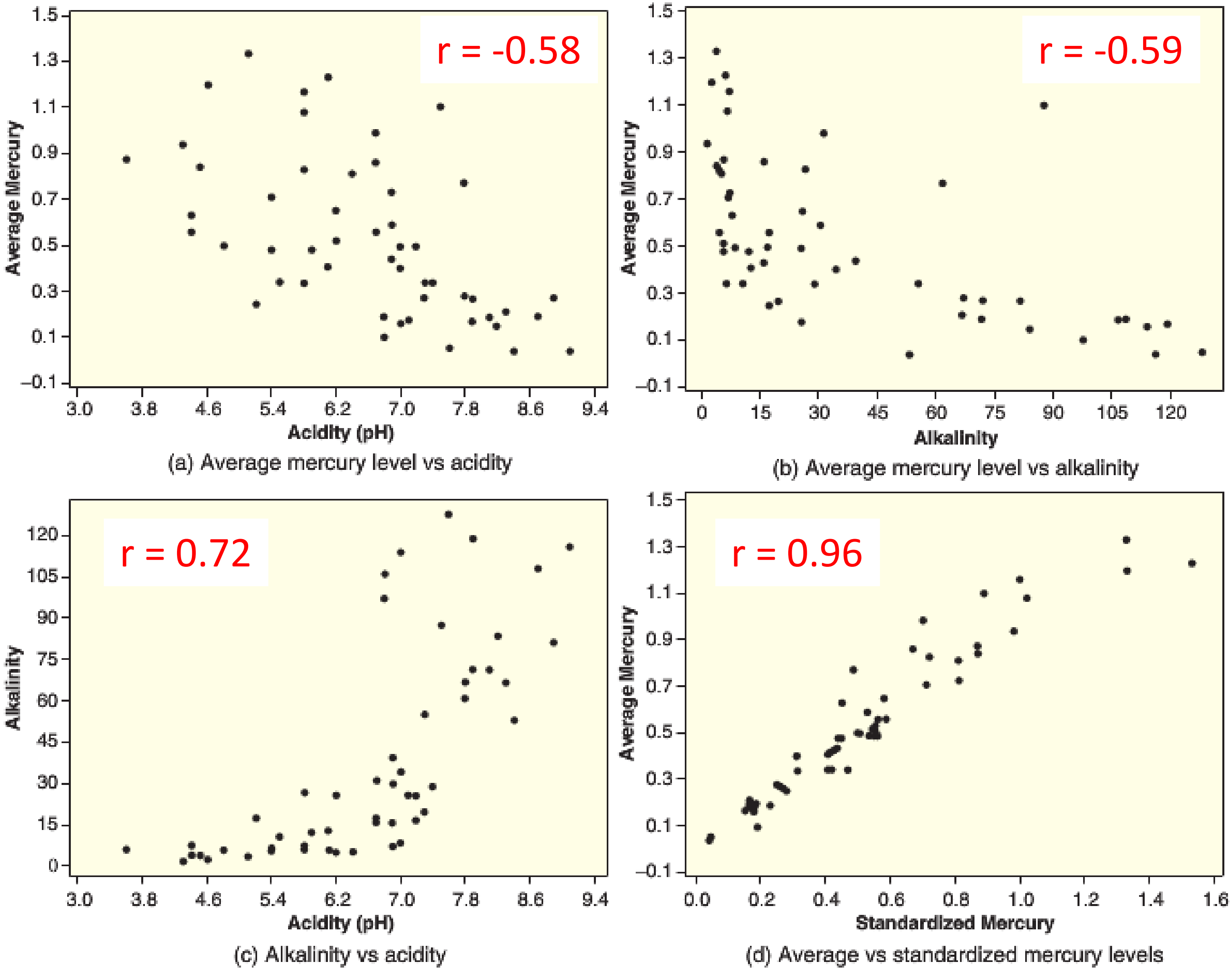
r = 0.72

(c) Alkalinity vs acidity

(d) Average vs standardized mercury levels

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?

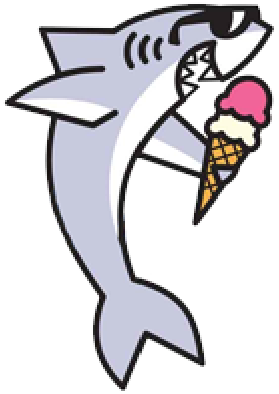
- Bladder cancer (BLAD)
- Kidney cancer (KID)
- Leukemia (LEUK)

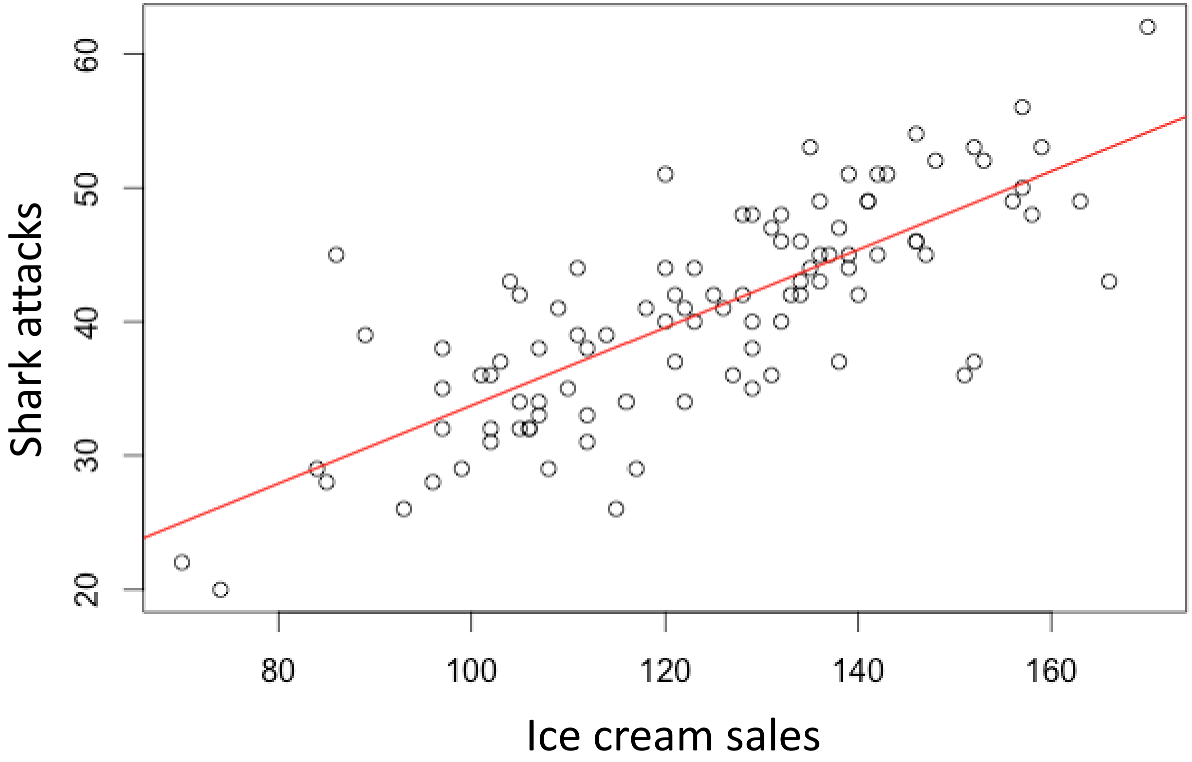
```
# load the data
> download_data("smoking_cancer.Rda")
> load("smoking_cancer.Rda")
```

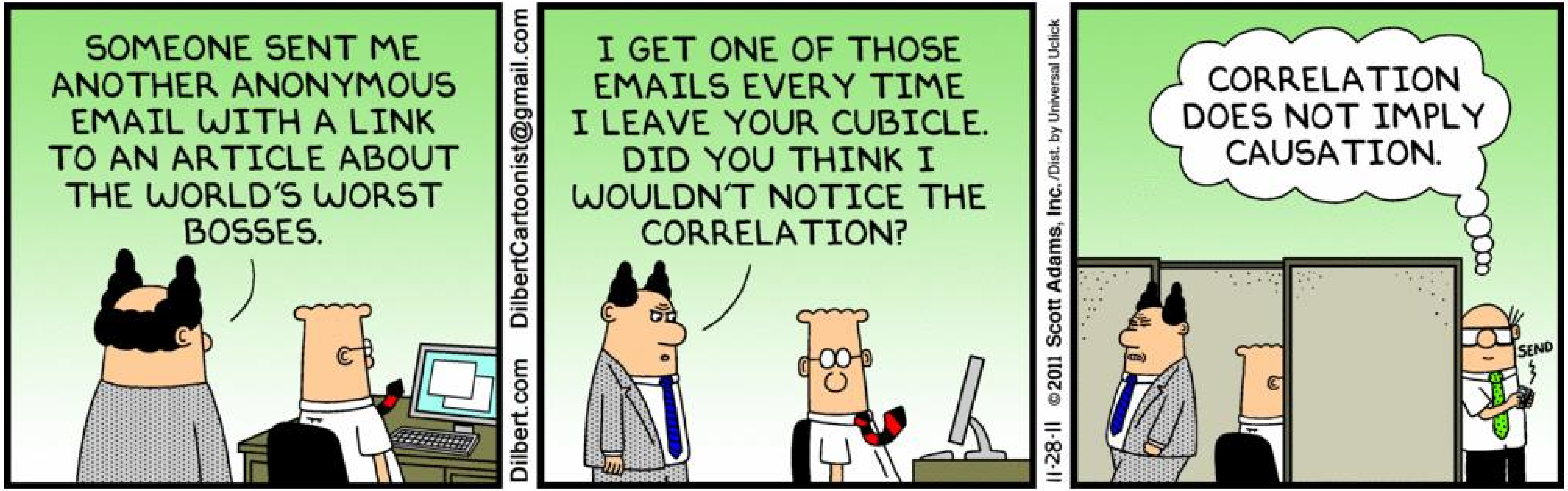
```
# create a scatter plot and calculate the correlation
> plot(smoking$CIG, smoking$LUNG)
> cor(smoking$CIG, smoking$LUNG)
```

Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



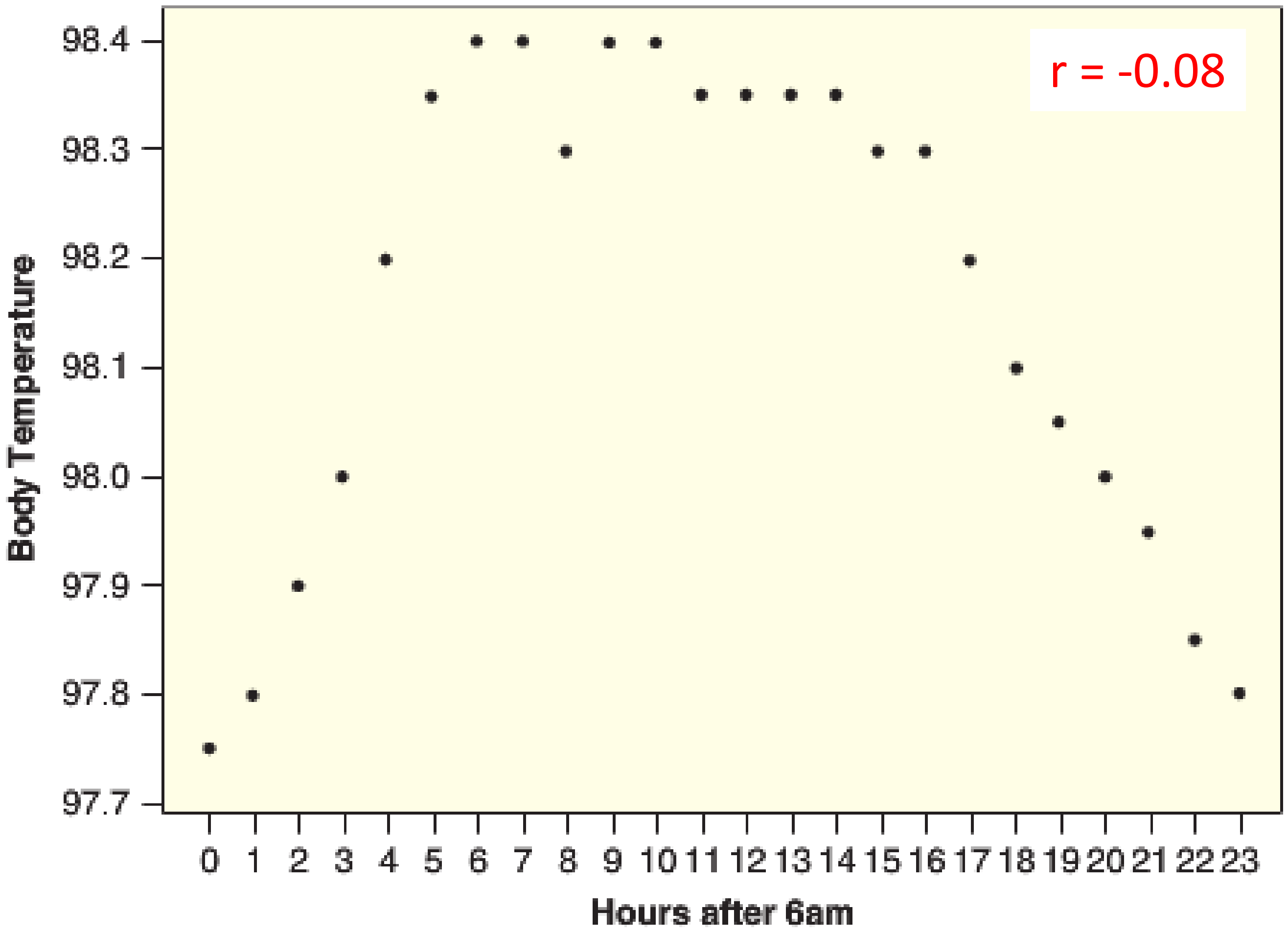


SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
CORRELATION DOES NOT IMPLY CAUSATION.
SEND
Dilbert.com DilbertCartoonist@gmail.com
11-28-11 ©2011 Scott Adams, Inc./Dist. by Universal Uclick

Correlation caution #2

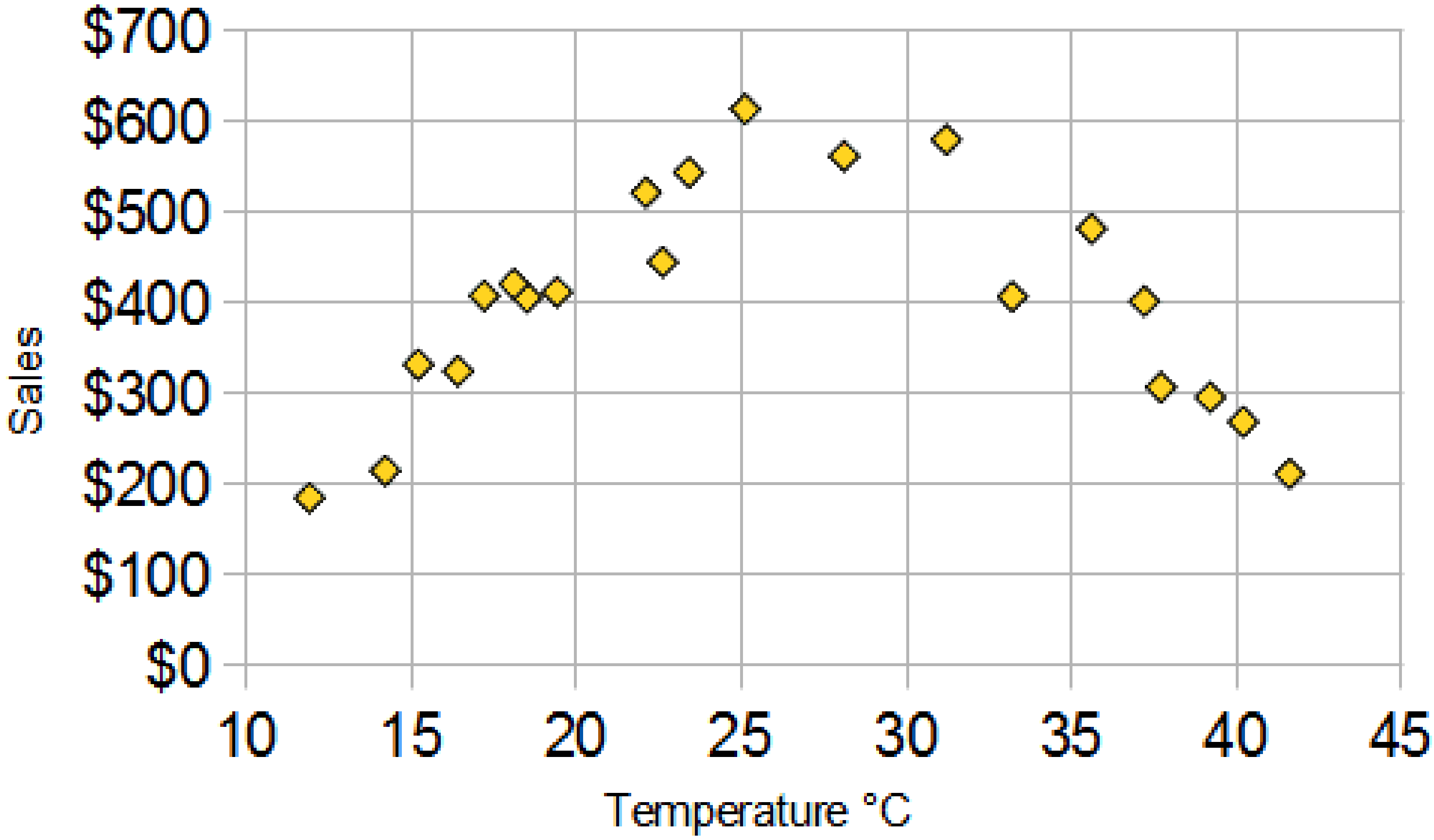
A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day



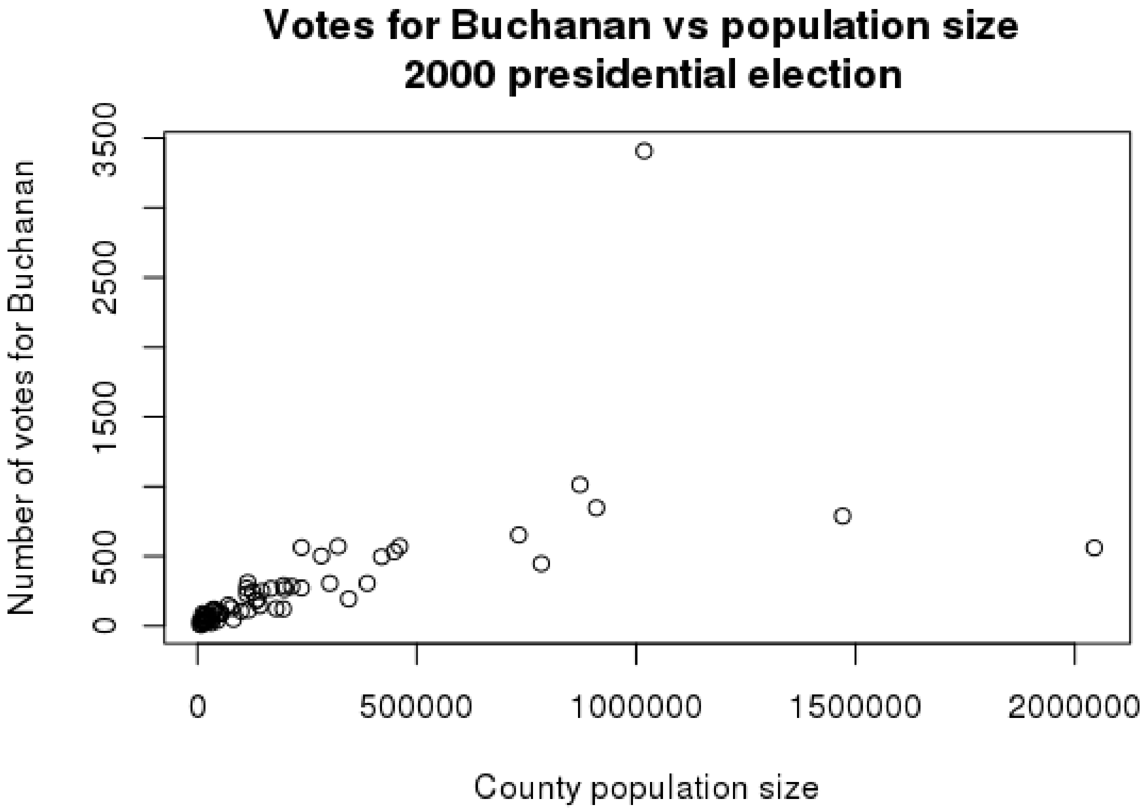
r = -0.08

Ice cream sales and temperature



Correlation caution #3

Correlation can be heavily influenced by outliers. Always plot your data!



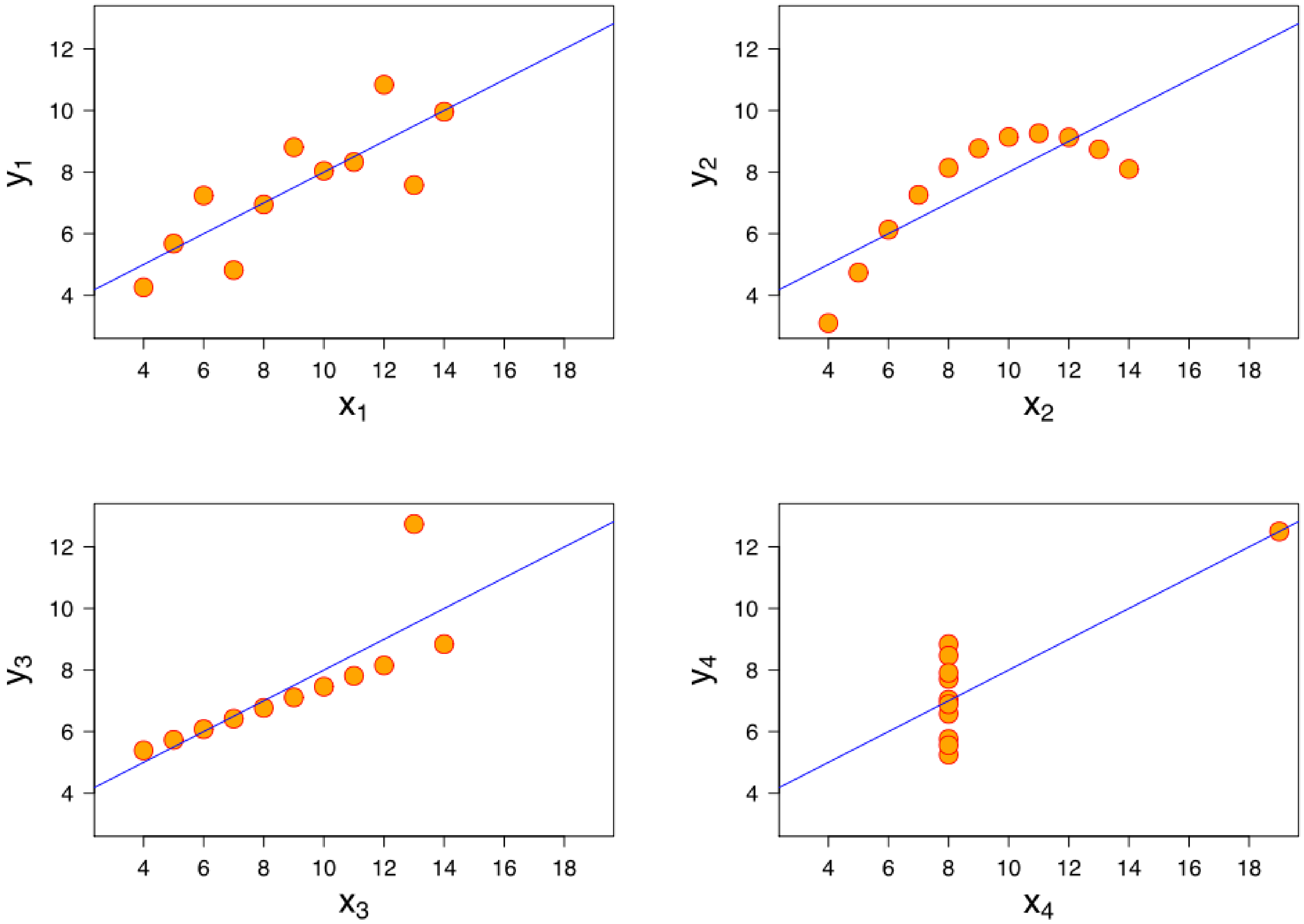
With Palm Beach

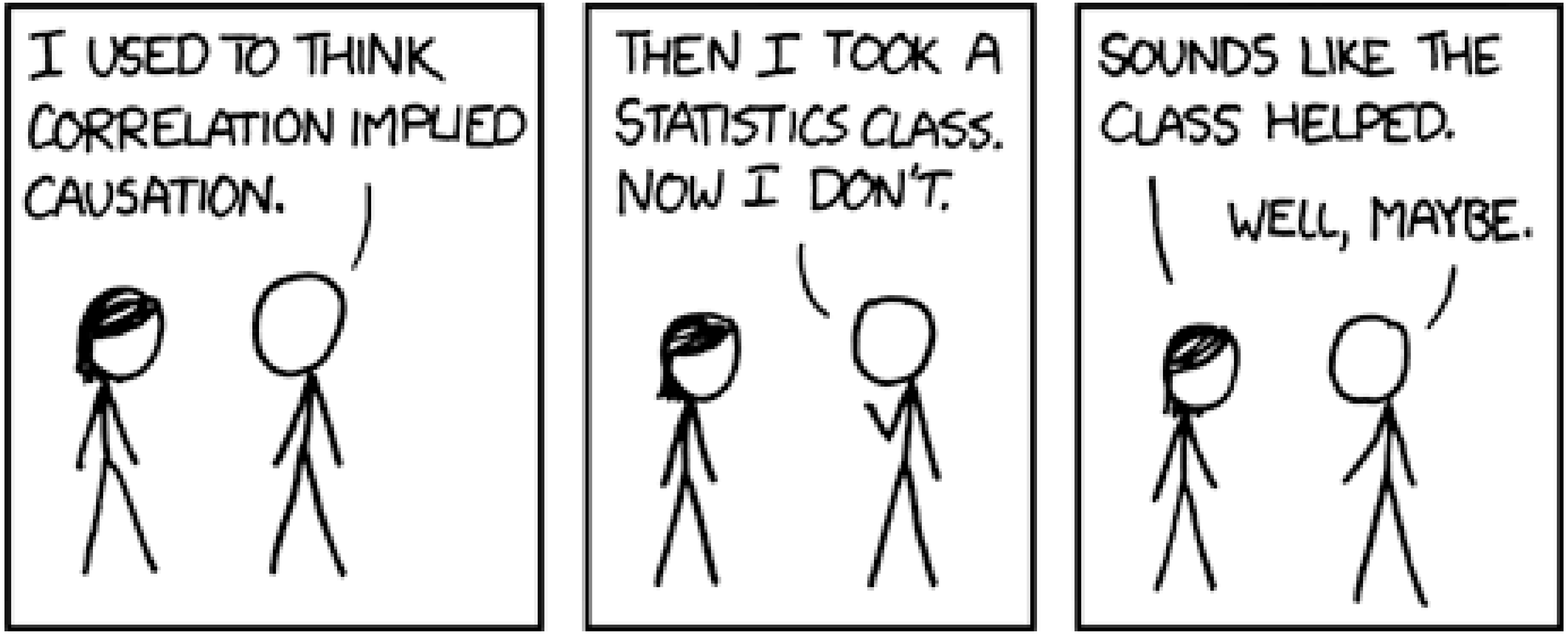
r = 0.61

Without Palm Beach

r = .78

Anscombe’s quartet (r = 0.81)




I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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