Simple linear regression

Overview

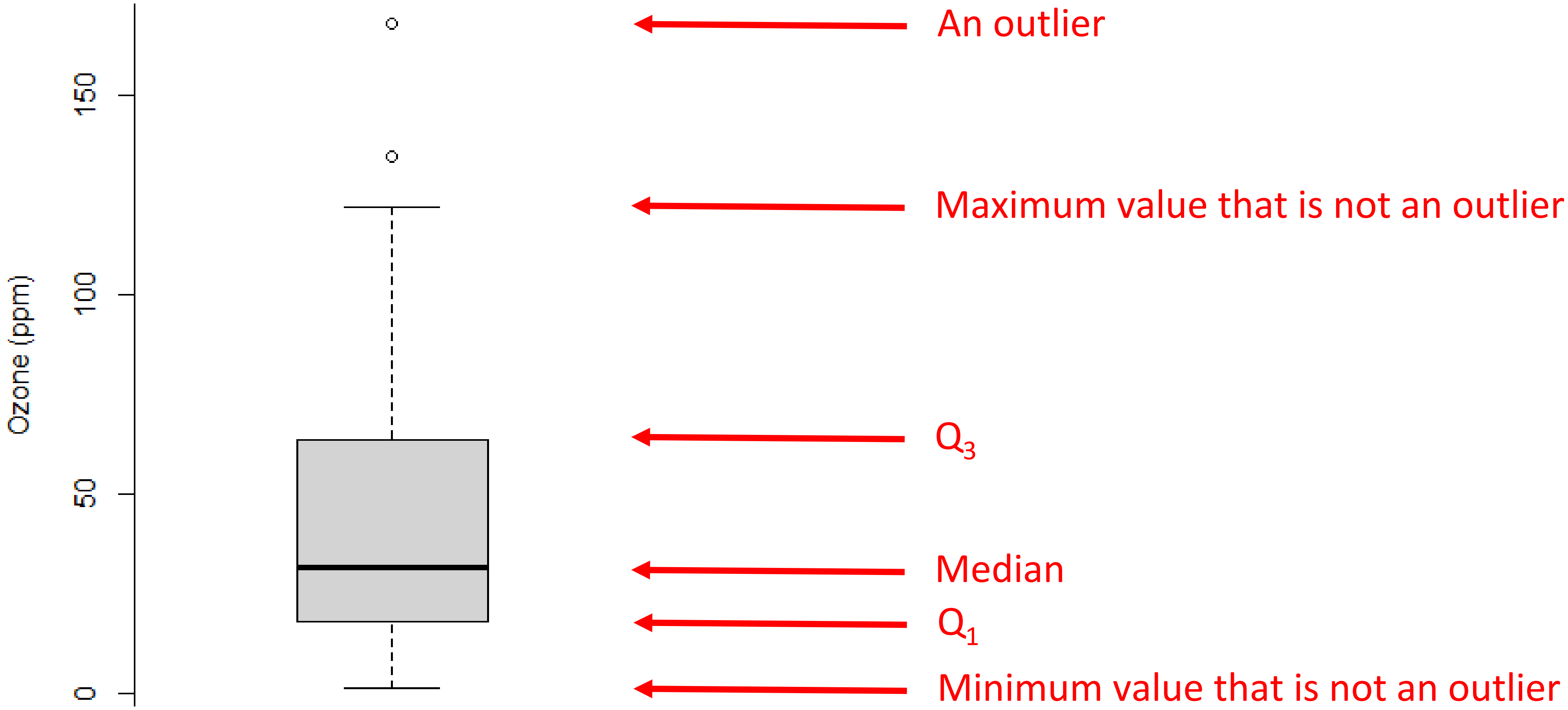
Quick review of box plots and correlation

Simple linear regression

Review of descriptive statistics

Quick review of box plots and correlation

Review Box Plots: Ozone in NYC (May-Sept 1973)



An outlier

Maximum value that is not an outlier

Q_3

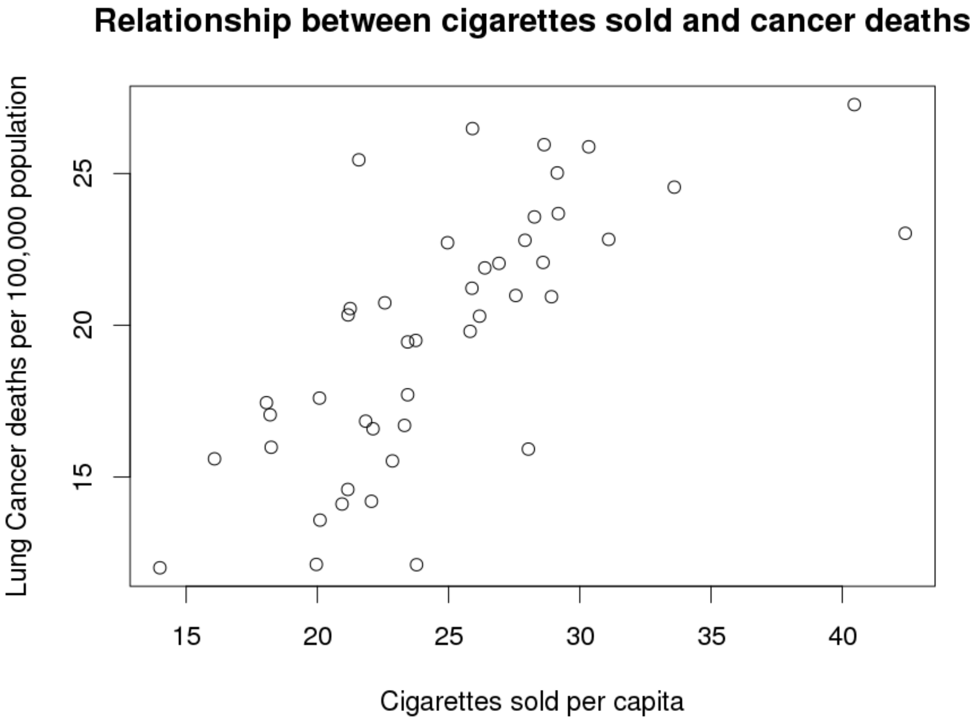
Median

Q_1

Minimum value that is not an outlier

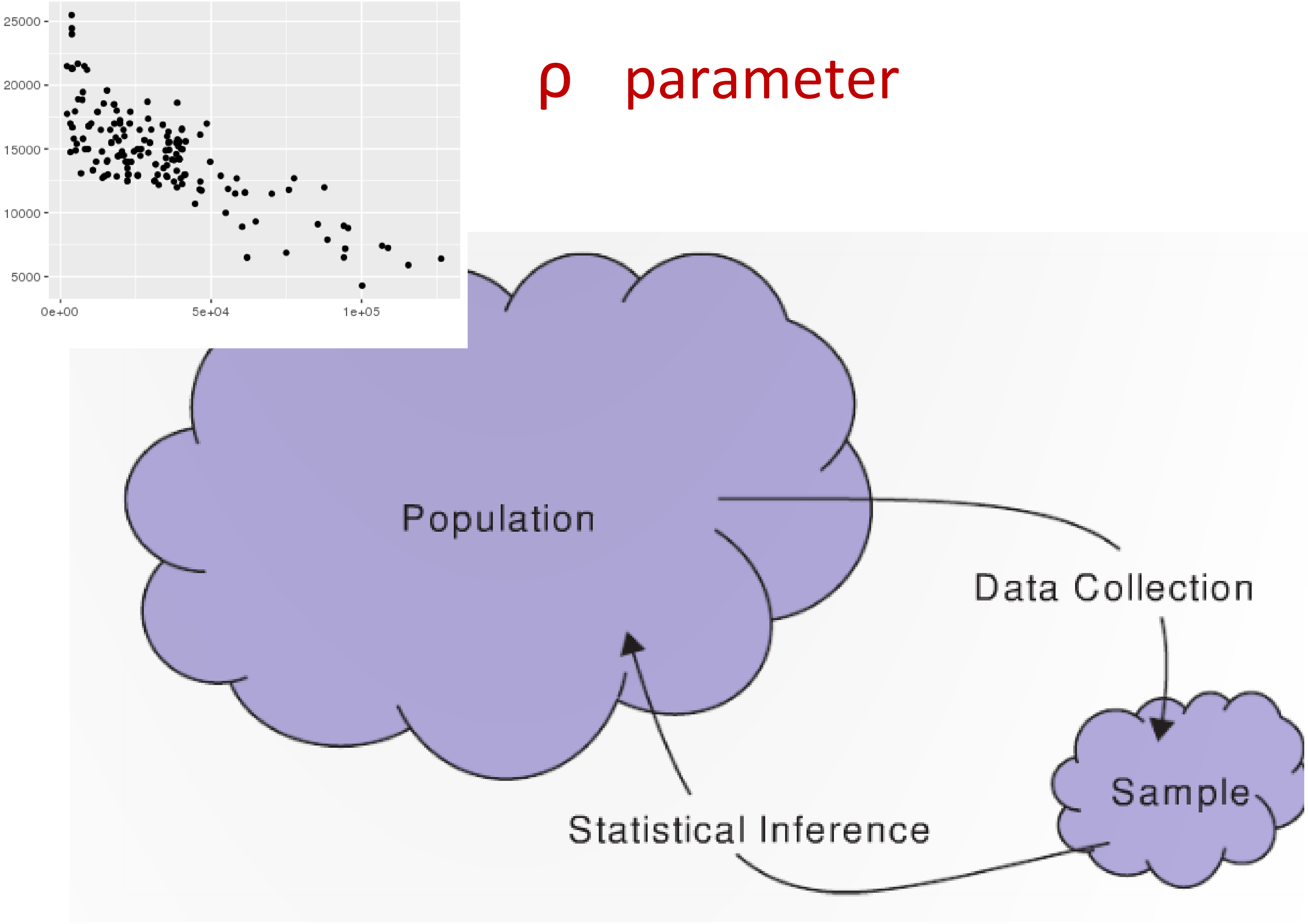
R: boxplot(airquality$Ozone)

Review: scatter plots and the correlation coefficient

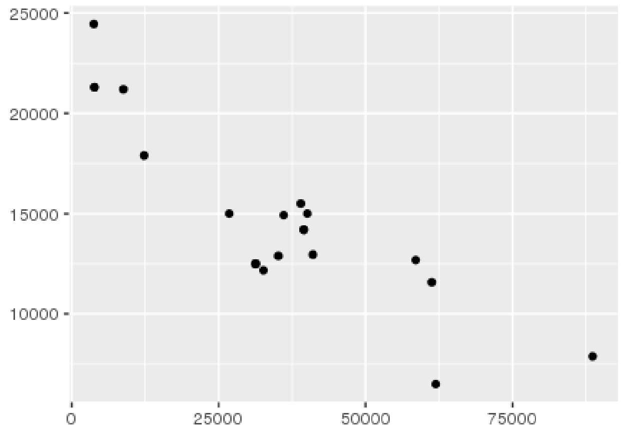


$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

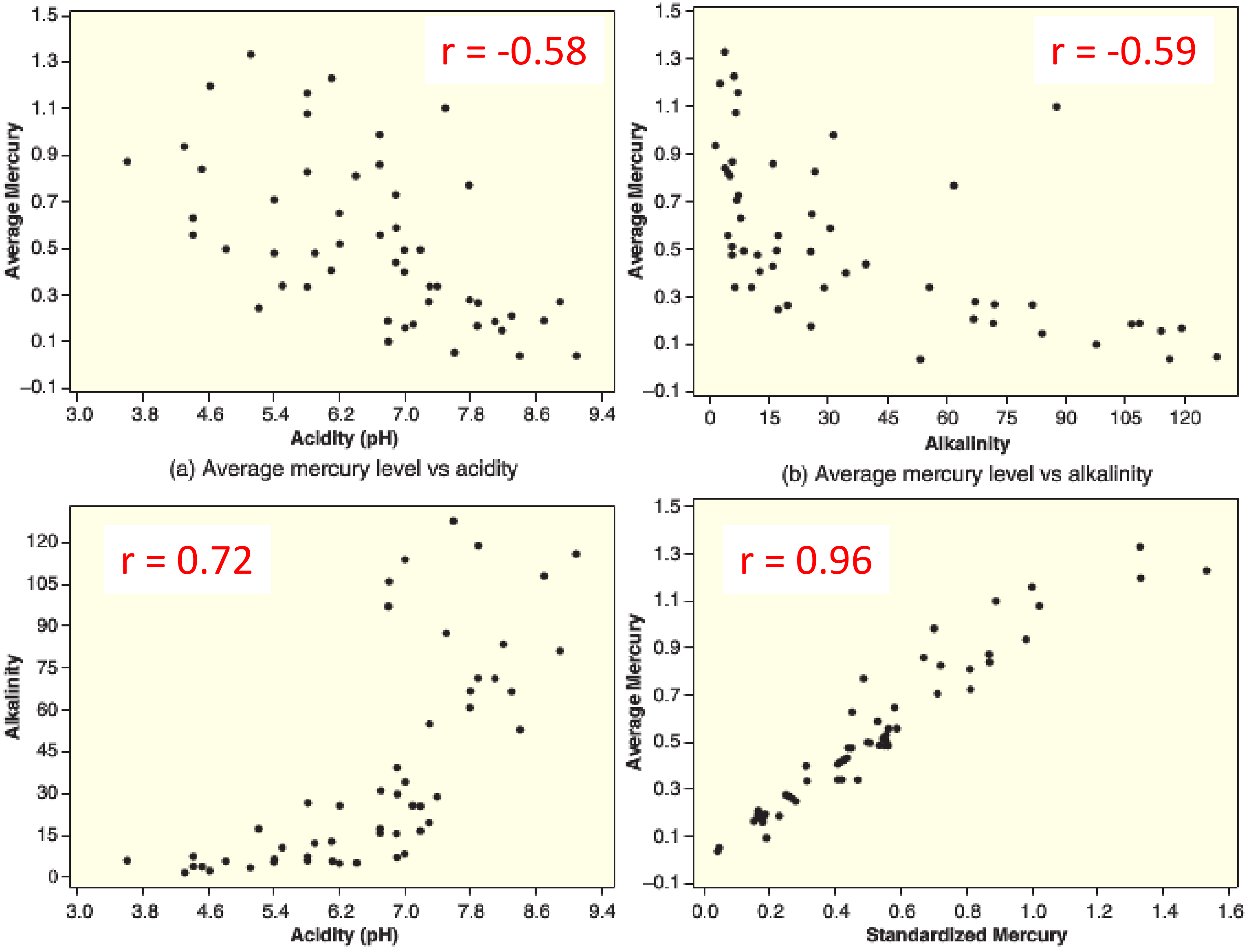
The **correlation** is measure of the strength and direction of a linear association between two variables ρ parameter



r statistic



Florida lakes



(a) Average mercury level vs acidity

(b) Average mercury level vs alkalinity

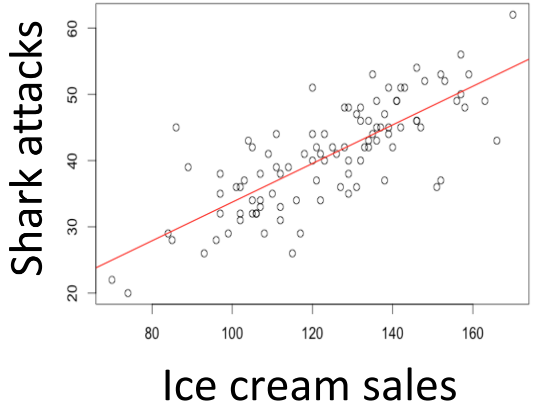
(c) Alkalinity vs acidity

(d) Average vs standardized mercury levels

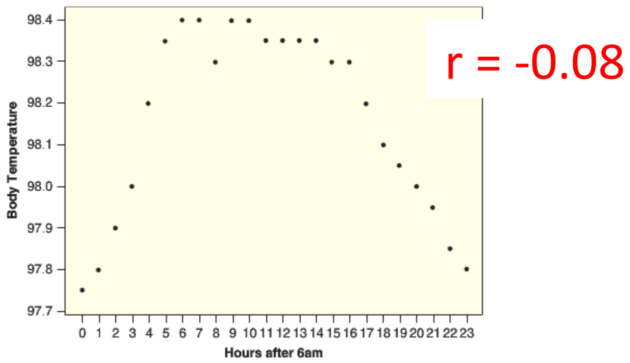
```
# create a scatter plot
plot(x, y)

# calculate the correlation
cor(x, y)
```

Correlation cautions

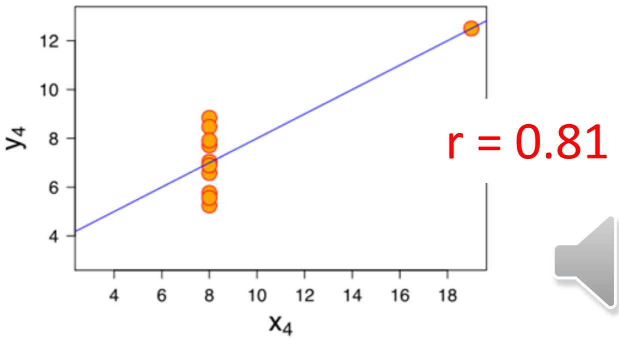
1. A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables.



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


r = -0.08

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


r = 0.81

Regression

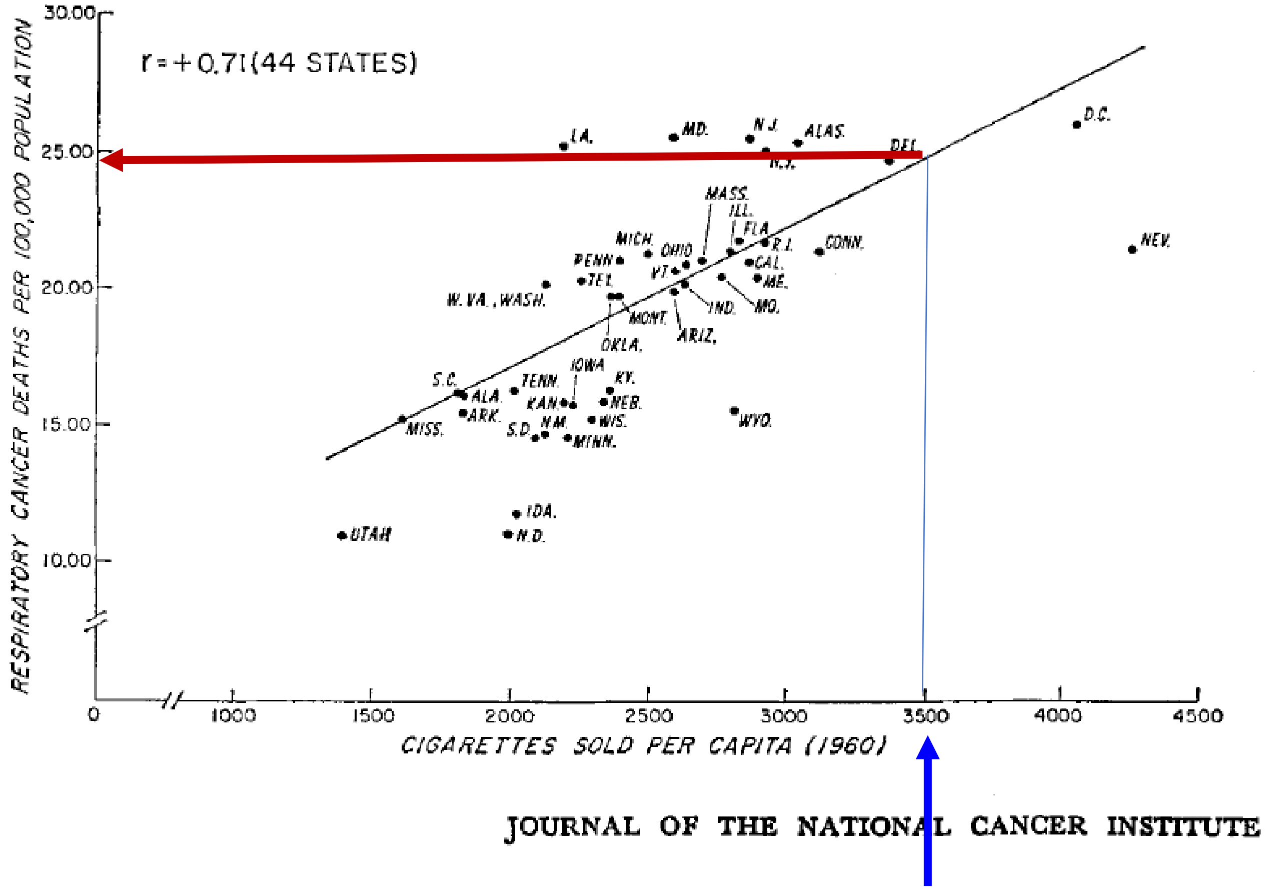
Regression is method of using one variable ***x*** *to predict* the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

Cigarette cancer regression line

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.



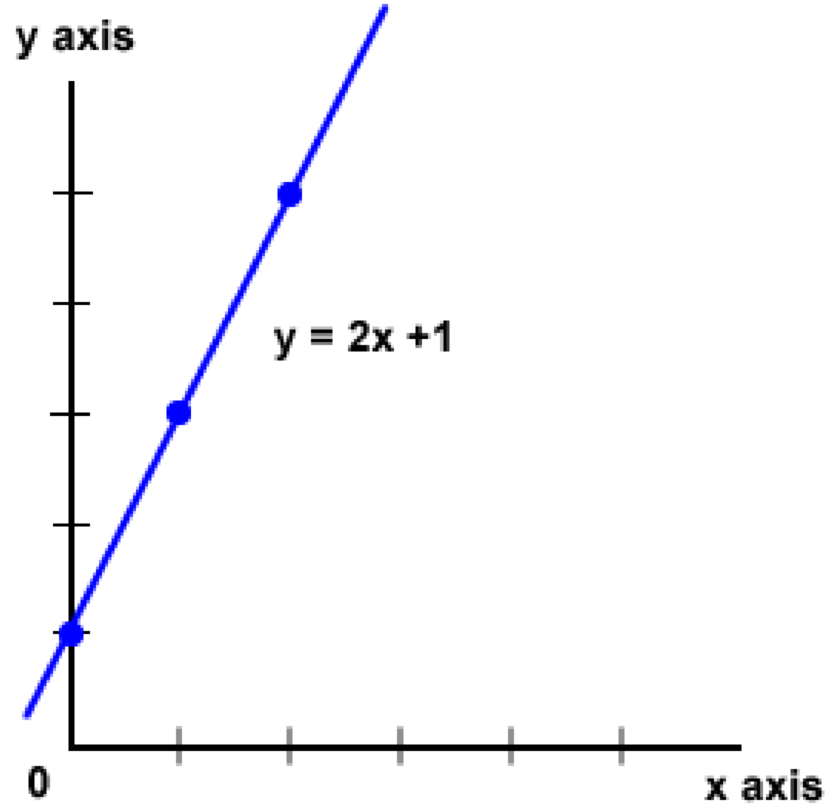
JOURNAL OF THE NATIONAL CANCER INSTITUTE

$x_i = 3500$

Equation for a line

What is the equation for a line?

$$\hat{y} = a + b \cdot x$$



Regression lines

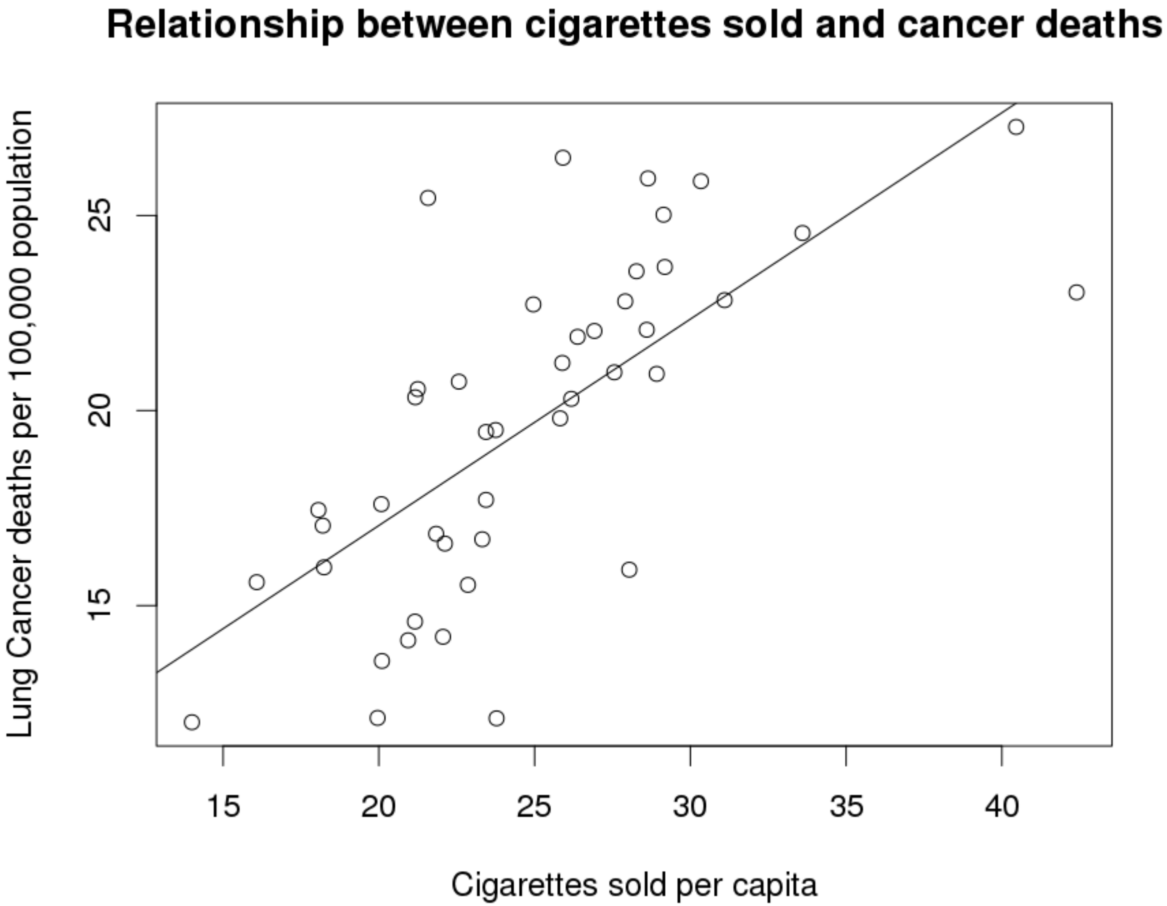
$\hat{y} = a + b \cdot x$

$\mathit{Response} = a + b \cdot \mathit{Explanatory}$

The slope ***b*** represents the predicted change in the response variable *y* given a one unit change in the explanatory variable *x*

The intercept ***a*** is the predicted value of the response variable y if the explanatory variable x were 0

Cancer smoking regression line



$\hat{y} = a + b \cdot x$

$a = 6.47$

$b = 0.53$

R: `lm(y ~ x)`

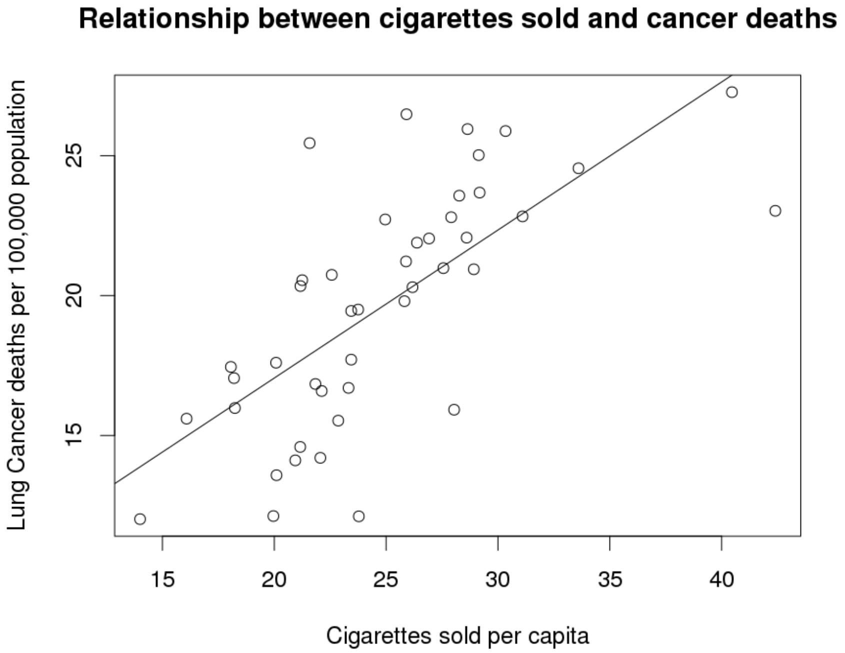
Using the regression line to make predictions

If a state sold 25 (hundred) cigarettes per person

How many cancer deaths (per 100,000 people) would you expect?

a = 6.47,   b = .53

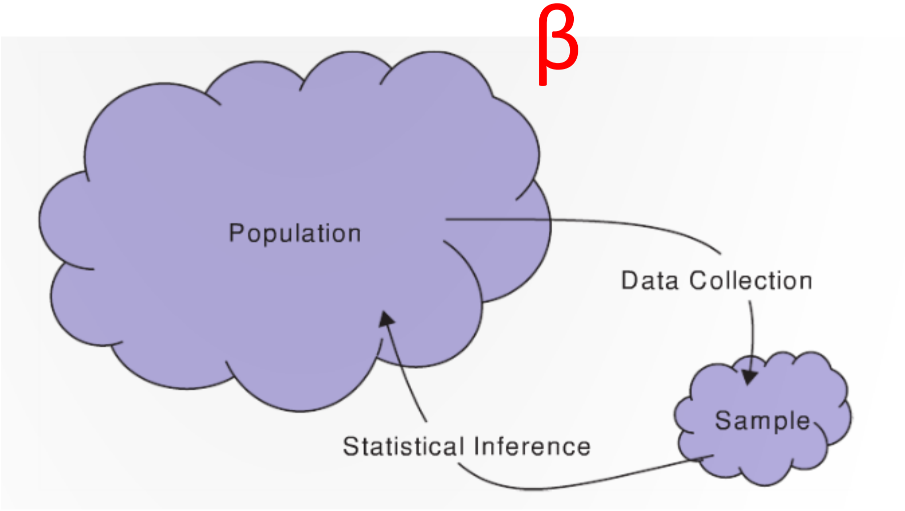
$\hat{y} = 6.47 + .53 \cdot x$



Notation

The letter **b** is typically used to denote the slope of the sample

The Greek letter **β** is used to denote the slope of the population

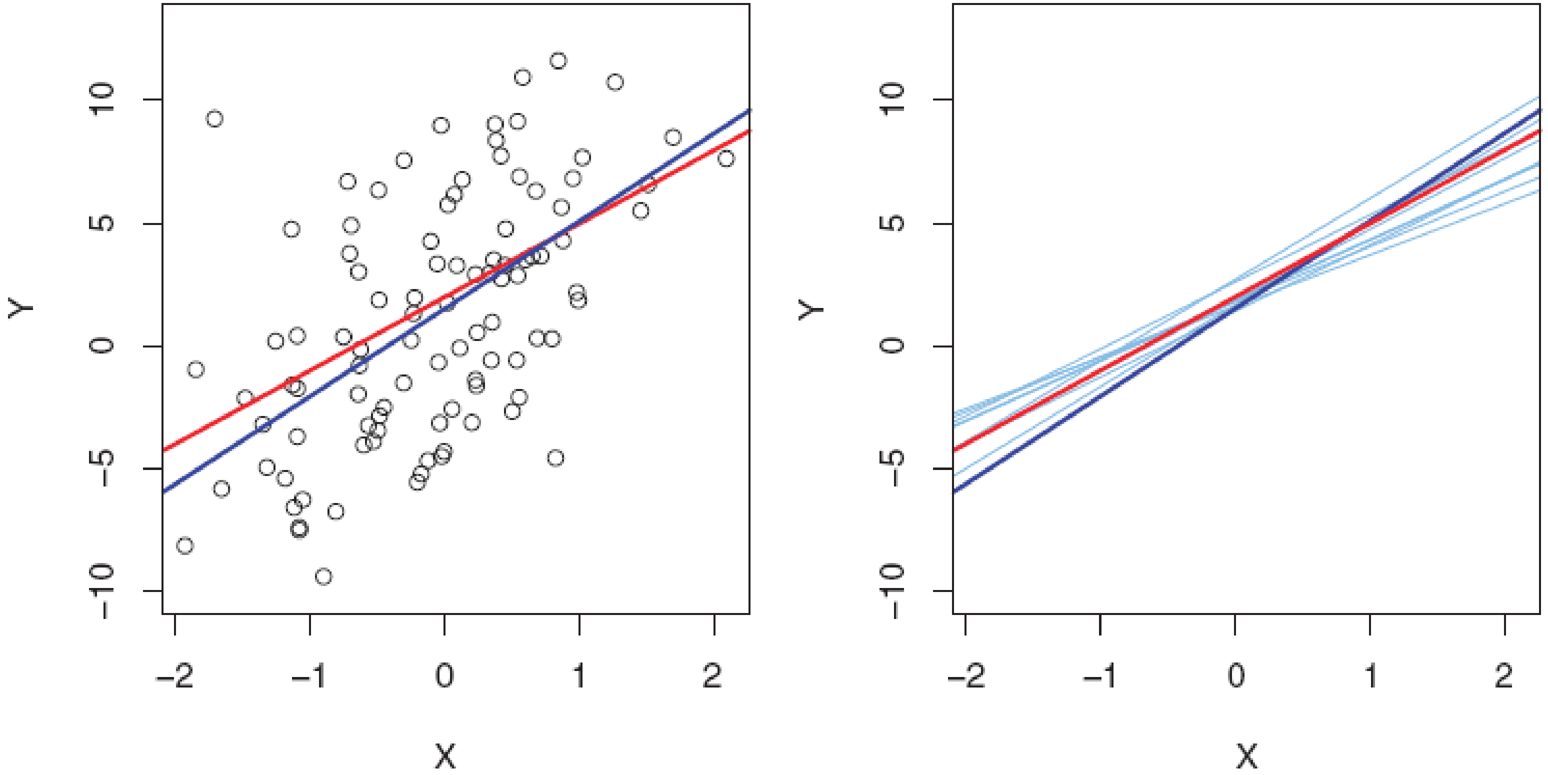


β

b

Population: β

Sample estimates: b

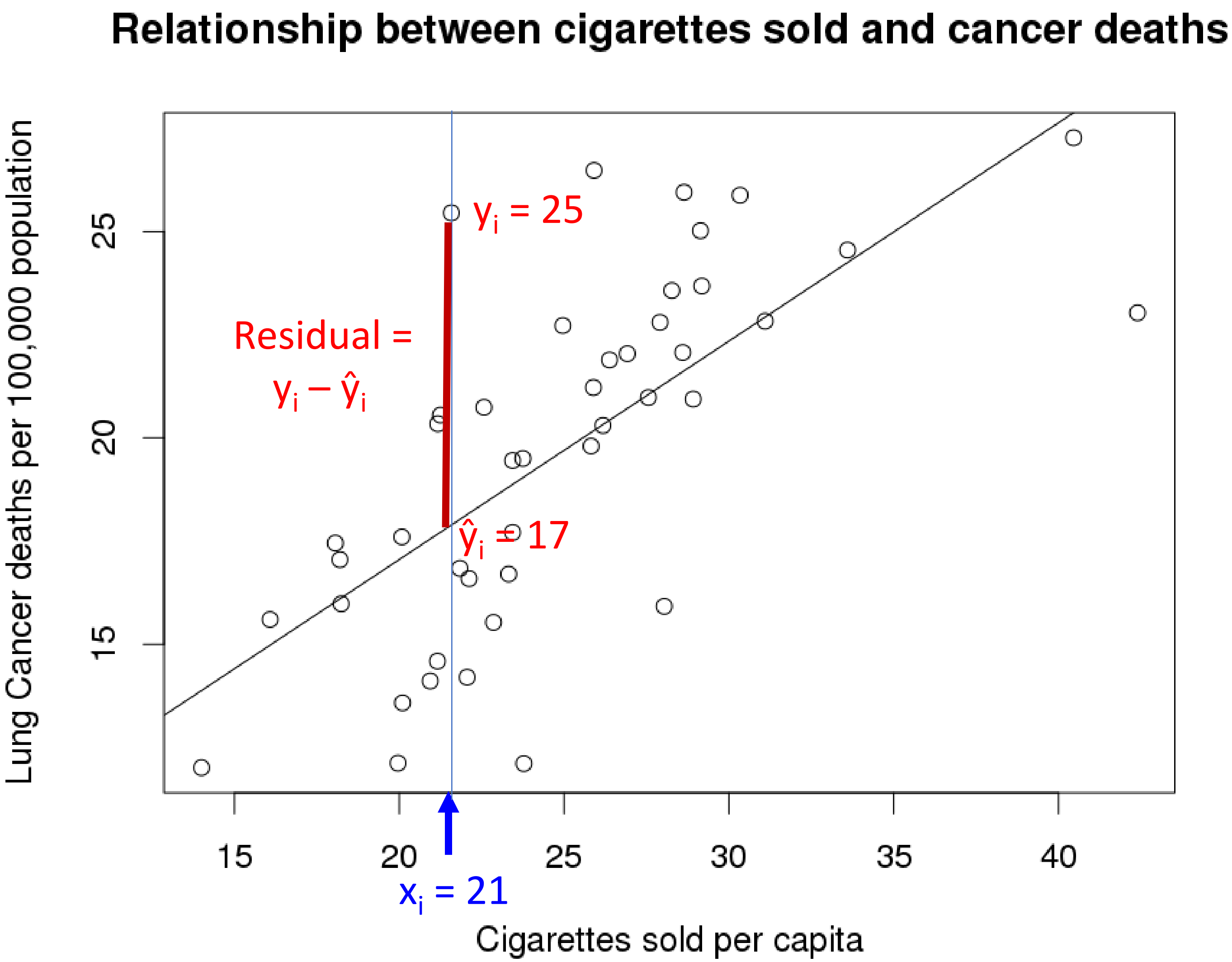


Residuals

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

$$Residual_i \;=\; Observed_i - Predicted_i \;=\; y_i - \hat{y}_i$$

Cancer smoking residuals

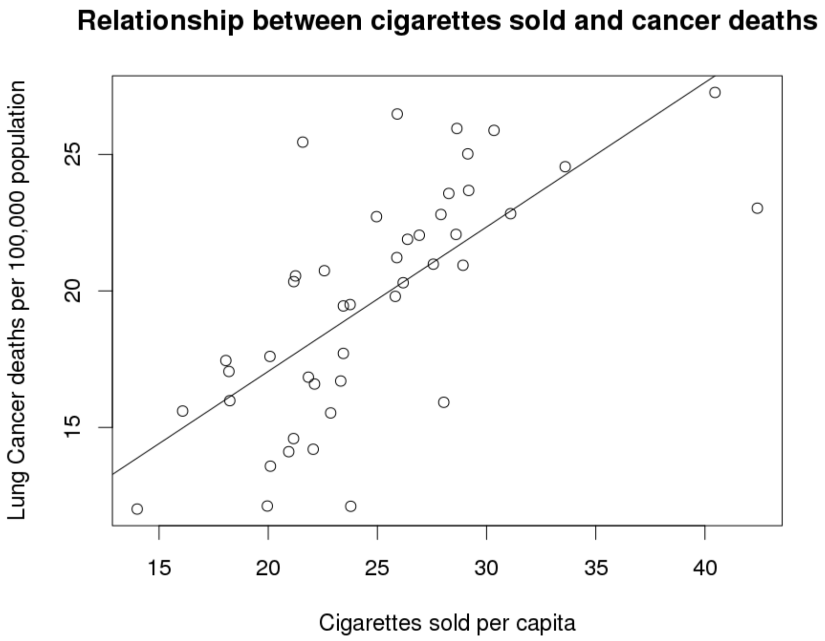


Cancer smoking residuals

Cancer obs (y)	**Cancer pred (ŷ)**	**Residuals (y - ŷ)**
17.05	16.10	0.95
19.80	20.13	-0.33
15.98	16.12	-0.14
22.07	21.60	0.47
22.83	22.93	-0.10
24.55	24.25	0.30
27.27	27.88	-0.61
23.57	21.24	2.14

Line of 'best fit'

The **least squares line**, also called **'the line of best fit'**, is the line which minimizes the sum of squared residuals



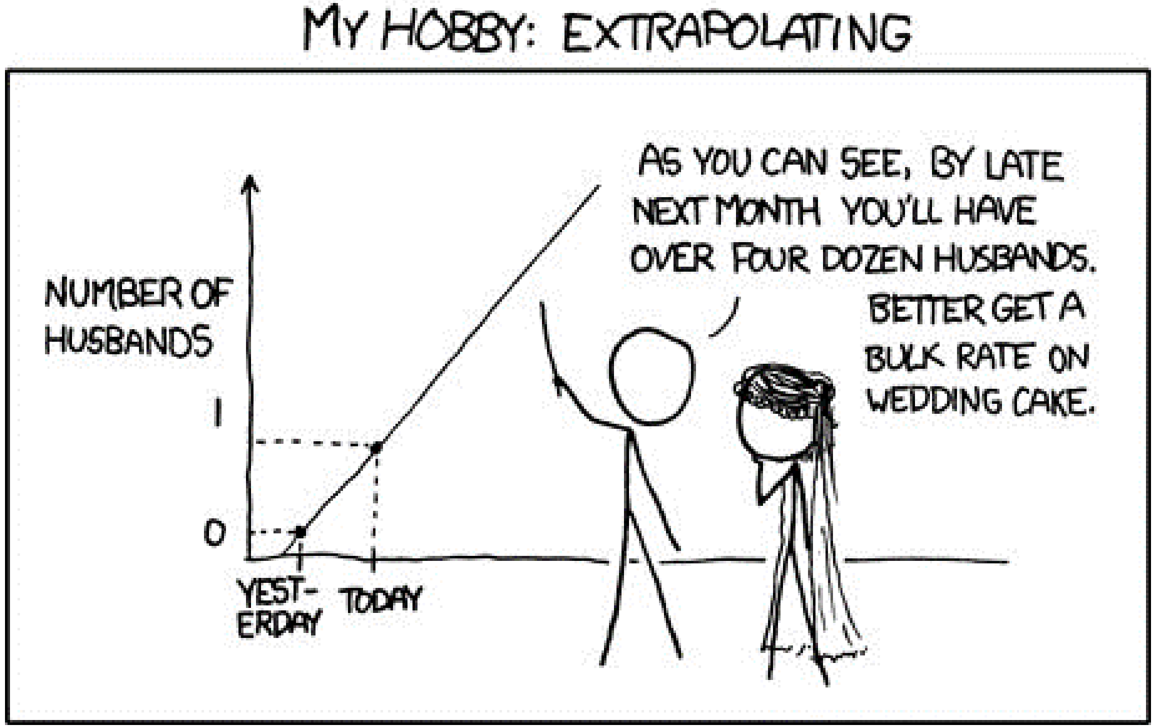
Try to find the line of best fit

Cancer smoking residuals

Cancer obs (y)	**Cancer pred ($\hat{y}$)**	**Residuals ($y - \hat{y}$)**	**Residuals^2 $(y - \hat{y})^2$**
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

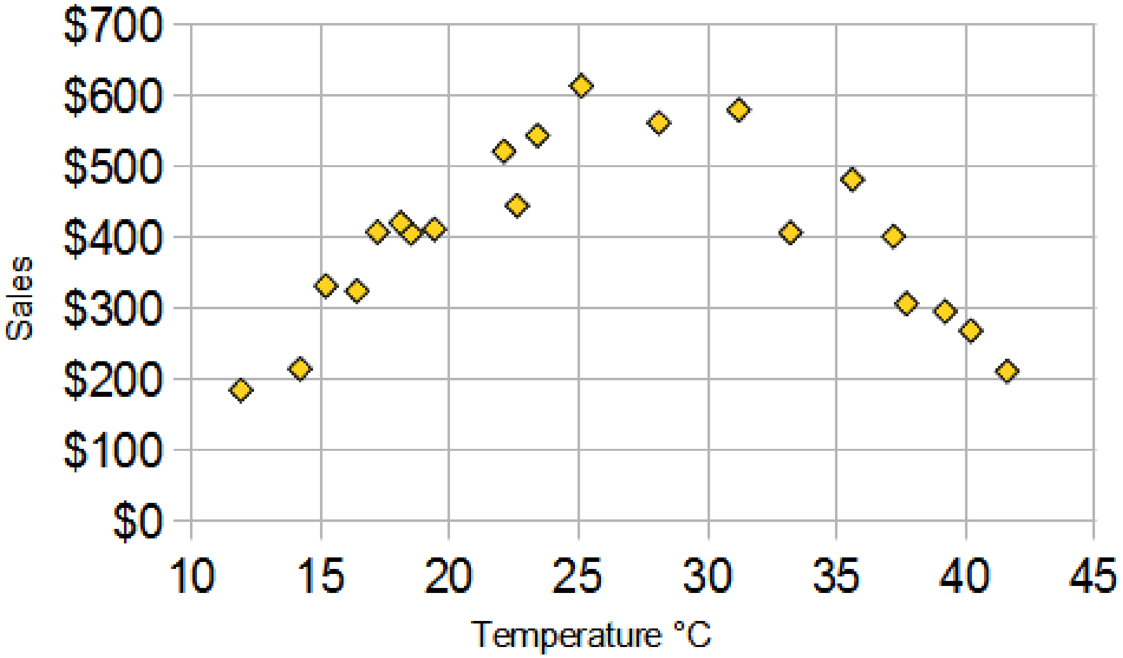
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

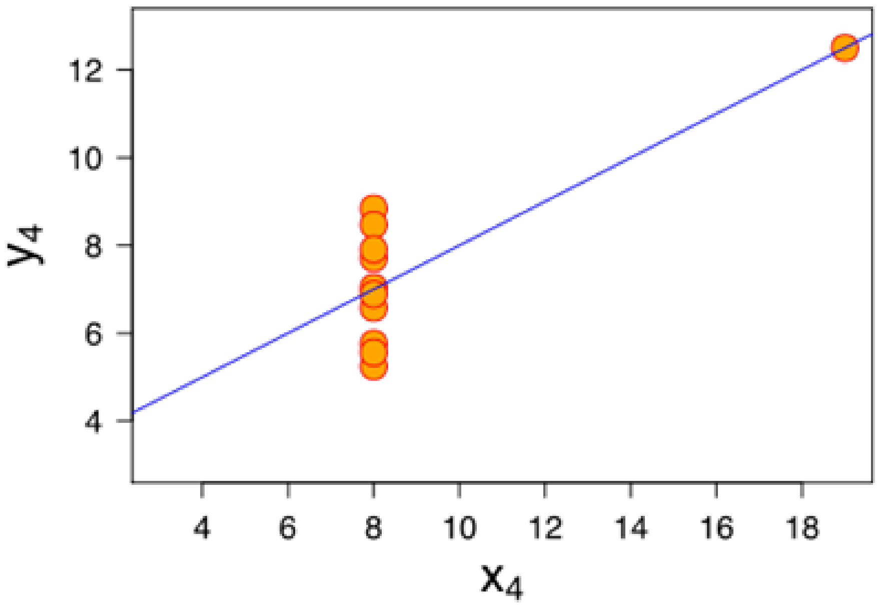
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3

Be aware of outliers – they can have an huge effect on the regression line.



Regression lines in R

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

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

# fit a regression model
> lm_fit <- lm(smoking$LUNG ~ smoking$CIG)

# examine the a and b coefficients
> coef(lm_fit)

# add the regression line to the plot
> abline(lm_fit)
```

Concepts for the relationship between two quantitative variables

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

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

- Value between -1 and 1

In **linear regression** we fit a line to the data, called the **regression line**

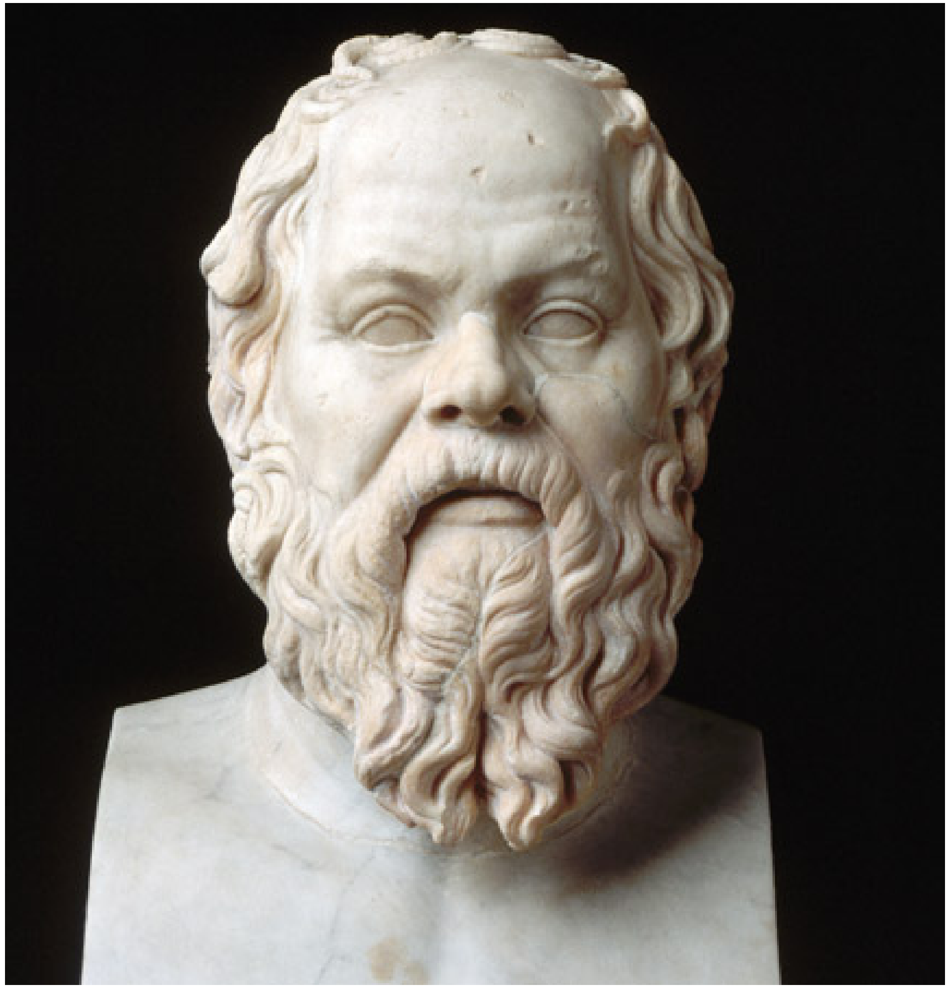
- We get coefficients for the slope (b) and the y-intercept (a)

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

- The regression line minimizes the sum of squared residuals

Review of descriptive statistics

Who is this?

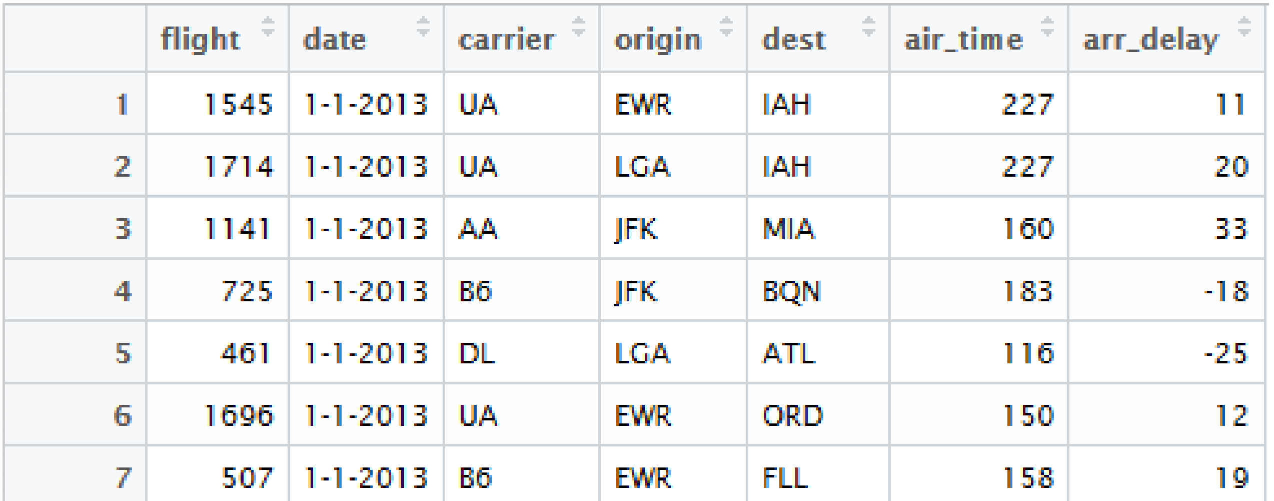


Intro to data

What is Statistics?

What are…

- Observational units?
- Variables?
- Categorical variables?
- Quantitative variables?



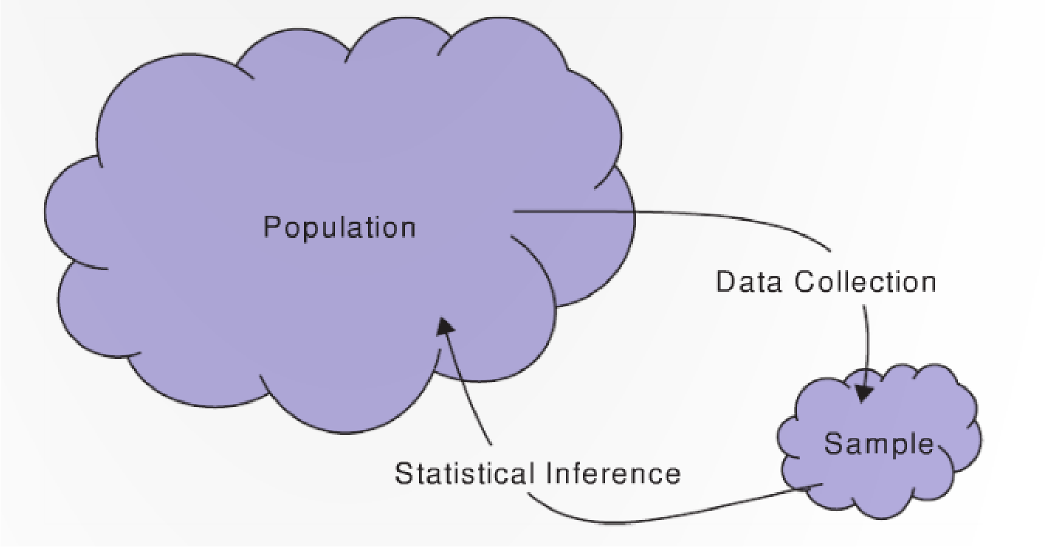
	flight	date	carrier	origin	dest	air_time	arr_delay
1	1545	1-1-2013	UA	EWR	IAH	227	11
2	1714	1-1-2013	UA	LGA	IAH	227	20
3	1141	1-1-2013	AA	JFK	MIA	160	33
4	725	1-1-2013	B6	JFK	BQN	183	-18
5	461	1-1-2013	DL	LGA	ATL	116	-25
6	1696	1-1-2013	UA	EWR	ORD	150	12
7	507	1-1-2013	B6	EWR	FLL	158	19

Sampling

What is a …?

- sample
- population
- statistic
- parameter

What is statistical inference?



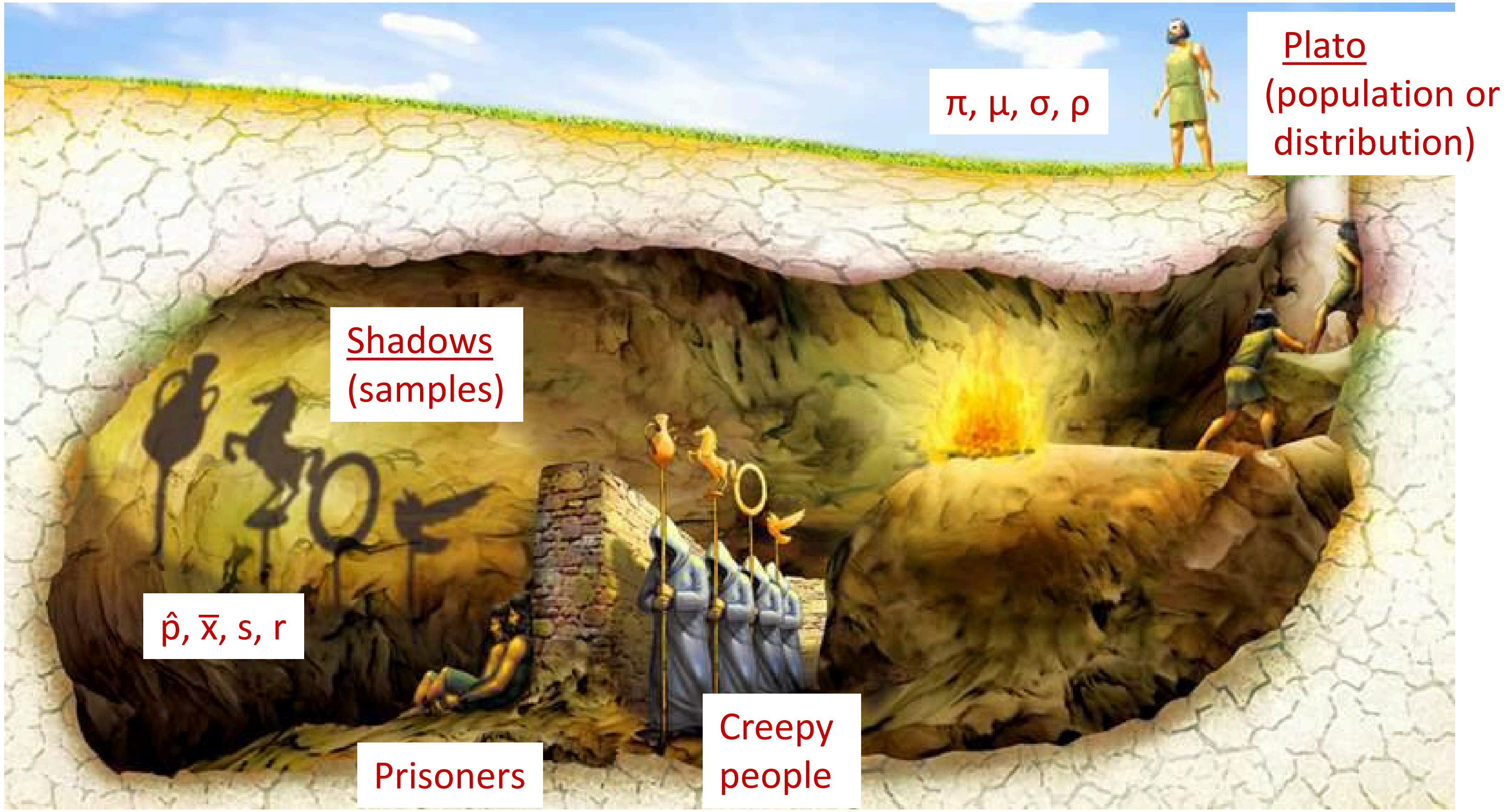
Population

Data Collection

Sample

Statistical Inference

Plato’s cave



π, μ, σ, ρ

Plato
(population or
distribution)

Shadows
(samples)

$\hat{p}$, $\bar{x}$, s, r

Prisoners

Creepy
people

From The Republic (~ 380 BCE)

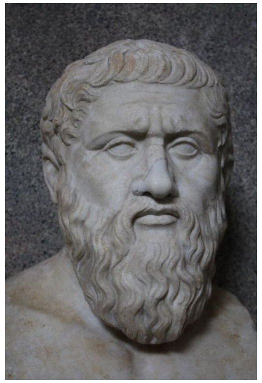
Quiz: parameters and statistics

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation		
Proportion		
Correlation		
Regression slope		

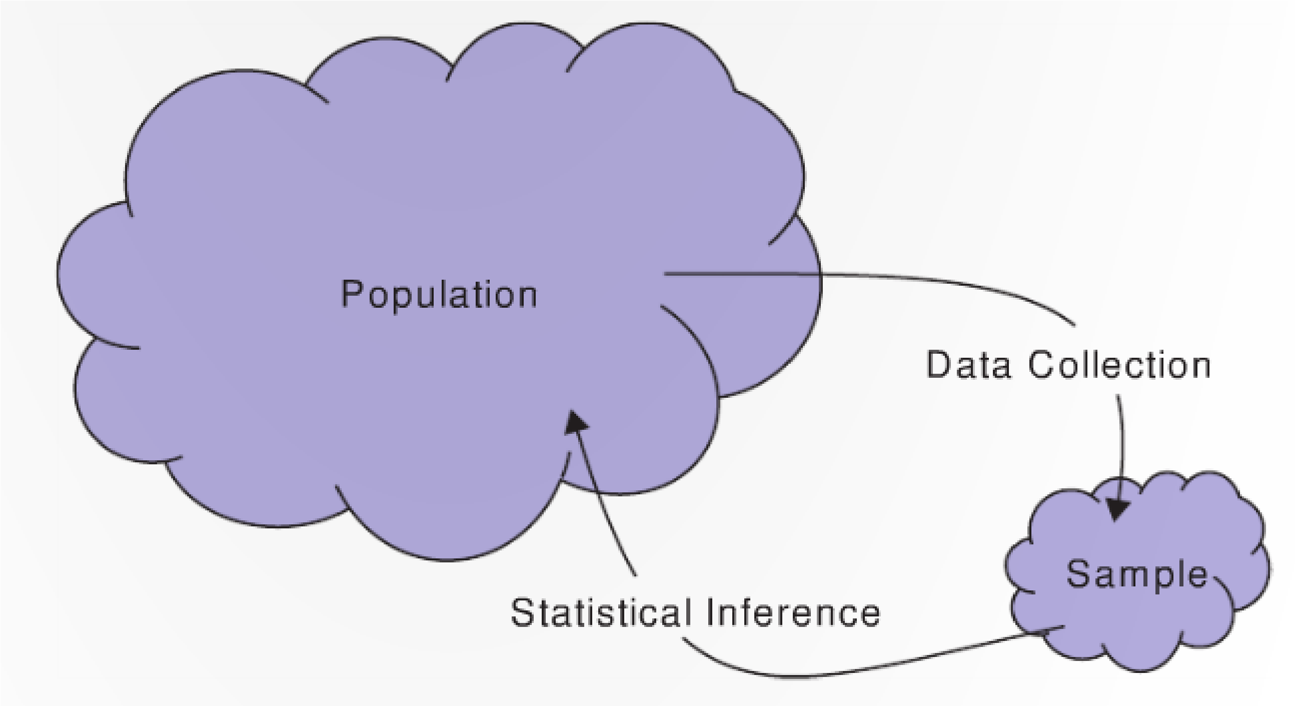
Quiz: parameters and statistics

	Sample Statistic	**Population Parameter**
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
regression slope	b	β

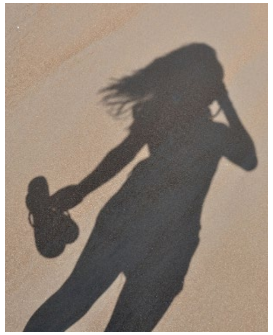
Population parameters vs. sample statistics



$\pi, \mu, \sigma, \rho, \beta$



$\hat{p}, \bar{x}, s, r, b$



Categorical data

What is the main statistic we discussed for categorical data?

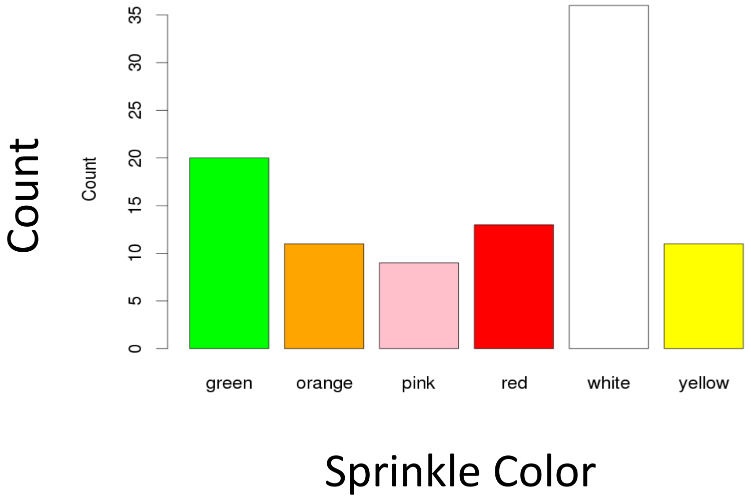
How can we plot categorical data?

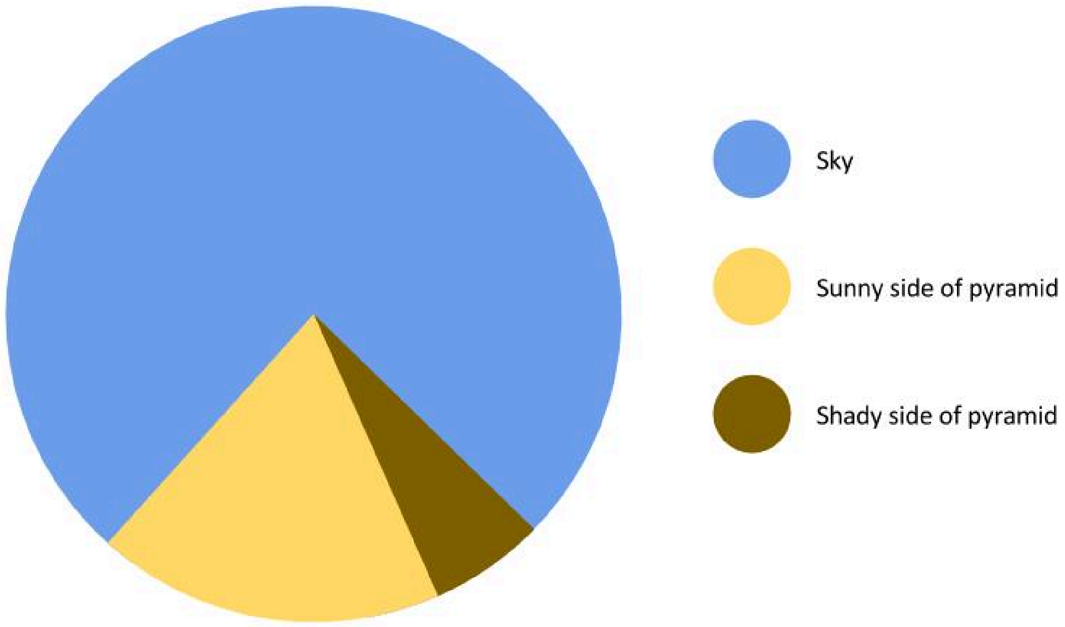
Categorical data

What is the main statistic we discussed for categorical data?

- π or p̂
- proportion = number in category/total

How can we plot categorical data?

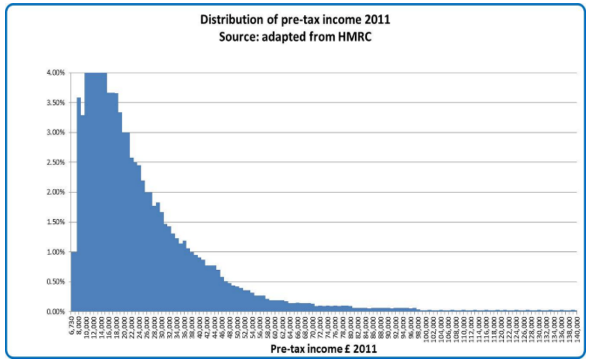


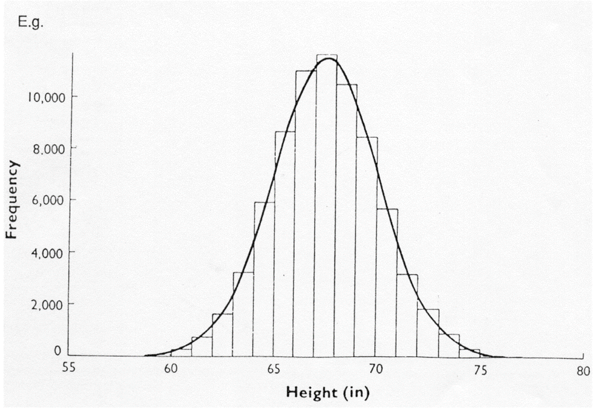


Quantitative data?

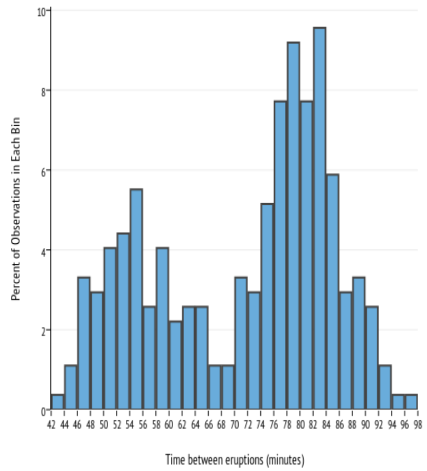
What is a good way to visualize the shape of quantitative data?

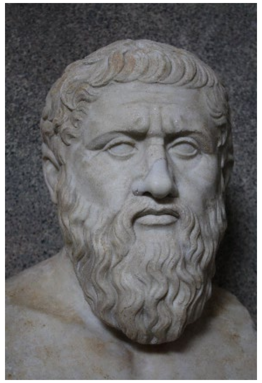
Income distribution

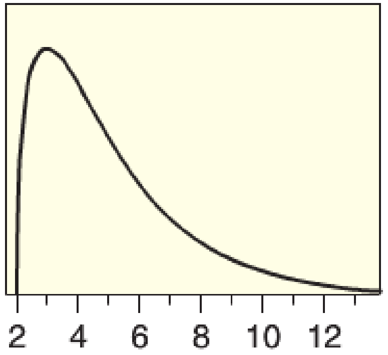


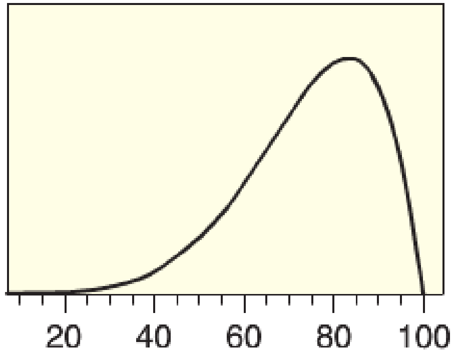


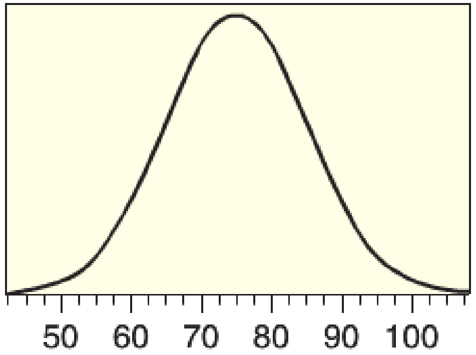
Old (Not So?) Faithful: A Bimodal Distribution

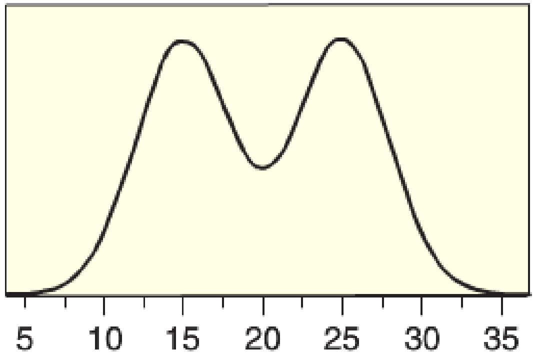




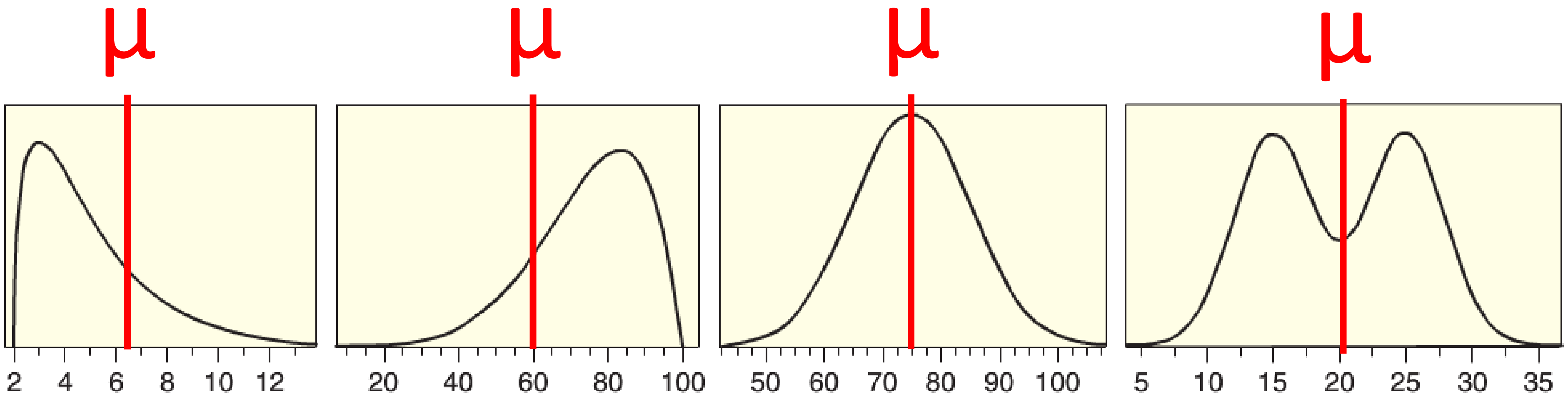






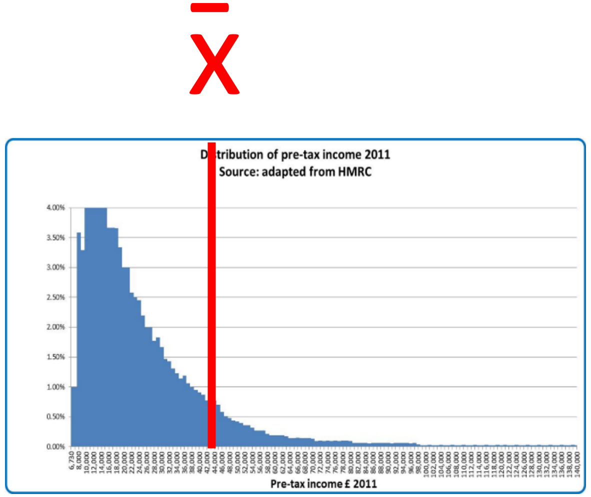


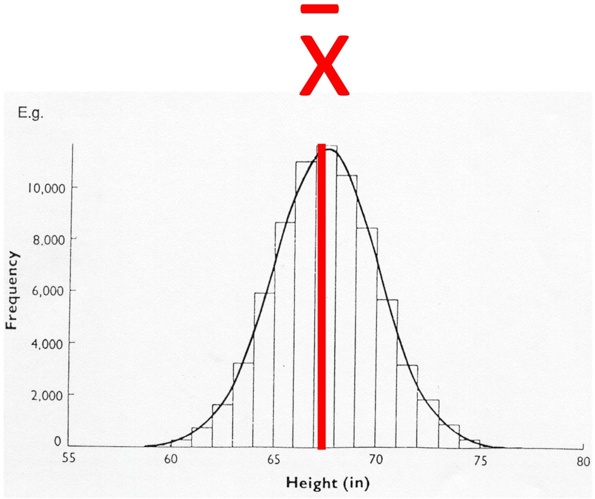
Measure of central tendency: the mean

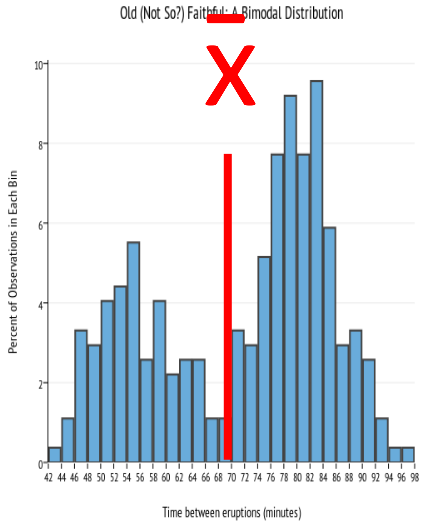


(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

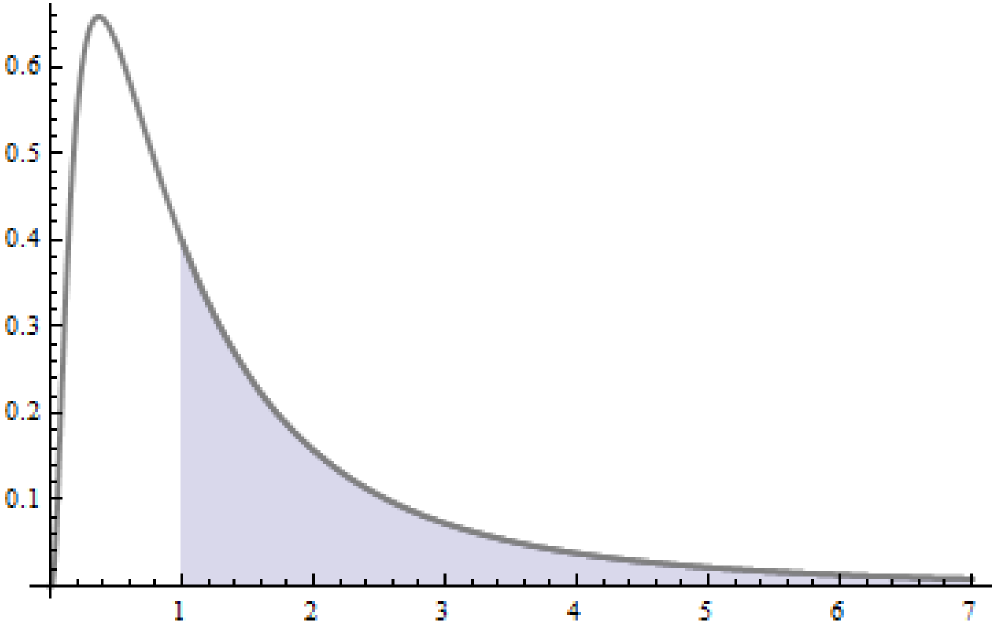
$$\frac{\Sigma_i^n x_i}{n}$$

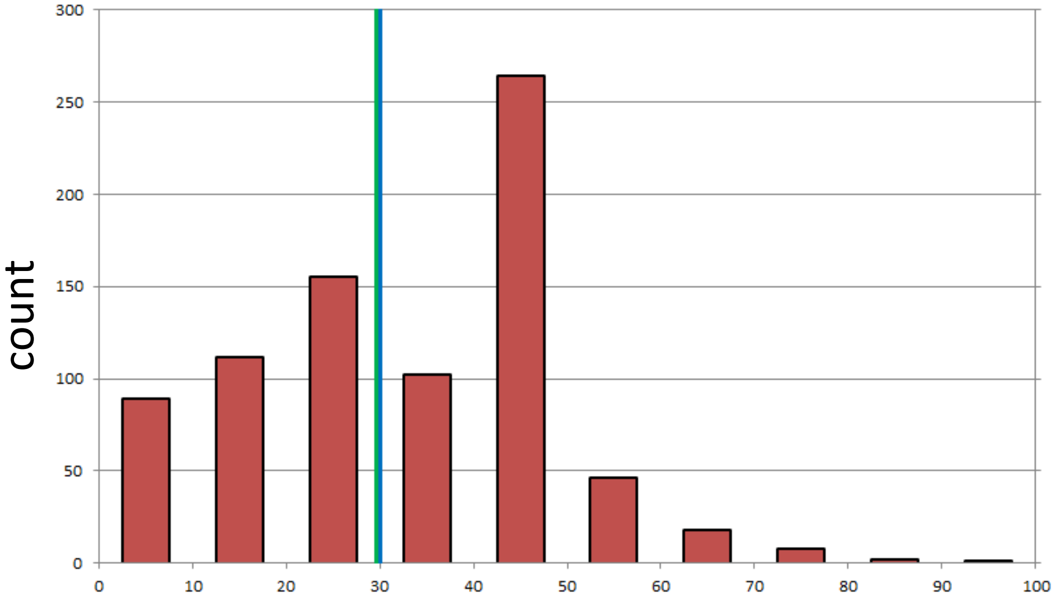






Measure of central tendency: the median



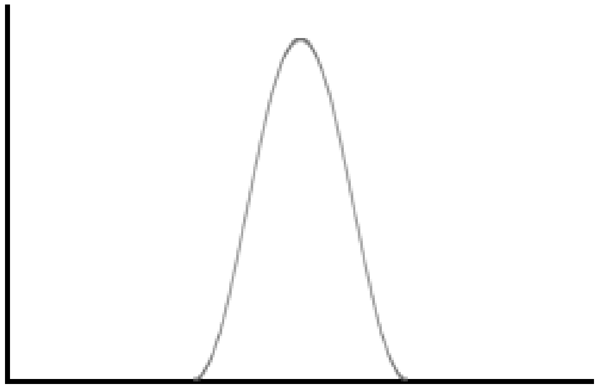


Which is resistant to outliers, the mean or the median?

The standard deviation

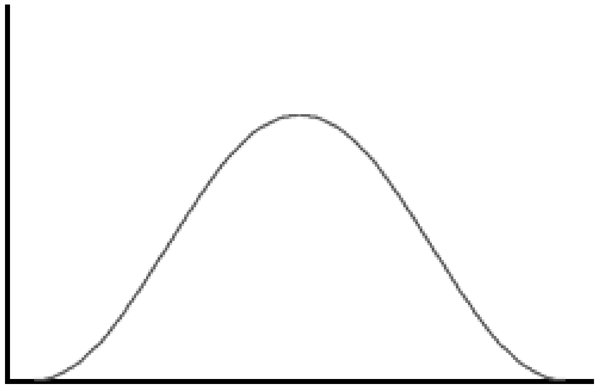
Which distribution has a larger standard deviation?

A



parameter σ

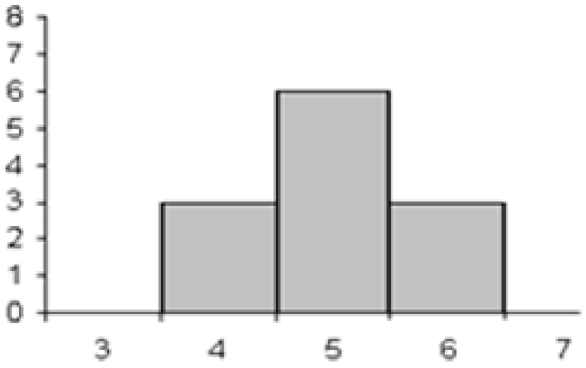
B

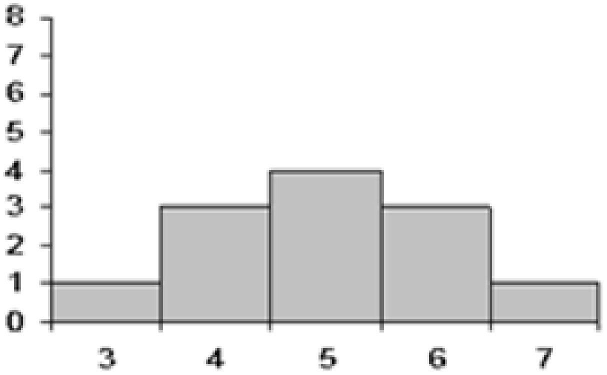


The standard deviation

Which distribution has a larger standard deviation?

statistic: s





What is the formula for the standard deviation?

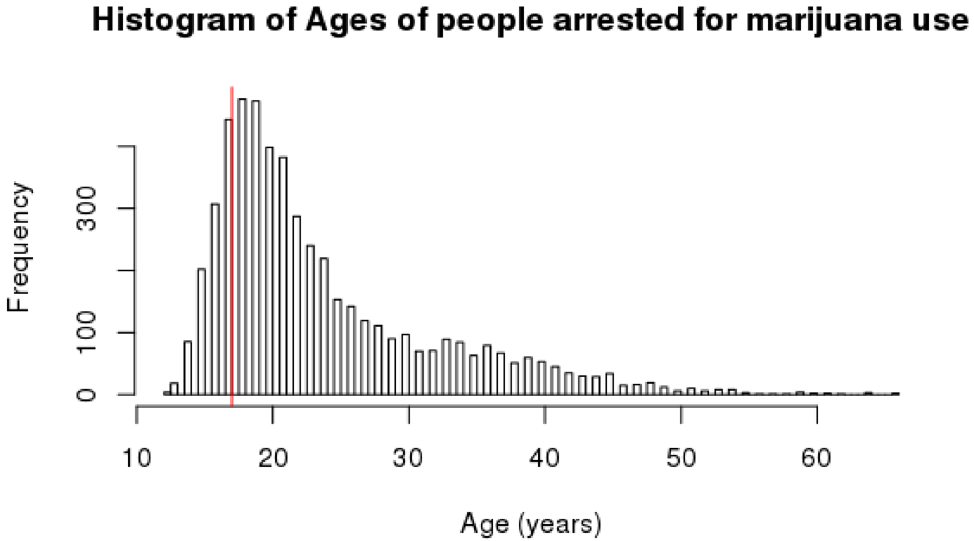
$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

z-scores and percentiles

What is a z-score and why is it useful?

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

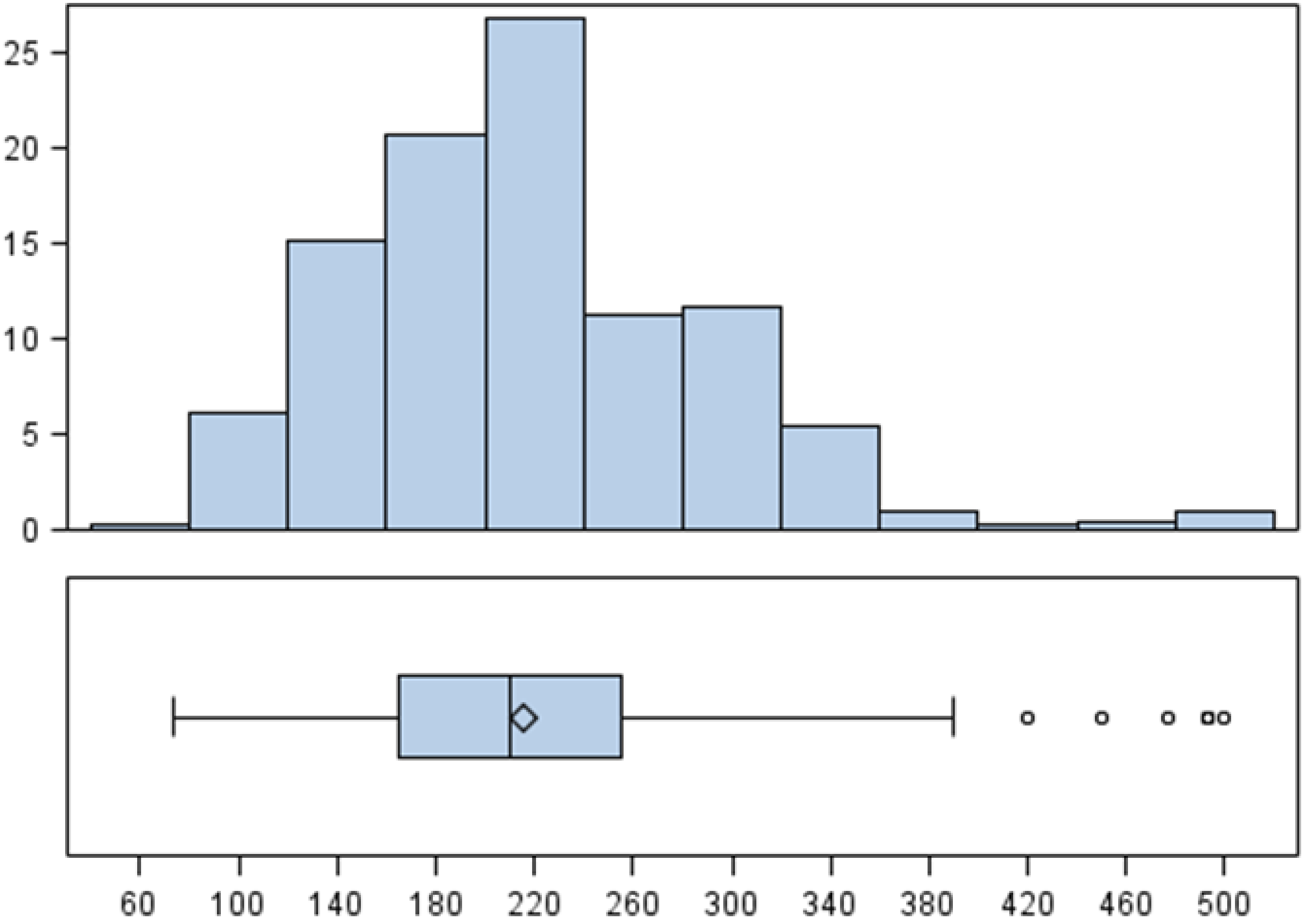
What is the pth percentile?



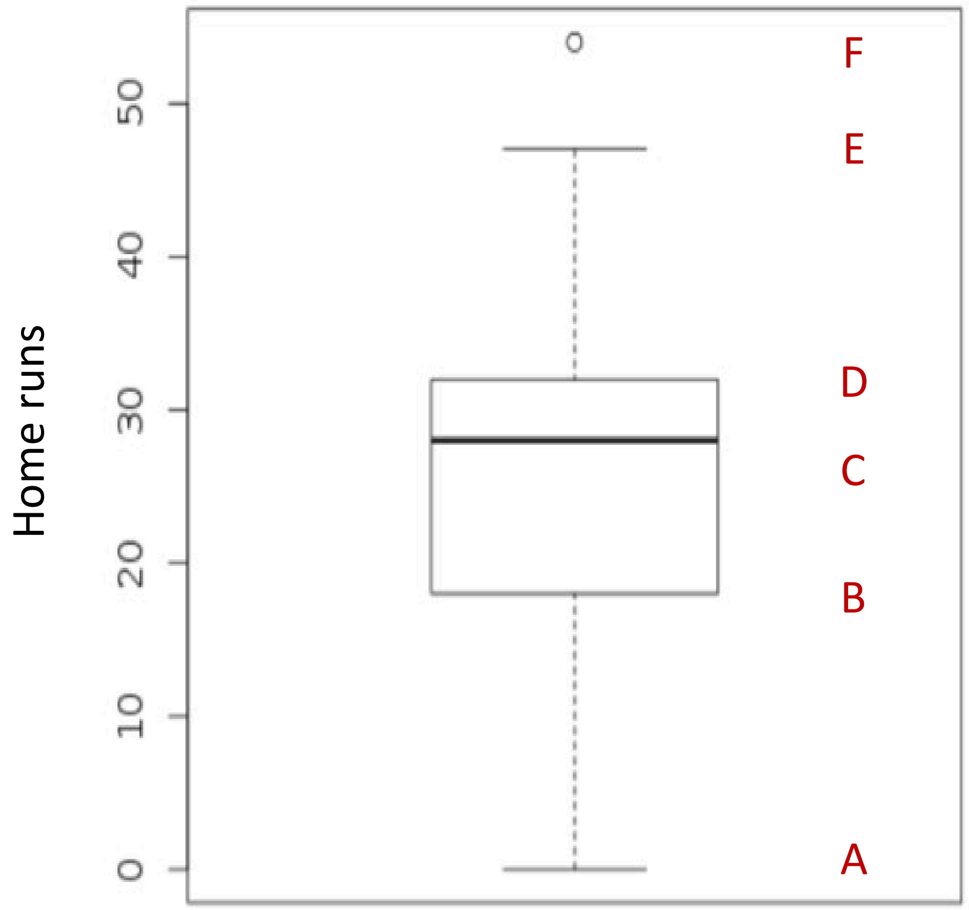
Normal pillow

What percent of the pillow’s mass is ± 1 standard deviations from the mean?

What is a five-number summary and a box plot?



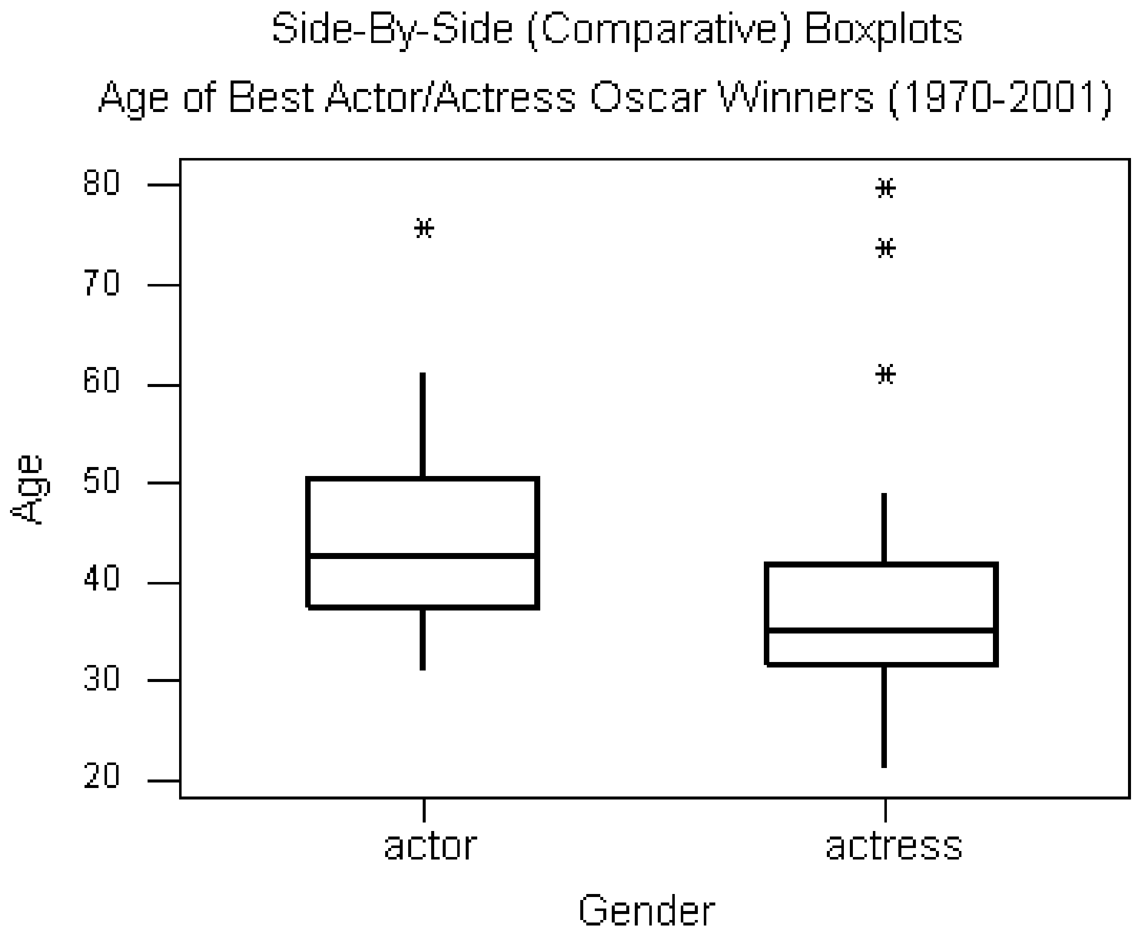
Box plot quiz



What is:

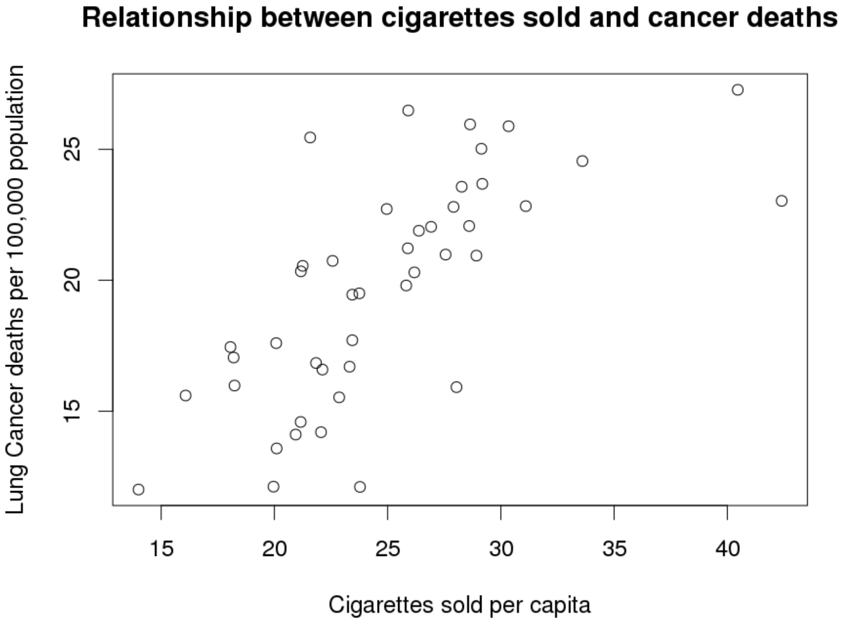
- Q1?
- Q3?
- The median?
- Most extreme values that are not outliers
- Outliers

Side-by-side boxplots



Relationships between measures

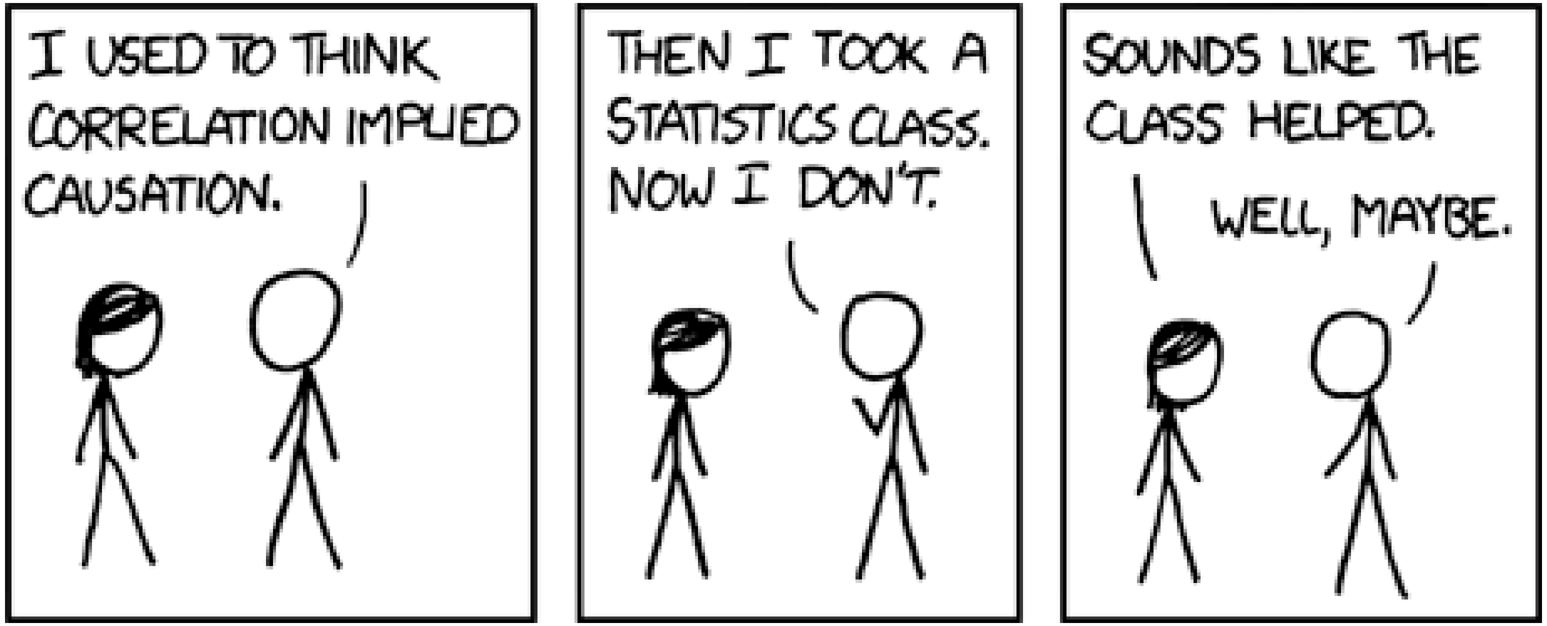
Q: What is this type of plot called?



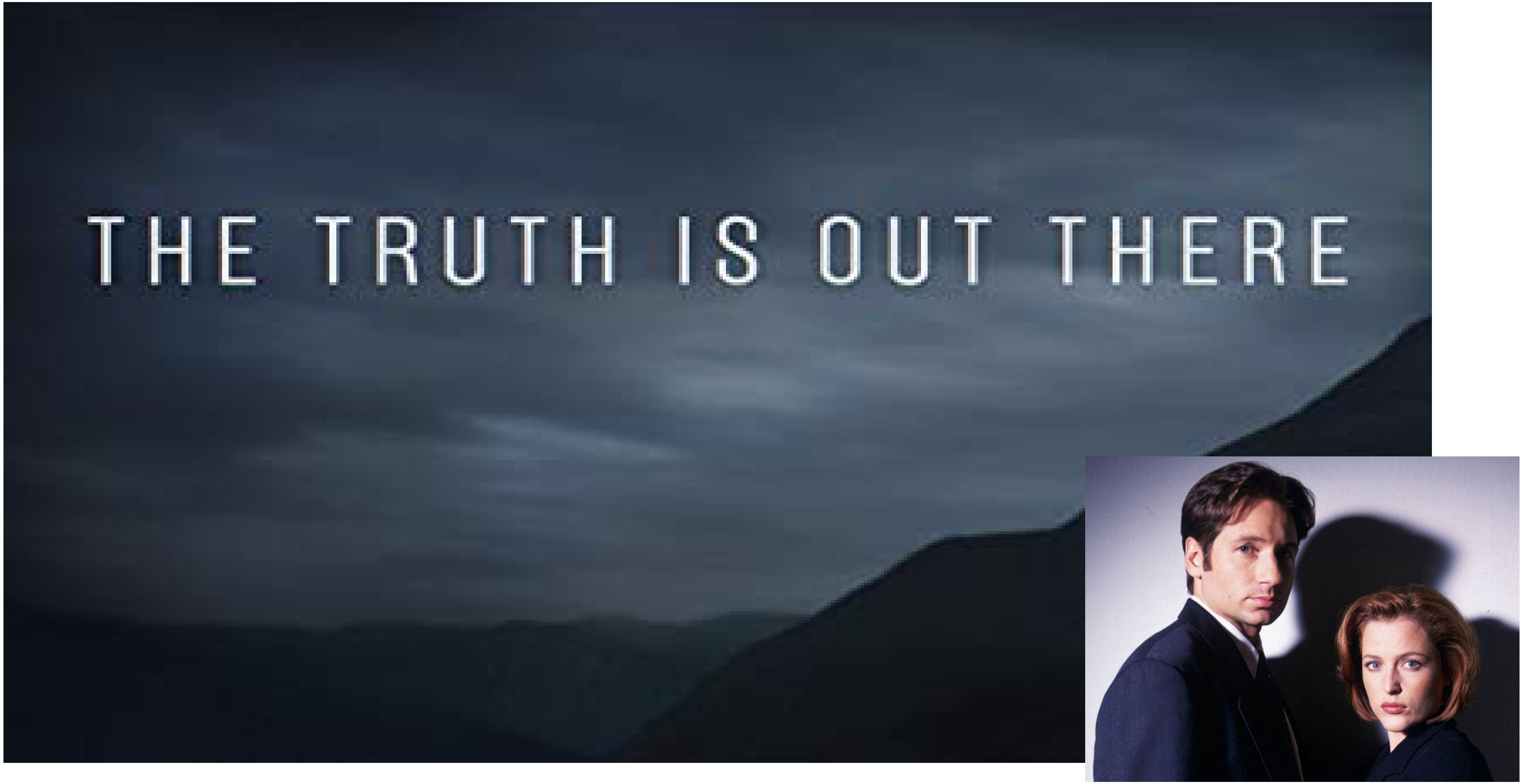
Q: What statistic have we used to describe the linear relationship between quantitative variables?

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

Does correlation imply causation?


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

What is our primary focus in Statistics?



Can you handle The TRUTH®?