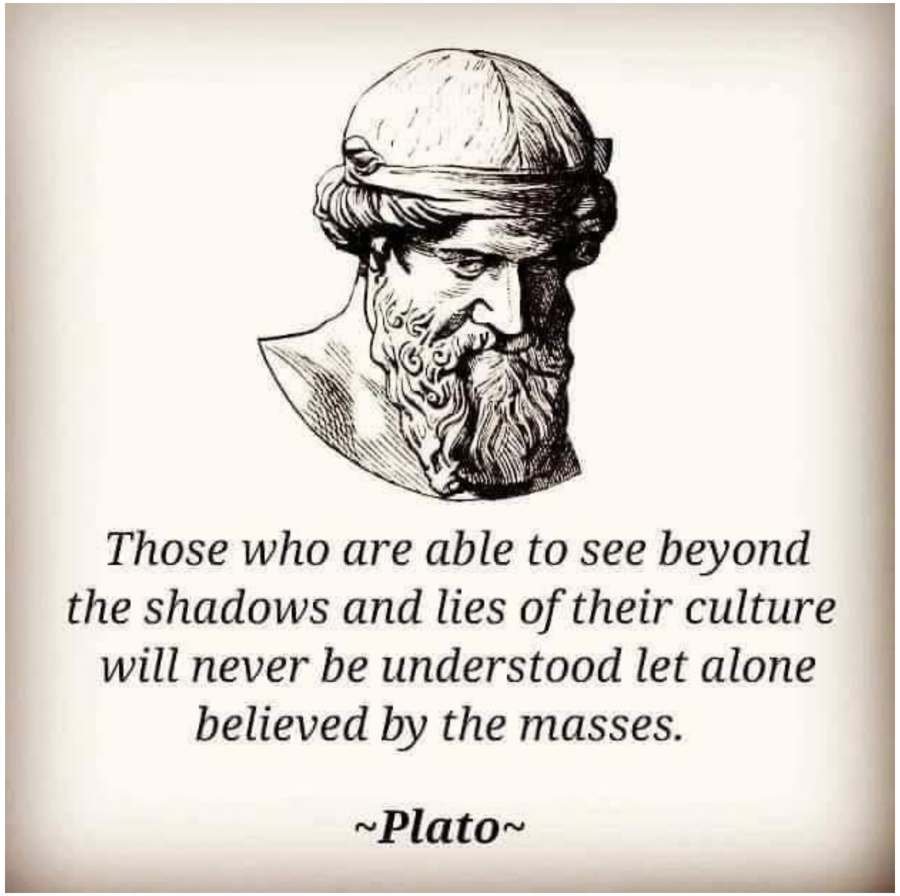
# Review!



*Those who are able to see beyond the shadows and lies of their culture will never be understood let alone believed by the masses.*

***~Plato~***

# Announcement: Midterm exam

Exam is during regular class time

- Exam is on paper

If you have accommodations, please schedule the exam with SAS

A practice exam (last year's exam) has been posted

# Midterm exam “cheat sheet”

You are allowed an exam “cheat sheet”

One page, double sided, that contains **only code and equations**

- No code comments allowed

Cheat sheet must be on a regular 8.5 x 11 piece of paper

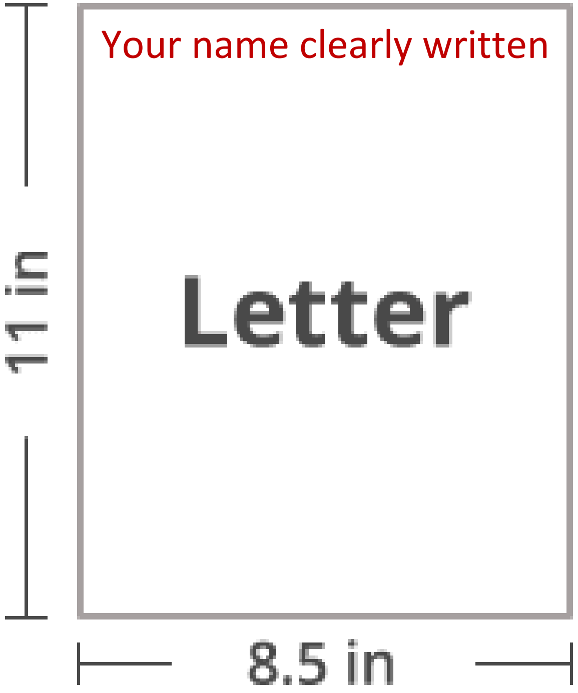
- Your name on the upper left of both sides of the paper

Recommend making a typed  list of all functions discussed in class and on the homework

- This will be useful beyond the exam

You must turn in your cheat sheet with the exam

- Failure to do so will result in a 20 point deduction



# Note about the review material

This review is not comprehensive of everything we have discussed so it is likely that there will be material on the exam that is not covered in this review material

Please review all class material (slides, class code, homework, etc.) to be well prepared for the exam!

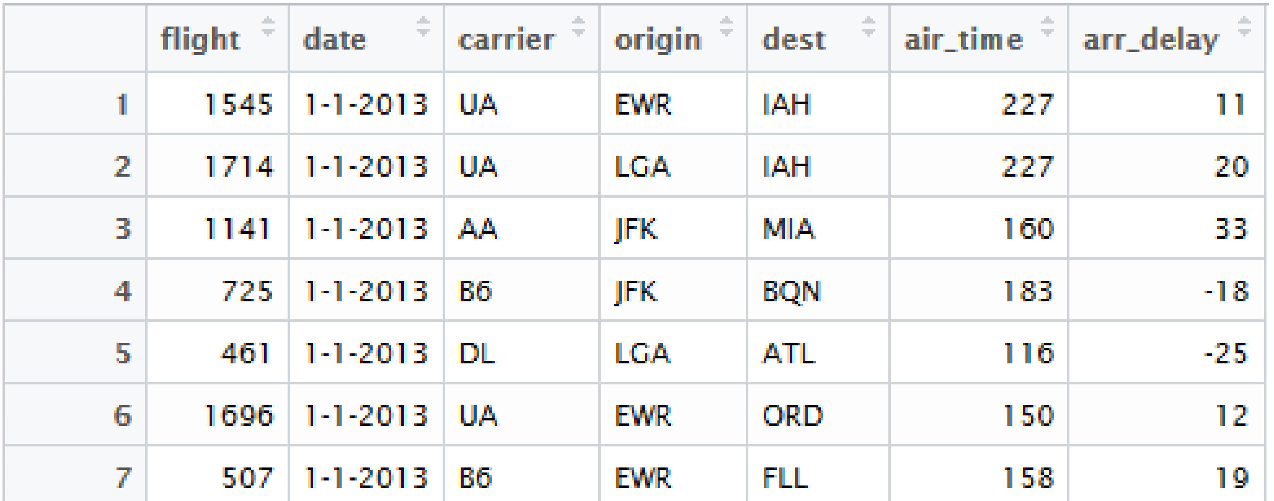
# Review of descriptive statistics

# 1. Intro to data

What is Statistics?

What are…

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



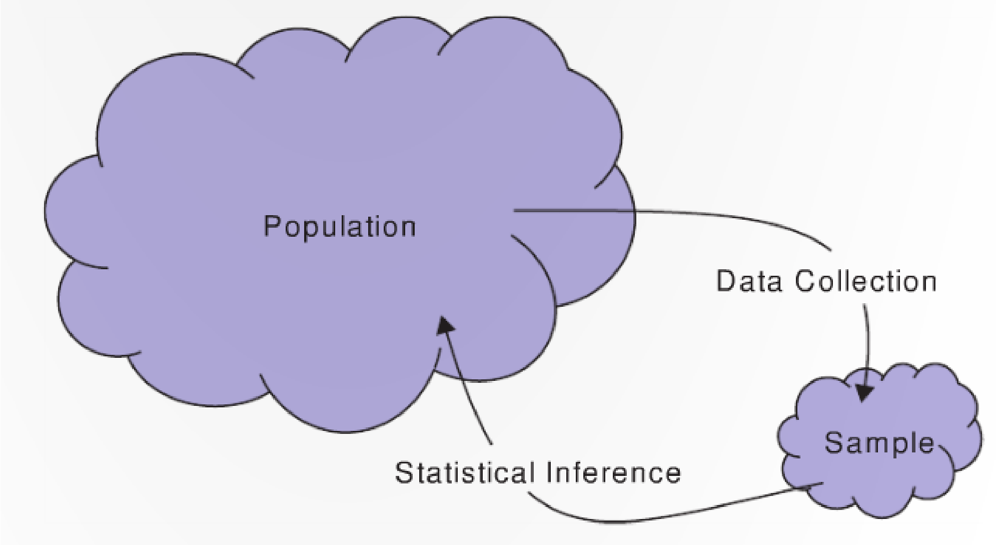
| | flight | date | carrier | origin | dest | air_time | arr_delay |
|---|---|---|---|---|---|---|---|
| 1 | 1545 | 1-1-2013 | UA | EWR | IAH | 227 | 11 |
| 2 | 1714 | 1-1-2013 | UA | LGA | IAH | 227 | 20 |
| 3 | 1141 | 1-1-2013 | AA | JFK | MIA | 160 | 33 |
| 4 | 725 | 1-1-2013 | B6 | JFK | BQN | 183 | -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 |

# 2. Sampling

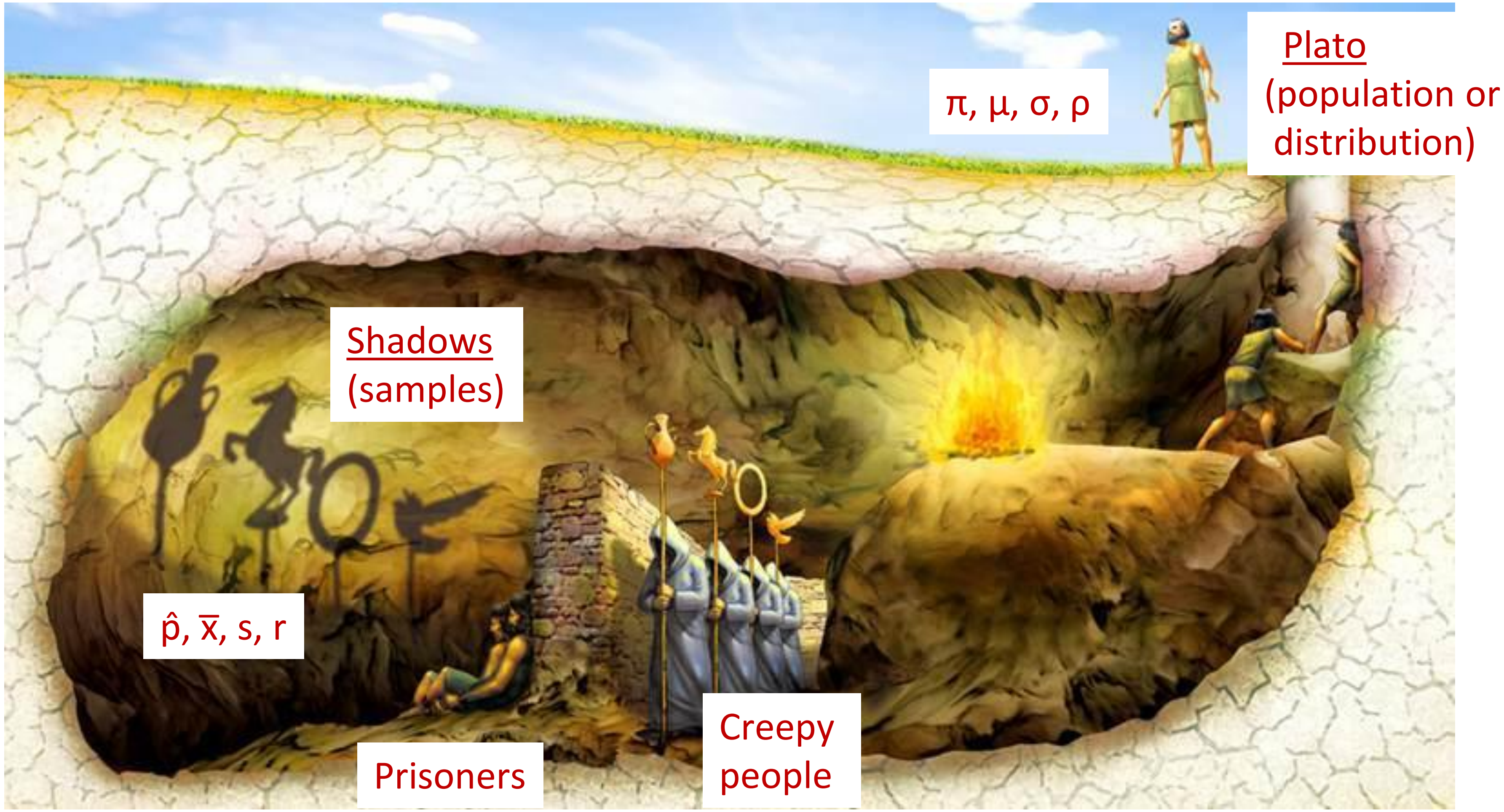
What is a …?

- sample
- population
- statistic
- parameter

What is statistical inference?

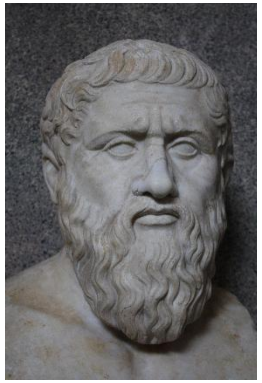


# Plato’s cave

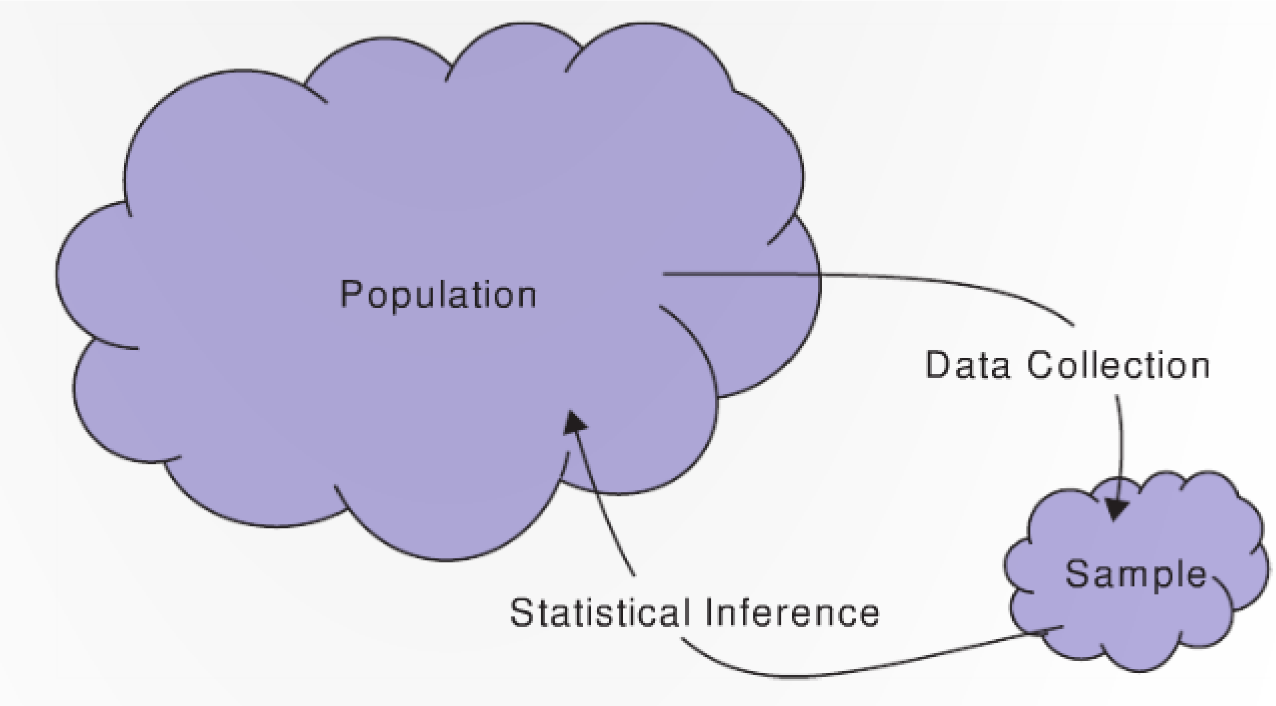


From The Republic (~ 380 BCE)

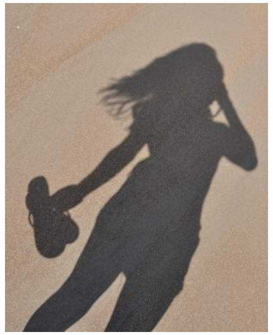
# Population parameters vs. sample statistics



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



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

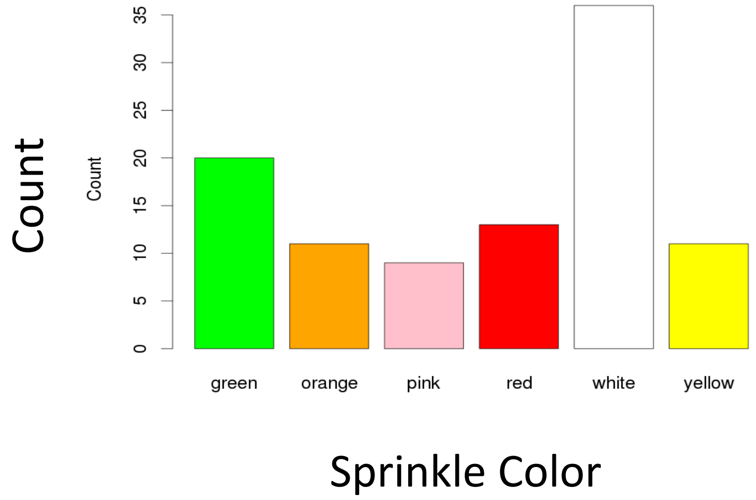


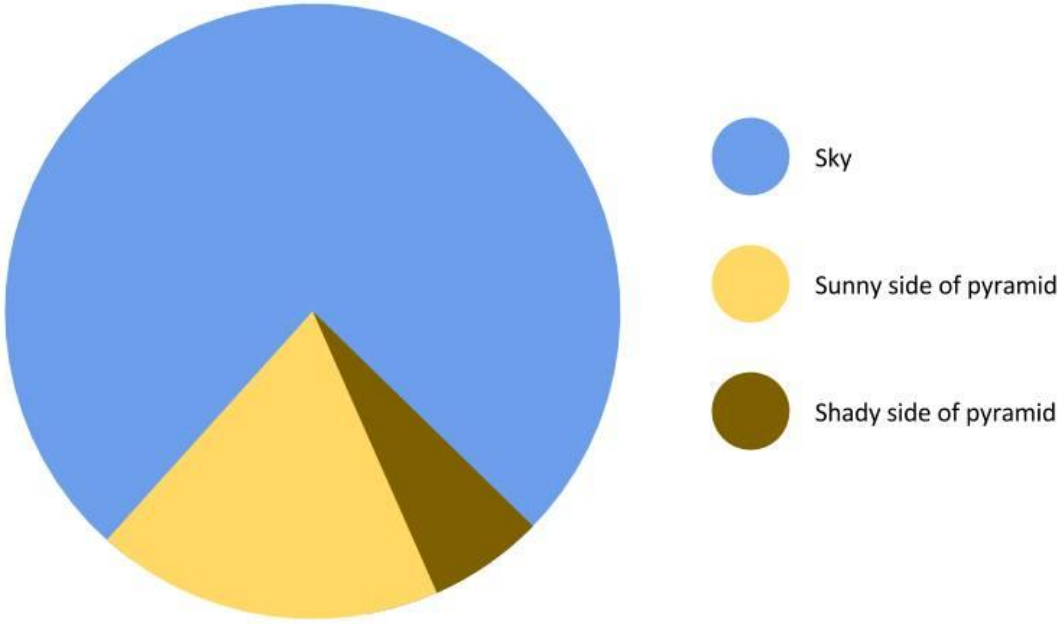
# Categorical data

What is the main statistic we discussed for categorical data?

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

How can we plot categorical data?

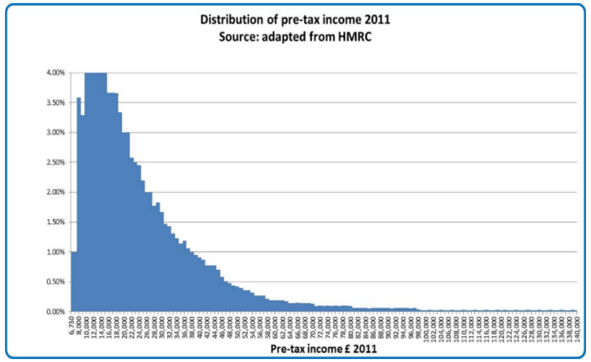


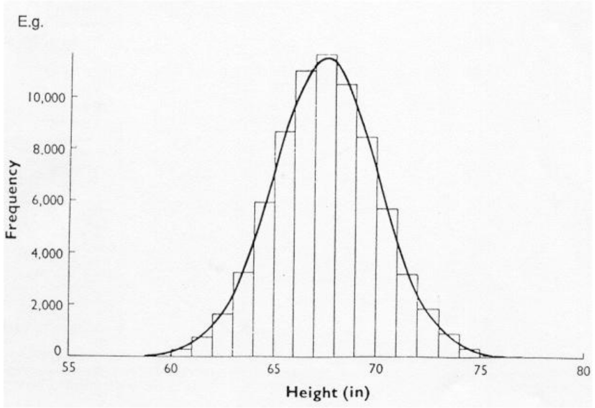


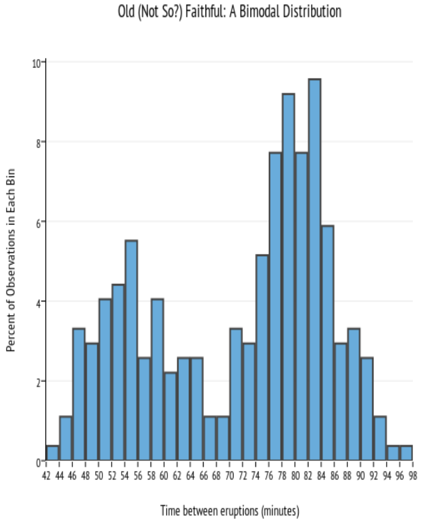
# Quantitative data?

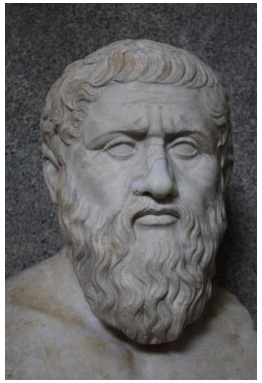
## What is a good way to visualize the shape of quantitative data?

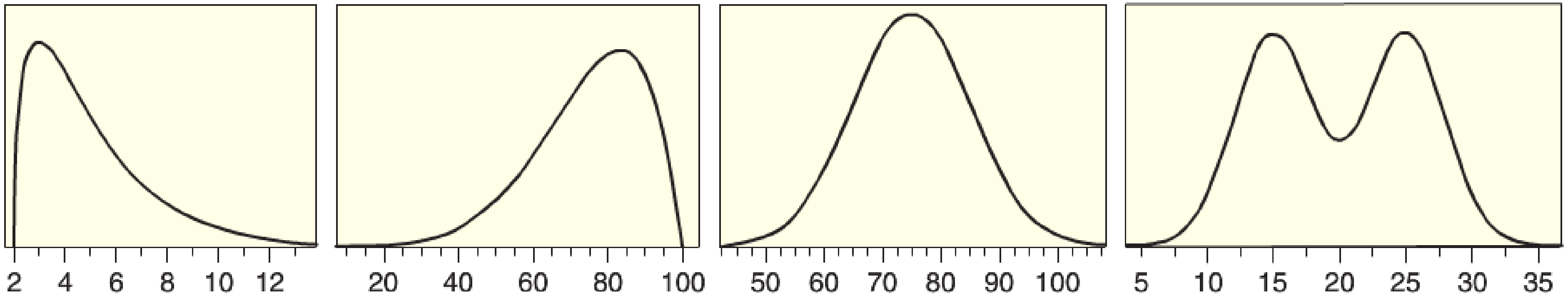
Income distribution



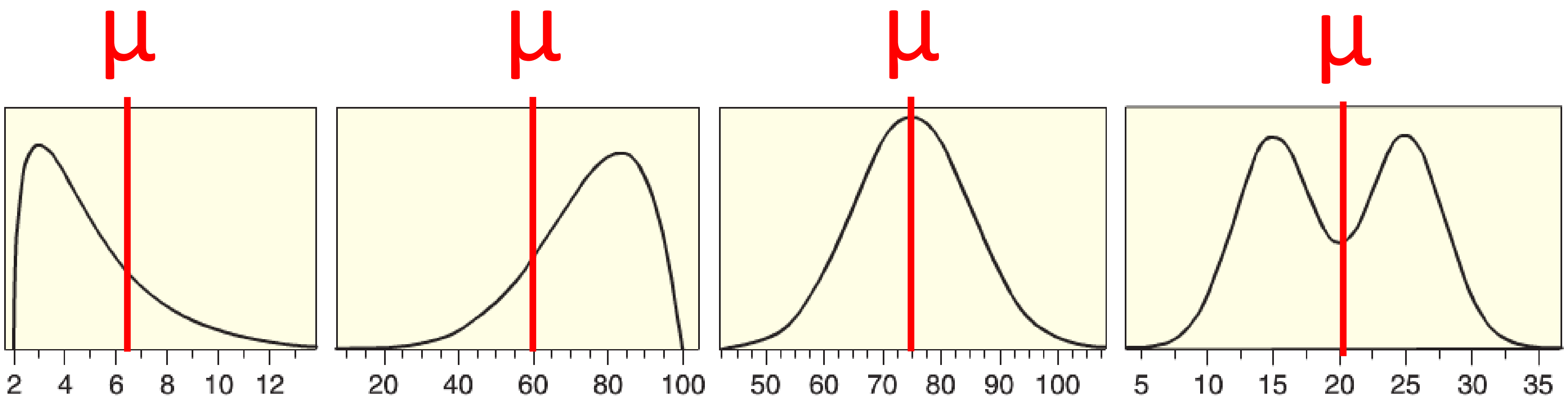






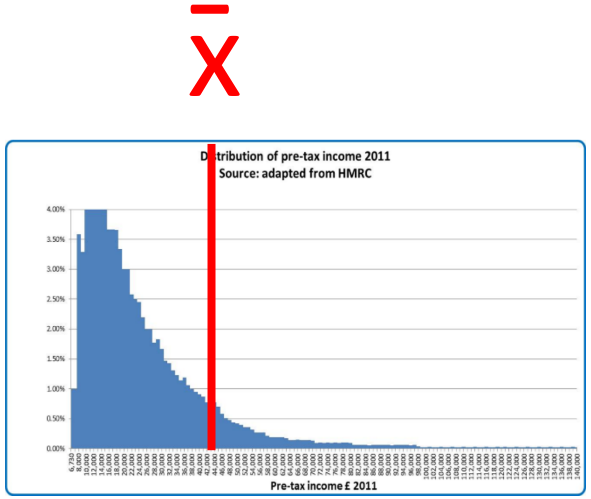


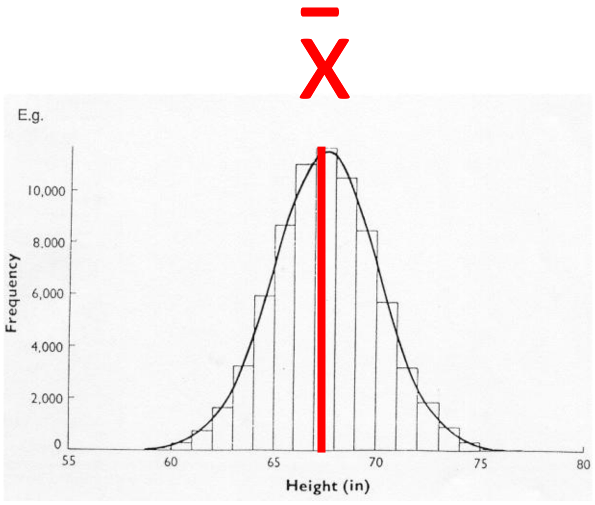
# Measure of central tendency: the mean

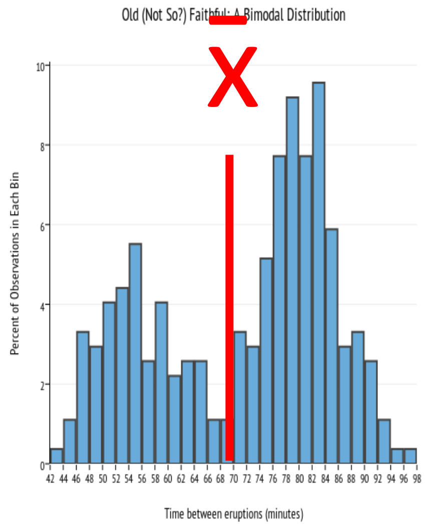


(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

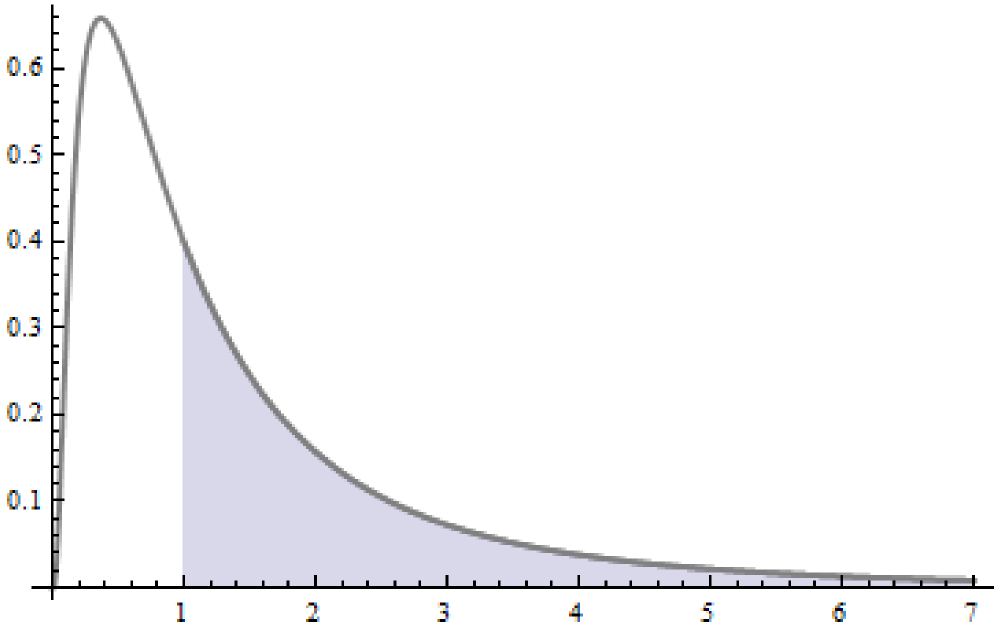
$$\frac{\Sigma_i^n x_i}{n}$$

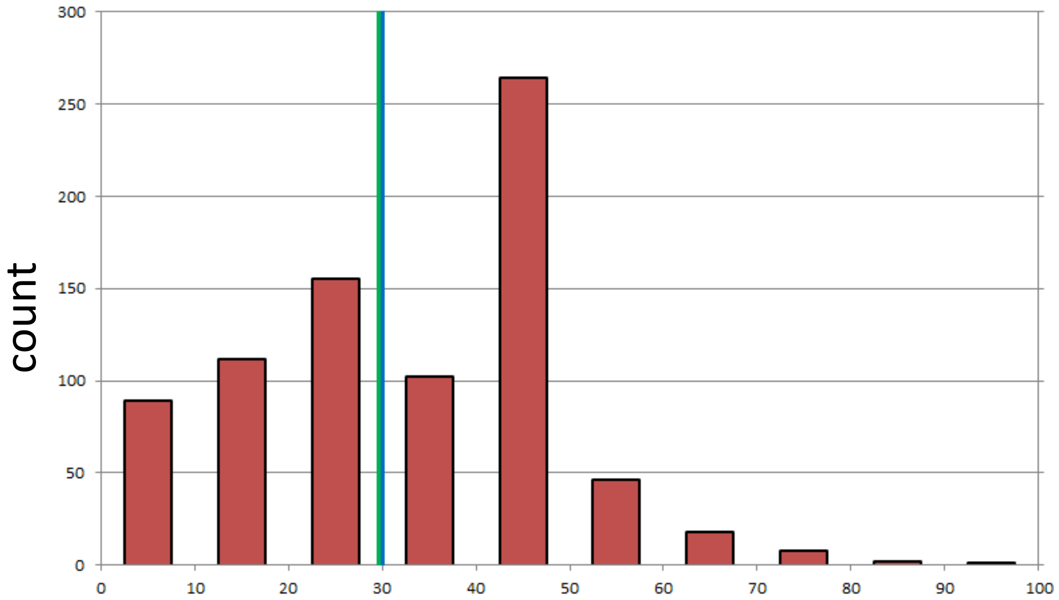






# Measure of central tendency: the median



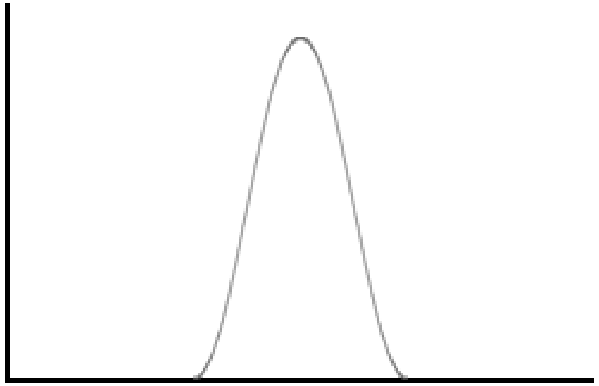


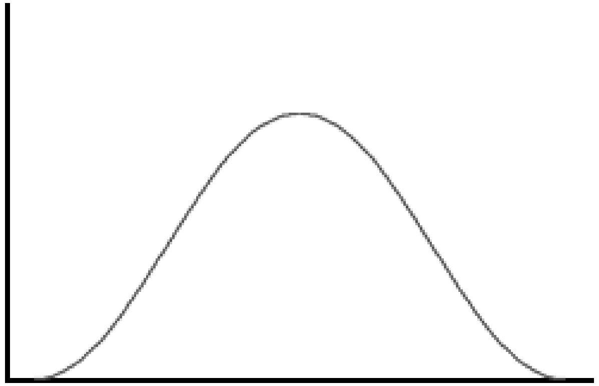
Which is resistant, the mean or the median?

# The standard deviation

## Which distribution has a larger standard deviation?

parameter σ

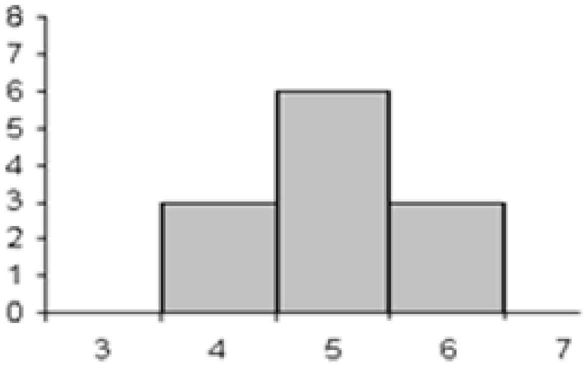


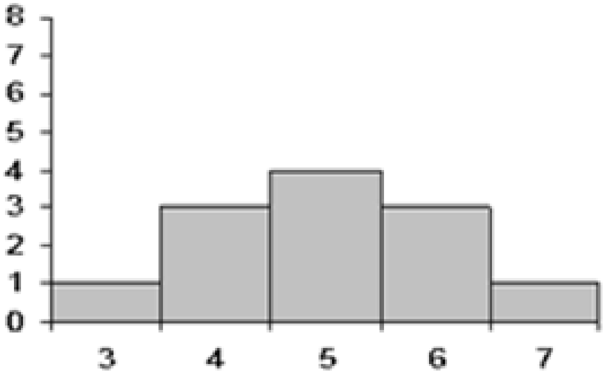


# The standard deviation

Which distribution has a larger standard deviation?

statistic: s





What is the formula for the standard deviation?

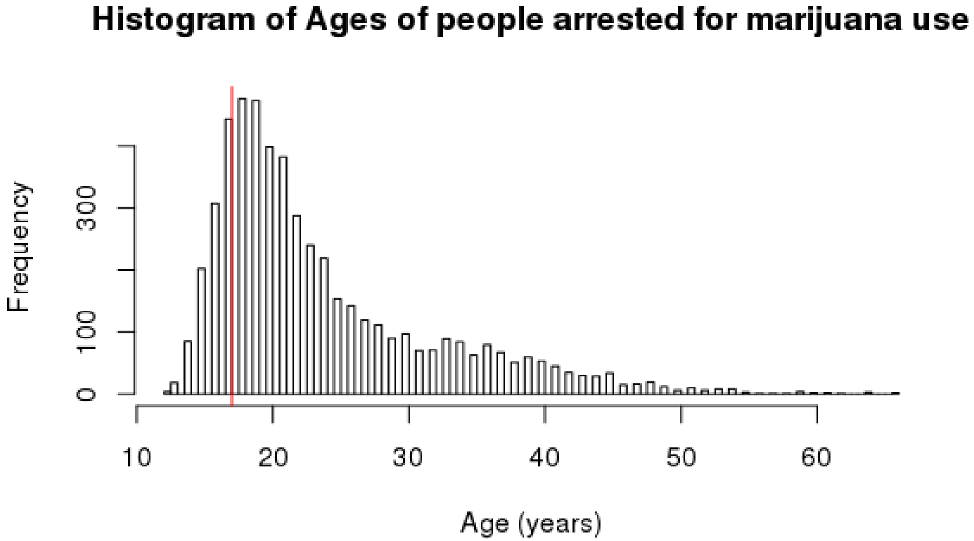
$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

# z-scores and percentiles

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

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

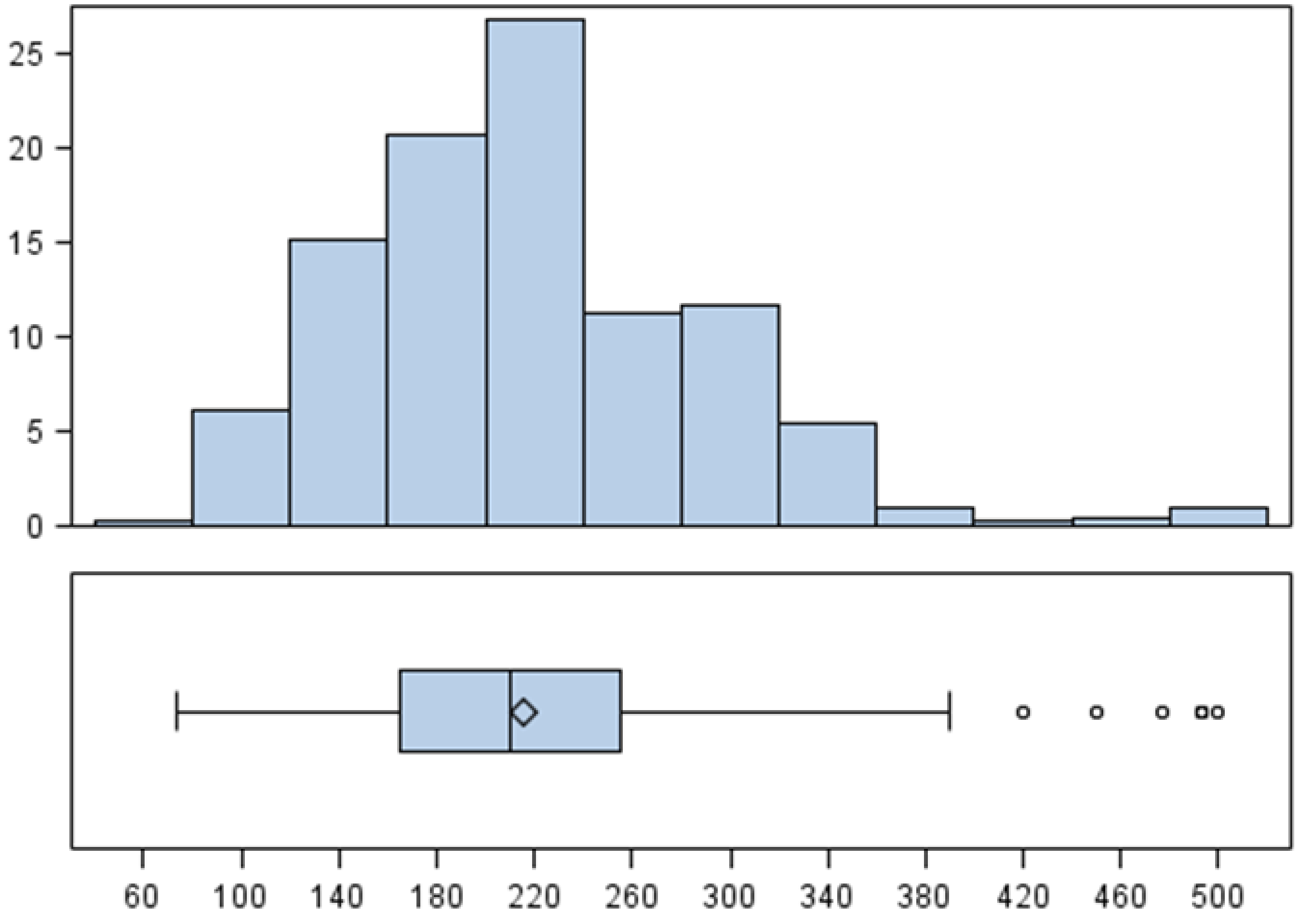
What is the $p^{th}$ percentile?



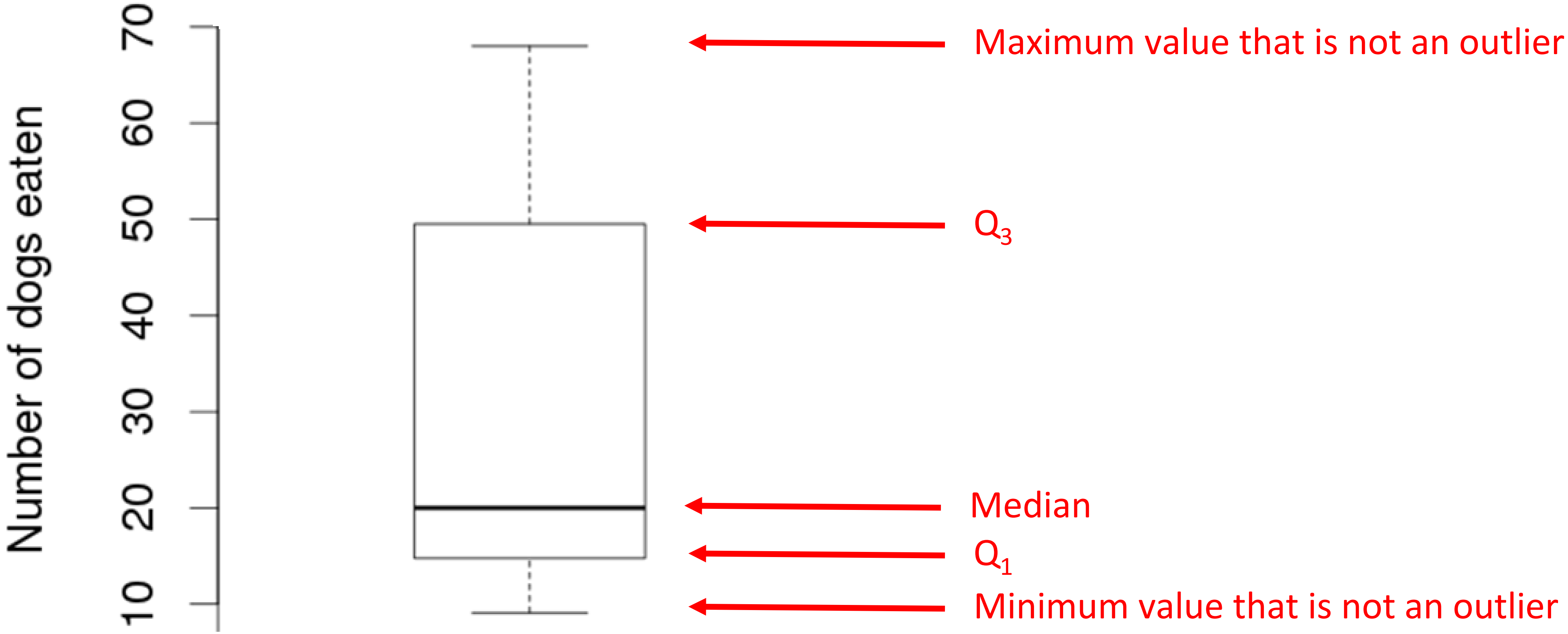
# Normal pillow

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

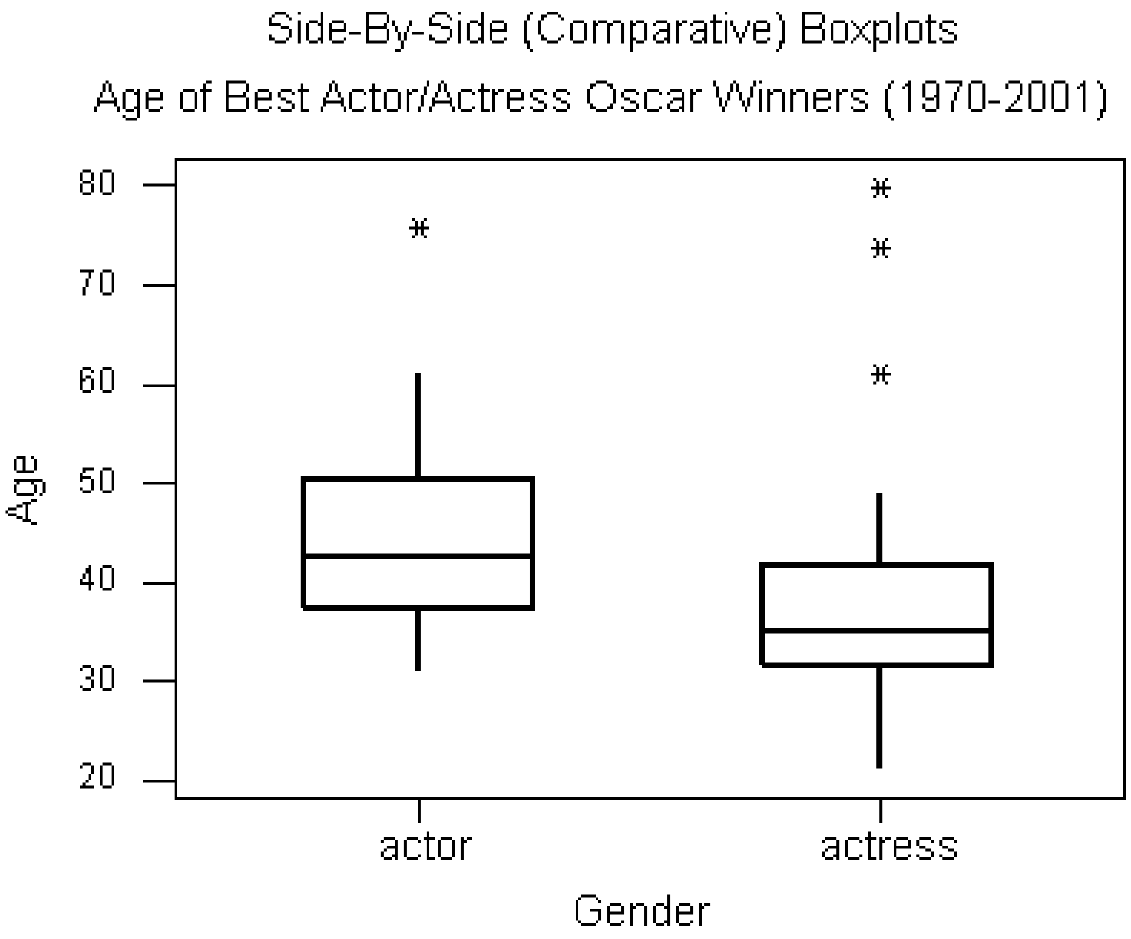
# What is a 5 number summary and a boxplot?



# What is a 5 number summary and a boxplot?

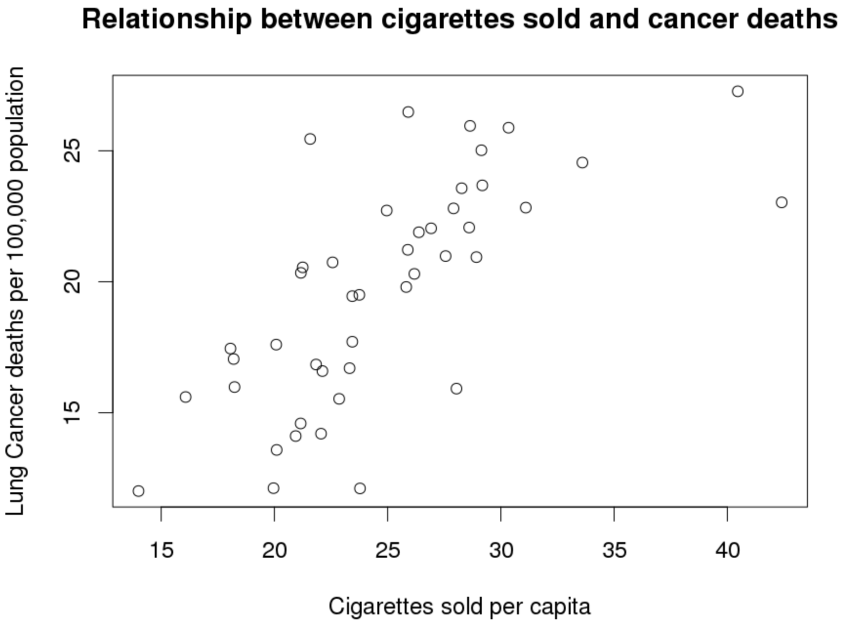


# Side-by-side boxplots



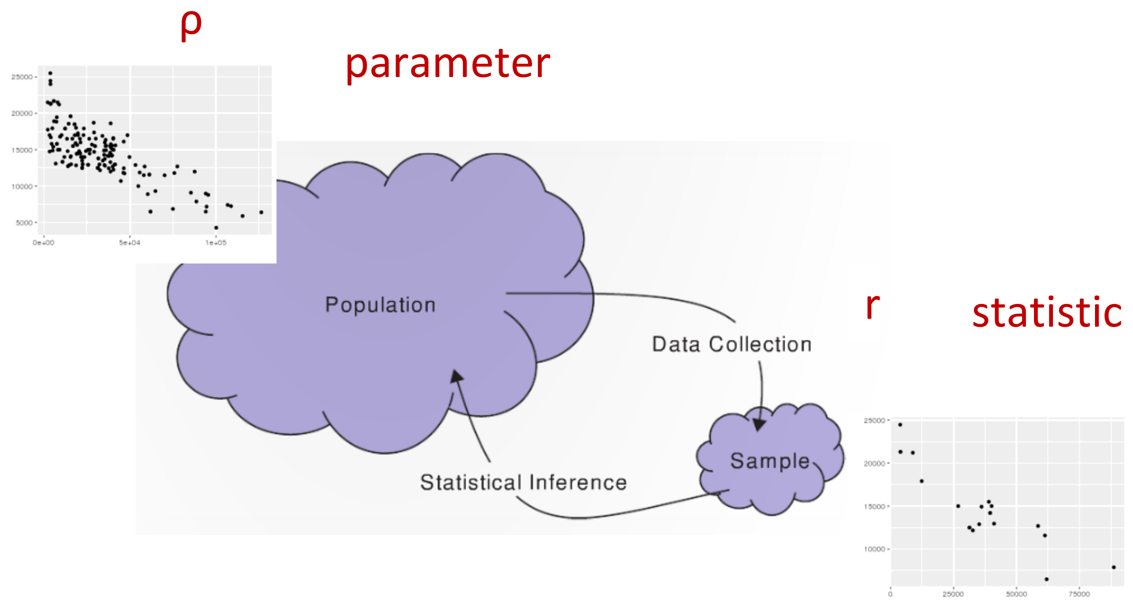
# Relationships between measures

Q: What is this type of plot called?



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

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



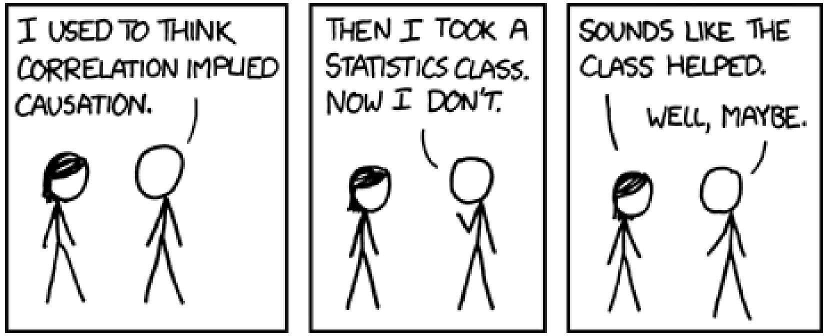
ρ

parameter

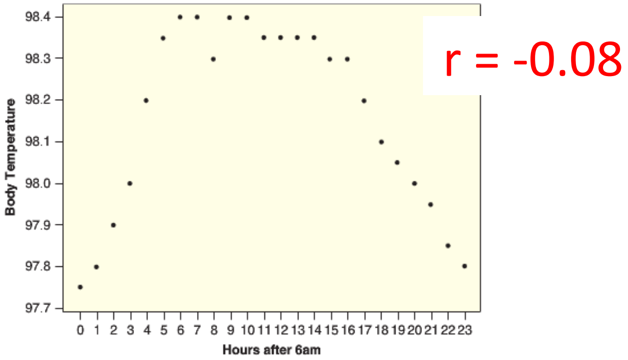
r

statistic

# Correlation cautions

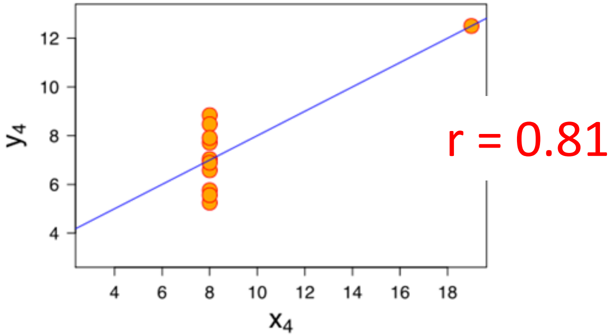

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

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



r = -0.08

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 = 0.81

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

# Review: Cancer smoking regression line

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

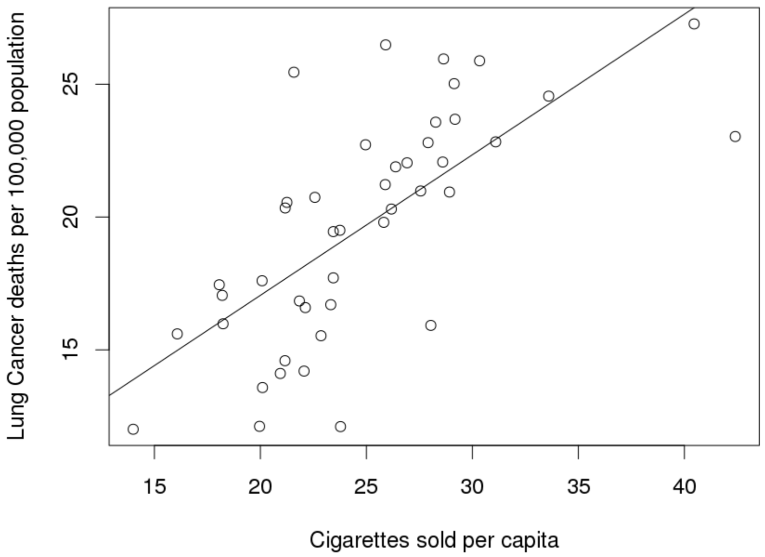
- i.e., $\hat{y} = f(x)$

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

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

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

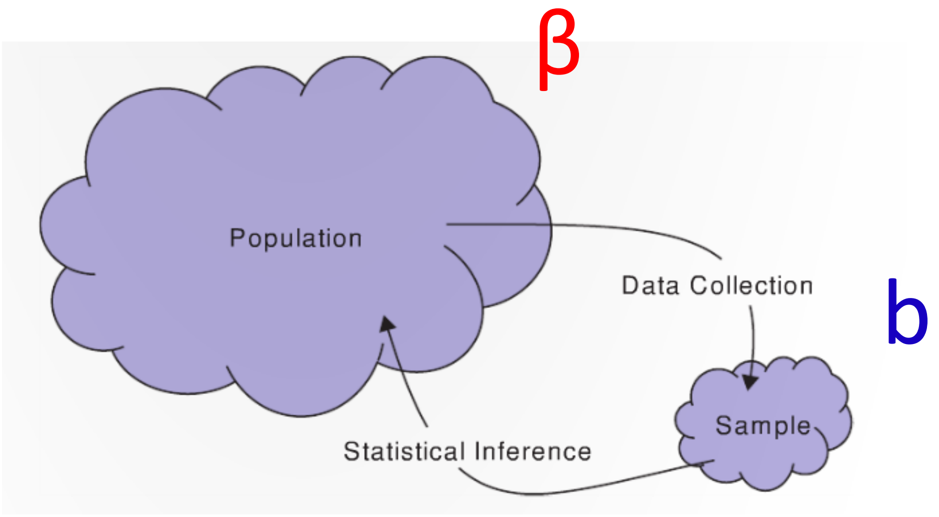
Relationship between cigarettes sold and cancer deaths



$a = 6.47 \quad b = 0.0053$

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

# Review: Notation

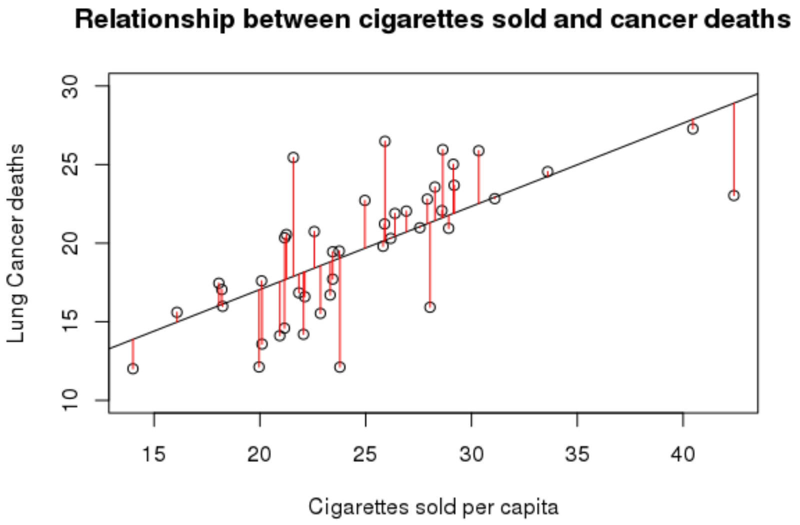


The **residual** is the difference between an observed ($y_i$) and a predicted value ($\hat{y}_i$) of the response variable

$$
\begin{aligned}
\mathit{Residual}_i &= \mathit{Observed}_i - \mathit{Predicted}_i \\
&= y_i - \hat{y}_i
\end{aligned}
$$

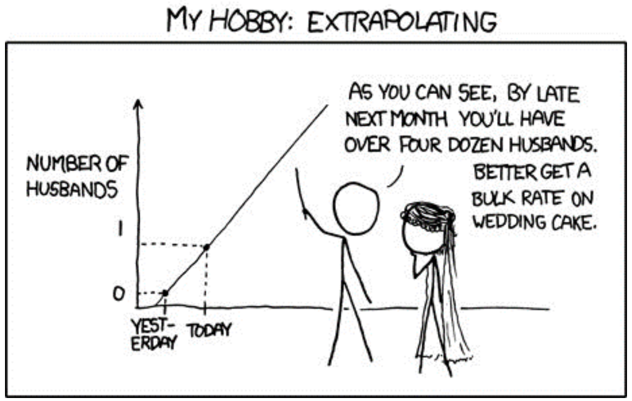
The **least squares line** minimizes the sum of squared residuals

- This is what lm(y ~ x) is doing

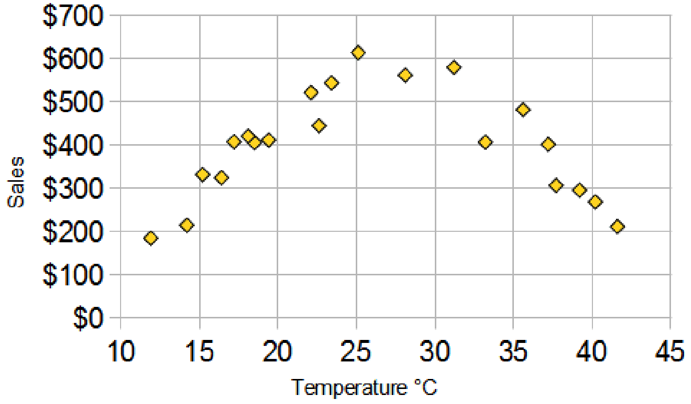


# Regression cautions

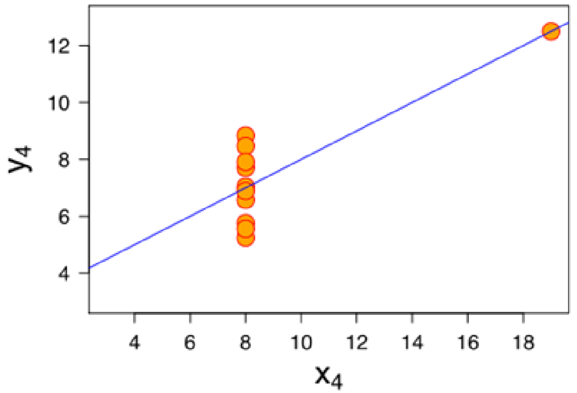
1. Avoid trying to apply the regression line to predict values far from those that were used to create the line.


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.

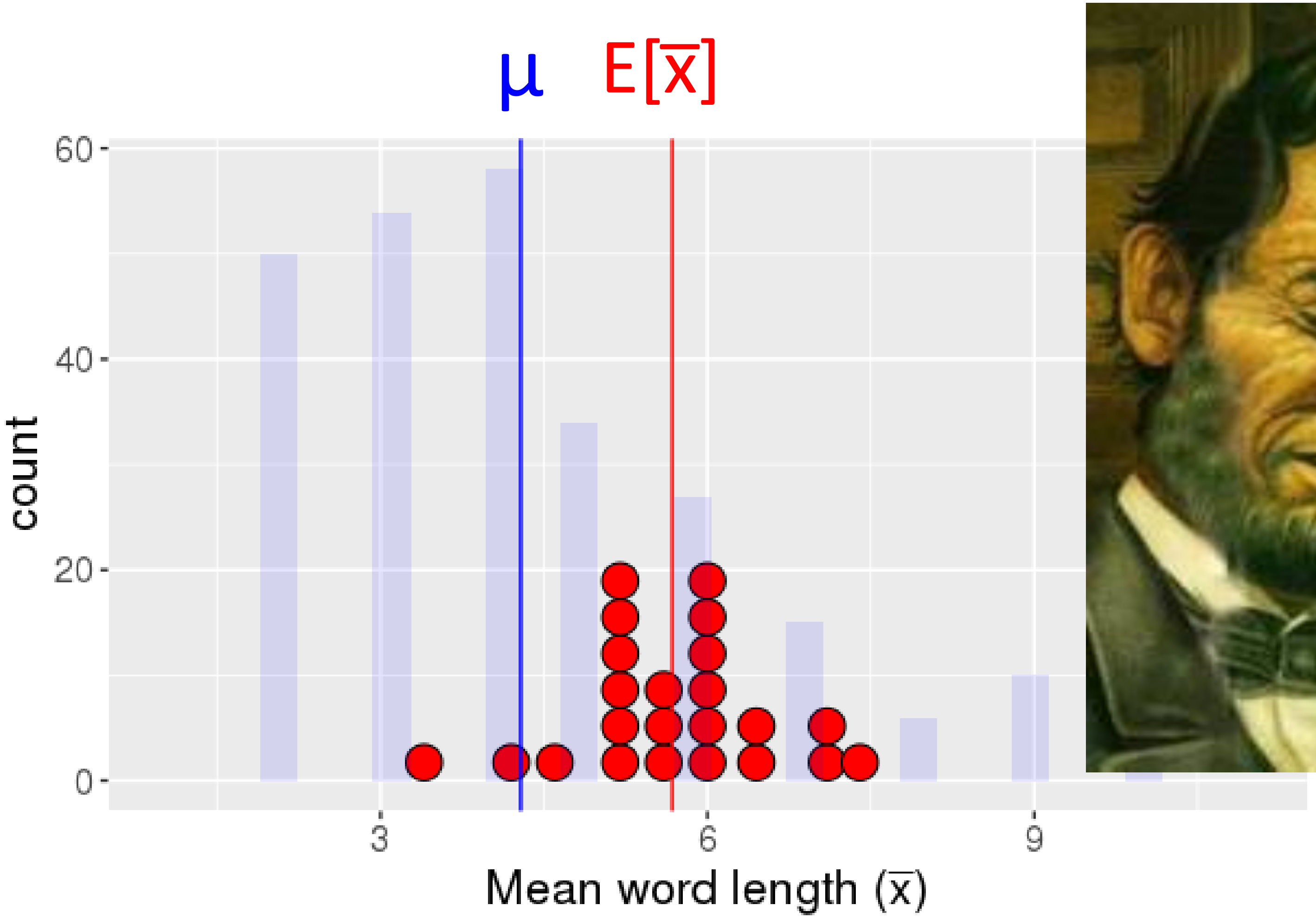
2. Plot the data! Regression lines are only appropriate when there is a linear trend in the data.

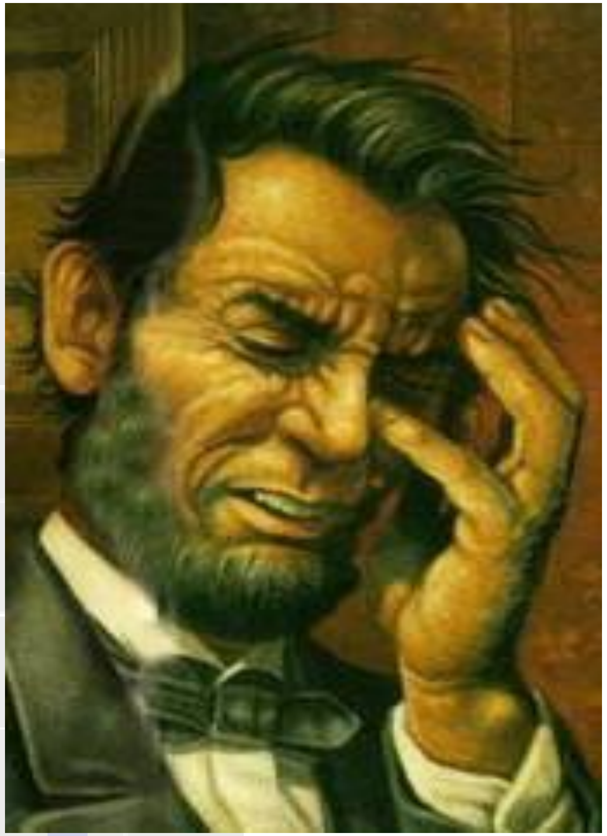


3. Be aware of outliers – they can have an huge effect on the regression line.



# Bias and the Gettysburg address word length distribution





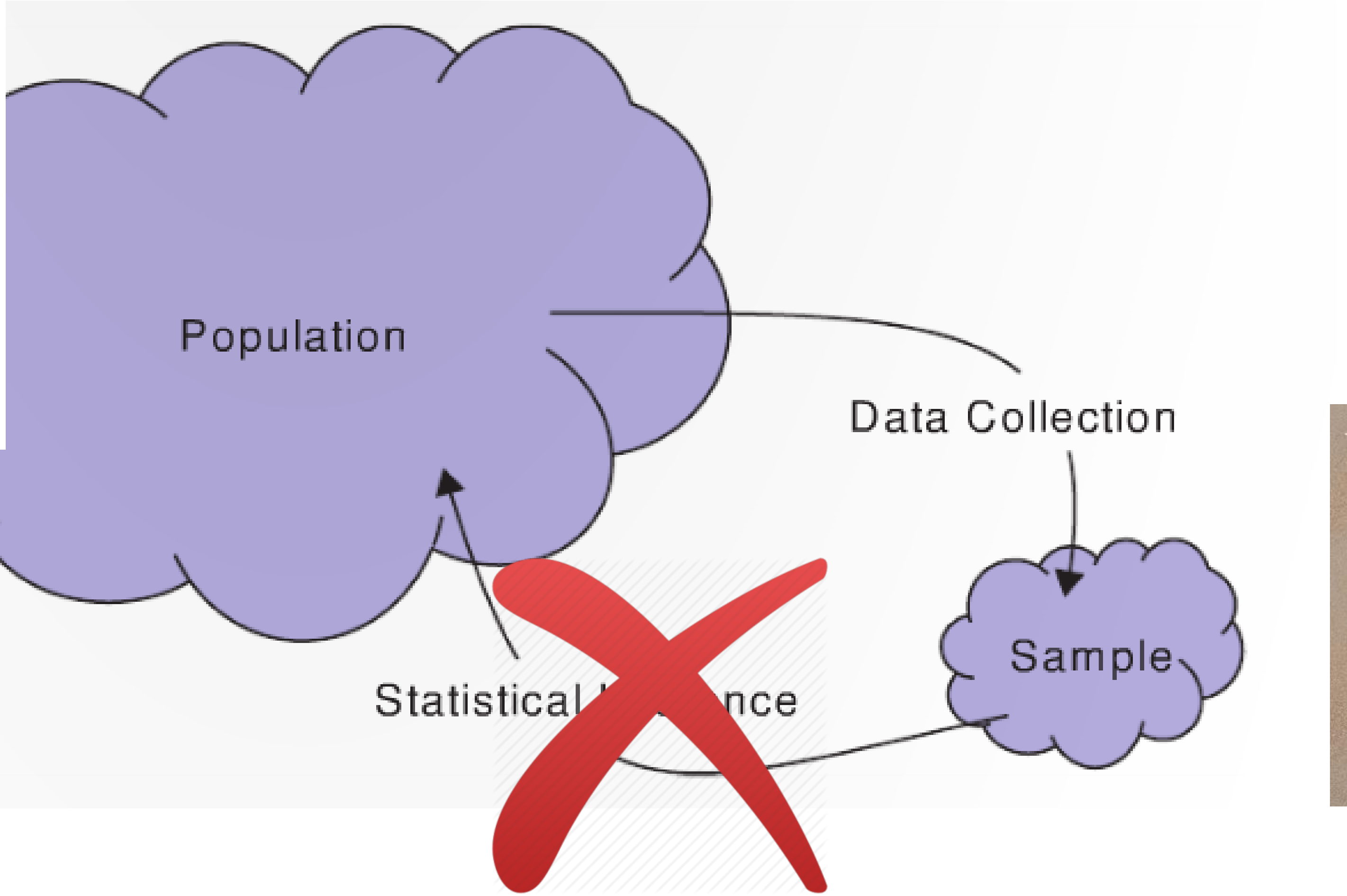
**Bias** is when our average statistic does not equal the population parameter

Here:

$E[\bar{x}] \neq \mu$

# Statistical bias

$\mu$



$\bar{x}$

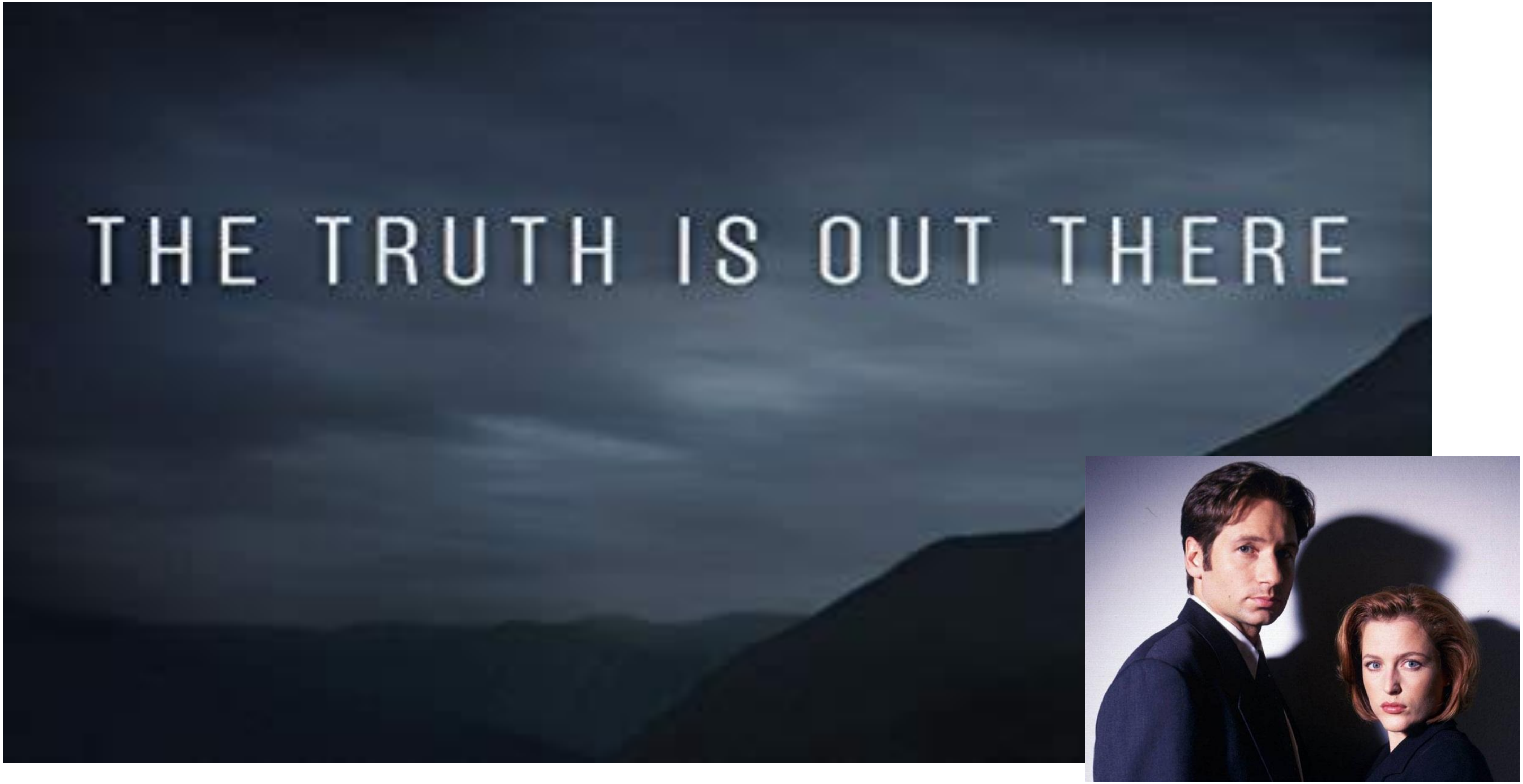
# To prevent bias: use simple random sample!

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

Allows for generalizations to the population!

Soup analogy!

# What is our primary focus in Statistics?

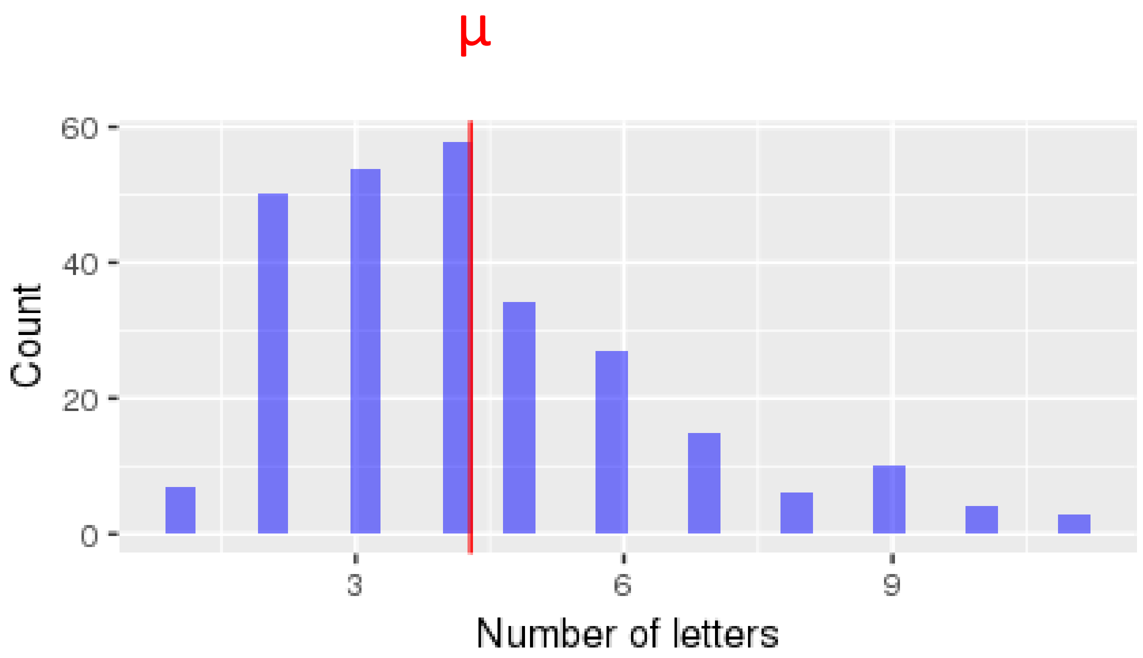


# Sampling distribution

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

# Gettysburg address word length sampling distribution



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

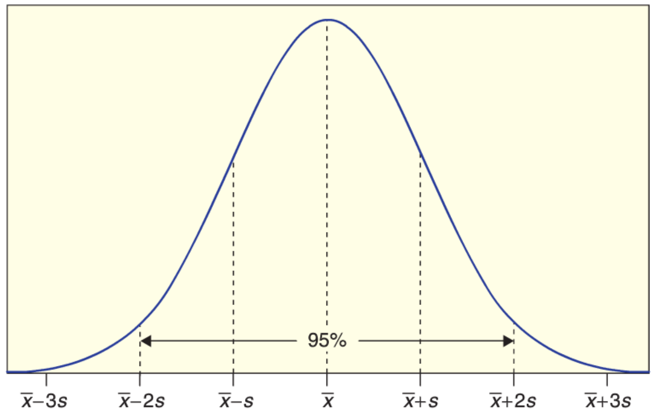
$\bar{x} = 5$

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

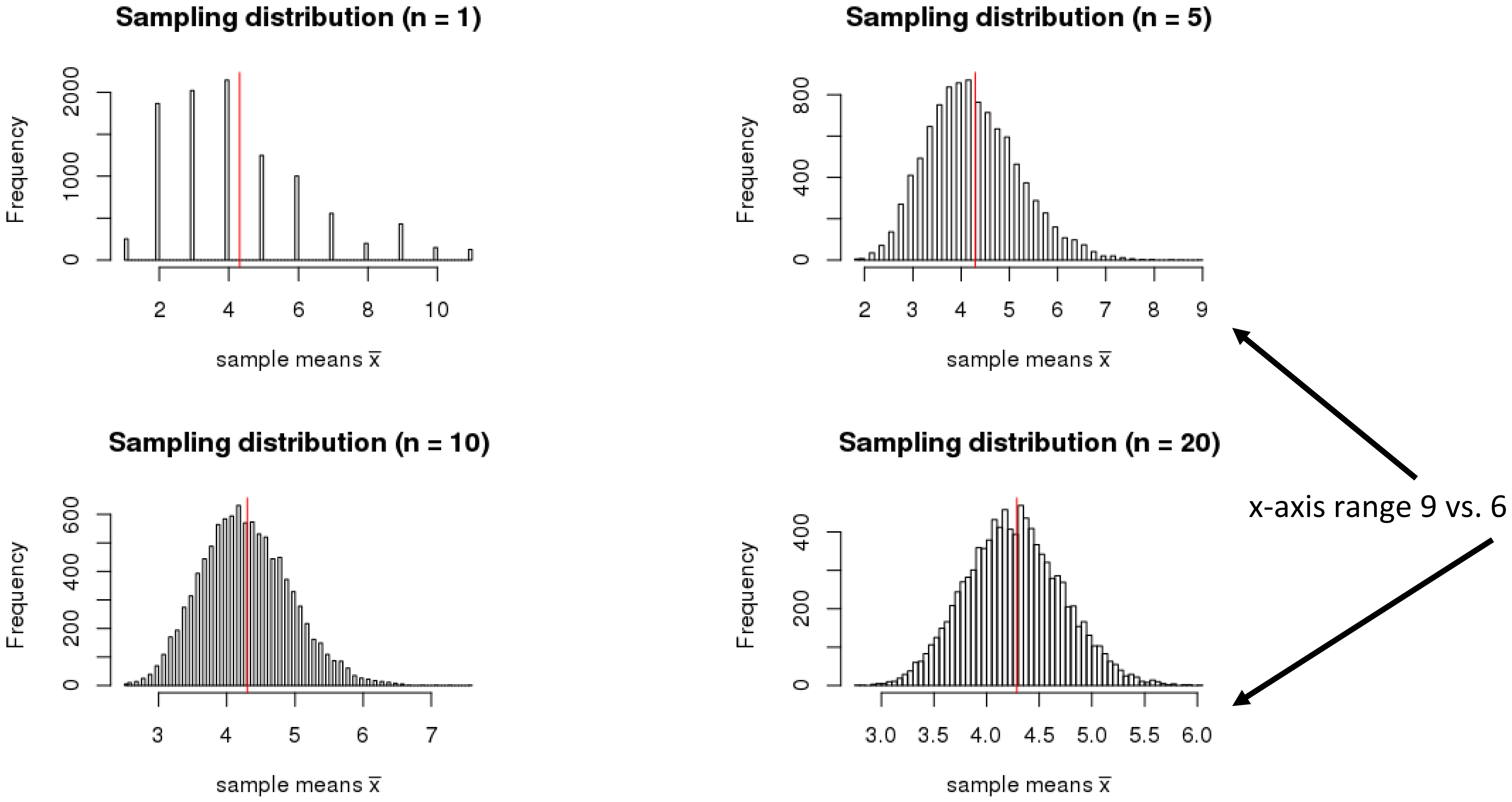
# Creating a sampling distribution in R

```
sampling_dist <- do_it(10000)  *  {

        curr_sample <- sample(word_lengths, 10)
        mean(curr_sample)

}

hist(sampling_dist)
```



Sampling distribution (n = 1)

Sampling distribution (n = 5)

Sampling distribution (n = 10)

Sampling distribution (n = 20)

x-axis range 9 vs. 6

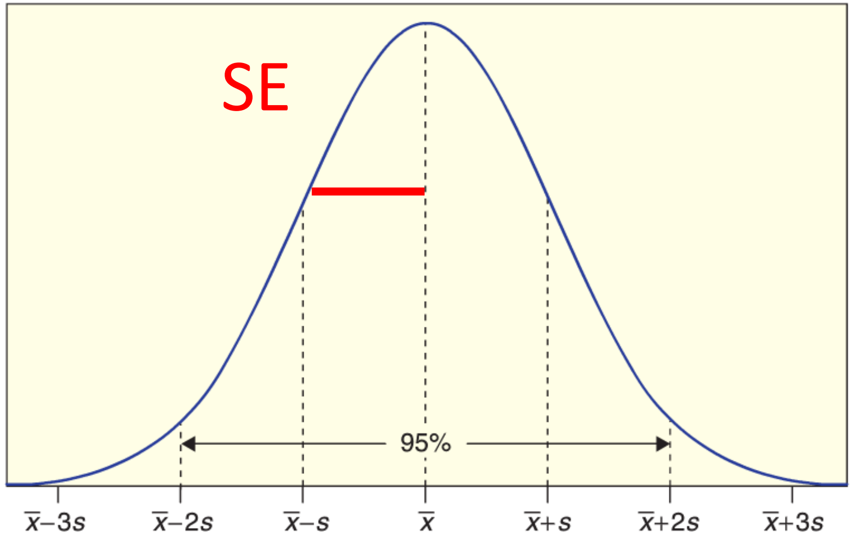
As the sample size n increases

1. The sampling distribution becomes more like a normal distribution
2. The sampling distribution points ($\bar{x}$'s) become more concentrated around the mean $E[\bar{x}] = \mu$

# The standard error

The **standard error** of a statistic, denoted SE, is the standard deviation of the sample statistic

- i.e., SE is the standard deviation of the *sampling distribution*



# Interval estimate based on a margin of error

We use the statistics from a sample as a **point estimate** for a population parameter

An **interval estimate** give a range of plausible values for a population parameter.

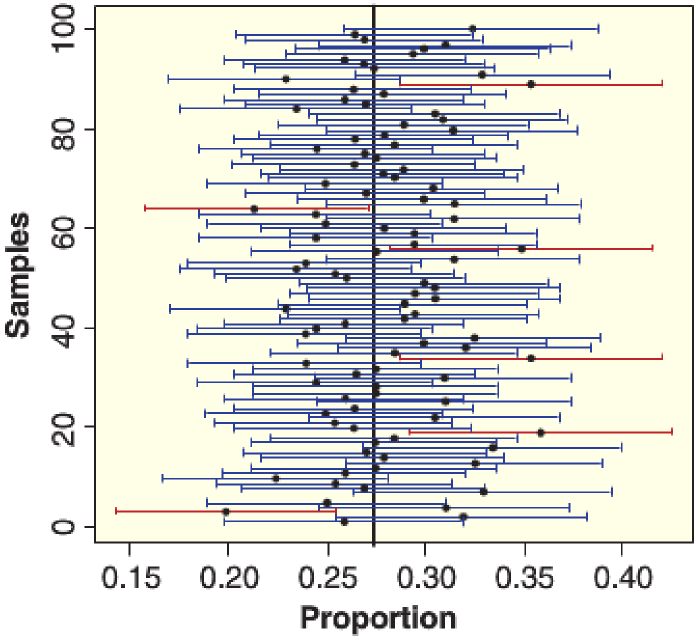
One common form of an interval estimate is:

*Point estimate ± margin of error*

Where the **margin of error** is a number that reflects the precision of the sample statistic as a point estimate for this parameter

# Confidence Intervals

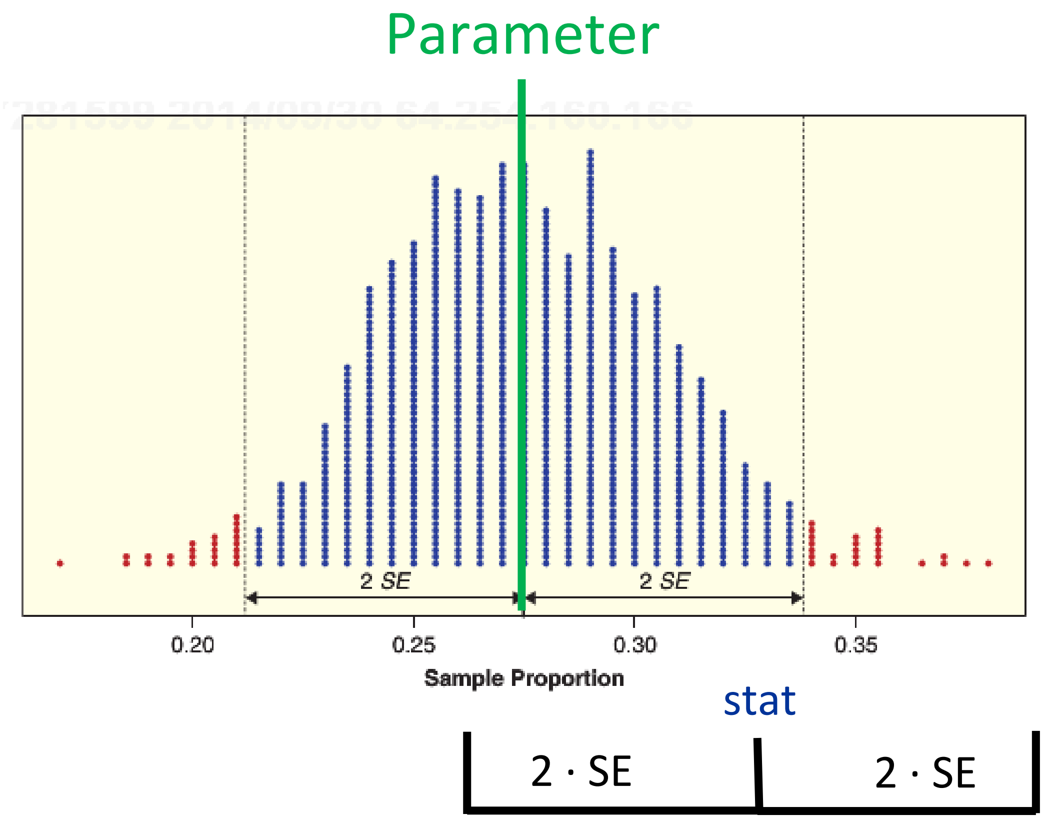
A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times



The **confidence level** is the percent of all intervals that contain the parameter

# Sampling distributions

For a sampling distribution that is a normal distribution, 95% of ***statistics*** lie within 2 standard deviations (SE) for the population mean



Confidence interval

Thus if we had:

- A statistics value
- The SE

We could compute a 95% confidence interval!

$$CI_{95} = \text{stat} \pm 2 \cdot SE$$

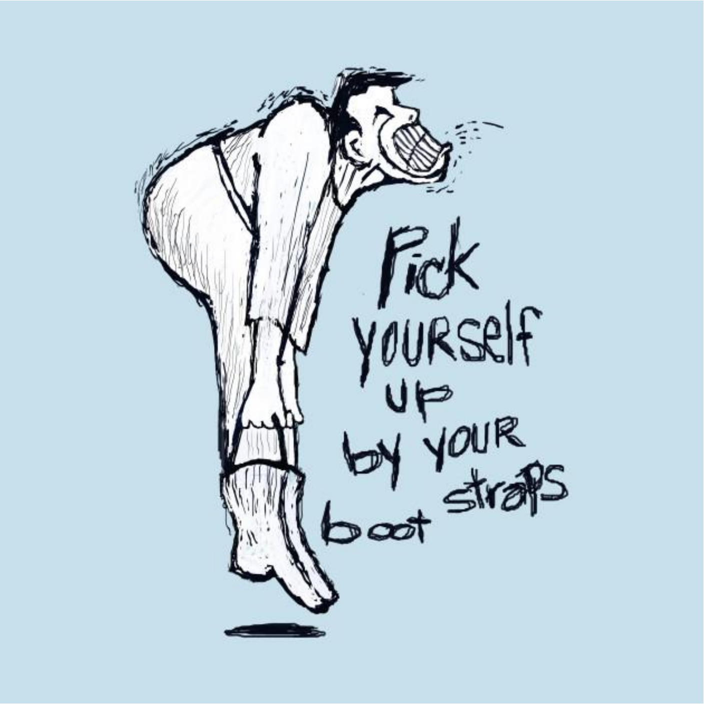
# Sampling distributions

Unfortunately we can't calculate the sampling distribution ☹

- Therefore we can't get the SE from the sampling distribution ☹

We have to pick ourselves up by the bootstraps!

1. Estimate SE with $\hat{SE}$
2. Then use $\overline{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI



# Plug-in principle

Suppose we get a sample from a population of size *n*

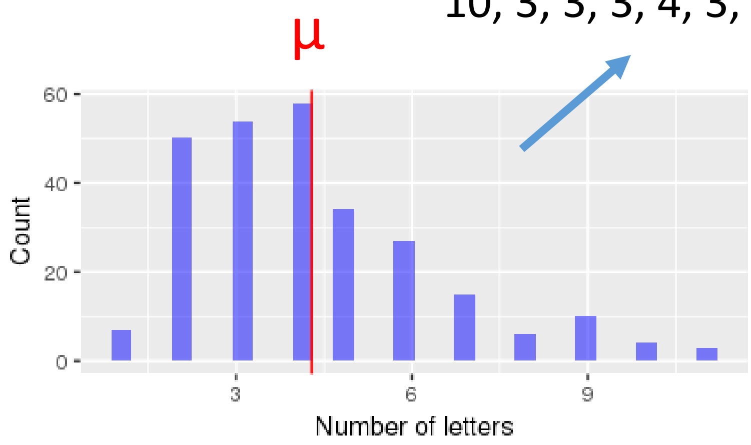
We pretend that *the sample is the population* (plug-in principle)

1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

2. We repeat this process 1000’s of times and get a ***bootstrap sample distribution***

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

# Bootstrap distribution illustration



**<u>The sample (n = 10)</u>**

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

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

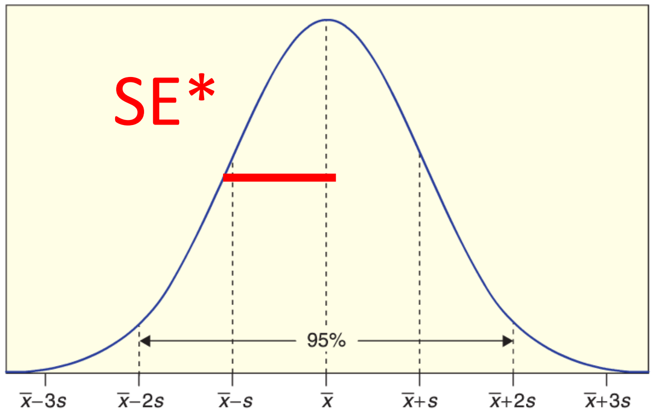
$\bar{x}^* = 4$

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

$\bar{x}^* = 4.1$

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

$\bar{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9’s in the bootstrap samples

# 95% Confidence Intervals

When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$\text{Statistic} \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

# Bootstrap distribution in R

```
my_sample <- c(21, 29, 25, 19, 24, 22, 25, 26, 25, 29)

bootstrap_dist <-  do_it(10000) * {

        curr_boot <- sample(my_sample , 10, replace = TRUE)
        mean(curr_boot)

}

SE_boot <- sd(bootstrap_dist)
```

# Bootstrap confidence interval in R

```
 obs_mean <- mean(my_sample)

CI_lower <-  obs_mean  - 2 * SE_boot

CI_upper <-  obs_mean  + 2 * SE_boot
```

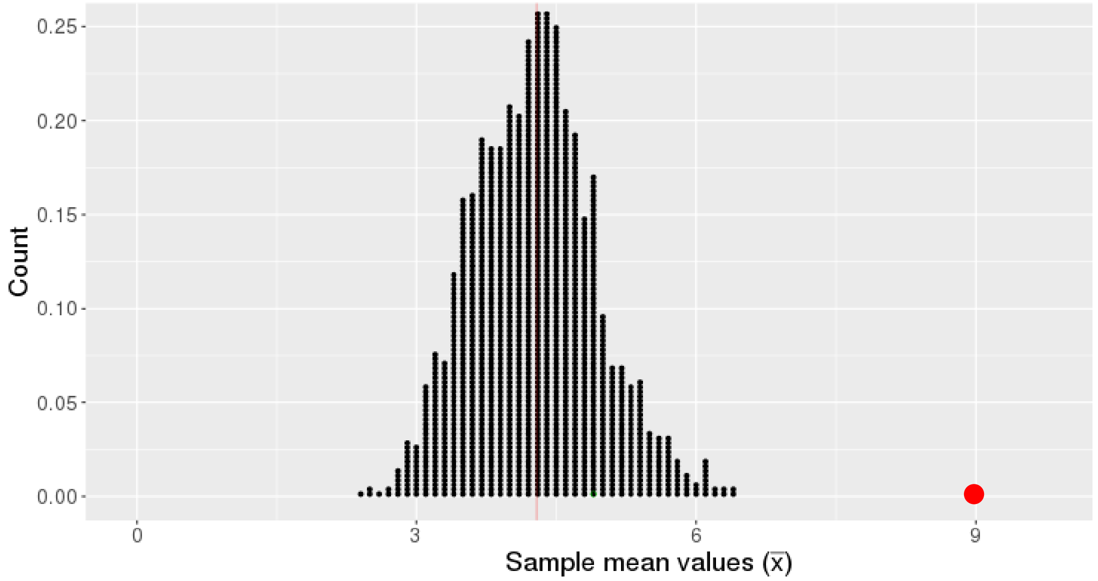
# Review of hypothesis tests

# Basic hypothesis test logic

We start with a claim about a population parameter

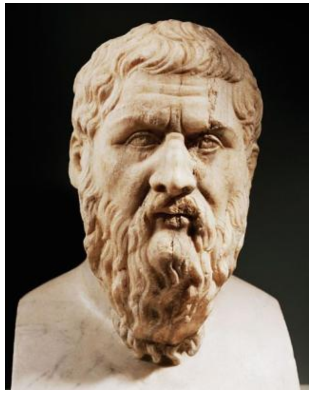
- E.g., μ = 4

This claim implies we should get a certain distribution of statistics



If our observed statistic is highly unlikely, we reject the claim

# Five steps of hypothesis testing

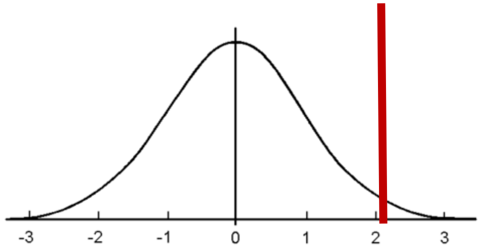


1. State $H_0$ and $H_A$
   - Assume Gorgias ($H_0$) was right

2. Calculate the actual observed statistic

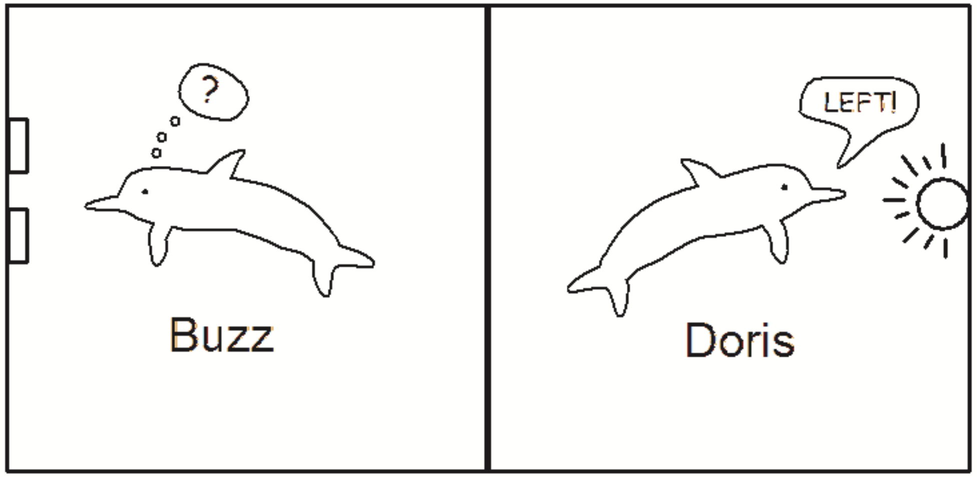
3. Create a **null distribution** of statistics that are consistent with $H_0$
   - i.e., a distribution of statistics that we would expect if Gorgias is right

4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value

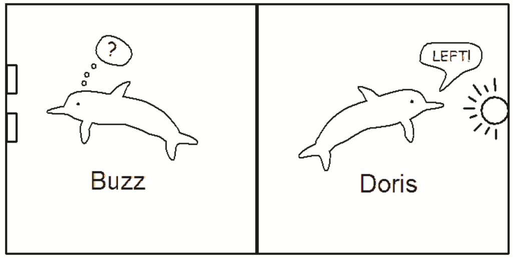


5. Make a judgement
   - Assess whether the results are statistically significant

# Hypothesis tests for a single proportion



# Hypothesis tests for a single proportion



1. State the null hypothesis… and the alternative hypothesis
   - Buzz is just guessing so the results are due to chance: $H_0: \pi = 0.5$
   - Buzz is getting more correct results than expected by chance: $H_A: \pi > 0.5$

2. Calculate the observed statistic
   - Buzz got 15 out of 16 guesses correct, or $\hat{p} = .973$

3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Buzz was just guessing

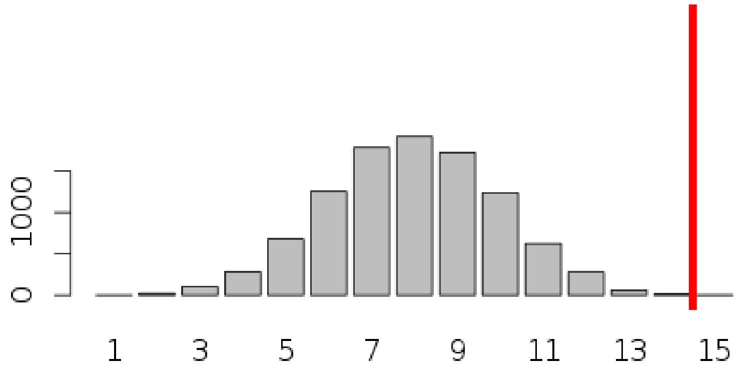
4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that the dolphins would guess 15 or more correct?
   - i.e., what is the p-value

5. Make a judgement
   - If we have a small p-value, this means that $\pi = .5$ is unlikely and so $\pi > .5$
   - i.e., we say our results are ‘statistically significant’

# Getting p-values using SDS1000 functions

Flipping coins many times:

```
null_dist <-  do_it(10000) * {
    rflip_count(16,  prob = .5)
}
```



We can get the number of values as or more extreme than an observed statistic (obs_stat) using the pnull() function:

```
obs_stat <-  ?
p_value  <- pnull(obs_stat,  null_dist,  lower.tail = FALSE)
```

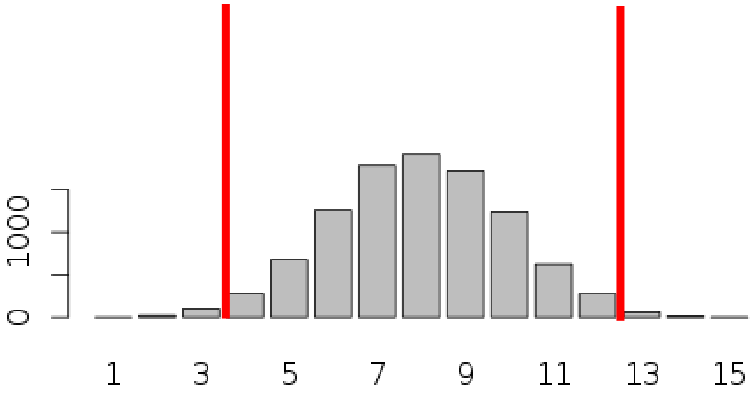
# Calculating a p-value from a null distribution

**For a one tailed alternative**: Find the proportion of statistics in the null distribution that equal or exceed the original statistic in the direction (tail) indicated by the alternative hypothesis

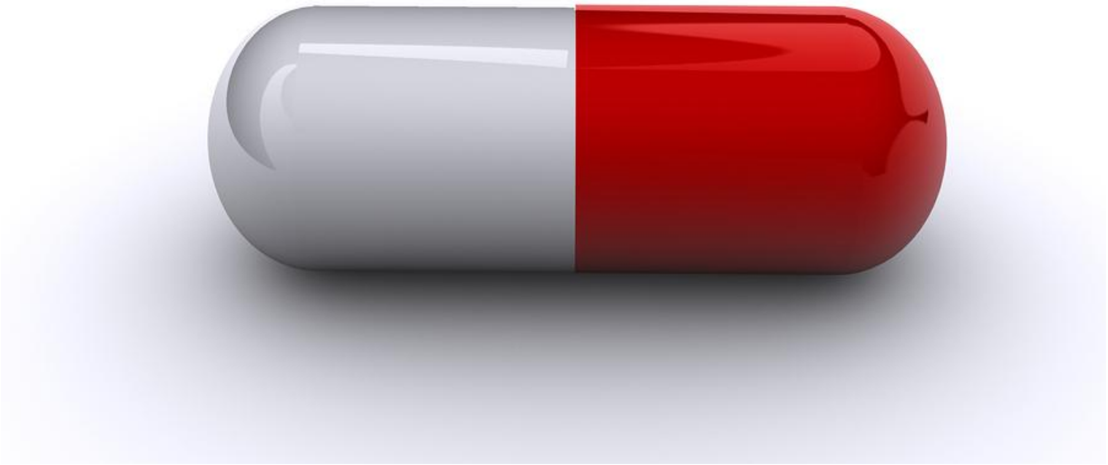
- E.g., $H_A$: $\pi > 0.5$

**For a two-tailed alternative**: Find the proportion of statistics in the null distribution in the tails beyond the observed statistic

- E.g., $H_A$: $\pi \neq 0.5$



# Hypothesis tests for comparing two means



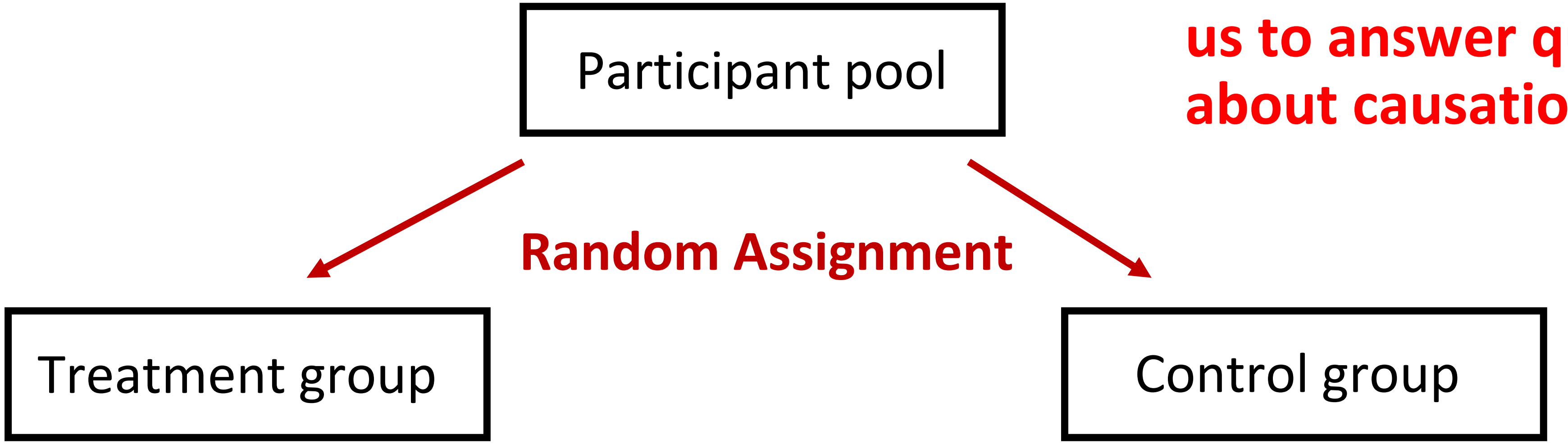
**Question**: Is this pill effective?

# Experimental design: randomized controlled trial

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get the pill
- Half in a *control group* where they get a fake pill (placebo)
- See if there is more improvement in the treatment group compared to the control group

**Random assignment allows us to answer questions about causation!**



Participant pool

**Random Assignment**

Treatment group

Control group

# Hypothesis tests for a single proportion

1. State the null hypothesis… and the alternative hypothesis
   - The means in the treatment and control group are the same: $H_0$: $\mu_T = \mu_C$
   - The means in the treatment and control group are not the same: $H_A$: $\mu_T \neq \mu_C$

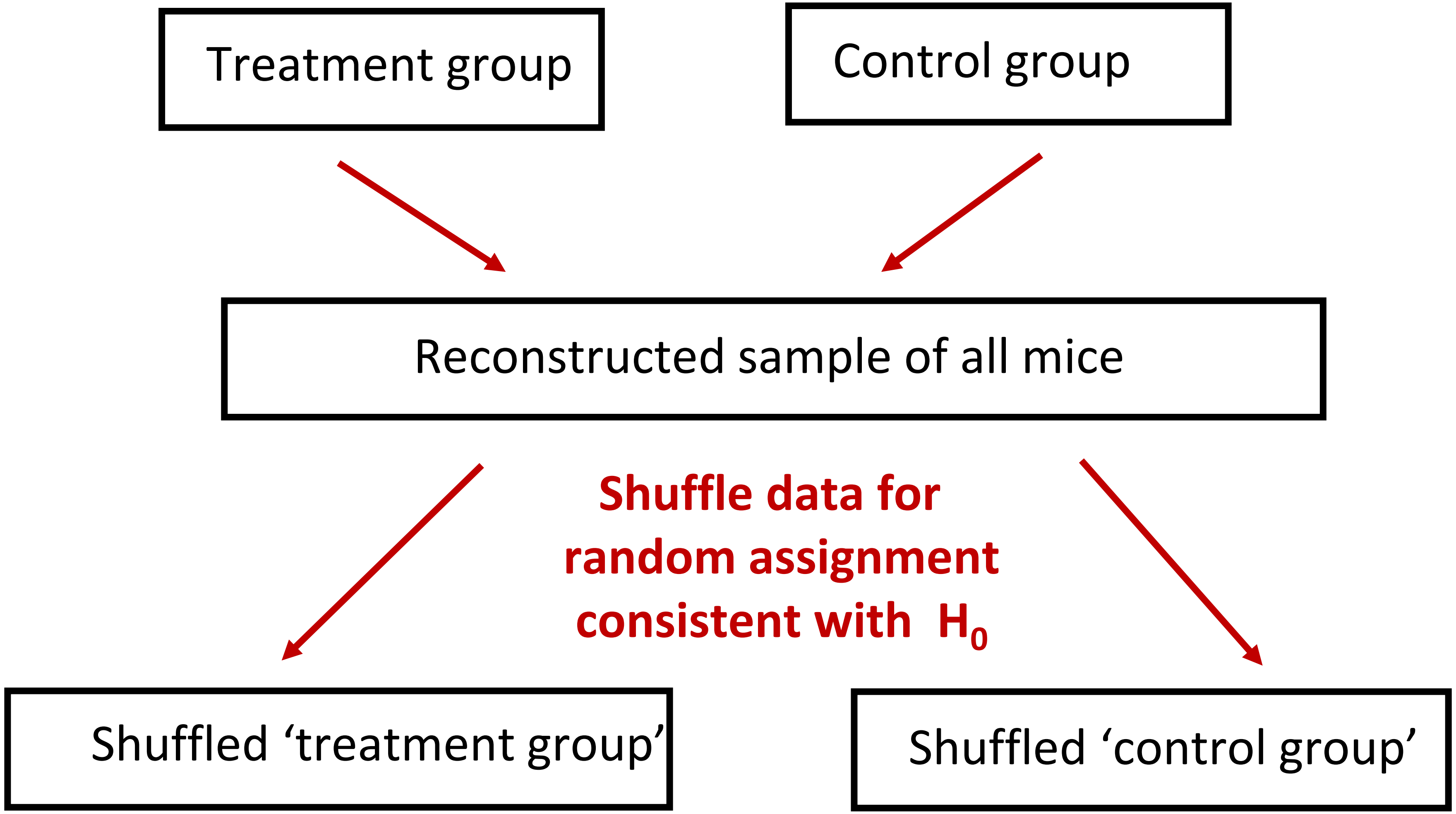
2. Calculate the observed statistic
   - $\overline{x}_{\text{Effect}} = \overline{x}_{\text{Treatment}} - \overline{x}_{\text{Control}}$

3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if there was no difference in the treatment and control groups

4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that we would get $\overline{x}_{\text{Effect}}$ if there was no differences in the groups?
   - i.e., what is the p-value

5. Make a judgement
   - If we have a small p-value, this means that $H_0$: $\mu_T = \mu_C$ is unlikely so we reject $H_0$ and conclude that $H_A$: $\mu_T \neq \mu_C$
   - i.e., we say our results are 'statistically significant'

# 3. Create the null distribution!



Treatment group

Control group

Reconstructed sample of all mice

**Shuffle data for random assignment consistent with $H_0$**

Shuffled ‘treatment group’

Shuffled ‘control group’

One null distribution statistic: $\bar{x}_{\text{Shuff\_Treat}} - \bar{x}_{\text{Shuff\_Control}}$

# 3. Creating a null distribution in R

```
# the data from the calcium study
treat <- c(7, -4, 18, 17, -3, -5,  1, 10, 11, -2)
control <- c(-1,  12,  -1,  -3,   3,  -5,   5,   2, -11,  -1,  -3)

# observed statistic
obs_stat <- mean(treat) - mean(control)

#  Combine data from both groups
combined_data <- c(treat, control)
```

# 3. Creating a null distribution in R

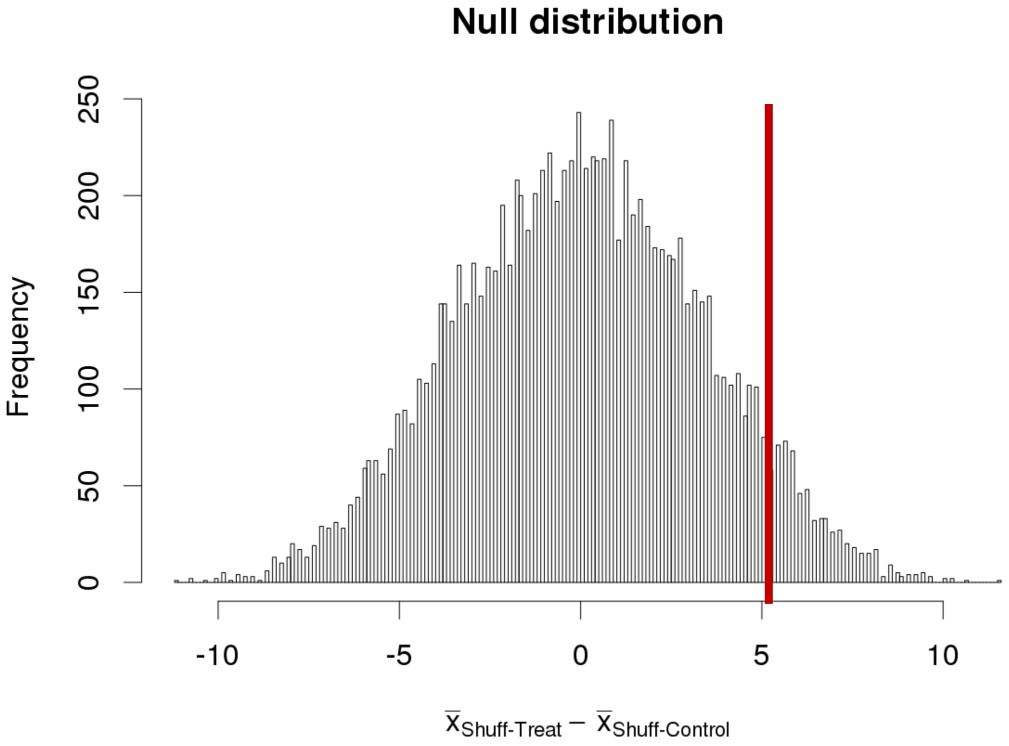
```
null_distribution <-  do_it(10000) * {

    # shuffle data
    shuff_data <- shuffle(combined_data)

    # create fake treatment and control groups
    shuff_treat   <-  shuff_data[1:10]
    shuff_control  <-  shuff_data[11:21]

    # save the statistic of interest
    mean(shuff_treat) - mean(shuff_control)

}
```



```
hist(null_distribution, breaks = 200)
```

```
# Calculate the p-value
p_value <- pnull(obs_stat,
                 null_distribution,
                 lower.tail = FALSE)
```

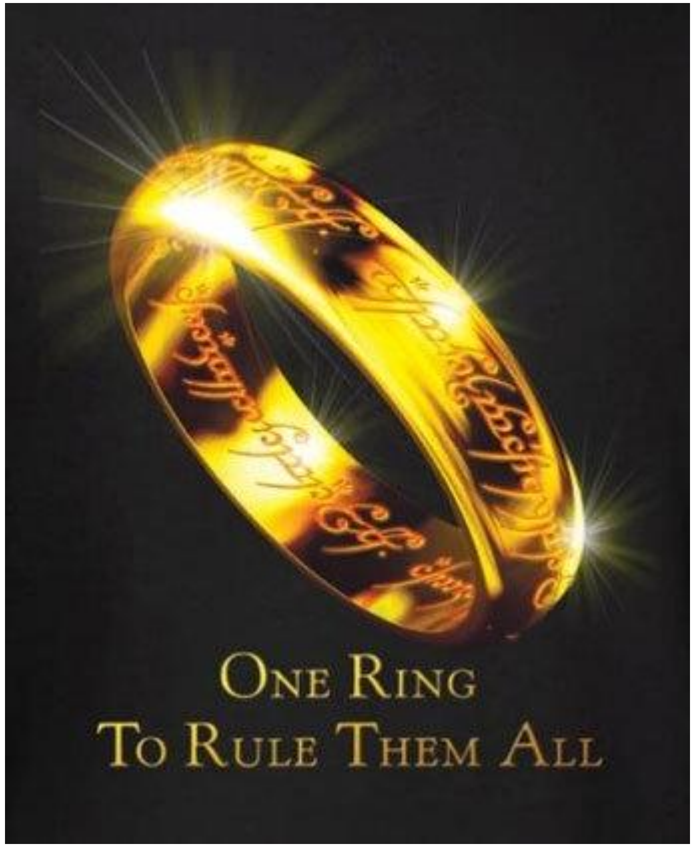
Suppose p-value = .064

# Hypothesis tests for more than 2 means

1. State the null and alternative hypotheses!

**$H_0$**: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$
**$H_A$**: $\mu_i \neq \mu_j$ for one pair of fields of study

We then continue with steps 2-5…



IF IT CAN BE DESTROYED BY THE TRUTH, IT DESERVES TO BE DESTROYED BY THE TRUTH.
SCHULZ