Sampling, bias and sampling distributions

Overview

Review of correlation and review/continuation of linear regression

Sampling

Sampling bias

If there is time: Sampling distributions

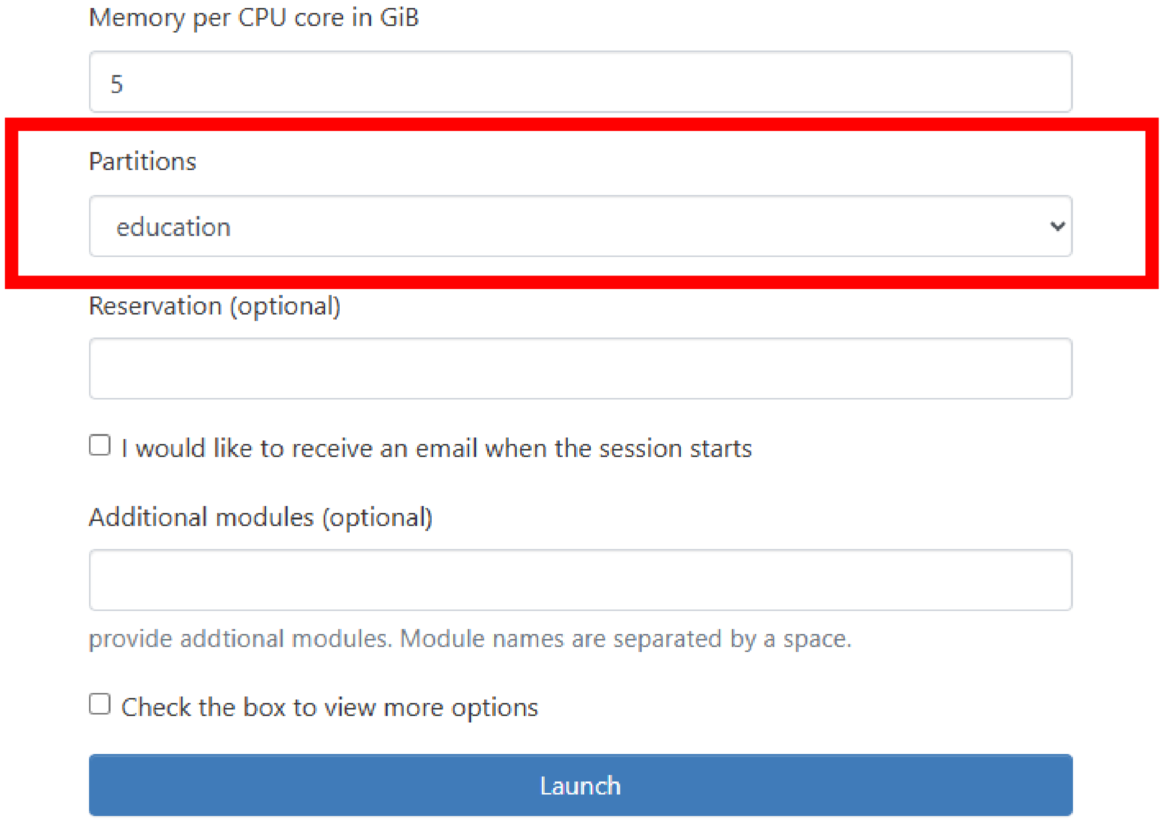
Announcement

Homework 3 has been posted!

It is due on Gradescope on Sunday September 21st at 11pm

- **Be sure to mark each question on Gradescope!**

Also, be sure to use the “education” partition on the YCRC server



Memory per CPU core in GiB

5

Partitions

education

Reservation (optional)

□ I would like to receive an email when the session starts

Additional modules (optional)

provide addtional modules. Module names are separated by a space.

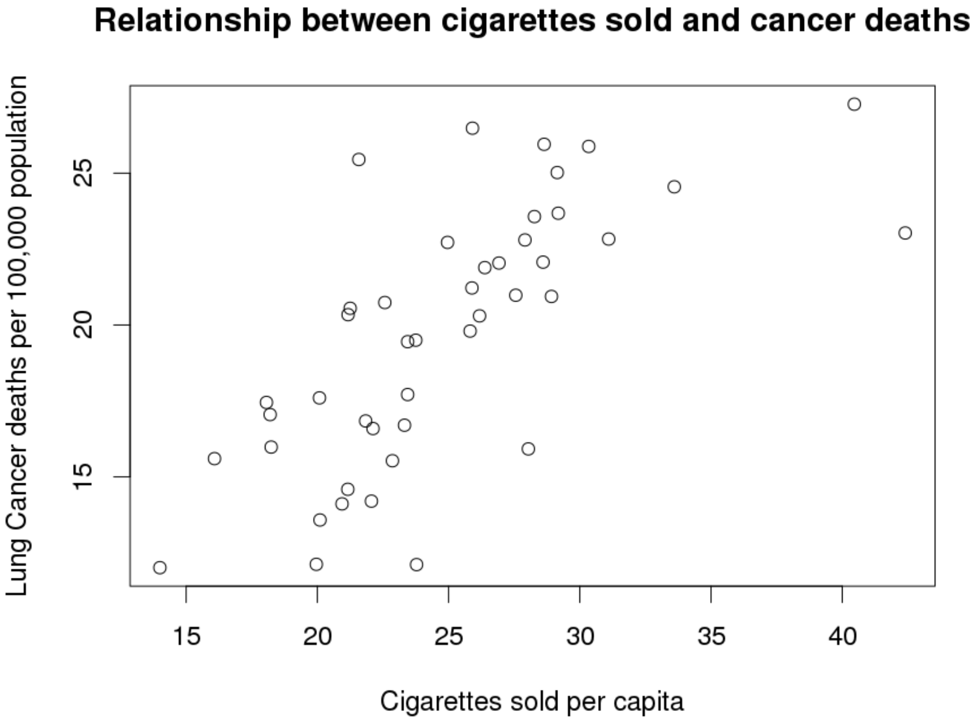
□ Check the box to view more options

Note: The homework involves taking a “Quiz” on Canvas to test your knowledge

- You should study for the Quiz before you take it, although you will only be graded on completion and not on how many questions you get right

Review of correlation and linear regression

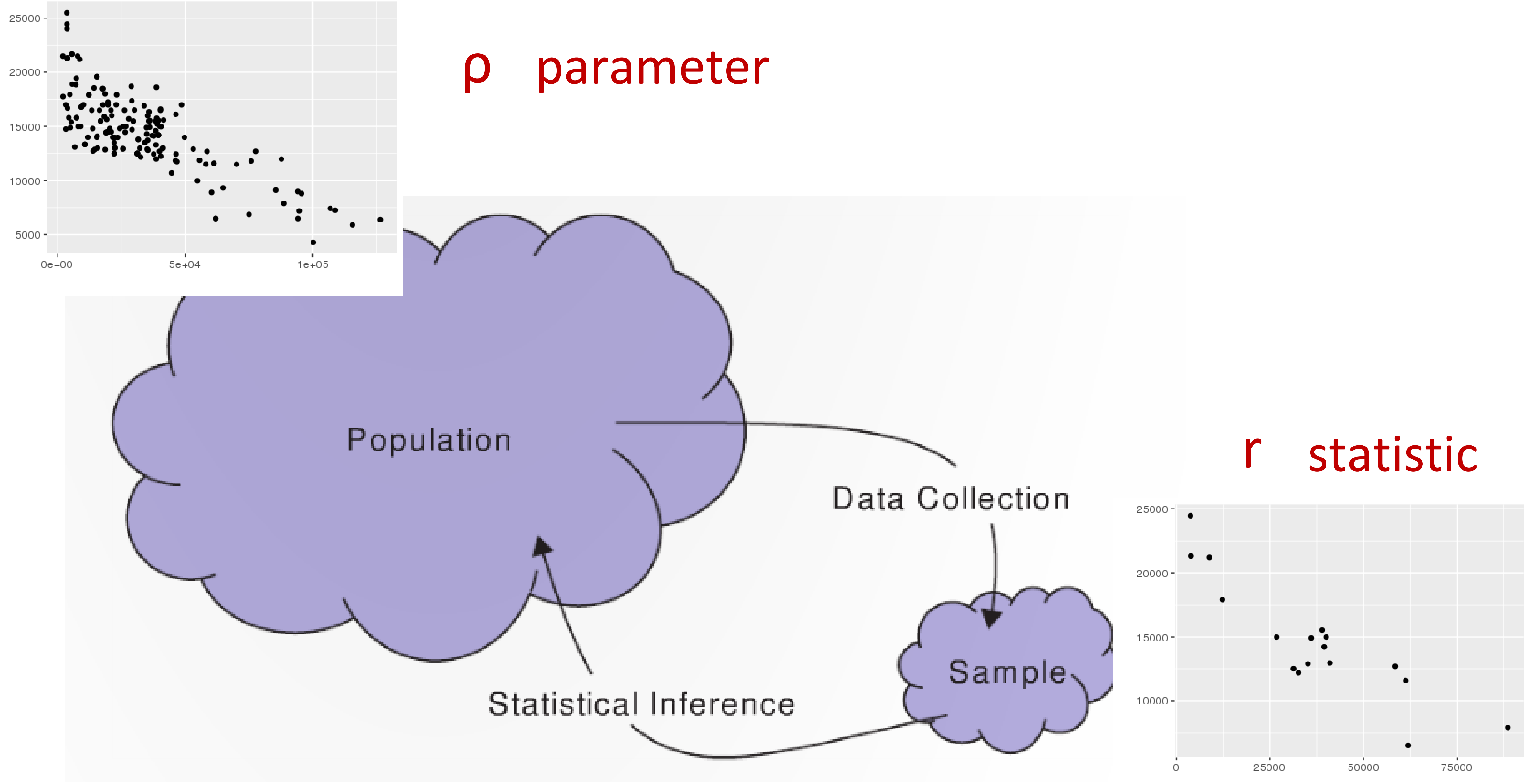
Review: scatter plots and the correlation coefficient



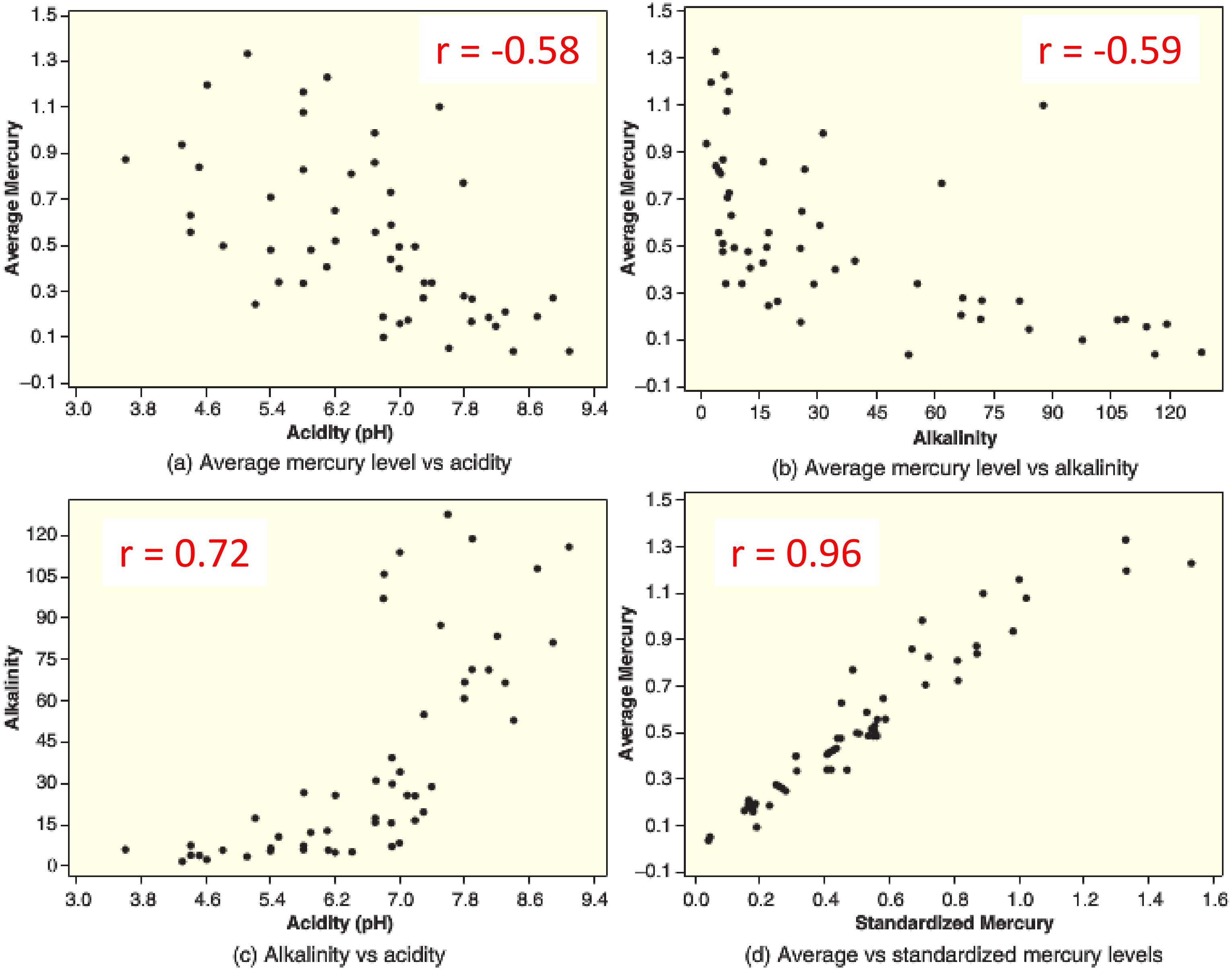
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

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

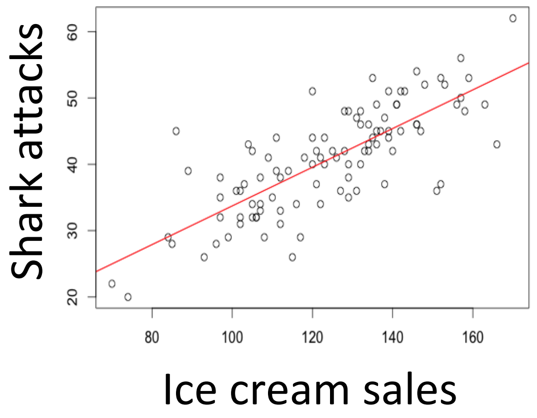
plot(x, y)

# calculate the correlation

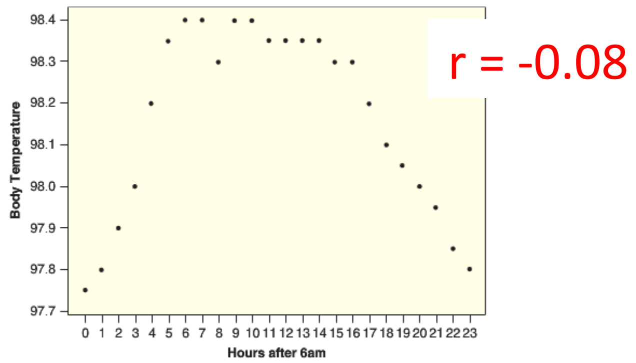
cor(x, y)

Correlation cautions

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

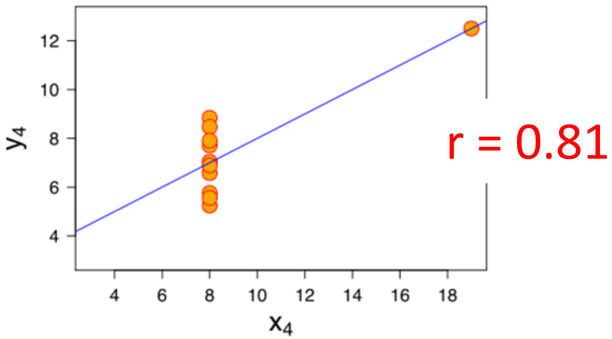


2. A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.



r = -0.08

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



r = 0.81

Review: Regression

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

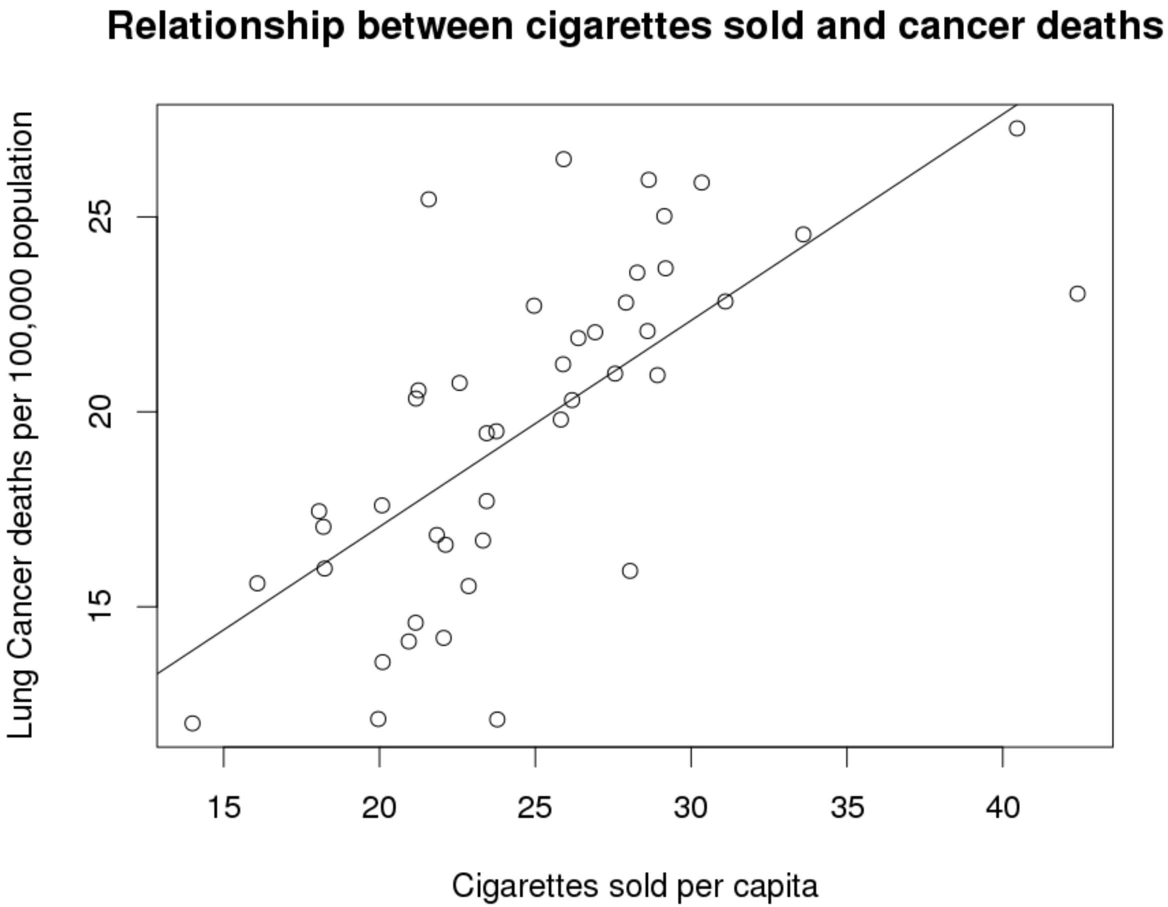
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a <u>line</u> to the data, called the **regression line**

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

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

Review: Cancer smoking regression line



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

R: `my_fit <- lm(y ~ x)`

```
coef(my_fit)
```

$a = 6.47 \quad b = 0.0053$

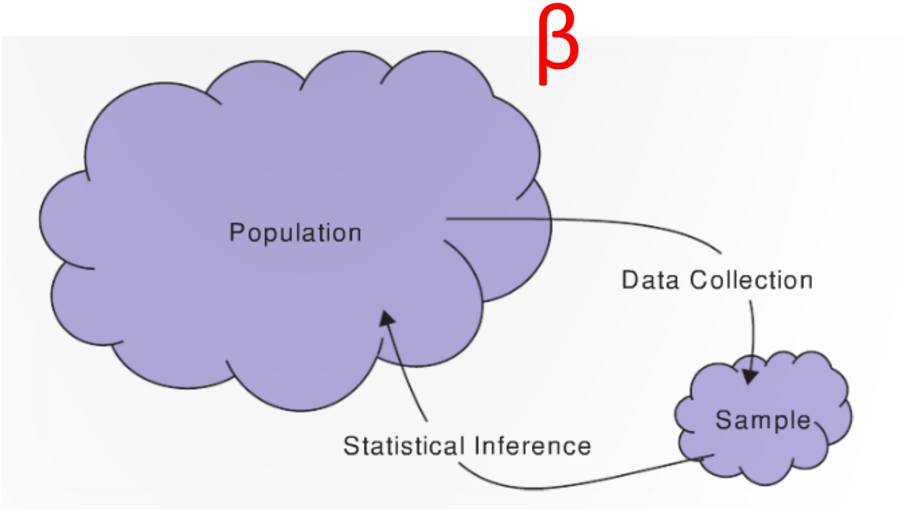
$$\hat{y} = 6.47 + .0053 \cdot x$$

Review: Notation

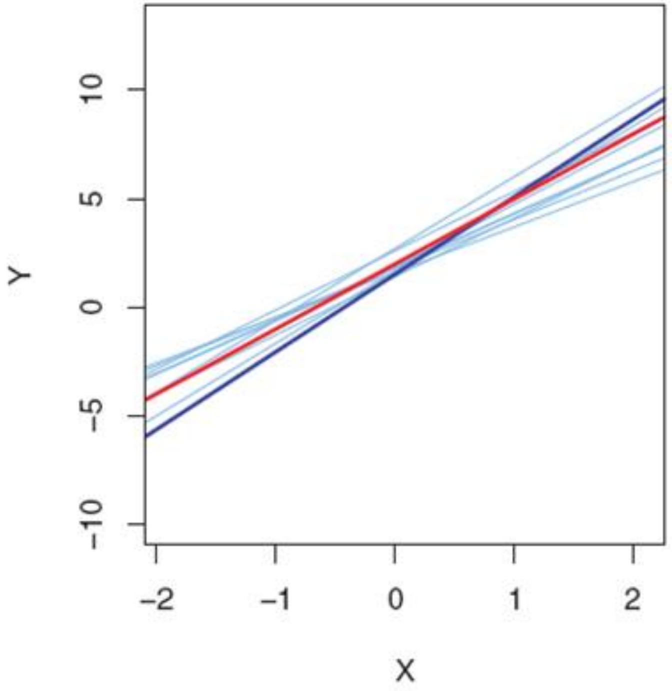
The Greek letter **β** is used to denote the slope of the **population**

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

β



b

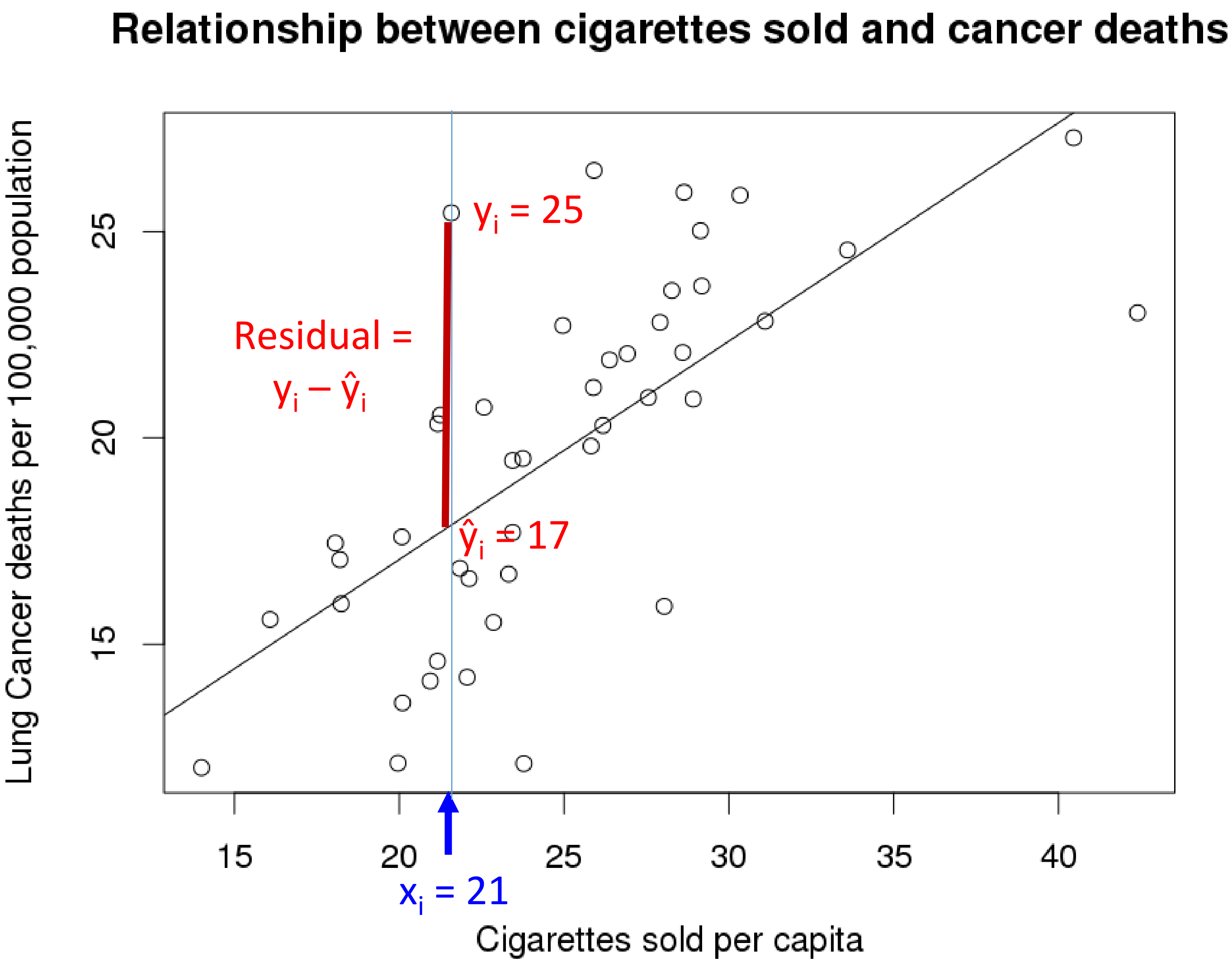


Residuals

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

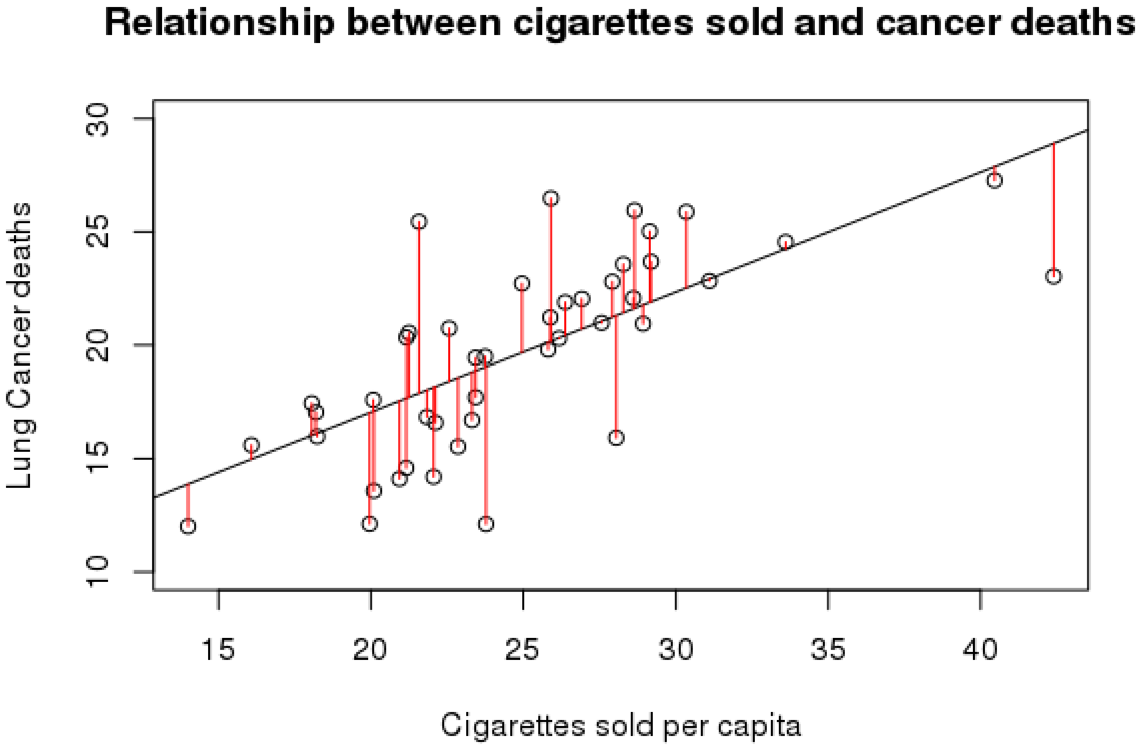
$$\mathit{Residual}_i \;=\; \mathit{Observed}_i - \mathit{Predicted}_i \;=\; y_i - \hat{y}_i$$

Cancer smoking residuals



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



Find the line of best fit

Cancer smoking residuals

Cancer obs	Cancer pred	Residuals	Residuals^2
(y)	($\hat{y}$)	($y - \hat{y}$)	$(y - \hat{y})^2$
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

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

Find the a and b

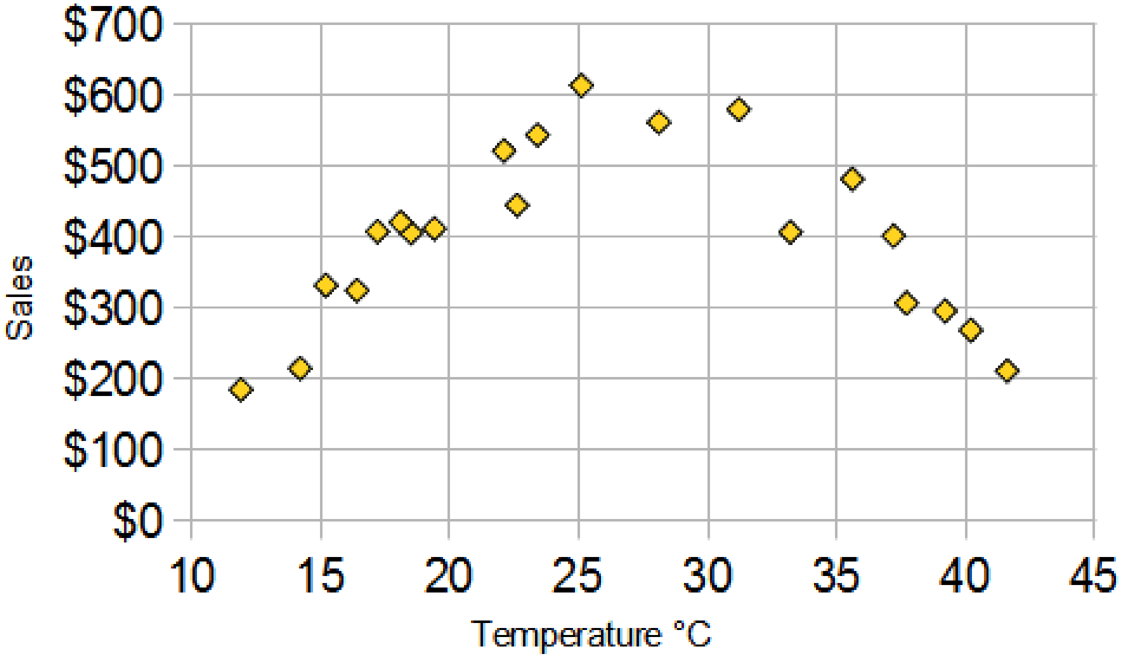
That minimizes the sum of the squared residuals

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

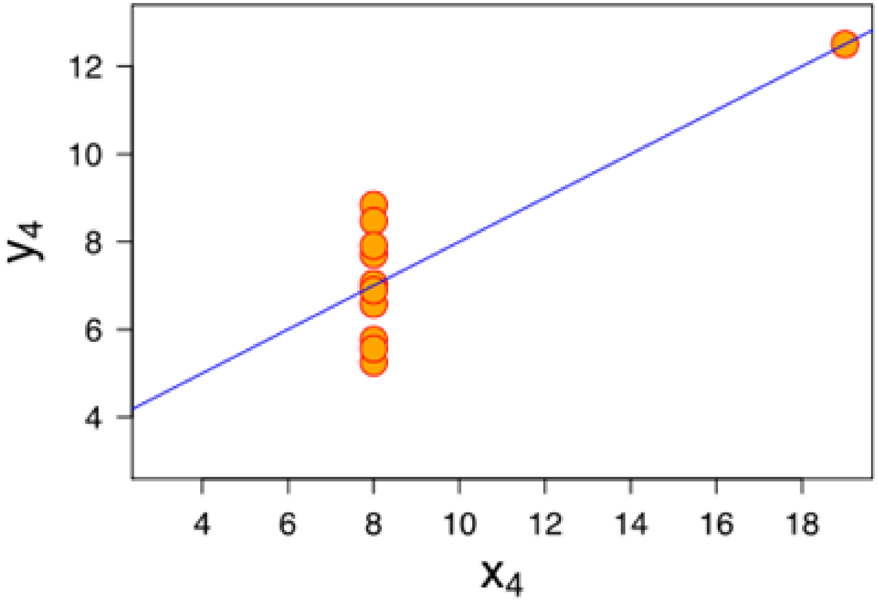
Regression caution # 2

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



Linear regression in R

Regression lines in R

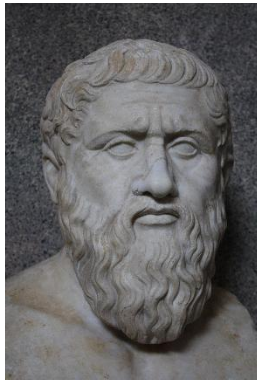
```
# load the data
    load("states_smoking.rda")

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

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

# get the a and b coefficients
   coef(lm_fit)

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

Try it in R!

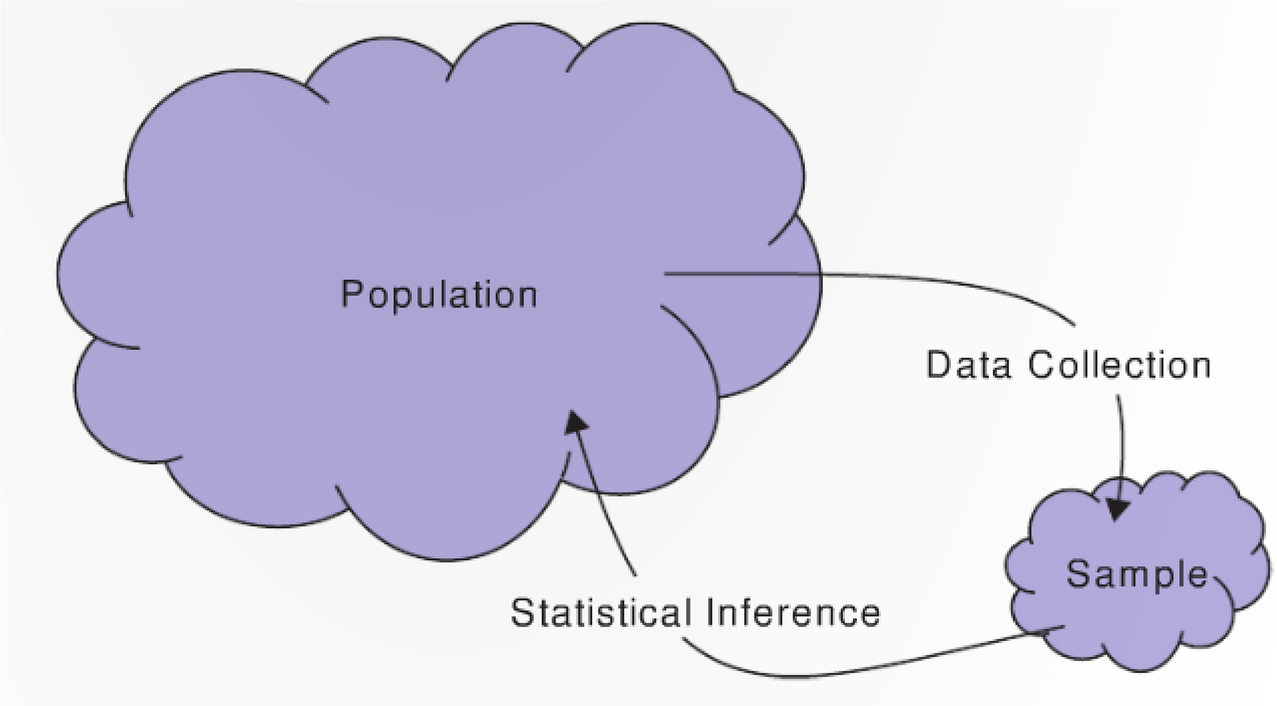
Review descriptive statistics

Once you’ve reviewed what we have covered, take the Descriptive statistics review “Quiz” on Canvas to test your knowledge

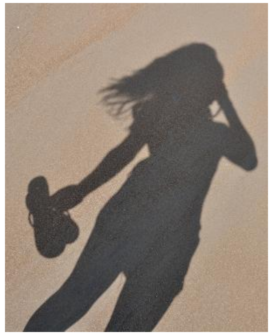
Parameters and statistics



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



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

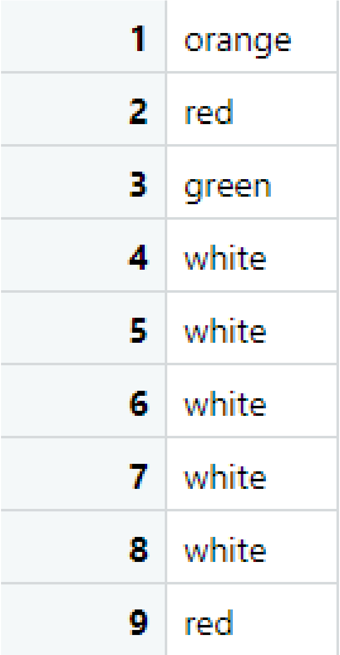


Can you handle the Truth®?

Sampling

Where do samples/data come from?

Example: sampling 100 sprinkles



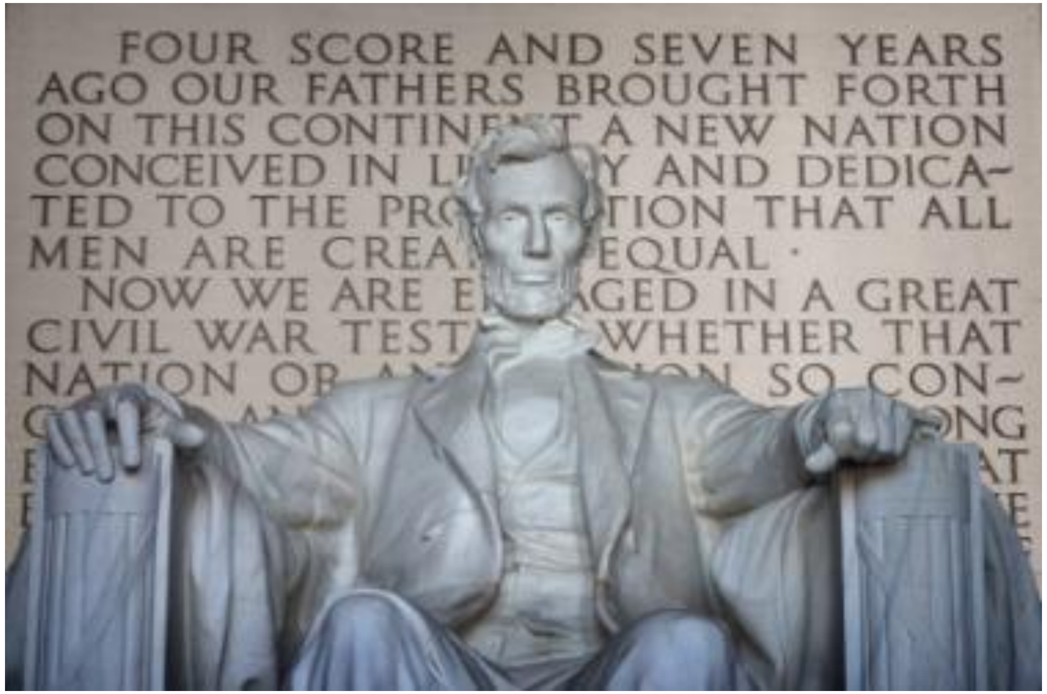
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

The ***sample size*** (n) is the number of items in the sample
What is ***n*** in the sprinkle example?

Let’s try some sampling ourselves…

Please fill out the Gettysburg sampling worksheet survey on Canvas where you will randomly sample 10 words from the Gettysburg address

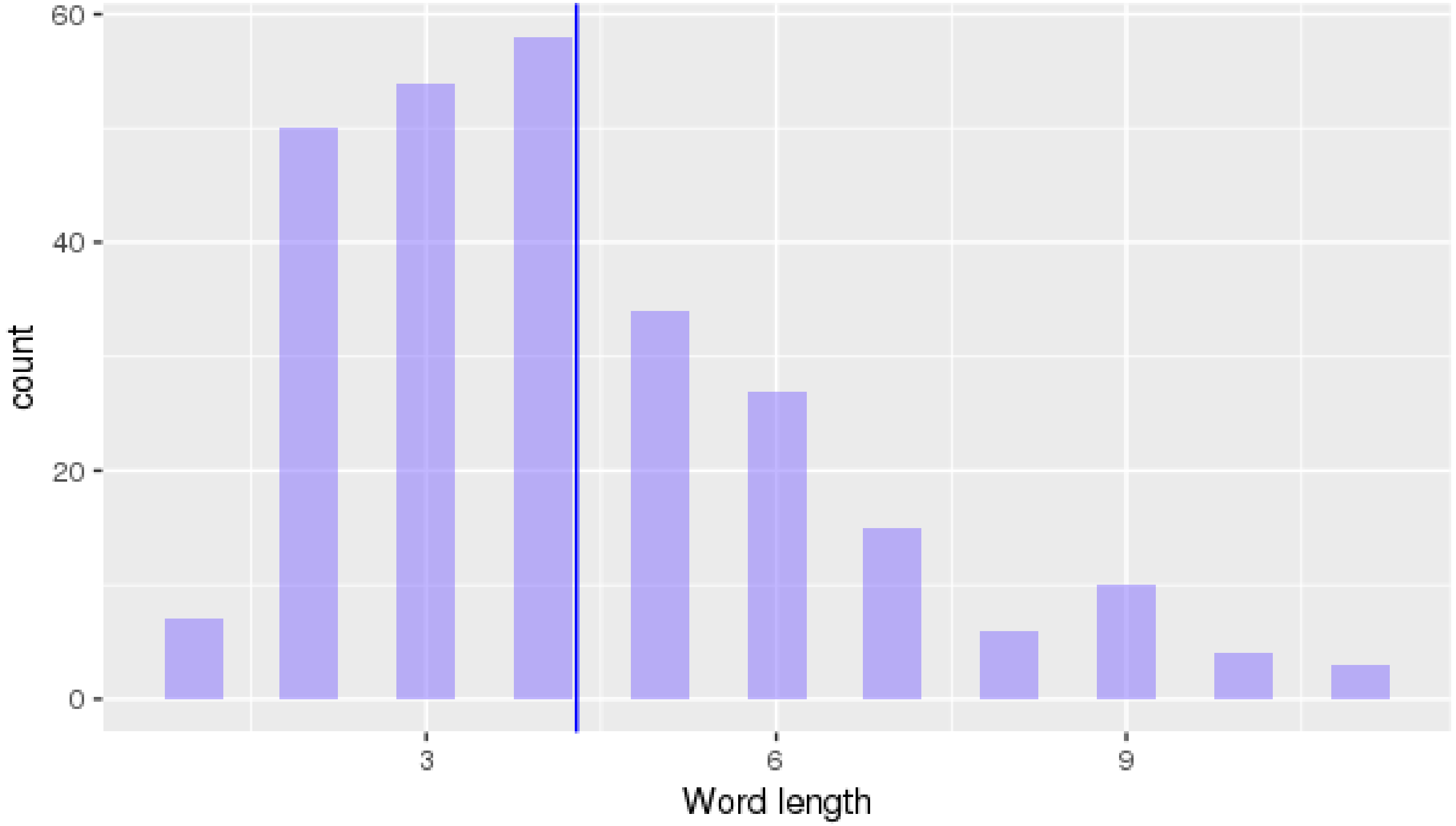
Let’s see if we can all fill it out in ~5 minutes!


FOUR SCORE AND SEVEN YEARS
AGO OUR FATHERS BROUGHT FORTH
ON THIS CONTIN T A NEW NATION
CONCEIVED IN LI Y AND DEDICA-
TED TO THE PR TION THAT ALL
MEN ARE CREA EQUAL ·
NOW WE ARE E GED IN A GREAT
CIVIL WAR TEST WHETHER THAT
NATION OR SO CON-

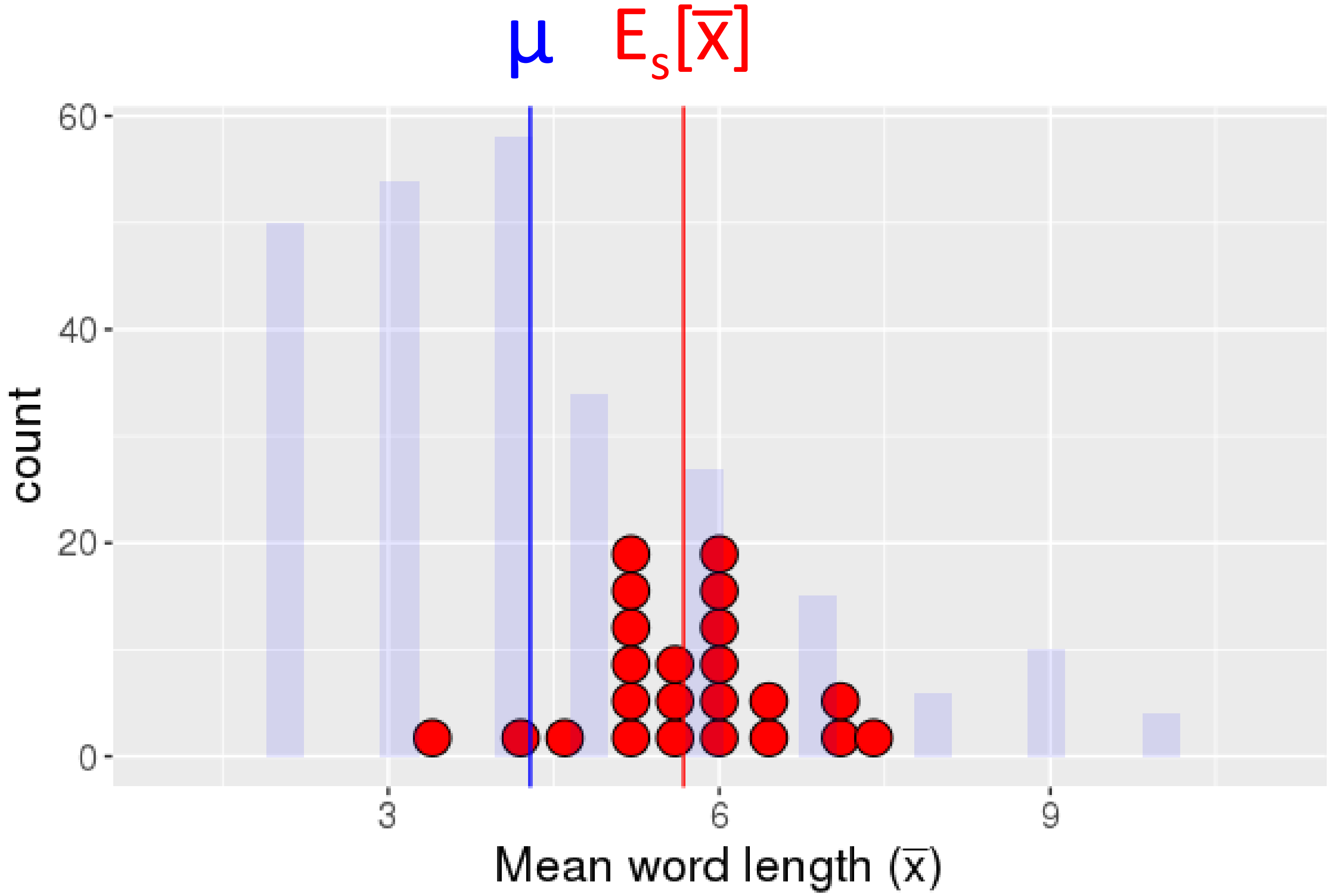
Sampling and bias

Gettysburg address: lengths of 268 words in the population

$\mu = 4.287$



Gettysburg address, mean word length



Bias

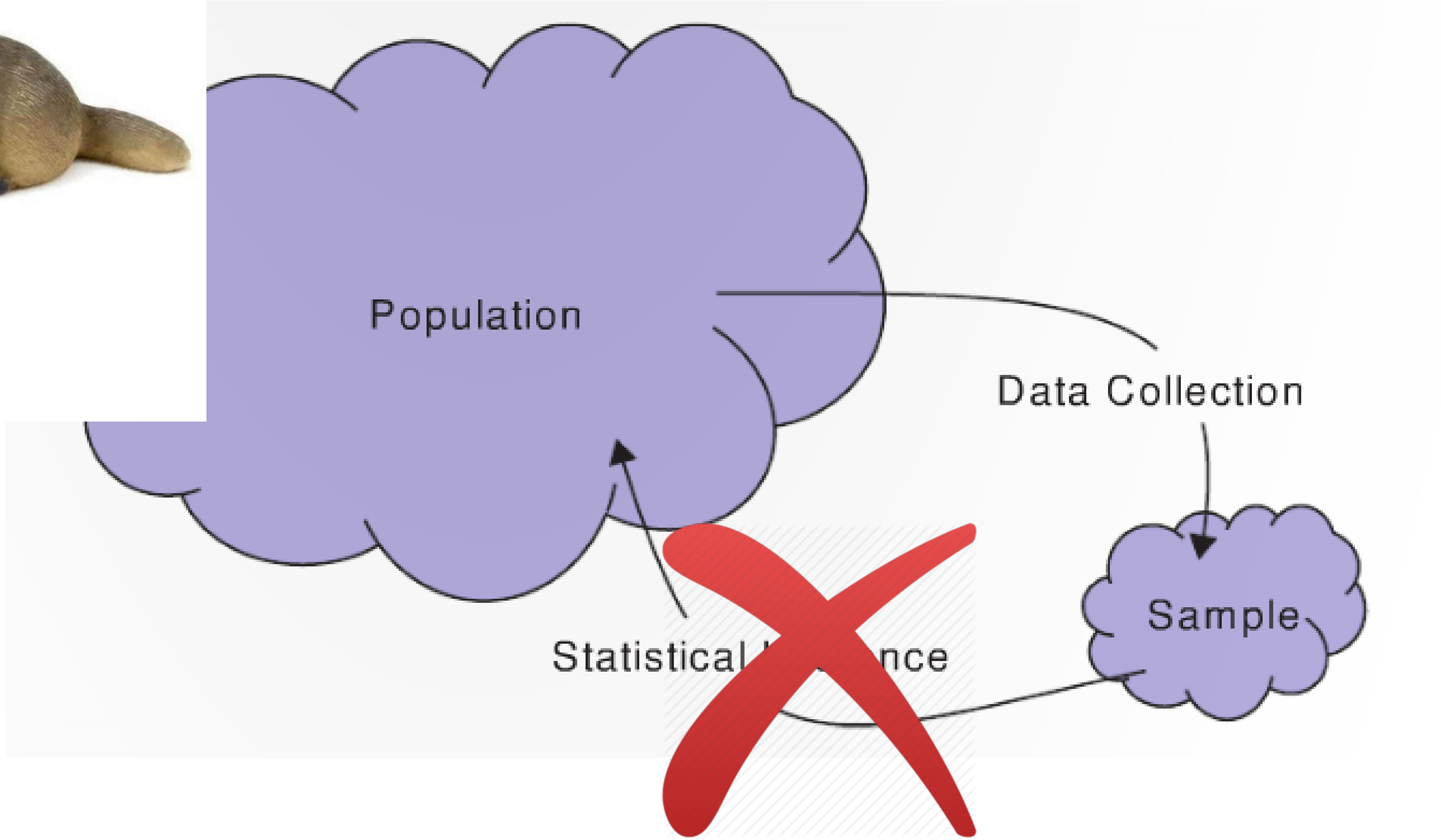
Sampling bias exists when the method of collecting the data causes the sample to inaccurately reflect the population.

This leads to ***biased statistics*** where our average statistic value does not equal the parameter value.

- E.g., $E[\bar{x}] \neq \mu$

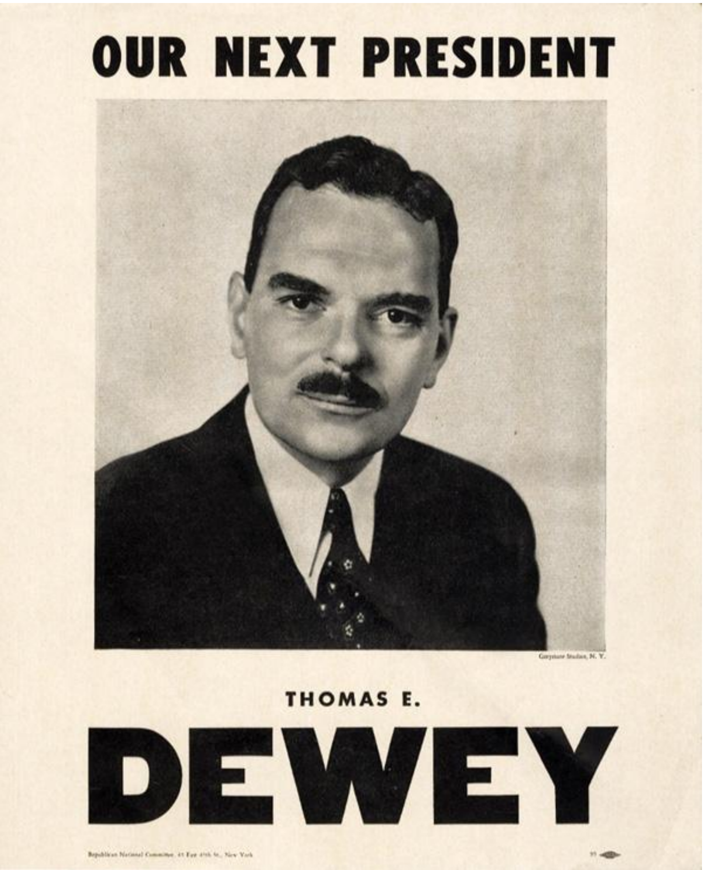
Statistical bias

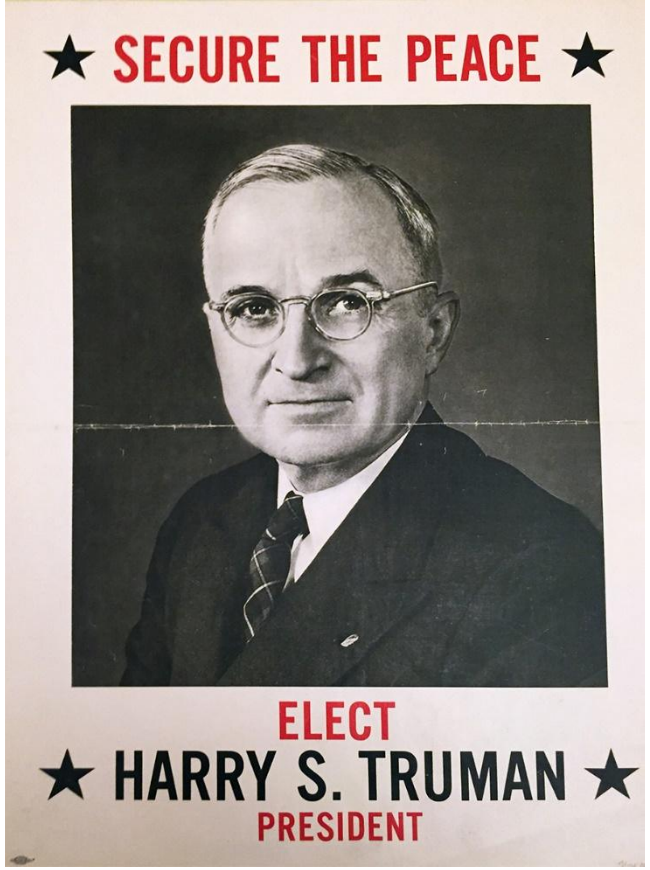
μ



$\bar{x}$

1948 US election





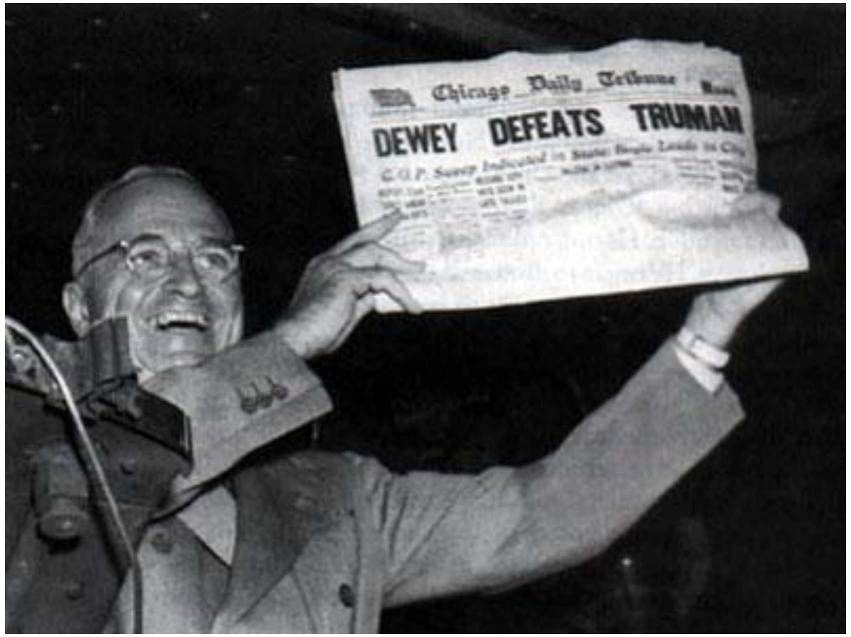
Newspaper title: Dewey Defeats Truman (1948)

The newspaper was published before the conclusion of the 1948 presidential election

The results were based on a large telephone poll which showed Dewey sweeping Truman

However, Harry S. Truman won the election

Q: What went wrong?



Basic questions for sampling

What is the population?

What is the sample?

Do they differ in a meaningful way?

To prevent bias: use simple random sample!

Simple random sample: each member in the population is equally likely to be in the sample.

Allows for generalizations to the population!

Soup analogy

How do we select a random sample?

Mechanically:

- Flip coins
- Pull balls from well mixed bins
- Deal out shuffled cards, etc.

Use computer programs

Bias or no bias?

1948 US election: Dewey vs. Truman

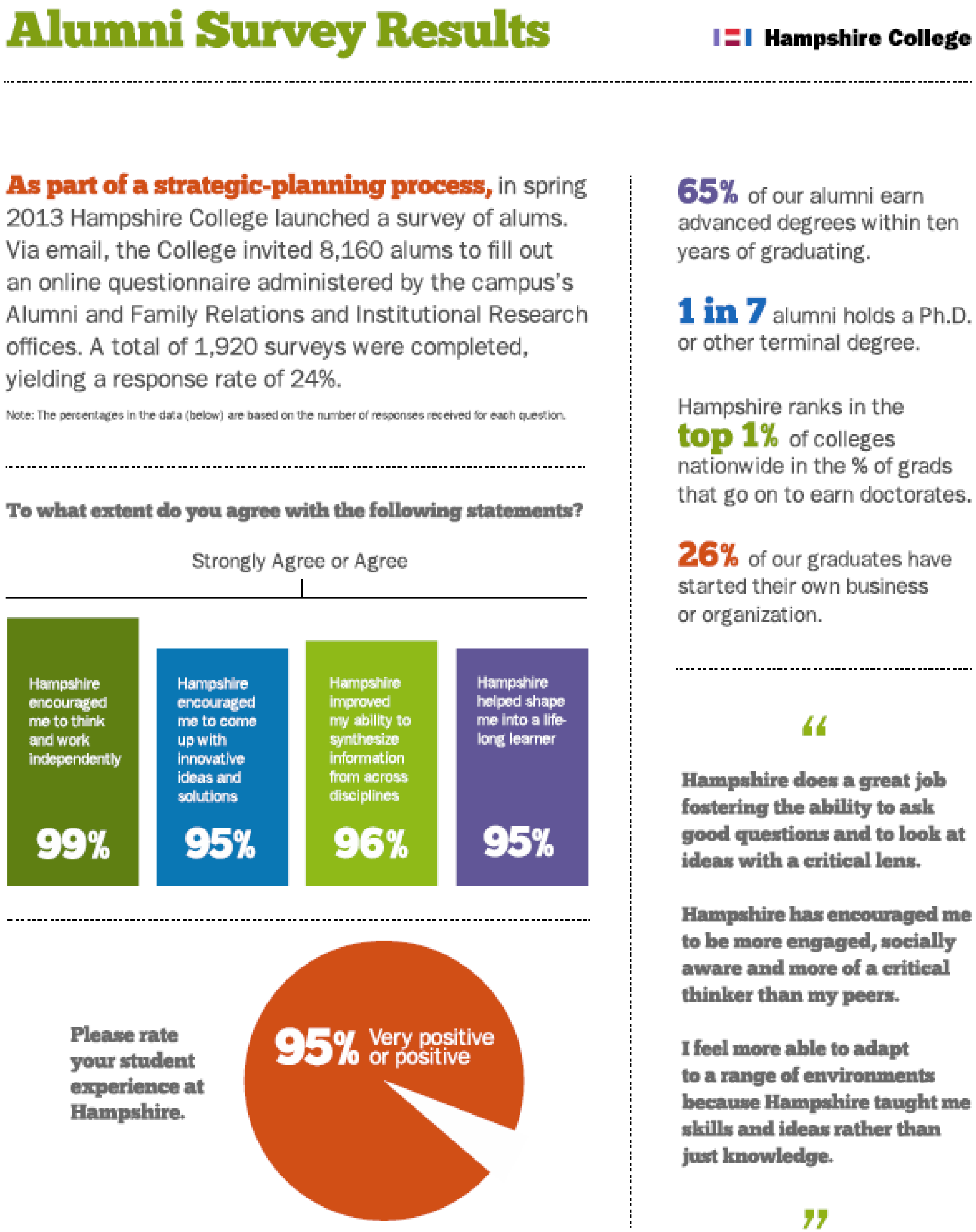
Suppose there was a poll for the Truman/Dewey election that randomly chose 6,000 people from all voters in the USA and calculated who they voted for.

Bias or no bias?

As part of a strategic-planning process, in spring 2013 Hampshire College launched a survey of alums.

Via email, the college invited 8,160 alums to fill out an online questionnaire administered by the campus's offices.

A total of 1,920 surveys were completed, yielding a response rate of 24%.



Alumni Survey Results

Hampshire College

As part of a strategic-planning process, in spring 2013 Hampshire College launched a survey of alums. Via email, the College invited 8,160 alums to fill out an online questionnaire administered by the campus's Alumni and Family Relations and Institutional Research offices. A total of 1,920 surveys were completed, yielding a response rate of 24%.

Note: The percentages in the data (below) are based on the number of responses received for each question.

To what extent do you agree with the following statements?

Strongly Agree or Agree

Hampshire encouraged me to think and work independently	Hampshire encouraged me to come up with innovative ideas and solutions	Hampshire improved my ability to synthesize information from across disciplines	Hampshire helped shape me into a life-long learner
99%	95%	96%	95%

Please rate your student experience at Hampshire.

95% Very positive or positive

65% of our alumni earn advanced degrees within ten years of graduating.

1 in 7 alumni holds a Ph.D. or other terminal degree.

Hampshire ranks in the **top 1%** of colleges nationwide in the % of grads that go on to earn doctorates.

26% of our graduates have started their own business or organization.

> **"Hampshire does a great job fostering the ability to ask good questions and to look at ideas with a critical lens.**
>
> **Hampshire has encouraged me to be more engaged, socially aware and more of a critical thinker than my peers.**
>
> **I feel more able to adapt to a range of environments because Hampshire taught me skills and ideas rather than just knowledge."**

Yelp reviews of restaurants?

An anonymous survey randomly select 6,000 people and asked them if have they used an elicit drug in the past month?

https://www.billoreilly.com/poll-center

The way you frame the question matters!

Quinnipiac University conducted two polls on November 5, 2015

First poll they asked: do you support "stricter gun control laws"?

- Yes = 46% No = 51% Difference = -5%

Second poll asked: do you support "stricter gun laws"?

- Yes = 52% No = 45% Difference = 7%

How could this affect the newspaper headlines?

- "Majority of Americans ***oppose*** stricter gun control laws" vs.
- "Majority of Americans ***support*** stricter gun laws"

Also see textbook section 1.2:

- "If you had to do it over again, would you have children?"

Practicalities…

It might not be feasible to randomly select equally from all members of a population.

This might not be a problem as long as the sample is representative of the population.

Example: If we wanted to know proportion of people left-handed in the US, randomly sampling Yale students might be good enough.

Need to think carefully to avoid bias!

Statistics requires thought!

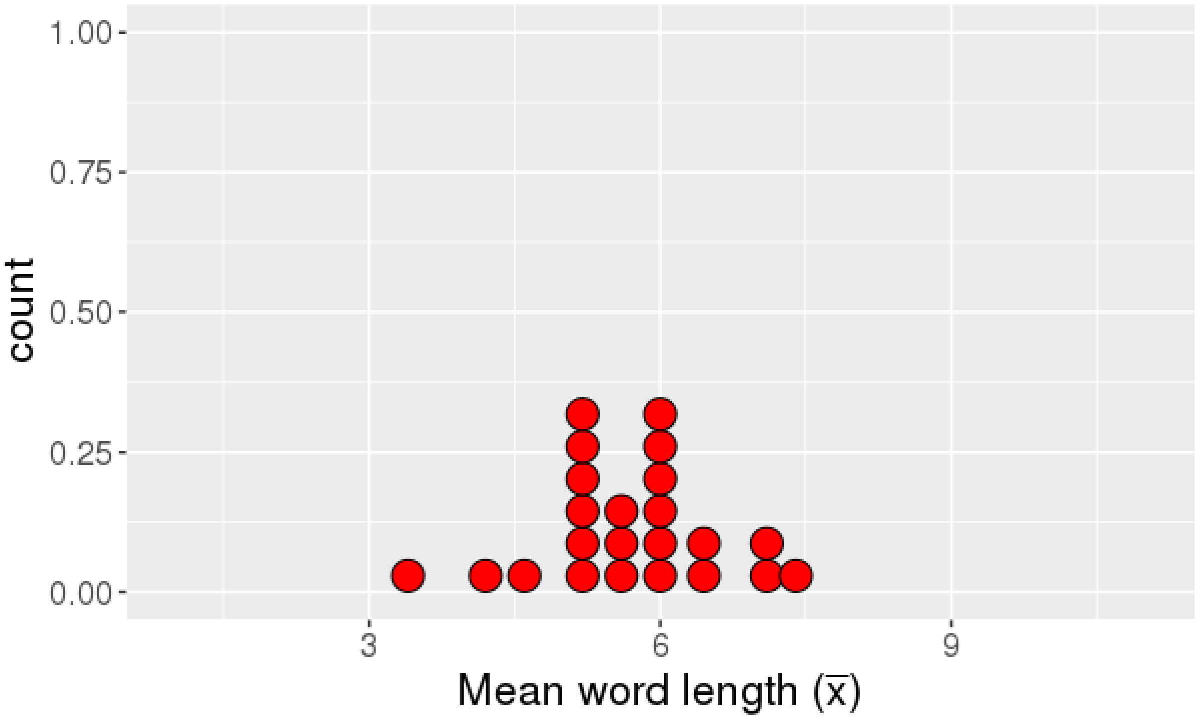
Use your own reasoning:

- What is the population I am interested in?
- Does the sample reflect the population of interest?
- Be your own worst critic!

Sampling distributions

Recall for our distribution of Gettysburg word lengths…

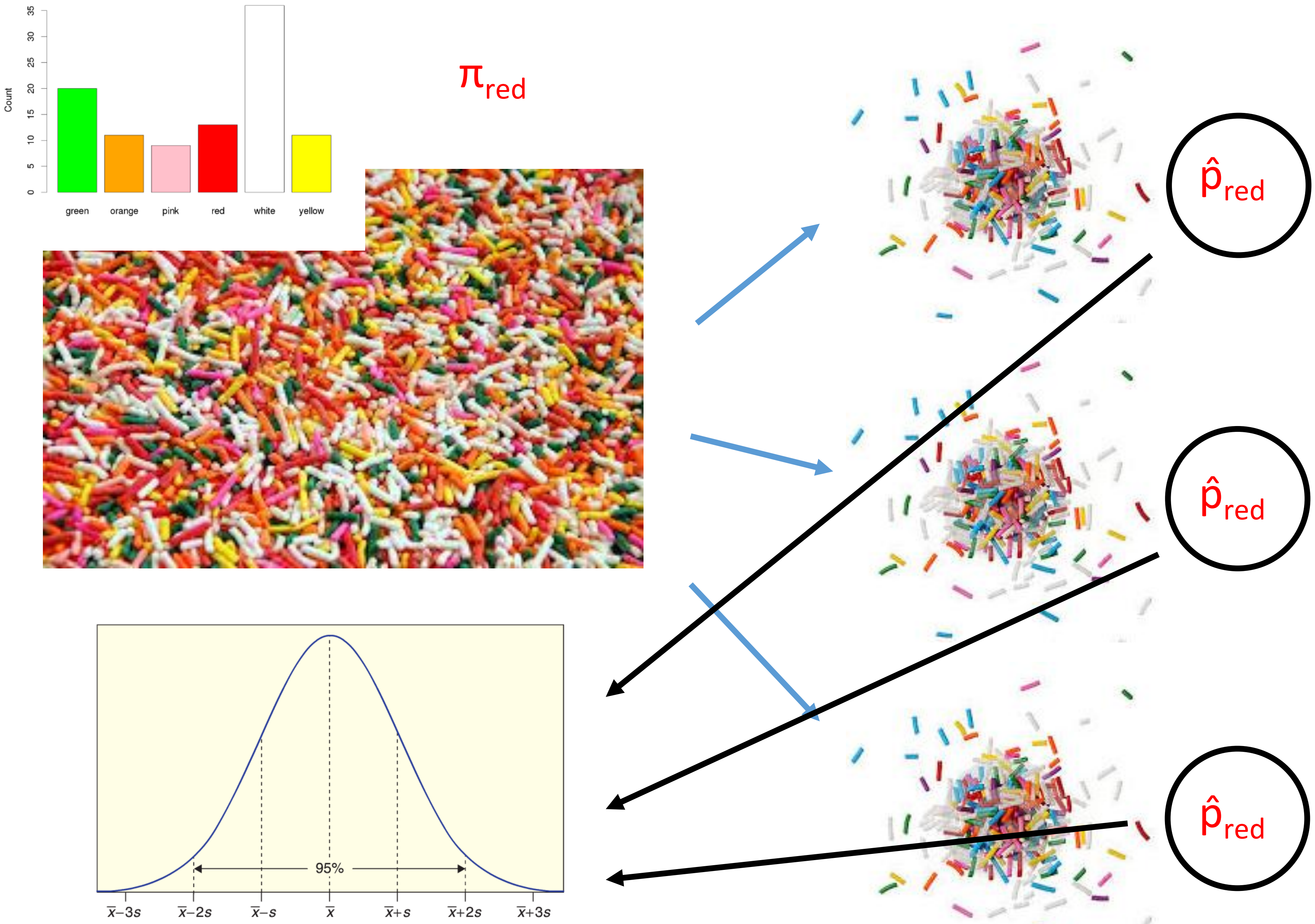
Q: What does each case that is plotted correspond to?



Sampling distribution

A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size (n) from the same population

A sampling distribution shows us how the sample statistic varies from sample to sample



π_{red}

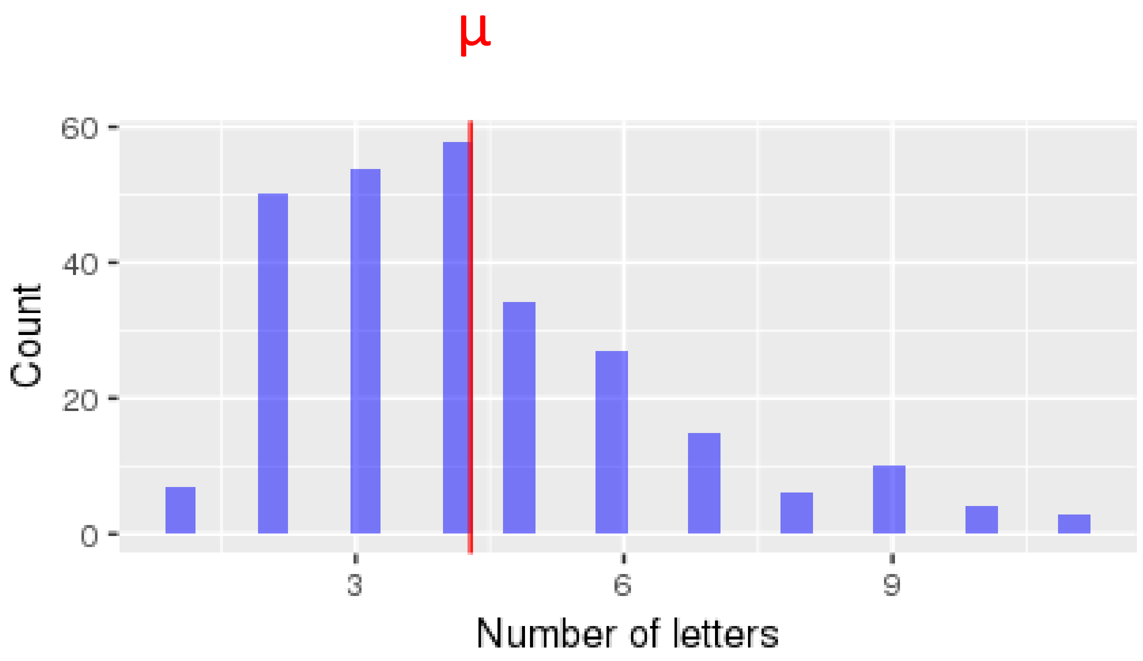
$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

Sampling distribution!

Gettysburg address word length sampling distribution



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

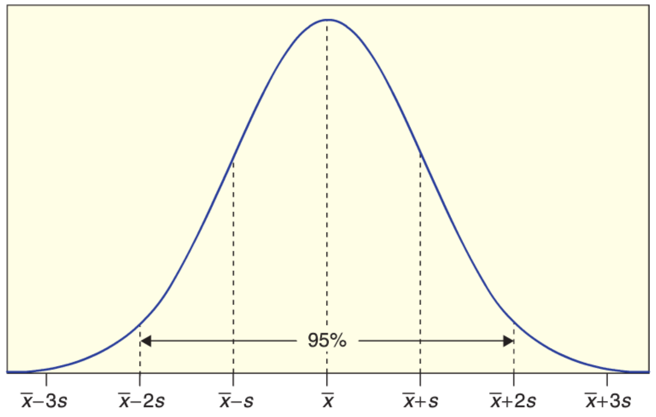
$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$



Sampling distribution!

Gettysburg sampling distribution app