Measures of spread continued

Overview

Quick review of shape, central tendency, and spread of quantitative data

Continuation of measures of variability

- Z-scores
- Percentiles
- Box plot

If there is time: correlation

Announcement: homework 2

Homework 2 is due on Gradescope on Sunday, February 1st at 11pm

```
library(SDS1000)
goto_homework(2)
```

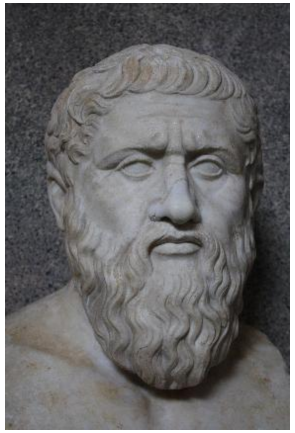
The TA office hours are on Canvas if you need help with the homework

Keep attending the practice sessions!

Quick review of…

Quantitative variables

Underlying concepts: the P’s and the S’s



P-Truth

Population or process

Parameter

Plato (Greek symbols)

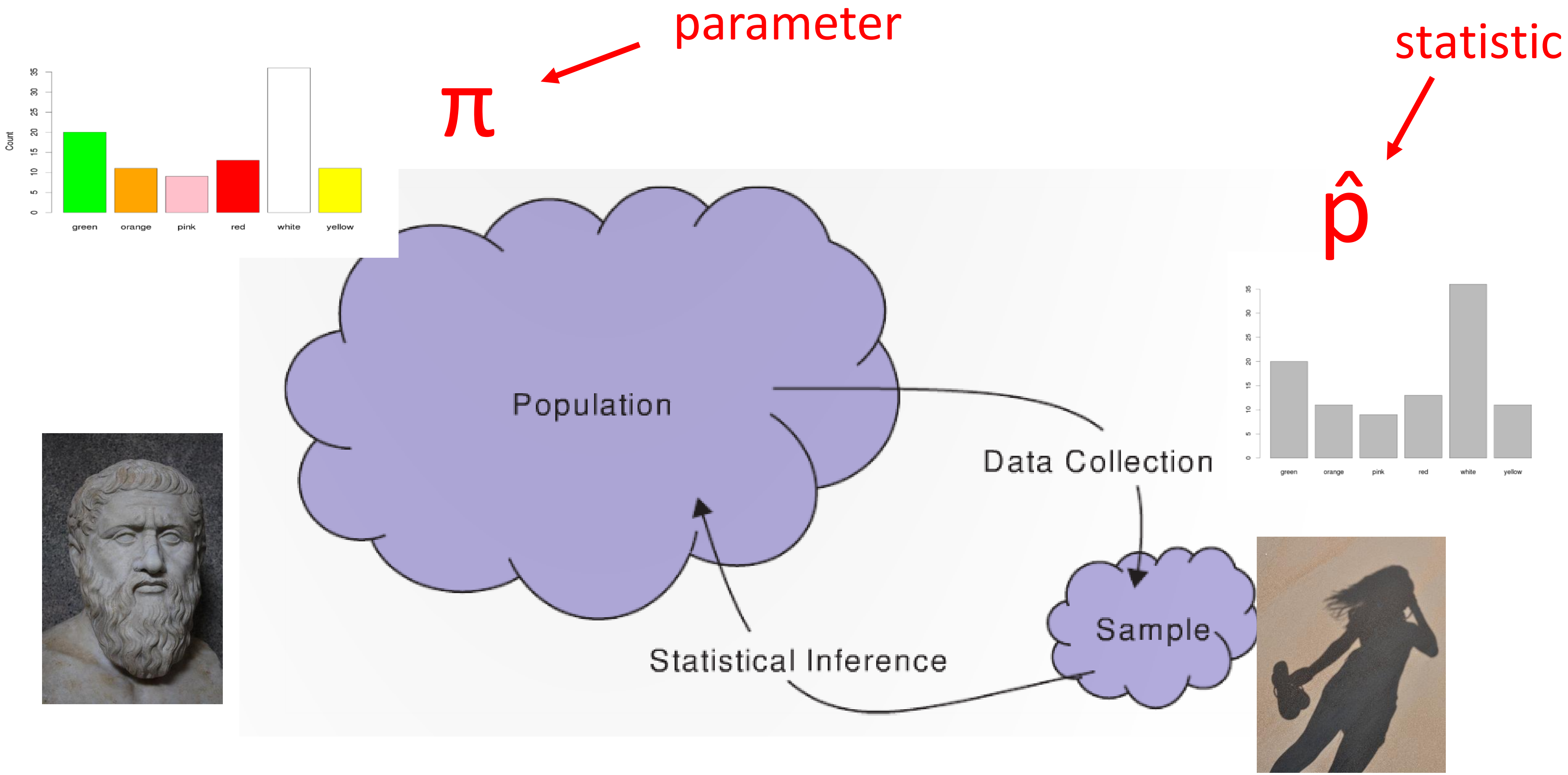
S-shadows

Sample

Statistic

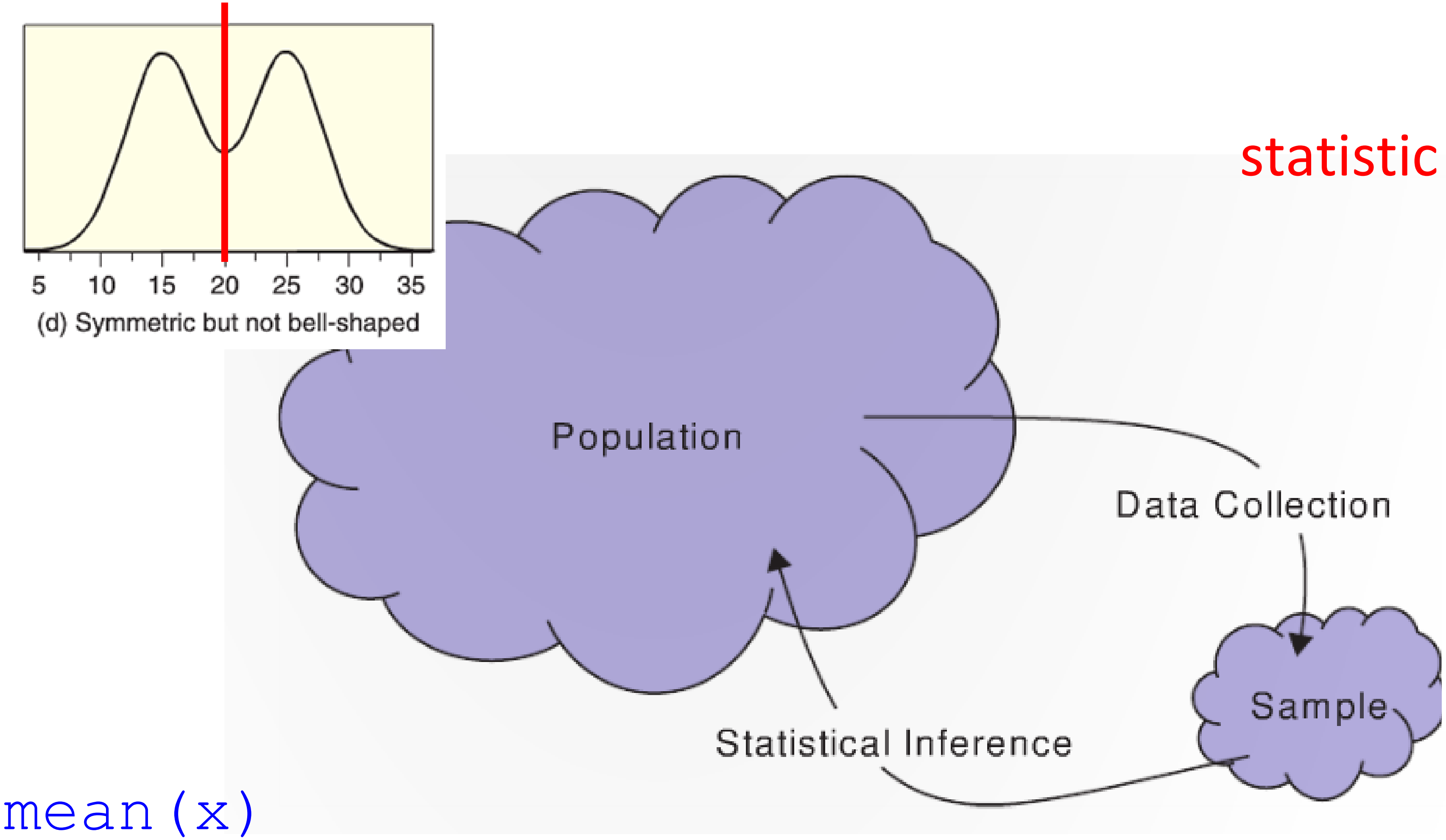
Shadow (Latin symbols)

Review: Categorical data and proportions



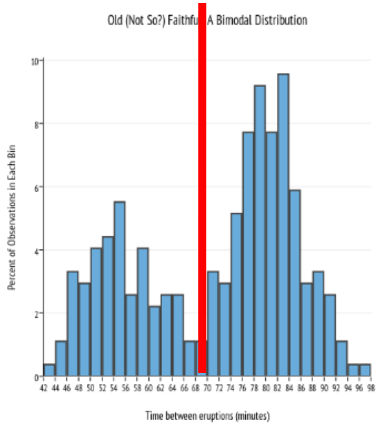
Review: Quantitative data and the mean

µ ← parameter



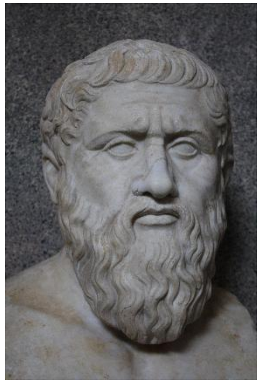
$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

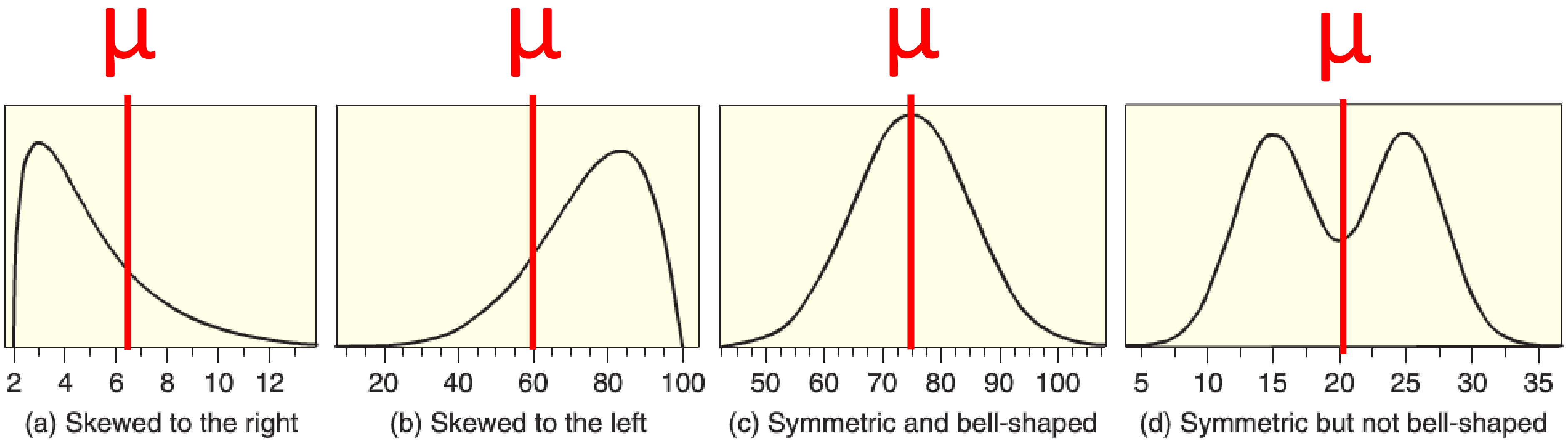
statistic → $\bar{x}$



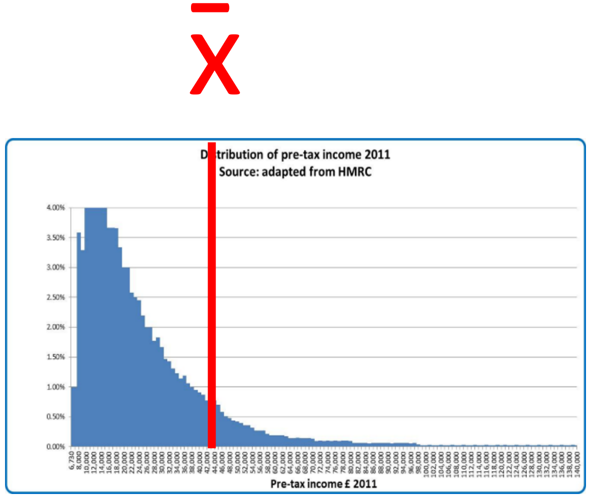
R: `mean(x)`

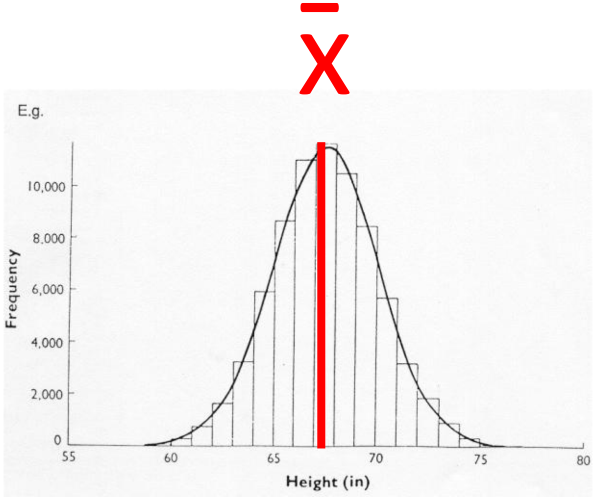
Means for differently shaped distributions

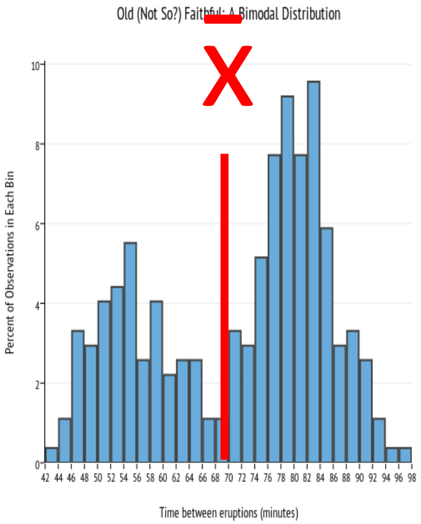




(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

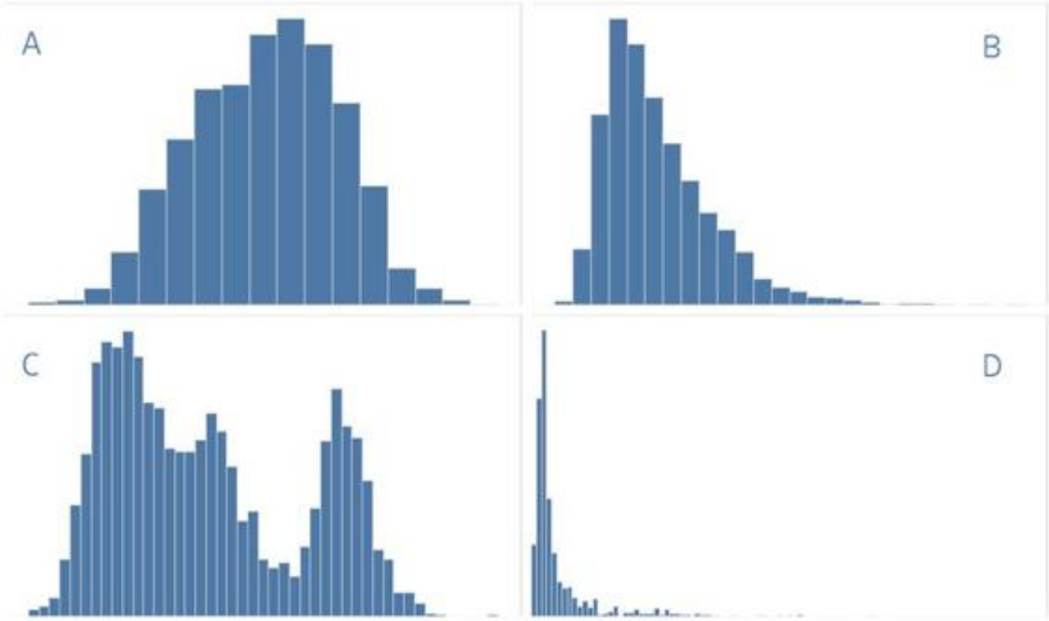






Neat facts – the average NFL player is:

1. **Age**: Is about 25 years old
2. **Height**: Is just over 6'2" in height
3. **Weight**: Weighs a little more than 244lbs
4. **Salary**: Makes slightly less than $1.5M in salary per year

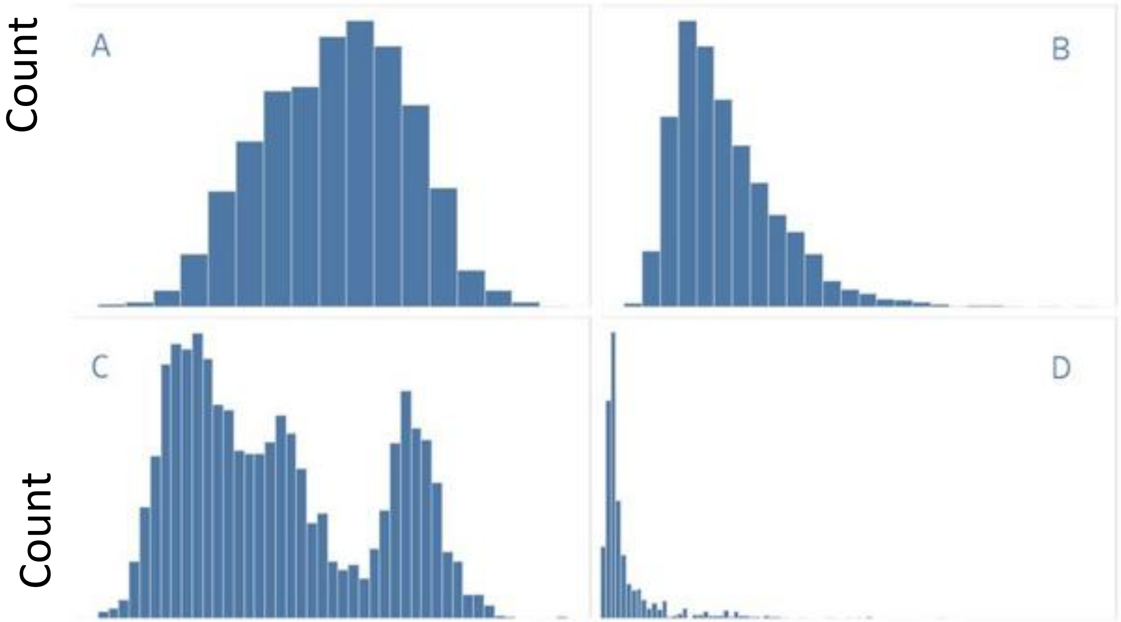


Question: Can you tell which histogram goes with which trait?

Task is to add the labels: **Age, Height, Weight, and Salary**

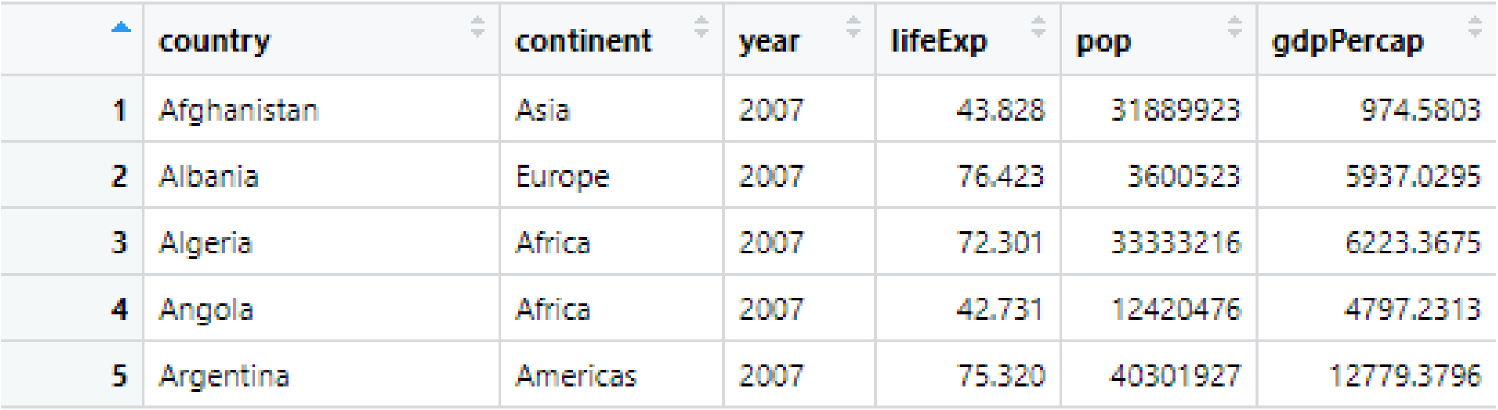
- Hint: There are a wide range of positions in football that have very different roles
  - E.g., placekickers only play for small factions of the game, while quarterbacks are essentially to a team’s success

First: what is the label for the y-axis?



Back to the Gapminder data...

```
# get a data frame with information about the countries in the world
    load("gapminder_2007.Rda")
```



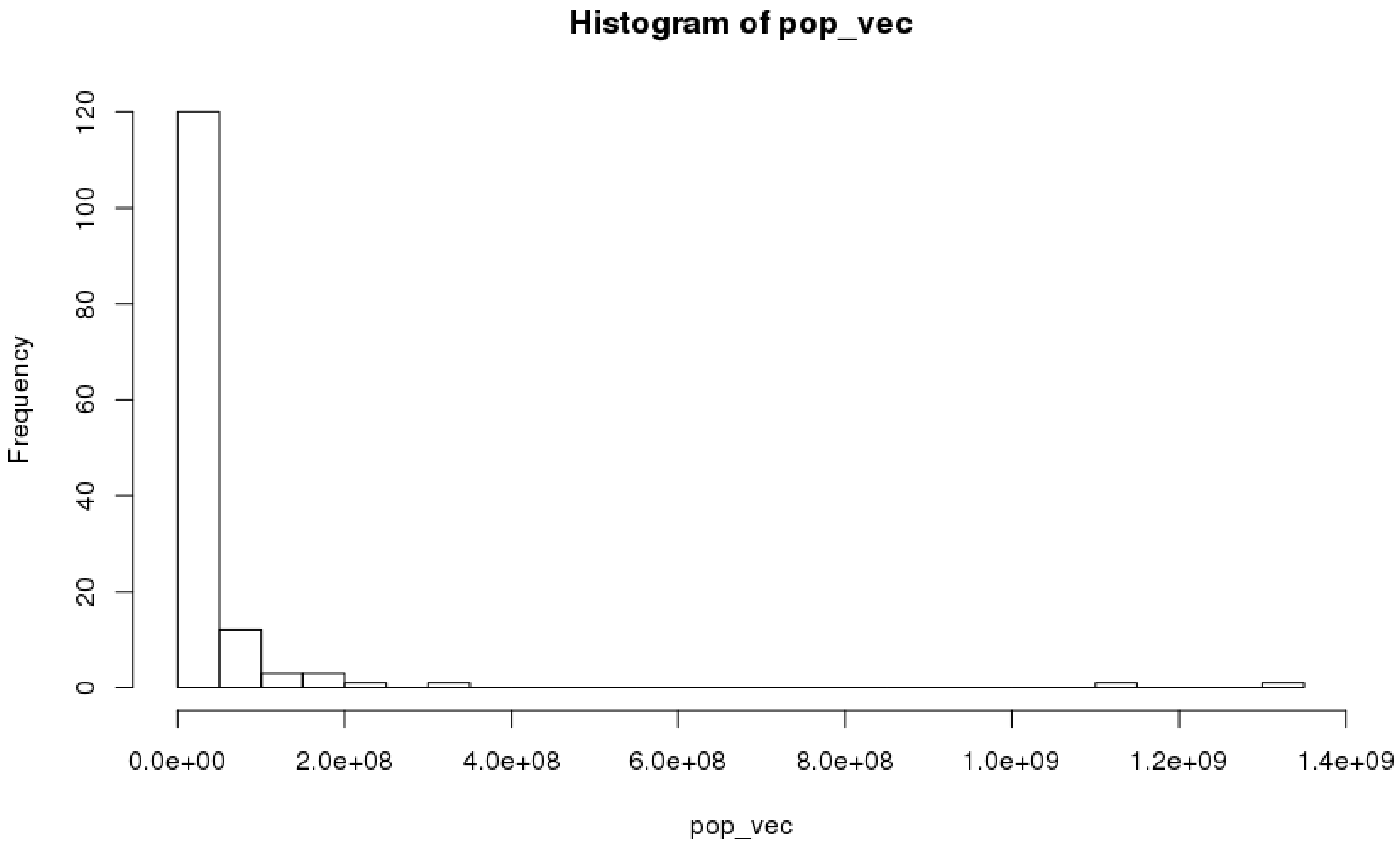
	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	2007	43.828	31889923	974.5803
2	Albania	Europe	2007	76.423	3600523	5937.0295
3	Algeria	Africa	2007	72.301	33333216	6223.3675
4	Angola	Africa	2007	42.731	12420476	4797.2313
5	Argentina	Americas	2007	75.320	40301927	12779.3796

Can you plot a histogram of the population of each country with 20 bins?

```
pop_vec <- gapminder_2007$pop   # create a vector with the pop  of each country

hist(pop_vec, breaks = 20)      # then create the histogram
```

What is missing from this histogram?



Axes labels could be more informative!

Labeling axes

Question: Can you figure out how to label the axes?

```
? hist
```

Answer: xlab and ylab!

```
hist(pop_vec, breaks = 20,
        ylab = "Frequency",
        xlab = "Population",
        main = "World countries population in 2007")
```

Review: The median

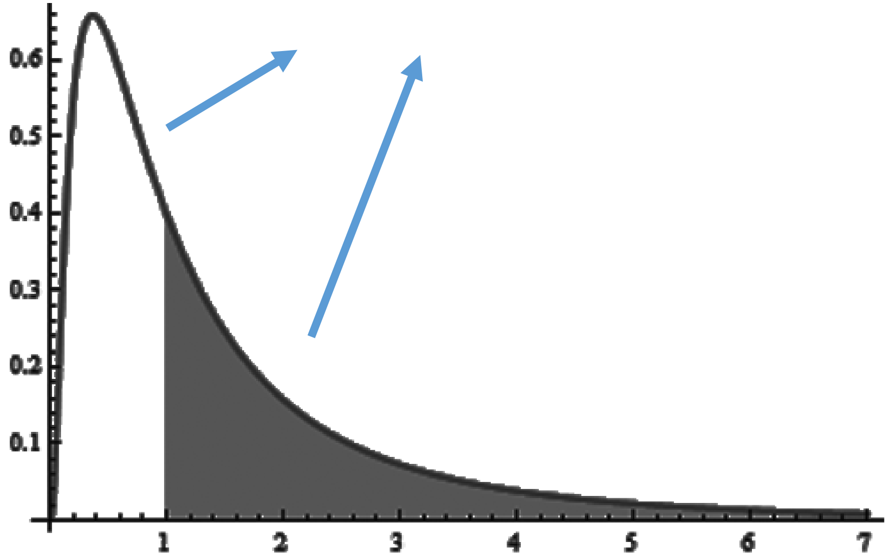
The **median** is a value that splits the data in half

- i.e., half the values in the data are smaller than the median and half are larger

To calculate the median for a data sample of size *n*, sort the data and then:

- If n is odd: The middle value of the sorted data

- If n is even: The average of the middle two values of the sorted data



The white area under the distribution is equal to the gray area

```
R: median(v)
   median(v, na.rm = TRUE)
```

Example of calculating the mean and median

When an individual visits a webpage a ‘ping’ is generated

Below is a random sample of ping counts from 7 people who pinged a website at least once:

4, 6, 10, 12, 45, 59, 158

Question: What is the mean and median ping count in this sample?

A: mean = 42

median = 12

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

1/7 · 294 = 42

Review: outliers

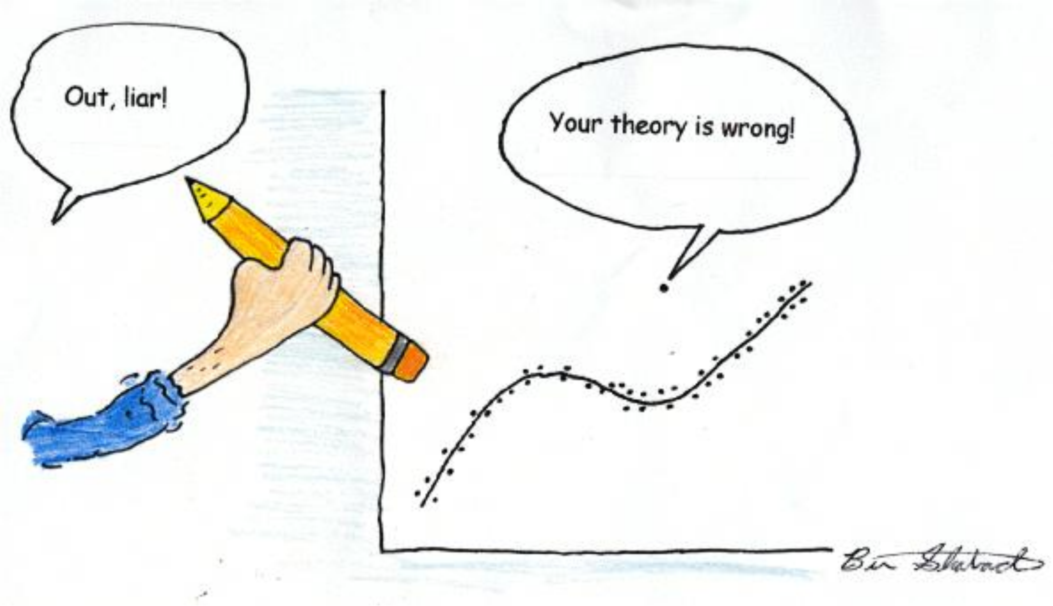
Q: What is an **outlier?**

Q: Why are they problematic?

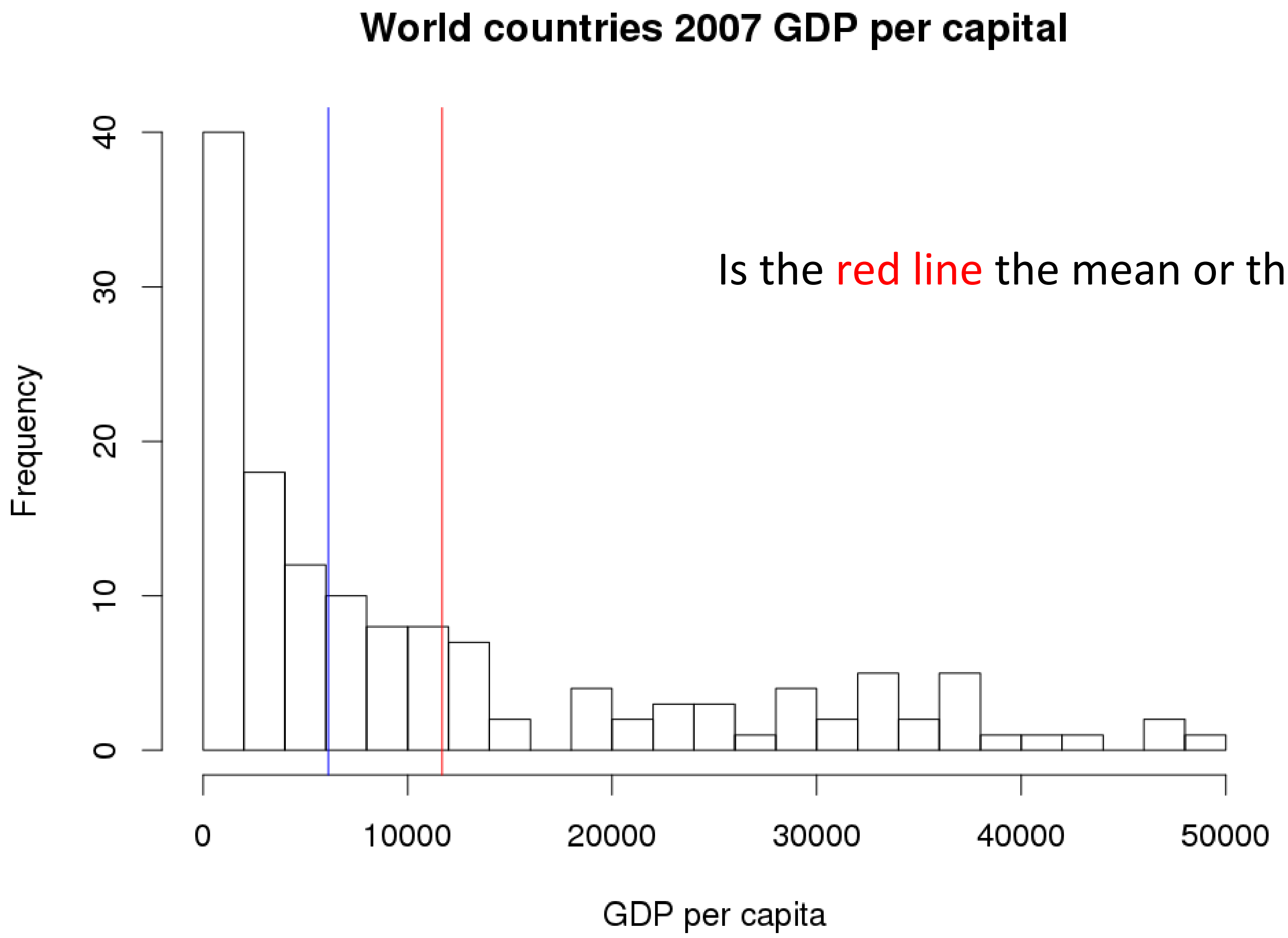
Q: What should you do if you have an outlier in your data?

Q: Is the mean and/or median resistant?

World countries population 2007

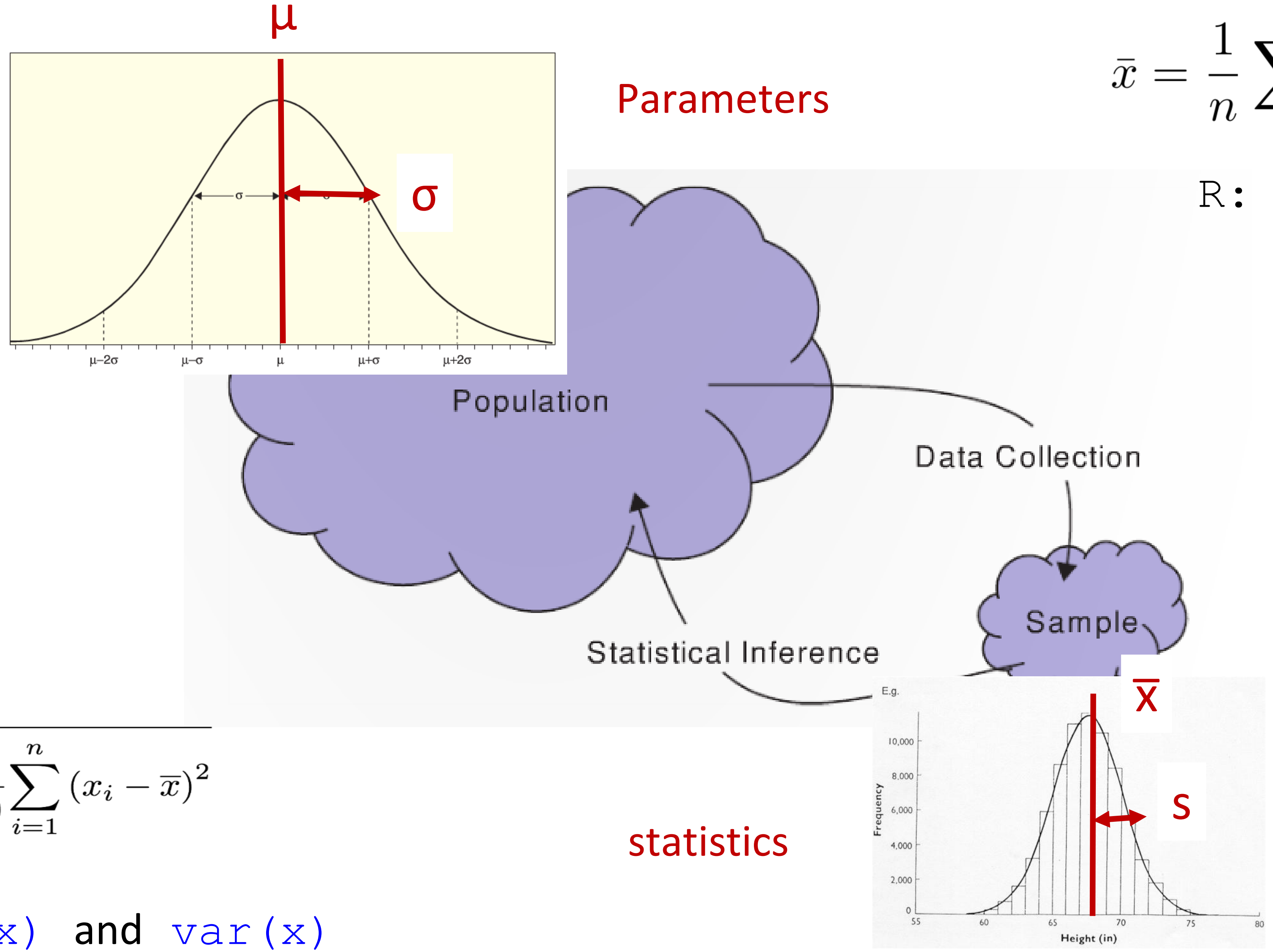


Measure of central tendency: mean and median



Is the red line the mean or the median?

Review measures of spread…



Parameters

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

R: `mean(x)`

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

R: `sd(x)` and `var(x)`

statistics

Normal distributions and z-scores

z-scores

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

- i.e., how far away a point x_i is from $\bar{x}$ in a way that is independent of the units of measurement

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

Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below

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

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

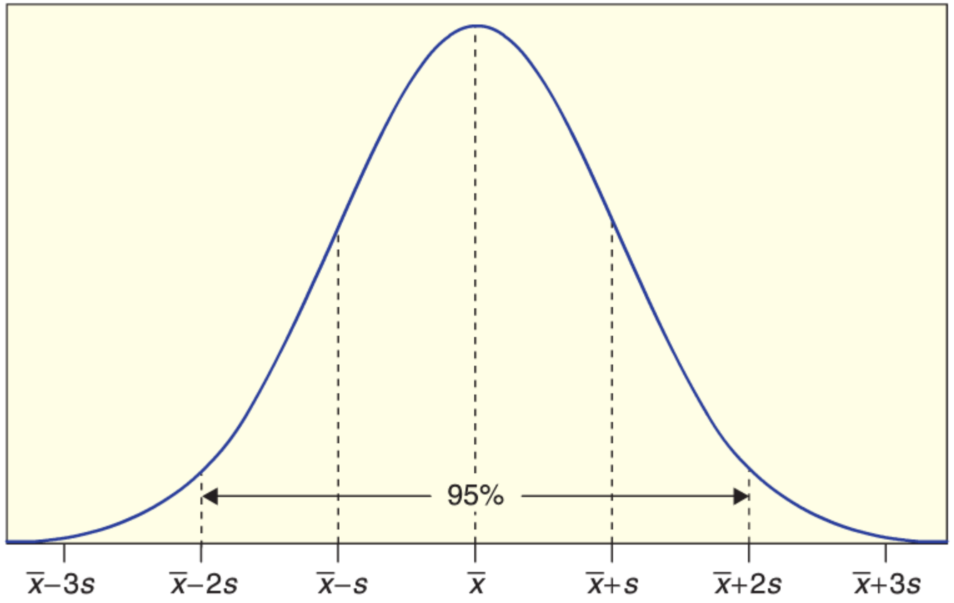
Question: Relative to his peers, which statistic is most and least impressive?

The 95% rule for *normal distributions*

A **normal distribution** is a common distribution that is symmetric and bell shaped

If a distribution of data is approximately normally distributed, about 95% of the data will fall within two standard deviations of the mean

i.e., 95% of the data is in the interval: x̄ -2s to x̄ +2s



If points are normally distributed, should we be impressed with LeBron’s z-score of 2.7?

The 95% rule for *normal distributions*

A **normal distribution** is a common distribution that is symmetric and bell shaped

If a distribution of data is approximately normally distributed, about 95% of the data will fall within two standard deviations of the mean

i.e., 95% of the data is in the interval: $\bar{x}$ -2s to $\bar{x}$ +2s

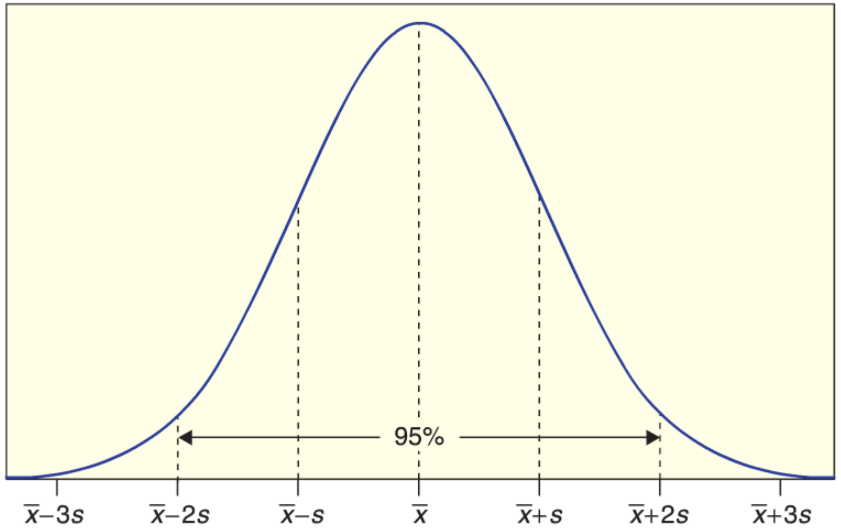
Example: IQ scores are normally distributed with a mean of 100 and a standard deviation of 15

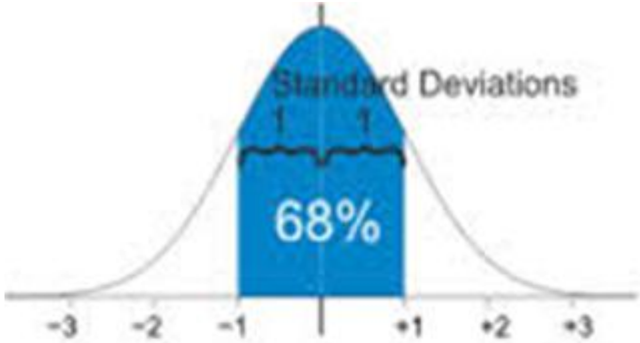
Question: what is the range of values that the middle 95% of IQ scores fall in?

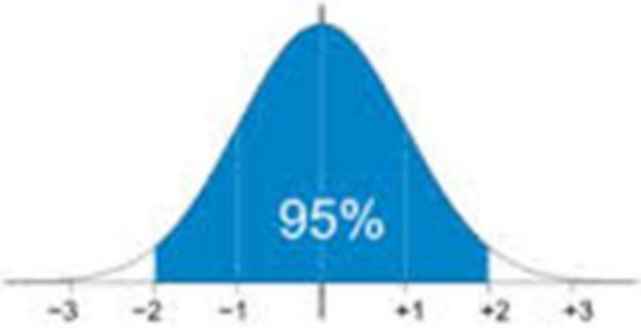
The 68%, 95% and 99.7% rules for *normal distributions*

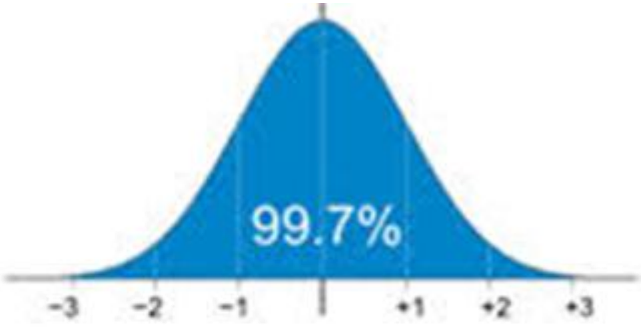
Properties of normal distributions are:

- 68% of the data falls within **one** standard deviations of the mean
- 95% of the data falls within **two** standard deviations of the mean
- 99.7% of the data falls within **three** standard deviations of the mean









Side note: Chebyshev's Inequality

Chebyshev's Inequality: No matter what the shape of the distribution, the proportion of values in the range "average ± $z \cdot$ SDs" is at least $1 - 1/z^2$

Range	**Proportion**
Average ± 2 SDs	at least 1 - 1/4 (75%)
Average ± 3 SDs	at least 1 - 1/9 (88.88…%)
Average ± 4 SDs	at least 1 - 1/16 (93.75%)
Average ± 5 SDs	at least 1 - 1/25 (96%)

Percentiles

Percentiles

The **Pth percentile** is the value (*v*) which is greater than P percent of the data

- i.e., p% of the data is less than *v*

For the US income distribution what are the 20th and 80th percentiles?

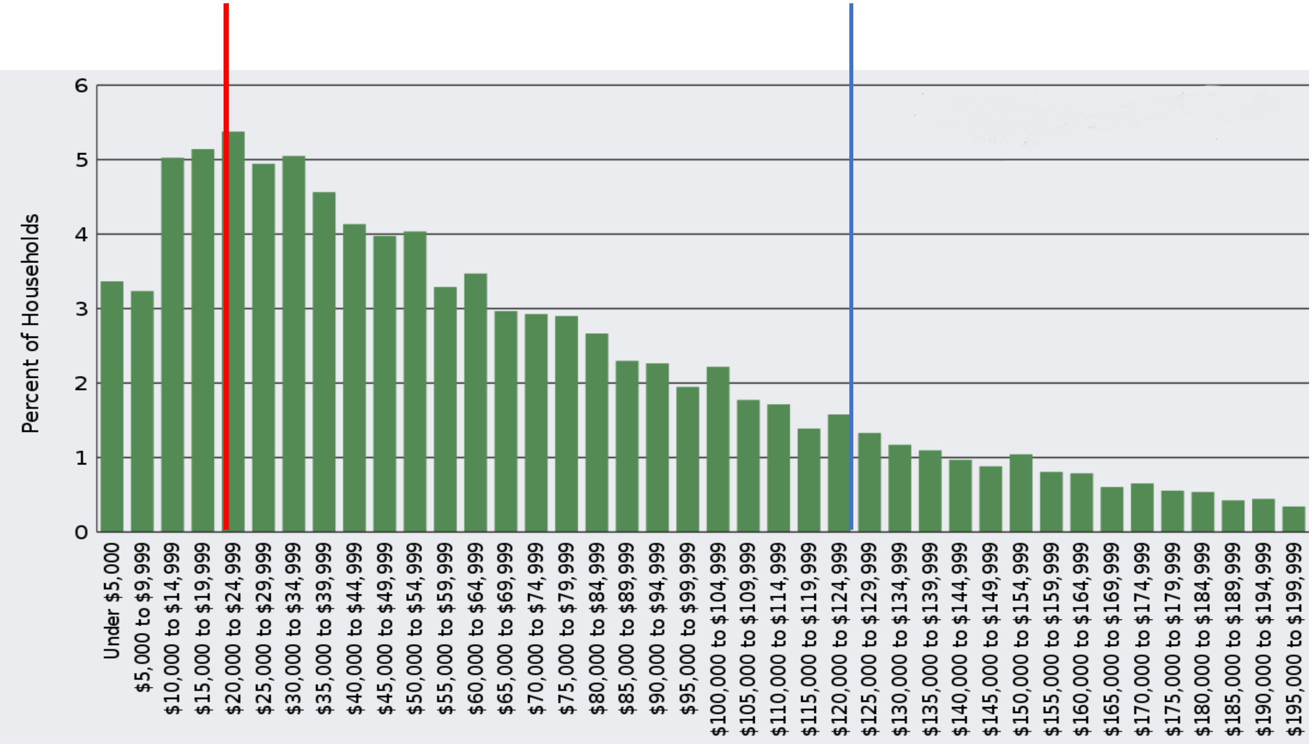
Note of caution: there is not just one way define a percentile

- Some definitions always use values in the data sample
- Other definitions interpolate between data points

R: `quantile(v, .95)`

20th percentile = $21,430

80th percentile = $112, 254



Age of best actors

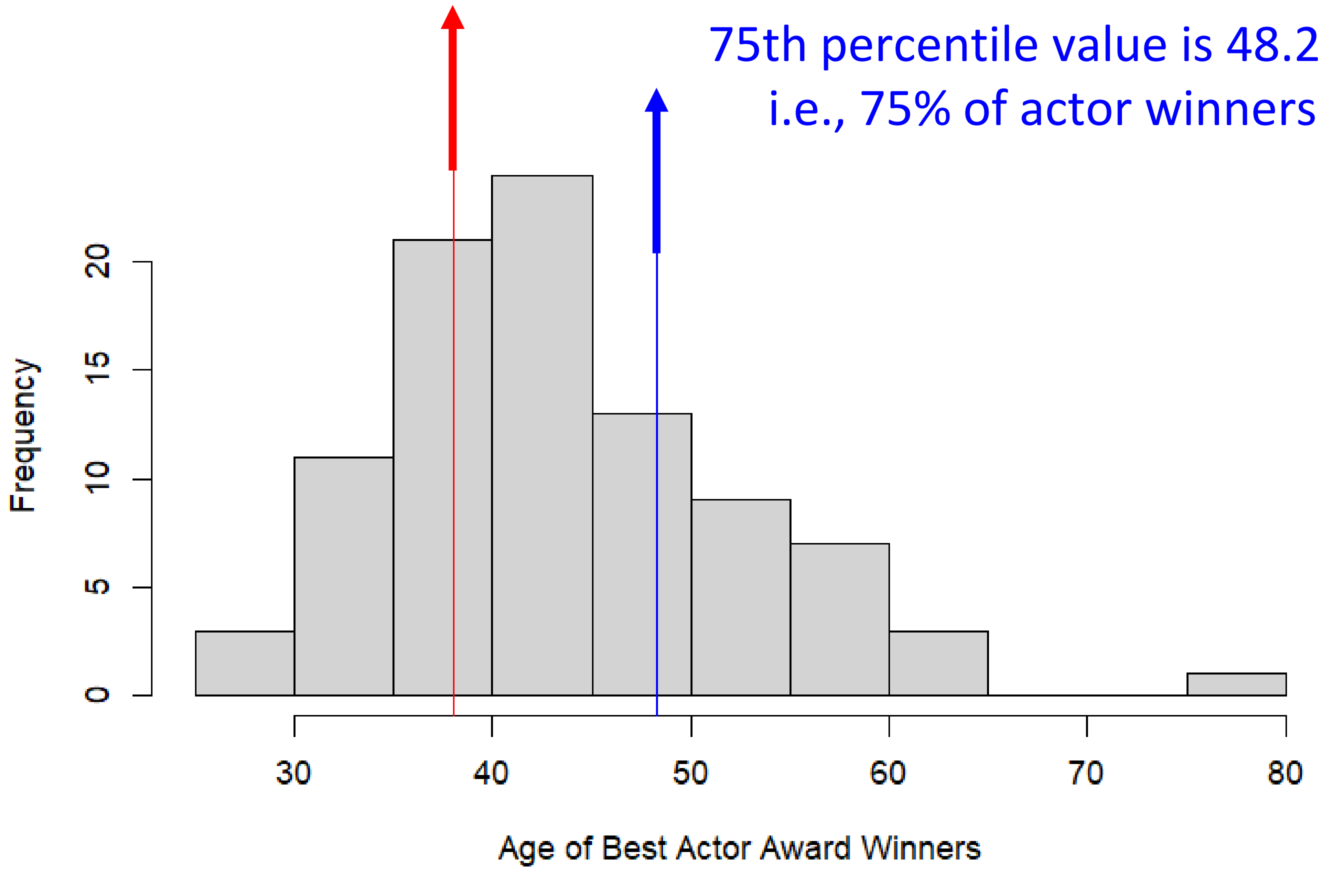
The **Academy Awards**, give out **Oscars** to the best actor and best actress each year

Let’s examine the age of these Oscar winners from the years 1929 to 2012

Age of best actors

25th percentile value is 38

i.e., 25% of actor winners were 38 or younger



75th percentile value is 48.25

i.e., 75% of actor winners were 48.25 or younger

Middle 50% of winners (25^{th} to 75^{th} percentiles)

In the age range:

- 38 to 48.25 years old

Five Number Summary

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

Q_1 = 25th percentile   (also called 1st quartile)

Q_3 = 75th percentile   (also called 3rd quartile)

Roughly divides the data into fourths

Calculating quartiles “by hand”

Our sorted ping data is: 4 6 10 12 45 59 158

1. Calculate the median as the middle of the sorted data

2. For all values less than the median, calculate the median of these values, which will give you Q_1

3. For all values greater than the median, calculate the median of these values, which will give you Q_3

Note of caution (again): there is not just one way define a percentile

- Some definitions always use values in the data sample
- Other definitions interpolate between data points

Range and Interquartile Range

Other measures of spread are:

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

Detecting of outliers

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

Box plots

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

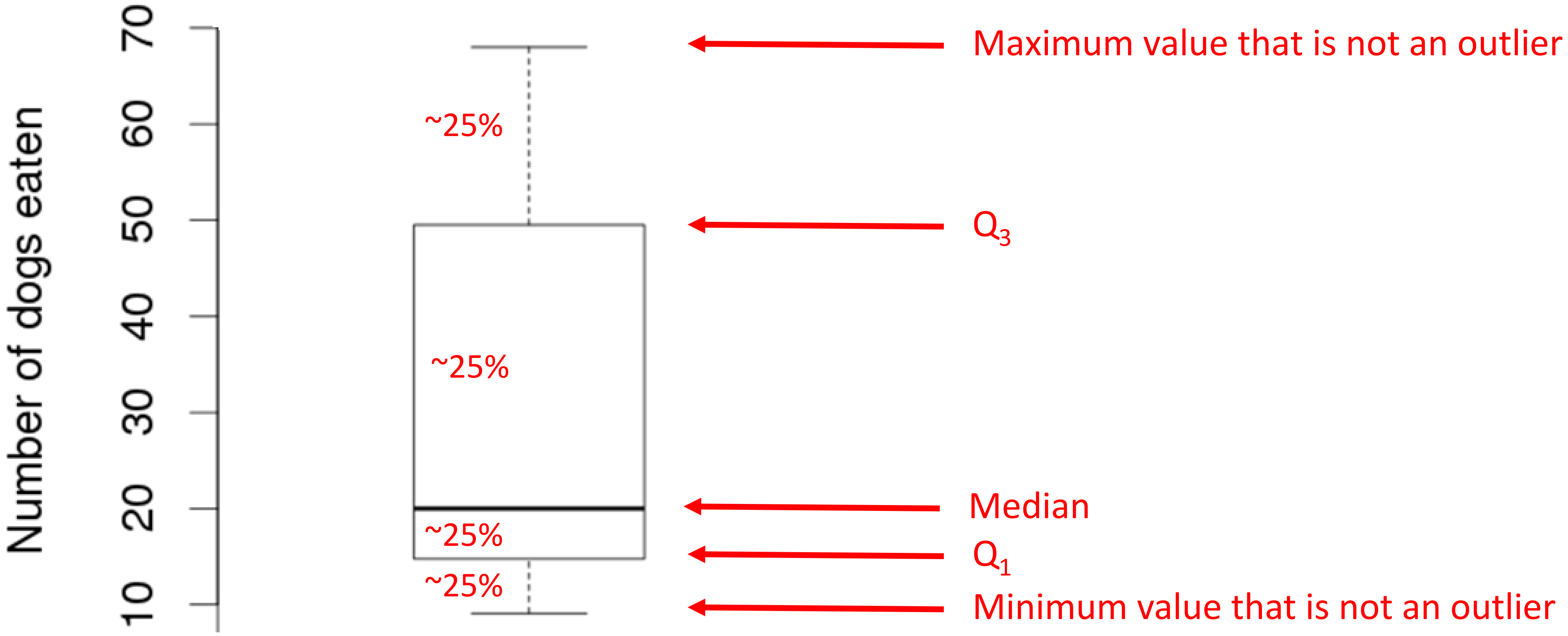
1. Drawing a box from Q_1 to Q_3

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

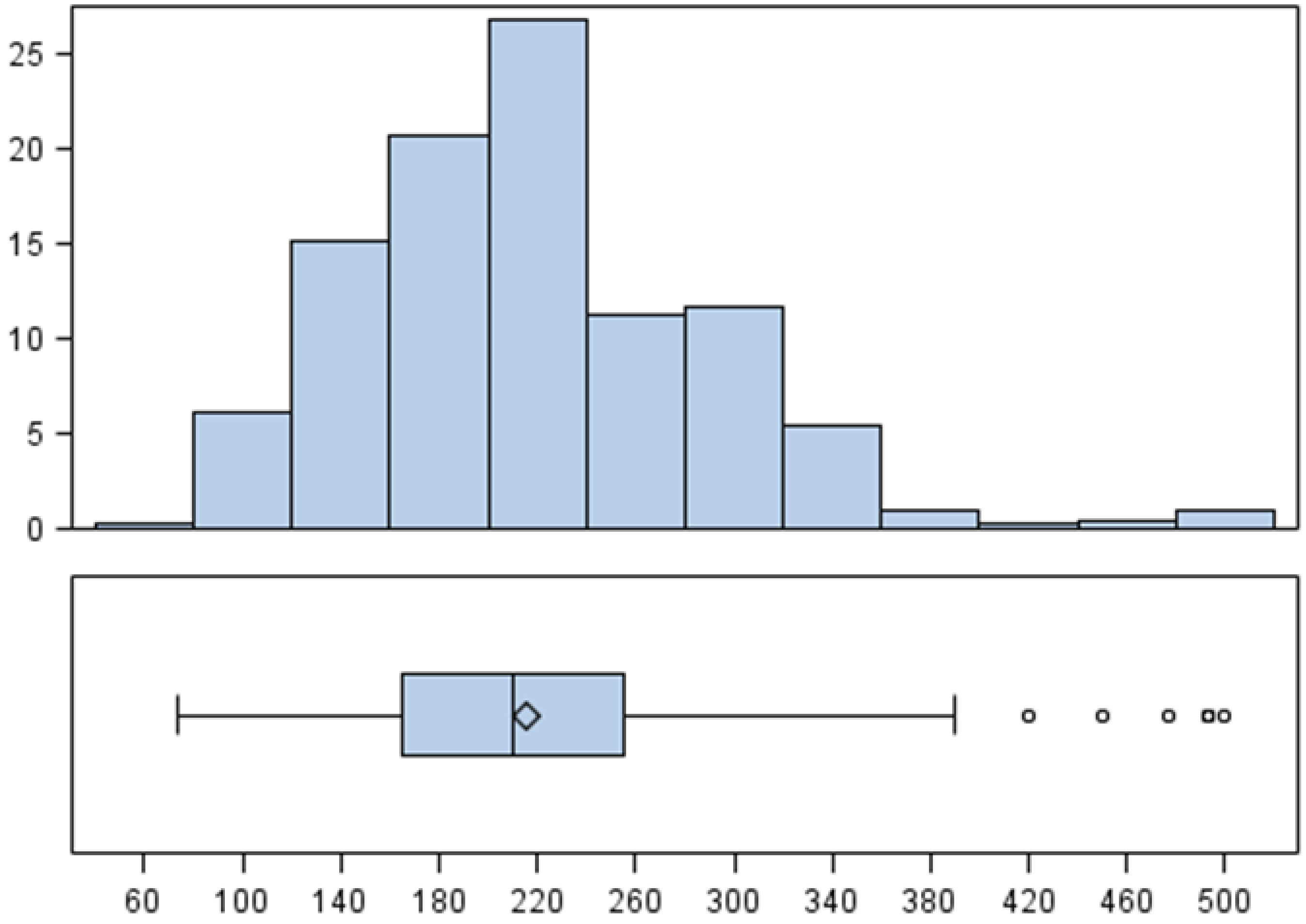
Median

Q_1

Minimum value that is not an outlier

R: `boxplot(v)`

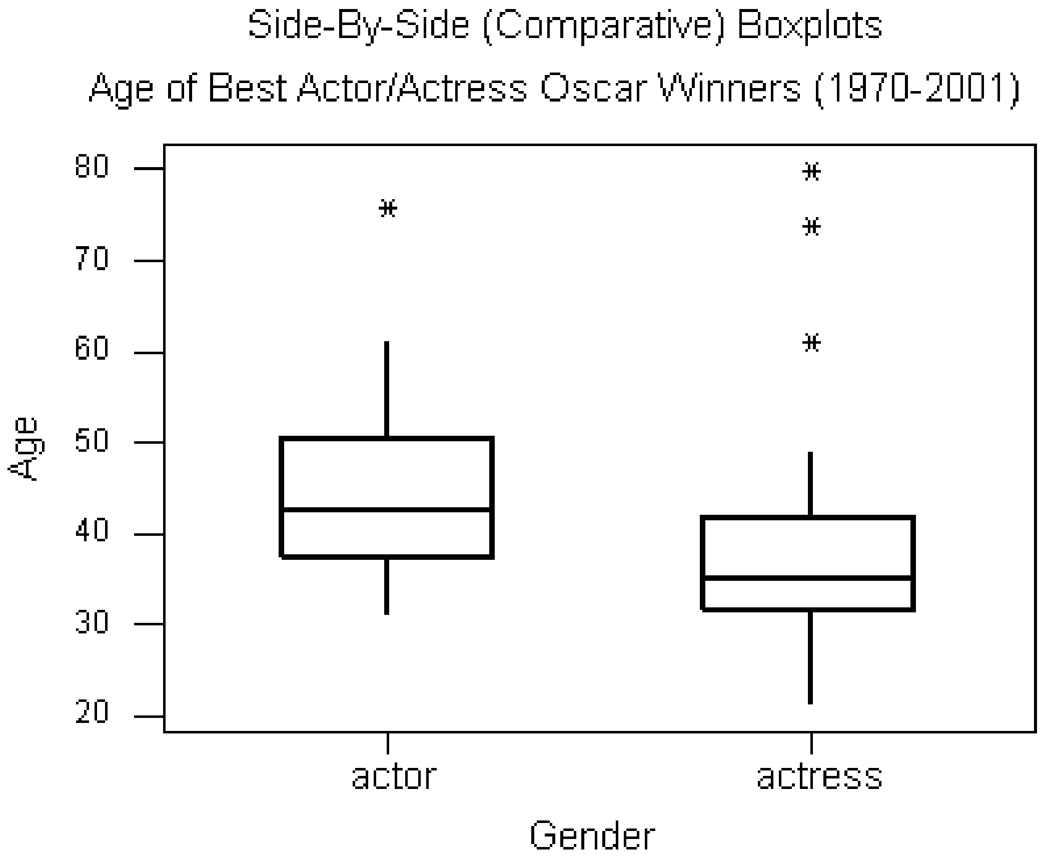
Box plots extract key statistics from histograms



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-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
)
```

Let’s try it in R!

Concepts for summarizing quantitative data

z-scores show how many standard deviations a point is from the mean

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

Quantiles show the value ***x,*** such that a fixed proportion of the data is less than x

- e.g., what is the value *x,* such that 20% of the data is less than *x*

Five Number Summary give key summary statistics of a data sample

- minimum, Q_1, median, Q_3, maximum

A **boxplot** is a visualization of the five number summary

- Side-by-side boxplots allow you to compare key summary statistics

Summary of R

We can compute a z-score for a value x, and a vector of values v using:

```
the_mean <- mean(v)
the_sd <- sd(v)
the_zscore <-   (x - the_mean)/the_sd
```

We can compute quantiles using the quantile() function:

`quantile(v, .2)` or `quantile(v,  c(.25, .4))`

We can compute a five number summary using the fivenum() function:

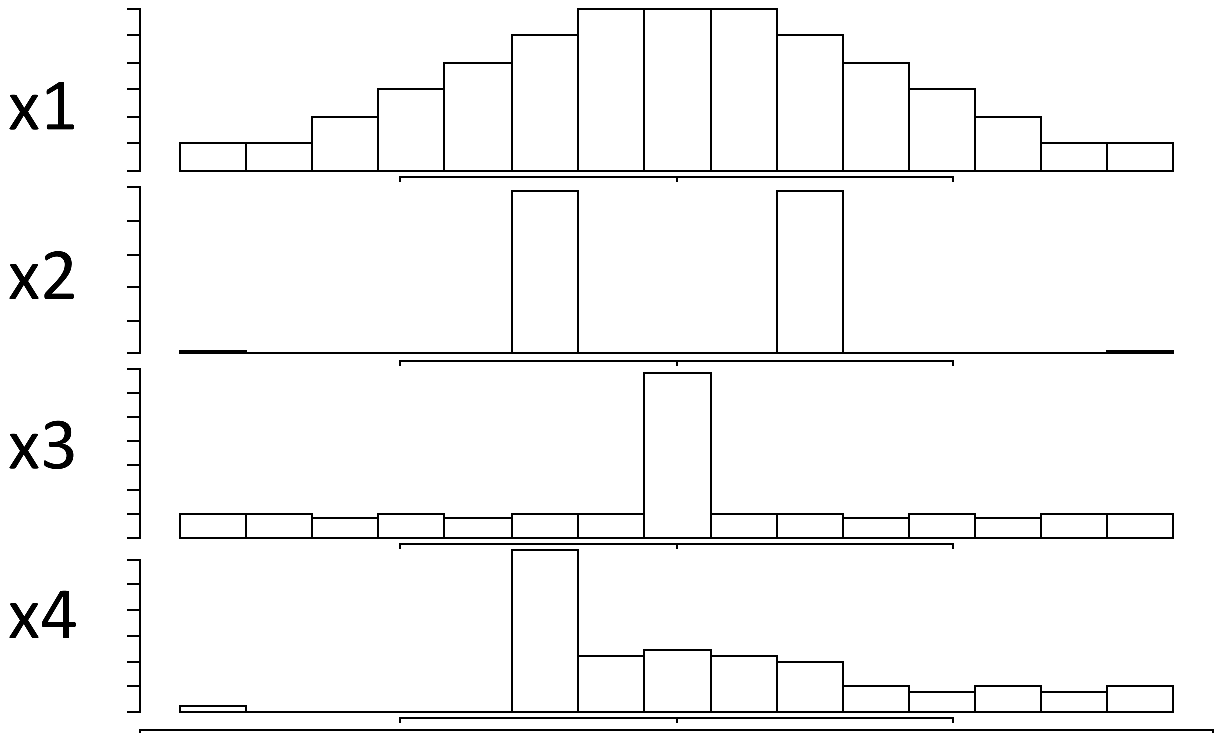
```
fivenum(v)
```

We can compute boxplots using the boxplot() function

```
boxplot(v)
```

Related histograms and boxplots

The image on the right shows histograms of four data sets



Do you think the boxplots of these for data sets will look similar?

Let’s try it in R!

```
load("hist_vs_boxplot.Rda")
```

If there is extra time…

Relationships between two quantitative variables

Two quantitative variables

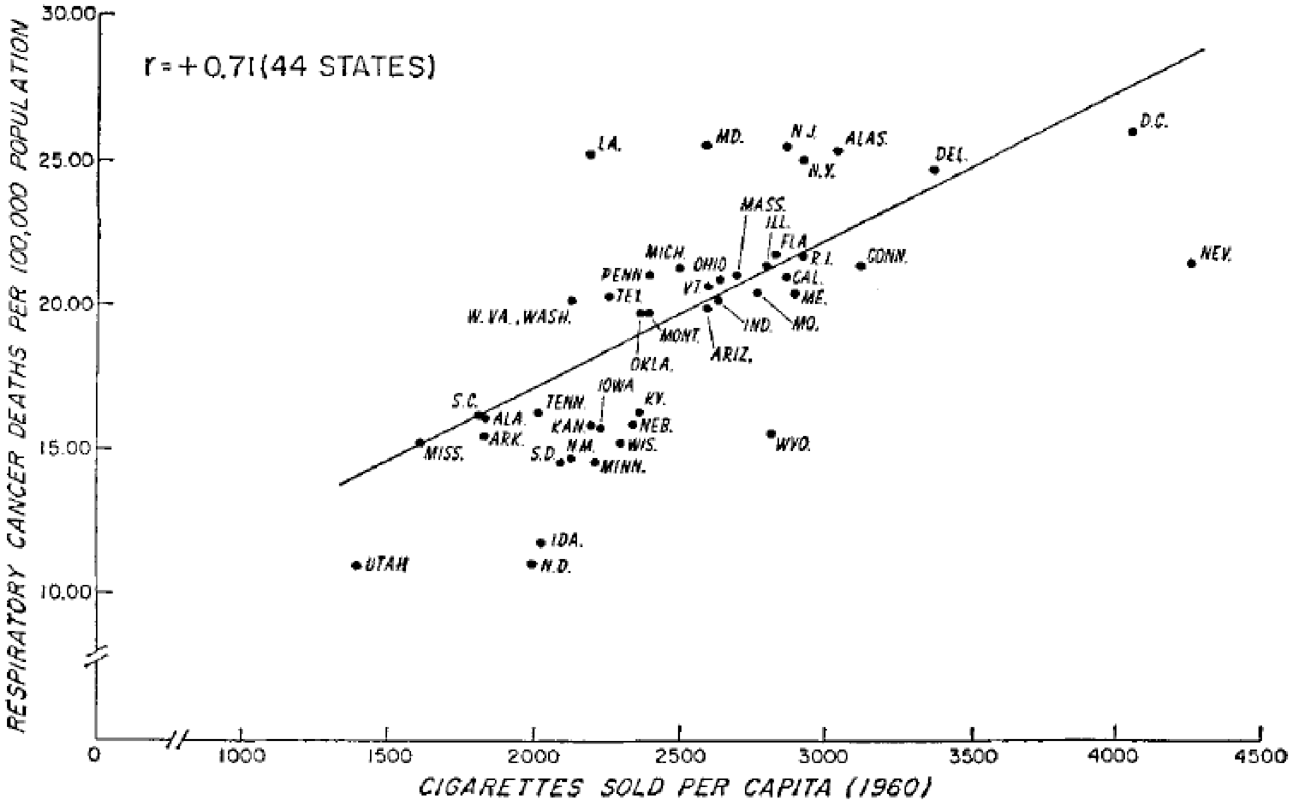
In 1968, Joseph Fraumeni published a paper in the Journal of the National Cancer Institute that examined the relationship between smoking and different types of cancer

Explanatory variable

Response variable

State	Cig per capita	Bladder	Lung	Kidney	Leukemia
AL	1,820	2.9	17.05	1.59	6.15
AZ	2,582	3.52	19.8	2.75	6.61
AR	1,824	2.99	15.98	2.02	6.94
CA	2,860	4.46	22.07	2.66	7.06
CT	3,110	5.11	22.83	3.35	7.2
DE	3,360	4.78	24.55	3.36	6.45
DC	40,460	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

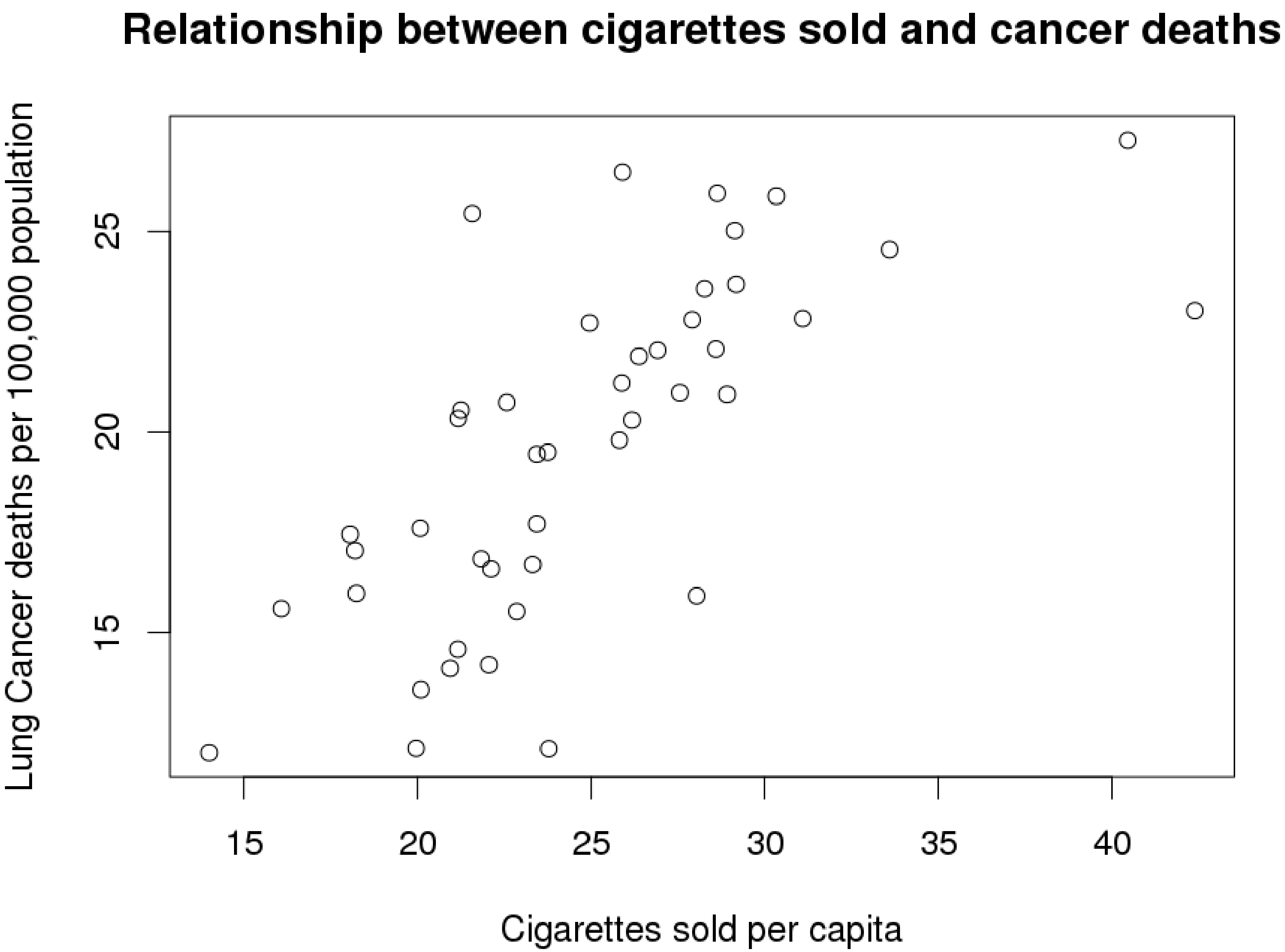
Scatterplot

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

- Each axis represents the value of one variables
- Each point 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?

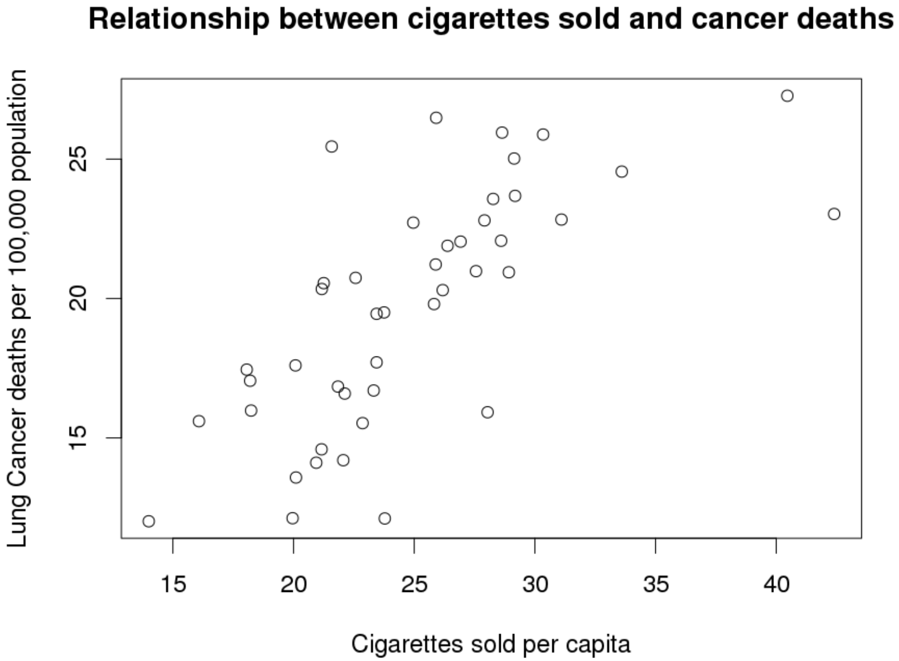
Does it go upward or downward?

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



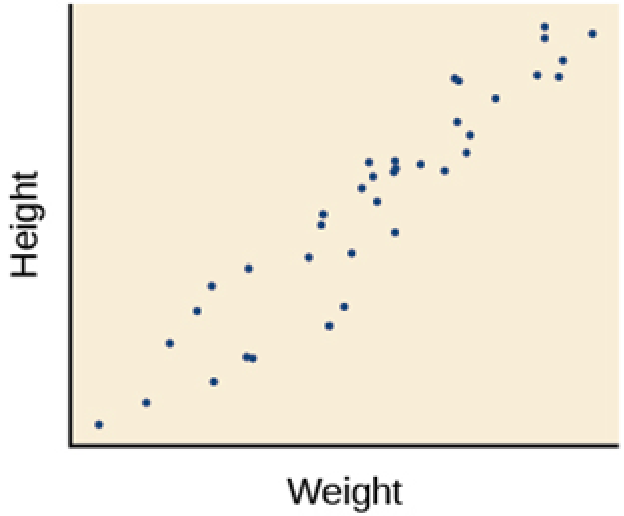
Positive, negative, no correlation

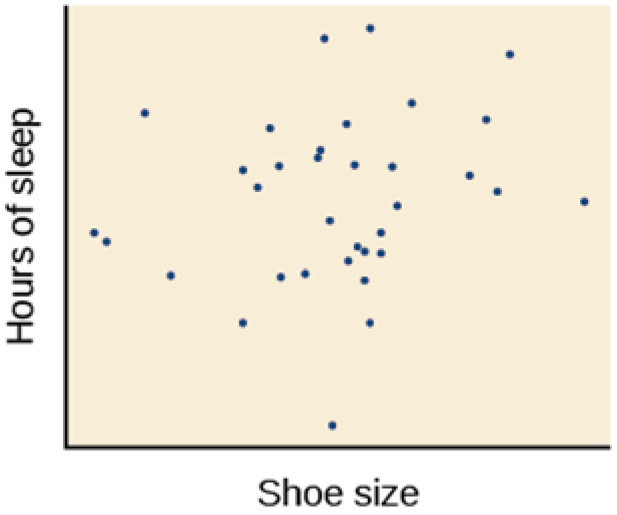
Do the points show a clear trend?

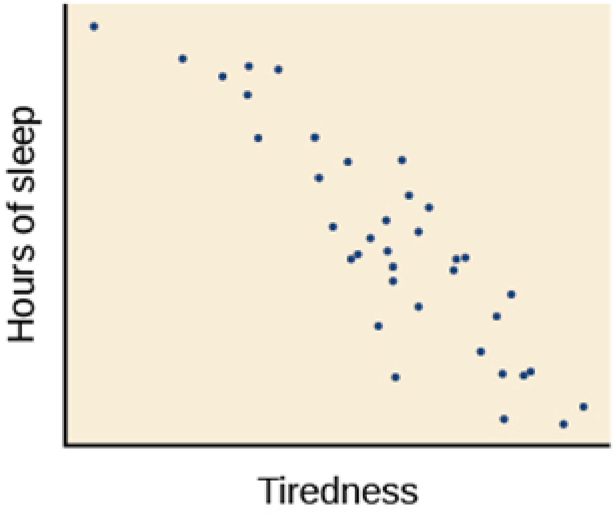
- Does it go upward or downward?
- How much scatter around the trend?

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

Are there any outlier points?







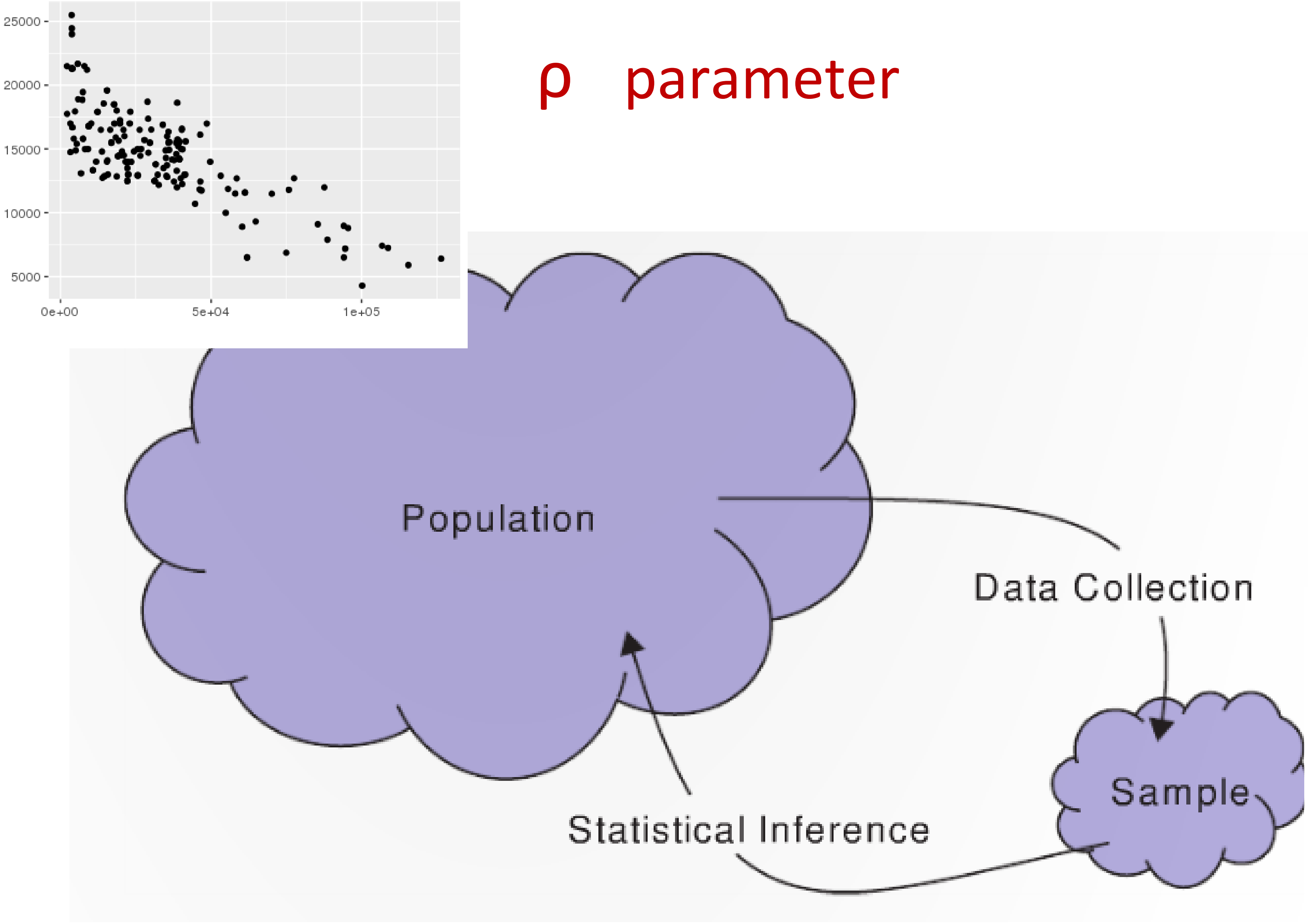
The correlation coefficient

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

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

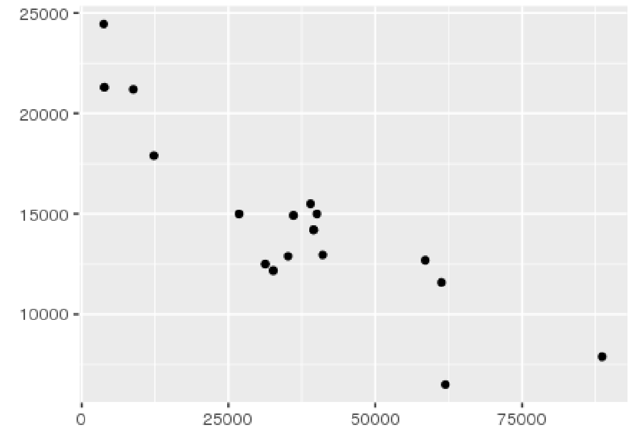
- The correlation for a sample is denoted with **r**
- The correlation in the population is denoted with **ρ**
  (the Greek letter rho)

```
R: cor(x, y)
```



ρ parameter

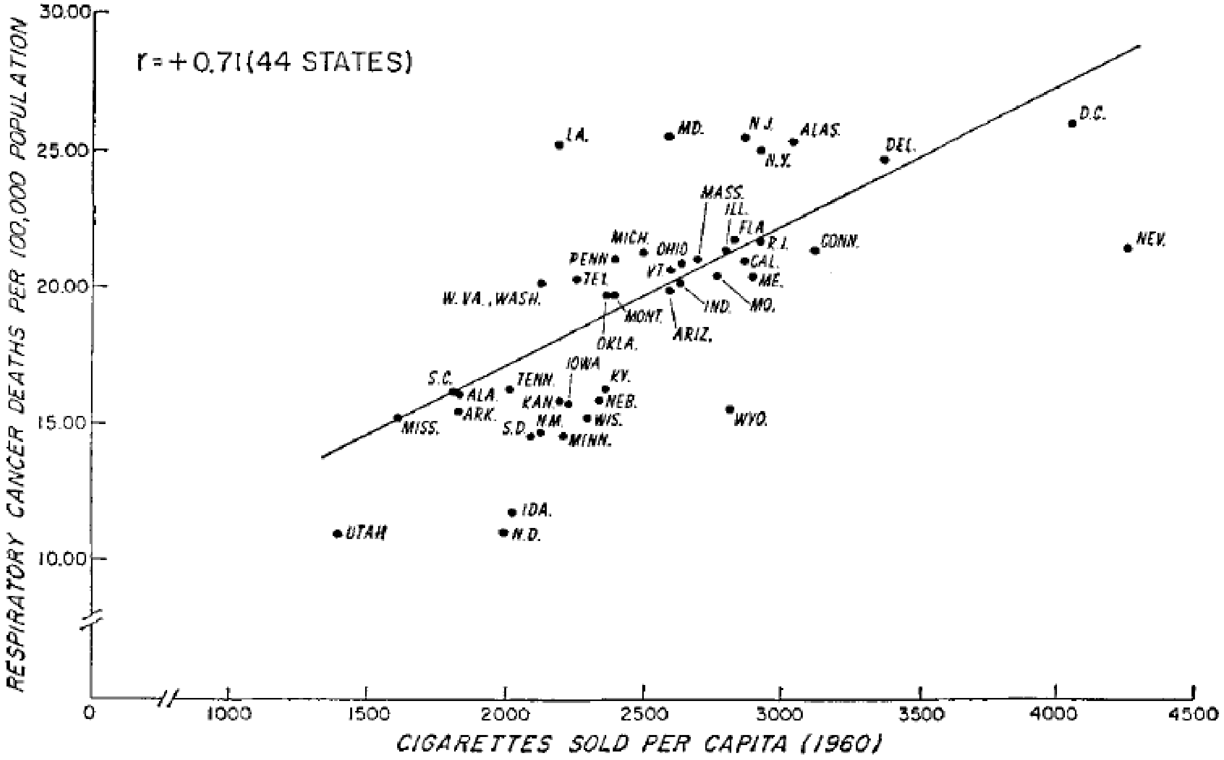
r statistic



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

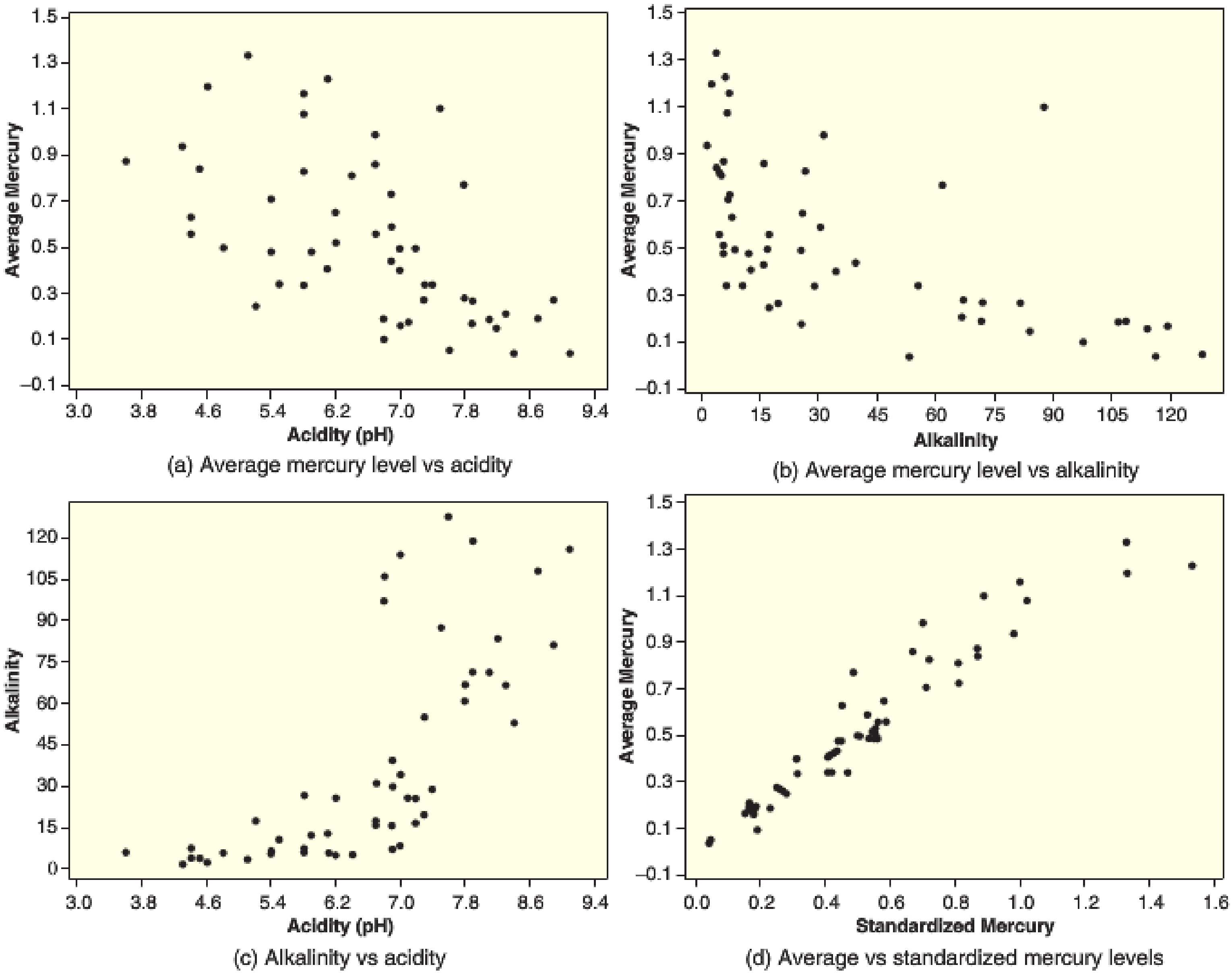
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

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



(a) Average mercury level vs acidity

(b) Average mercury level vs alkalinity

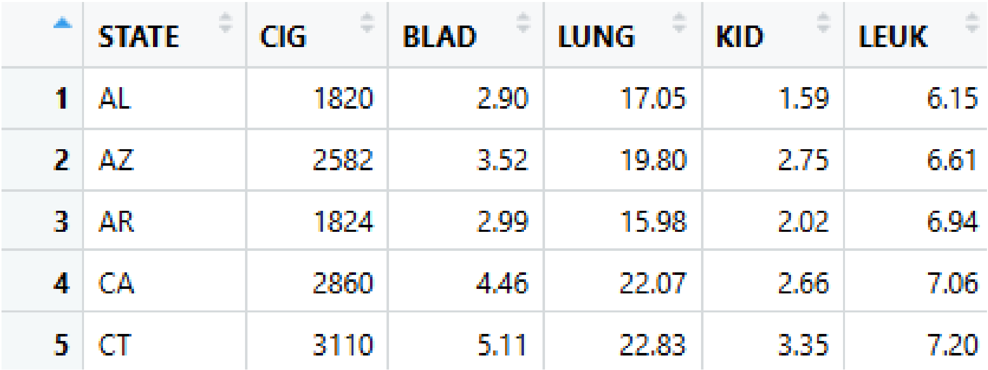
(c) Alkalinity vs acidity

(d) Average vs standardized mercury levels

Let’s calculate some correlations in R

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

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



	STATE	CIG	BLAD	LUNG	KID	LEUK
1	AL	1820	2.90	17.05	1.59	6.15
2	AZ	2582	3.52	19.80	2.75	6.61
3	AR	1824	2.99	15.98	2.02	6.94
4	CA	2860	4.46	22.07	2.66	7.06
5	CT	3110	5.11	22.83	3.35	7.20

Number of cigarette’s sold per capita

Cancer rates per 10k for bladder, lung, kidney and leukemia

We will try it in R next class…