INTRO STATS

BRIAN STOCK

03.25.17

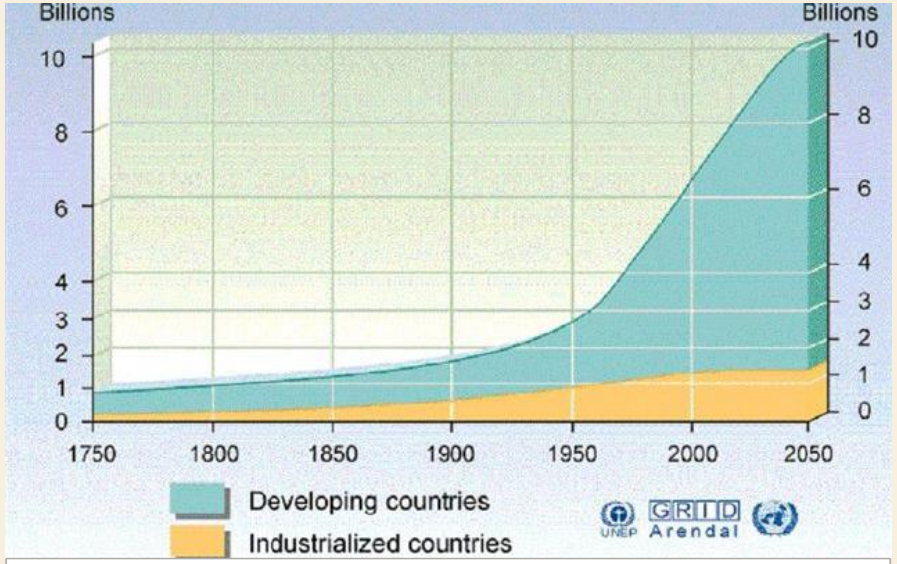
GOALS OF STATISTICS

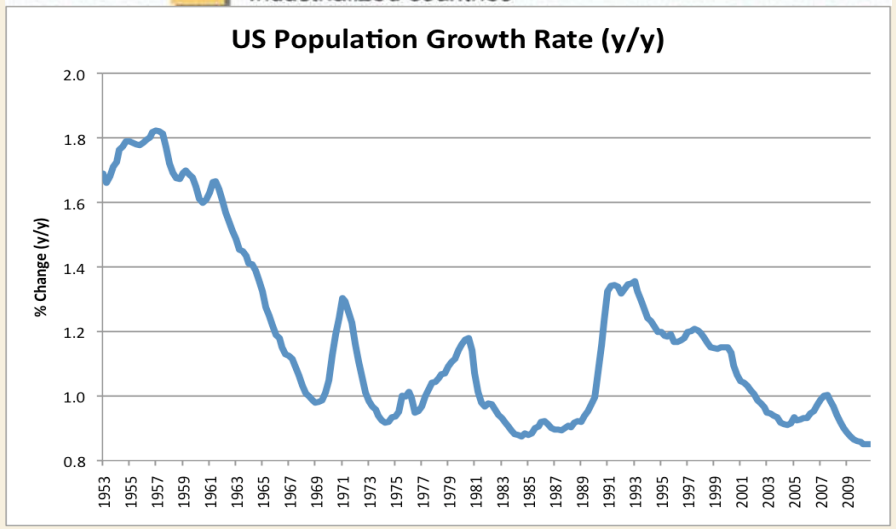
1. Estimate the values of important *parameters*

   "What is the global population growth rate?"

2. Test *hypotheses* about those parameters

   "Is the US pop growth rate **declining**?"





CHECK 1.
PARAMETERS VS. ESTIMATES

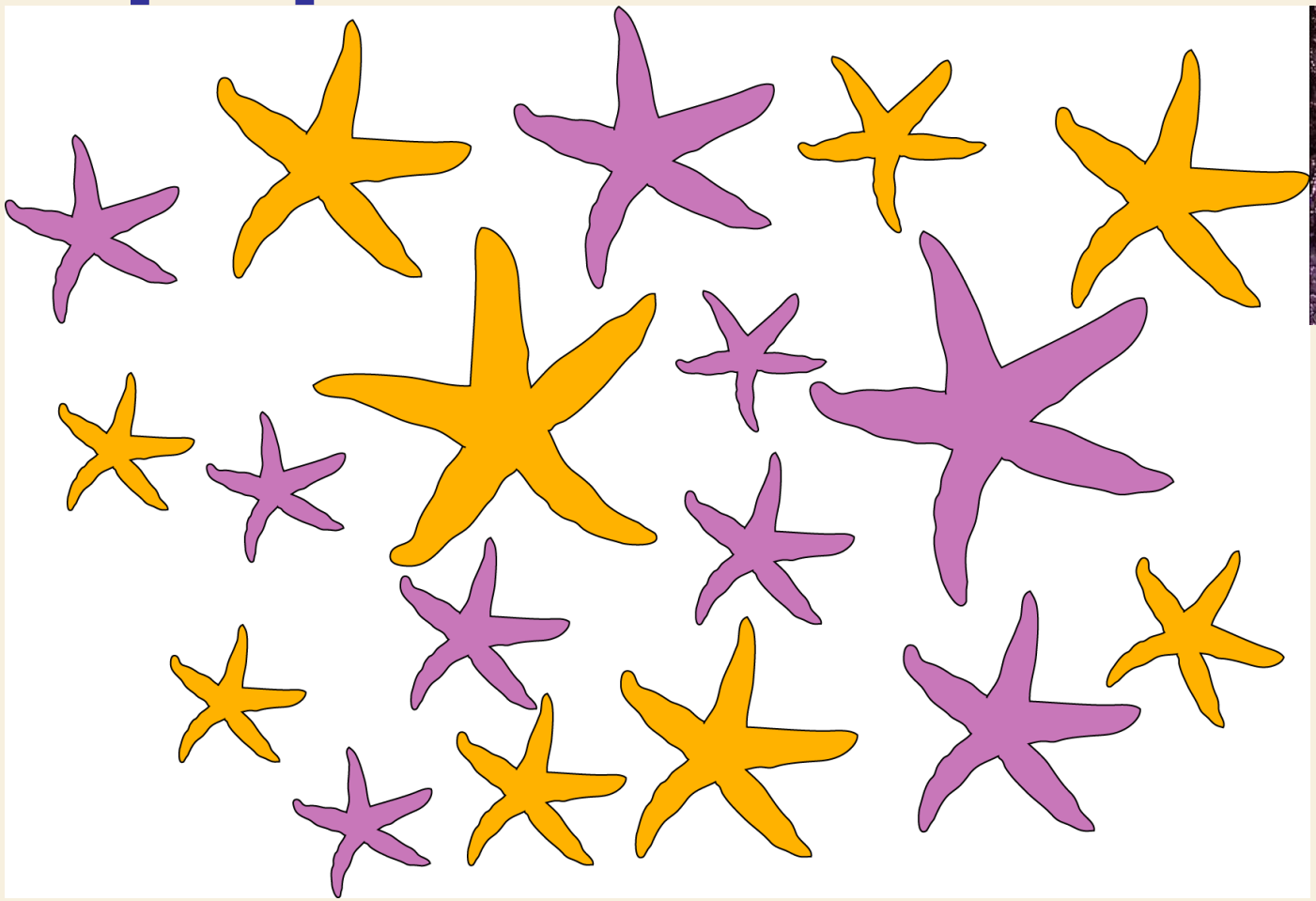
Write on a piece of paper:

1. What is the difference between a *parameter* and an *estimate*?
2. What is the difference between a *population* and a *sample*?
3. What makes a sample good?

Individually

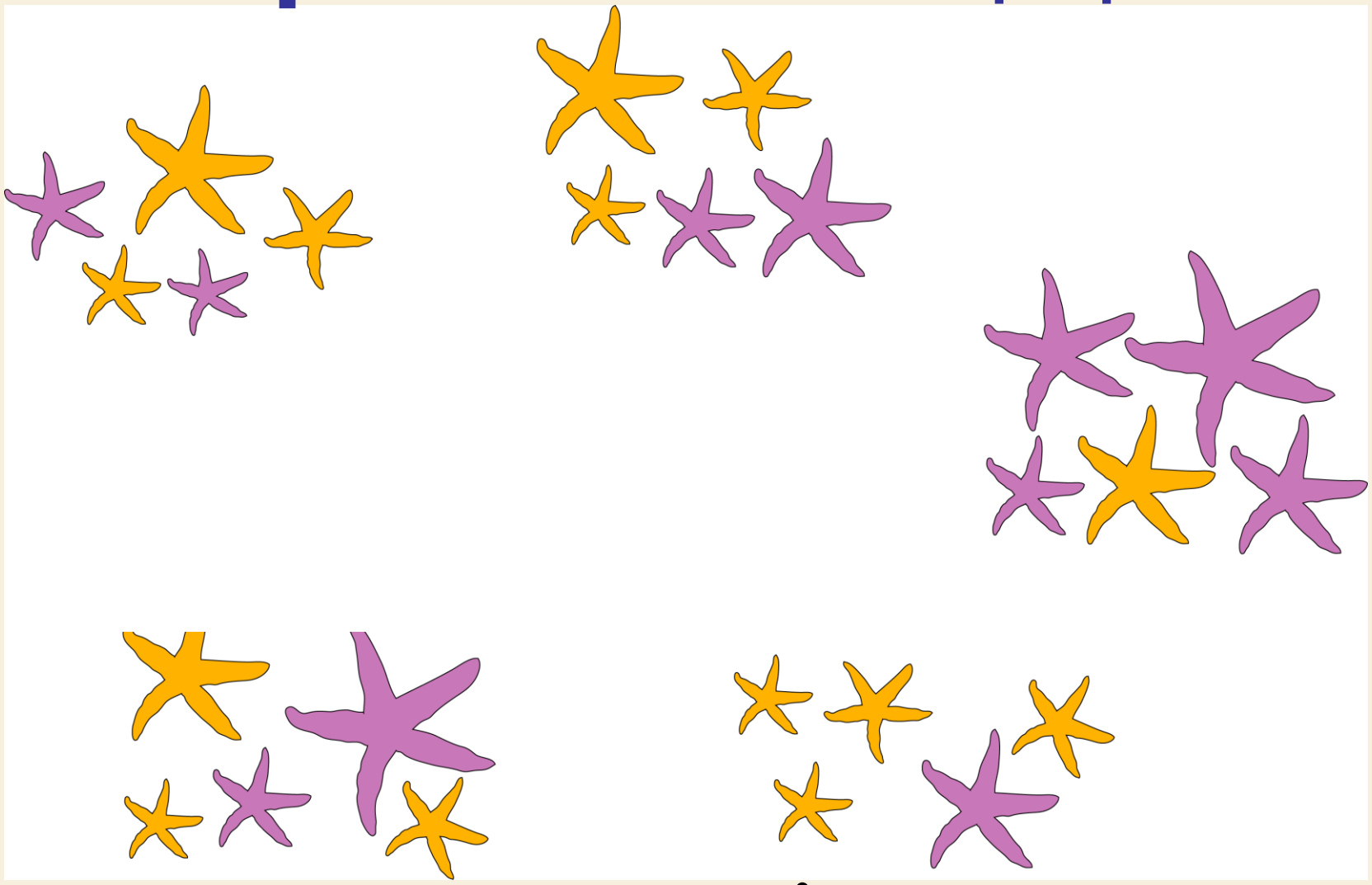
1-2 minutes

A **population** of starfish



What **parameters** may we be interested in?

Samples of starfish population



How do our **estimates** vary by each sample?

Populations --- **P**arameters Want to know

- Constant/fixed

Samples --- E**s**timates What we have

- Random variables

Populations --- Parameters

Want to know

- Constant/fixed

STATISTICS

Samples --- Estimates

What we have

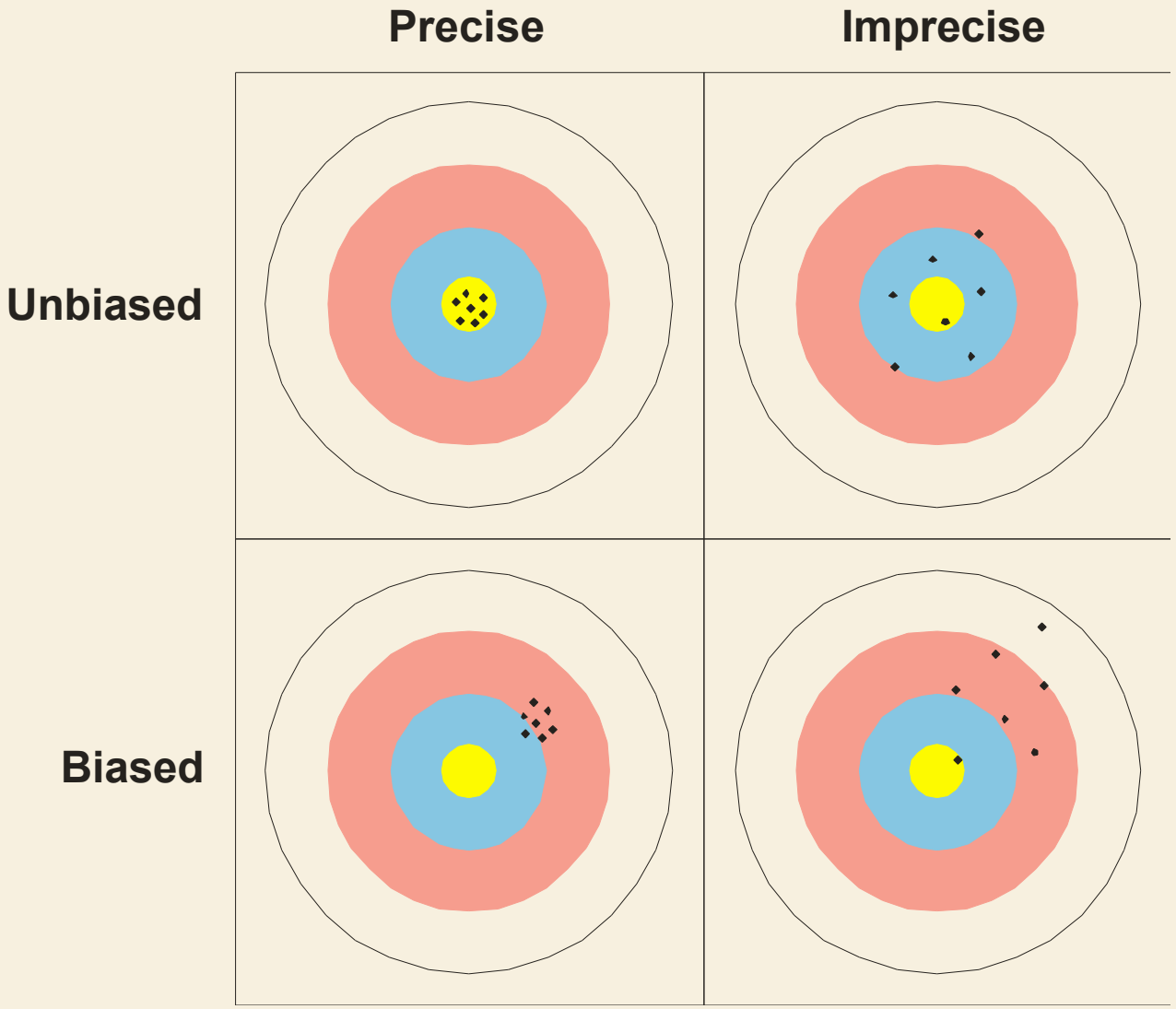
- Random variables

What makes a good sample?

What makes a good sample?

- *Independent* selection of individuals
- *Random* selection of individuals
  - Each individual has equal chance of being selected
- Sufficiently *large*

Sample **bias** vs. **precision**



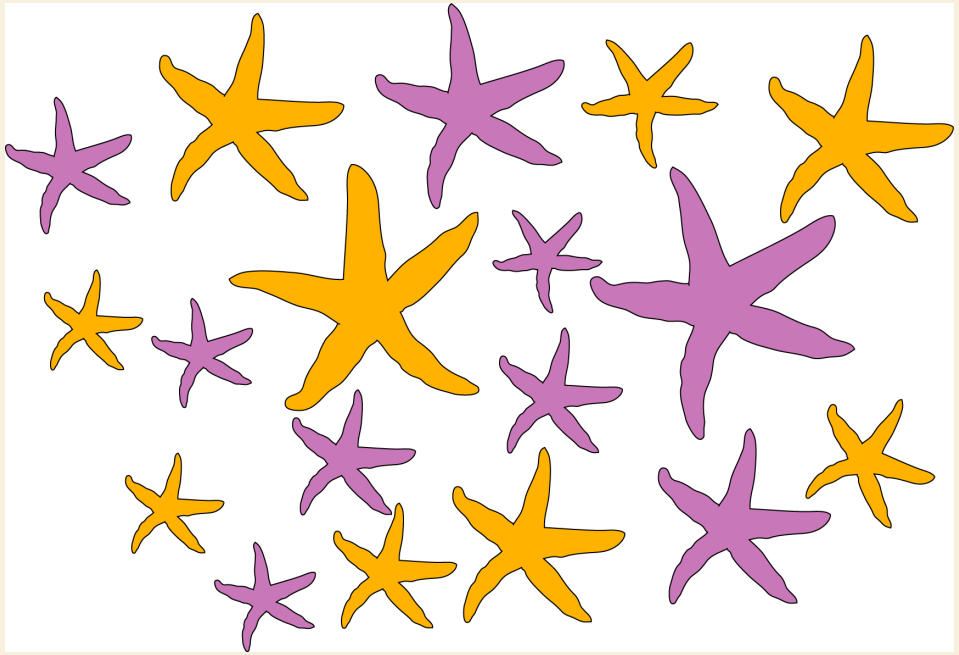
THE STUMBLING BLOCKS

1. What statistical test should I use?
2. How to do it (using R)

CHECK 2.
SCIENTIFIC METHOD / PROCESS

In groups of 4, discuss:

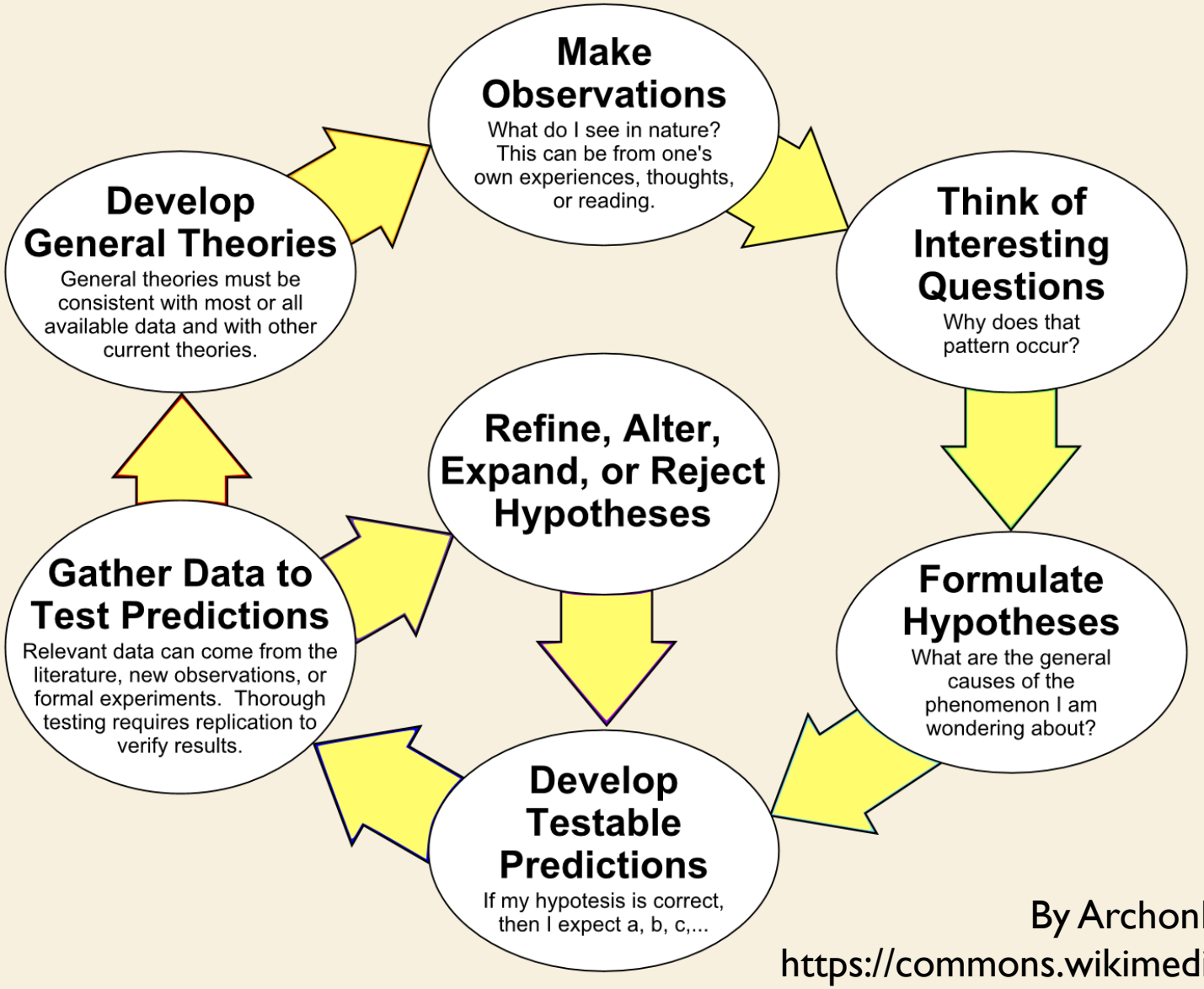
- What are the steps in the scientific method/process?
- Choose a simple biological question. Describe the steps you would take to find the answer.



~10 minutes

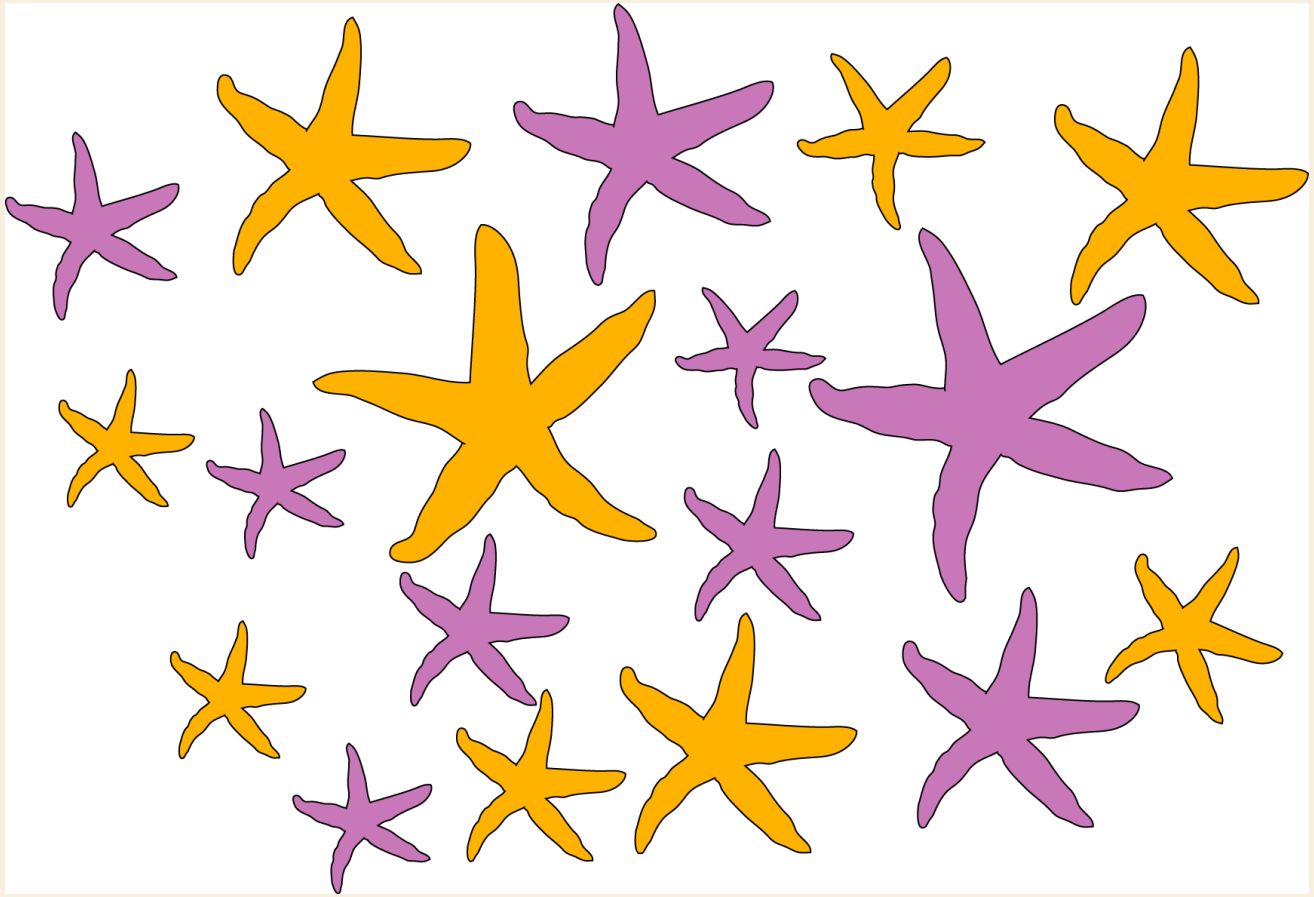
Share with class

SCIENTIFIC METHOD / PROCESS



1. **Biological question**. Be specific!
2. Put the question in the form of a **biological null hypothesis** and alternate hypothesis.
3. Translate the biological hypotheses into **statistical hypotheses**
4. **Which variables** are relevant to your question?
5. Determine what **kind of variable** each one is.
6. **Design an experiment** that controls or randomizes the confounding variables.
7. **Choose the best statistical test** to use. Depends on:
   - number of variables
   - kinds of variables
   - expected fit to the parametric assumptions
   - hypothesis to be tested
8. Determine a good **sample size** for the experiment. Power analysis.
9. **Collect data** (do the experiment).
10. See if your data meet the **assumptions of the statistical test** you chose. If it doesn't, choose a more appropriate test.
11. **Apply the statistical test,** and interpret the results.
12. **Communicate your results**, often with a graph or table

CHECK 3.
NULL VS ALTERNATE HYPOTHESES



Did your group list null and alternate hypotheses?

- If not, do so now

Groups

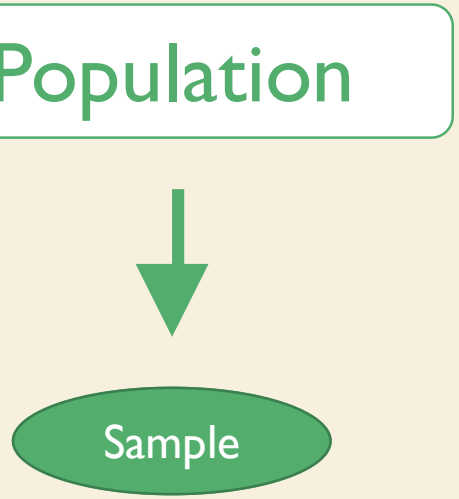
1-2 minutes

CHECK 4.
HYPOTHESIS TESTING

Population

Define our **population parameter** of interest.

CHECK 4. HYPOTHESIS TESTING

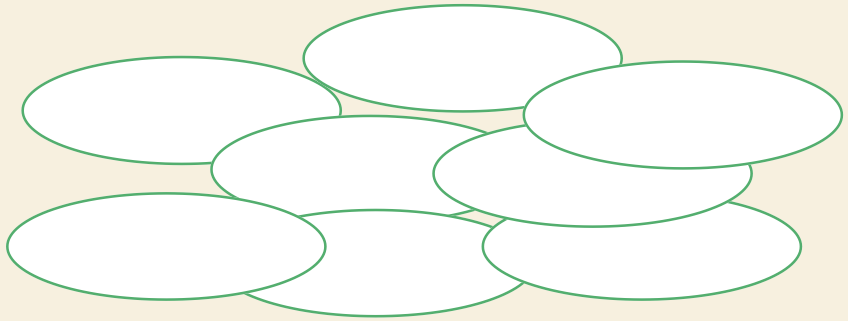


We can't measure EVERY starfish in the population. Instead, take a **random sample**. Calculate the **estimate** using our sample.

CHECK 4. HYPOTHESIS TESTING

Population

Sample

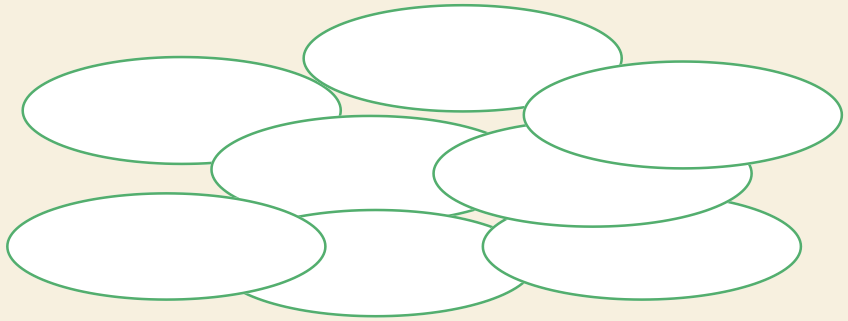


But our sample/estimate is random – imagine all of the other samples we COULD have gotten!

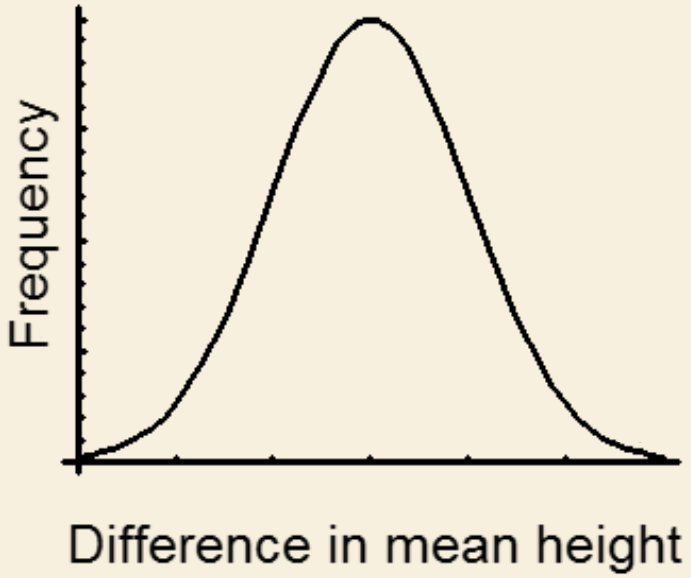
CHECK 4. HYPOTHESIS TESTING

Population

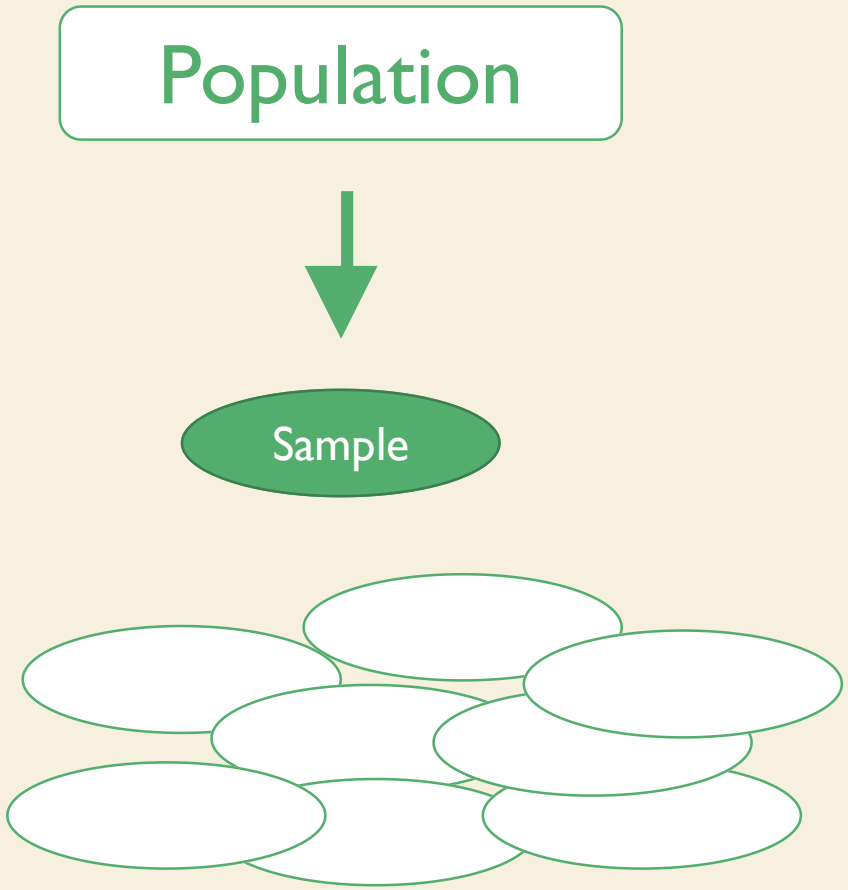
Sample



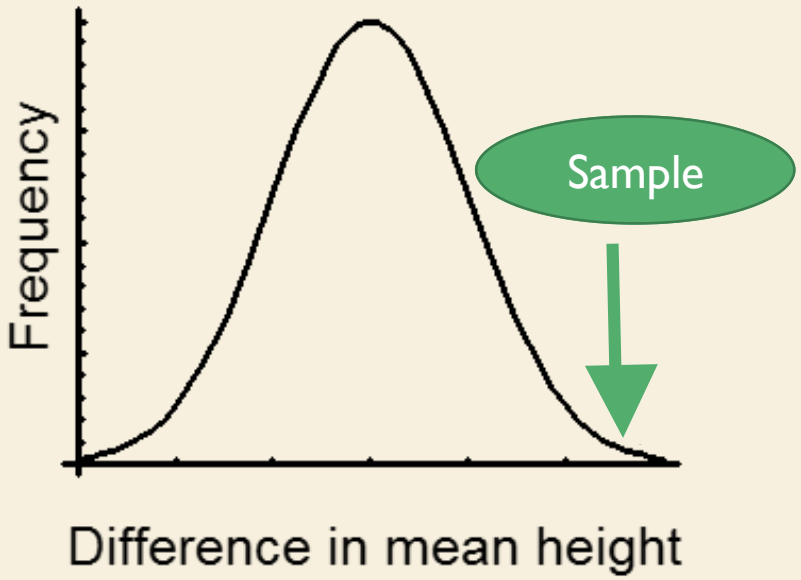
If we make an estimate (ex: calculate mean) from each of these samples, we have the **sampling distribution** under the **null hypothesis**.



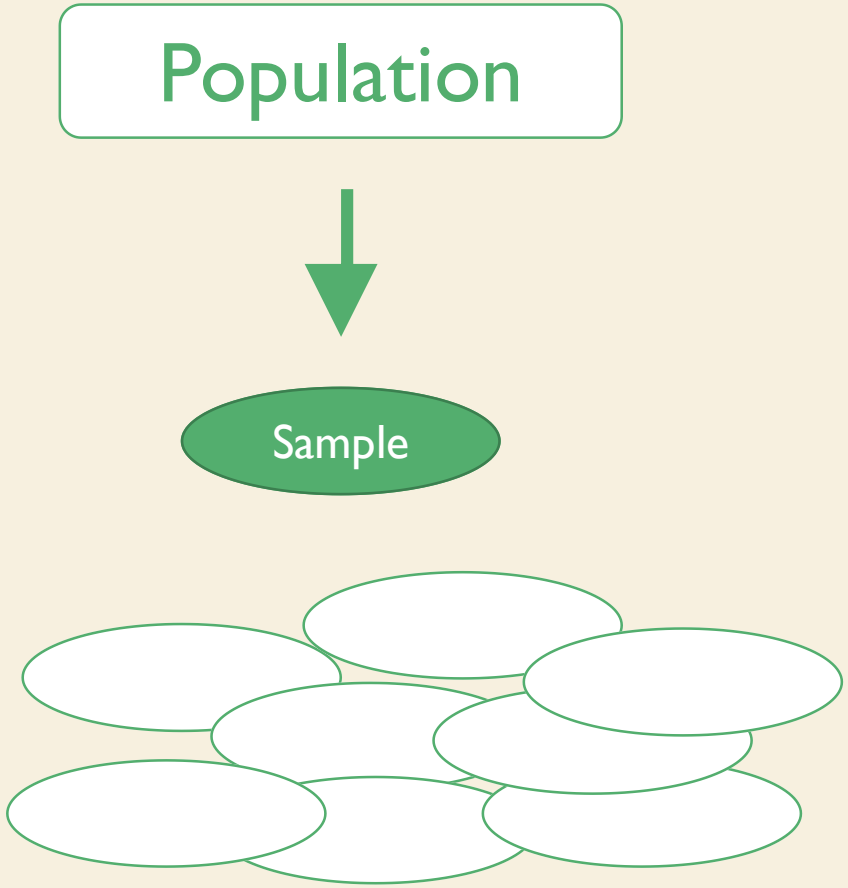
CHECK 4. HYPOTHESIS TESTING



If our sample is **very different** from what we expect under the null hypothesis, (calculate the probability, or **p-value**) then we **reject the null** (conclude that our population parameter is different).

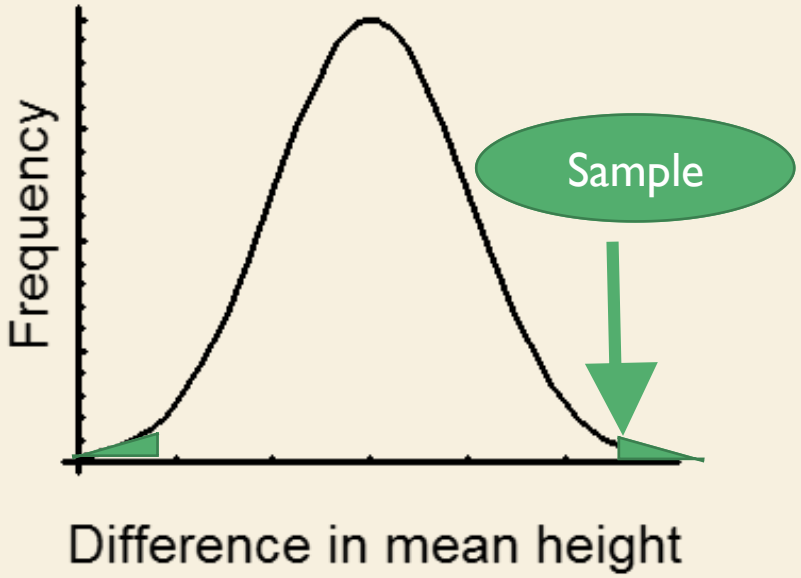


CHECK 4. HYPOTHESIS TESTING

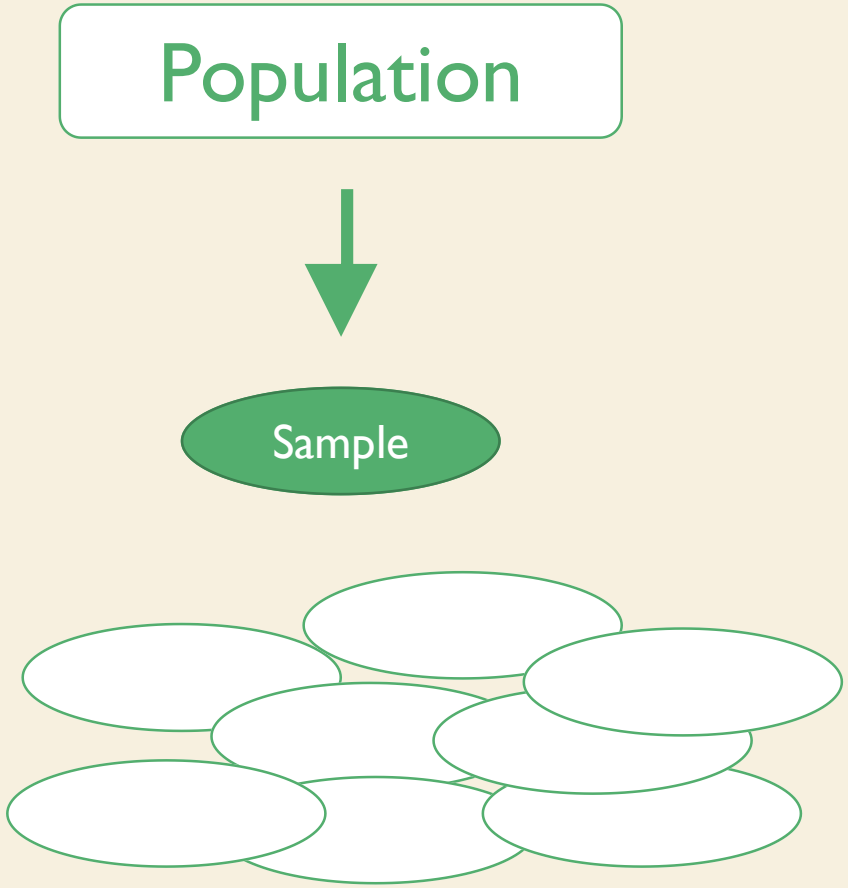


If our sample is **very different** from what we expect under the null hypothesis, (calculate the probability, or **p-value**) then we **reject the null** (conclude that our population parameter is different).

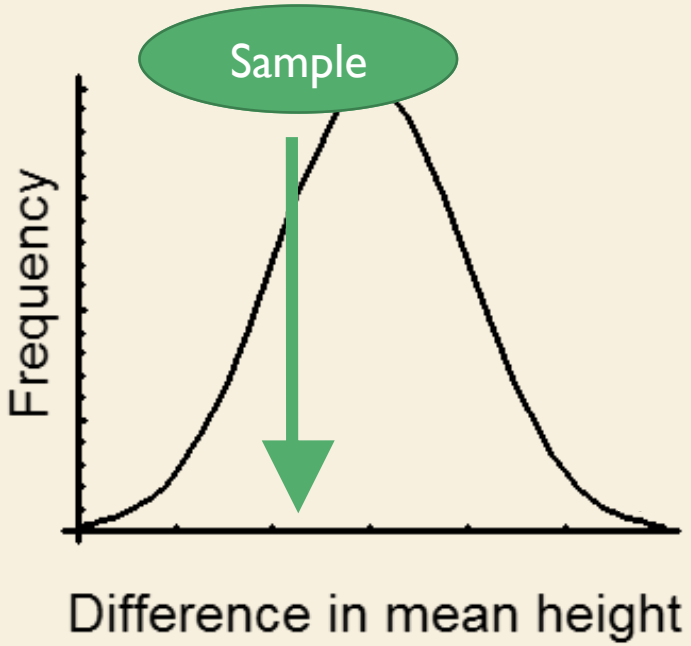
$p = 0.01$



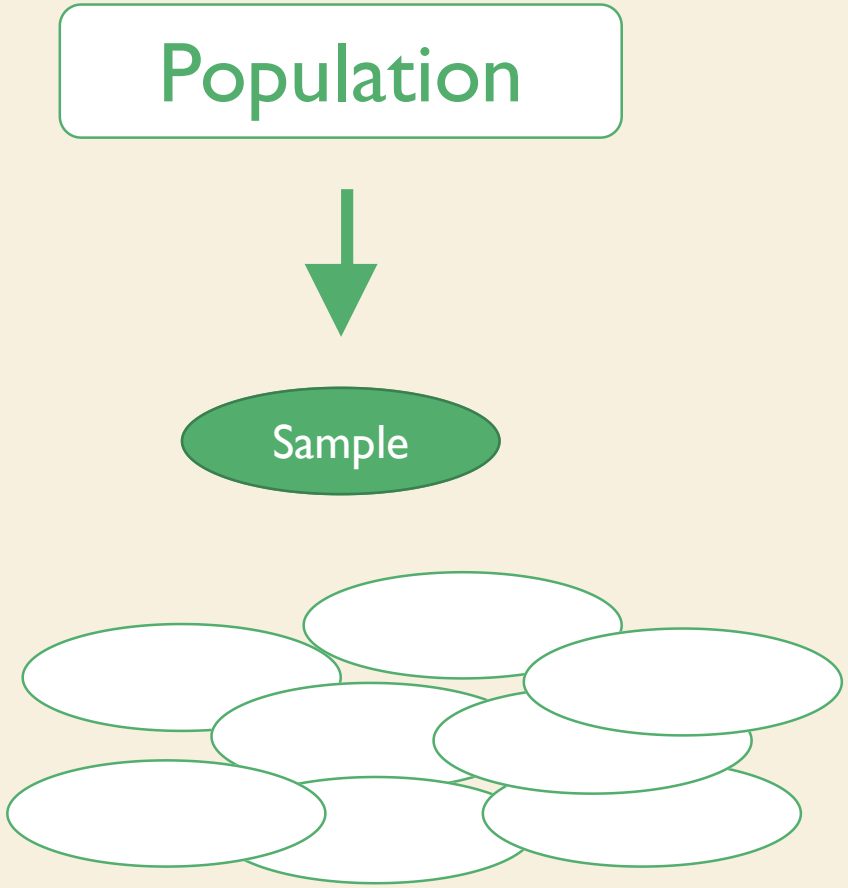
CHECK 4. HYPOTHESIS TESTING



If our sample is **not different** from what we expect under the null hypothesis, (calculate the probability, or **p-value**) then we **fail to reject the null** (conclude that our population parameter is not different).

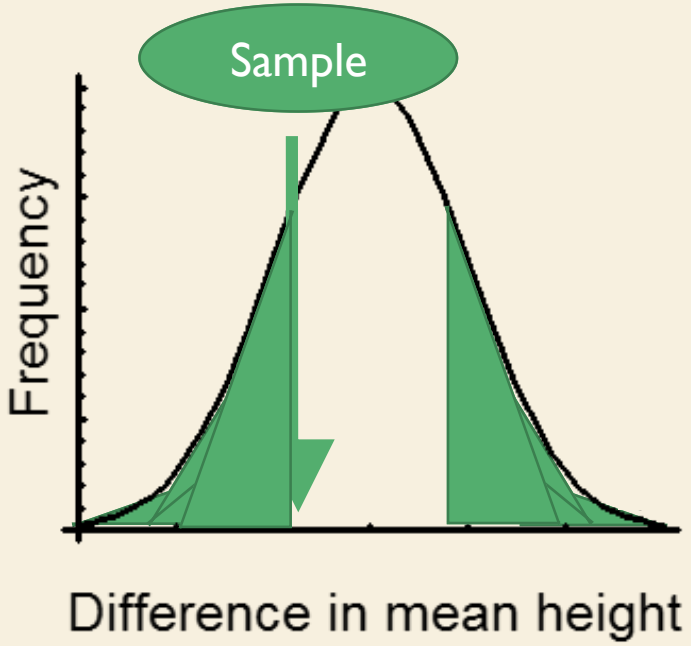


CHECK 4. HYPOTHESIS TESTING



If our sample is **not different** from what we expect under the null hypothesis, (calculate the probability, or **p-value**) then we **fail to reject the null** (conclude that our population parameter is not different).

$p = 0.48$



STRETCH!

CHECK 5.
TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

3) At a port, you track the number of every species of fish caught by the fishermen every day for 3 months

Fecha	Mero	Manta	Pargo	Mojarra
03.06.17	25	0	67	18
04.06.17	34	2	55	17
05.06.17				
06.06.17				

How could you display the data?

CHECK 5. TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

4) Beaufort sea state

Beaufort state	Wind speed	Wave height	Description
0	< 1	0	mirror-like
1	1-5	0-0.2	ripples
2	6-11	0.2-0.5	wavelets
3	12-19	0.5-1	crests break
4	20-28	1-2	whitecaps
5	29-38	2-3	spray
6	39-49	3-4	foam crests

How could you display the data?

CHECK 5. TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

5) For one fishing port in 2012 and 2013 you have the total number of fish caught. You want to see if there is a difference between 2012 and 2013.

Year	Catch
2012	2168
2013	3087

How could you display the data?

CHECK 5. TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

6) You study a population of grouper and collect length measurements of 500 fish

Fish	Length (mm)
1	466
2	680
3	512
…	
500	717

How could you display the data?

CHECK 5. TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

7) During a species' spawning season, you sample 200 females and record whether or not each female has eggs (sexual maturity)

Female	Mature?
1	Yes
2	No
3	No
…	
200	Yes

How could you display the data?

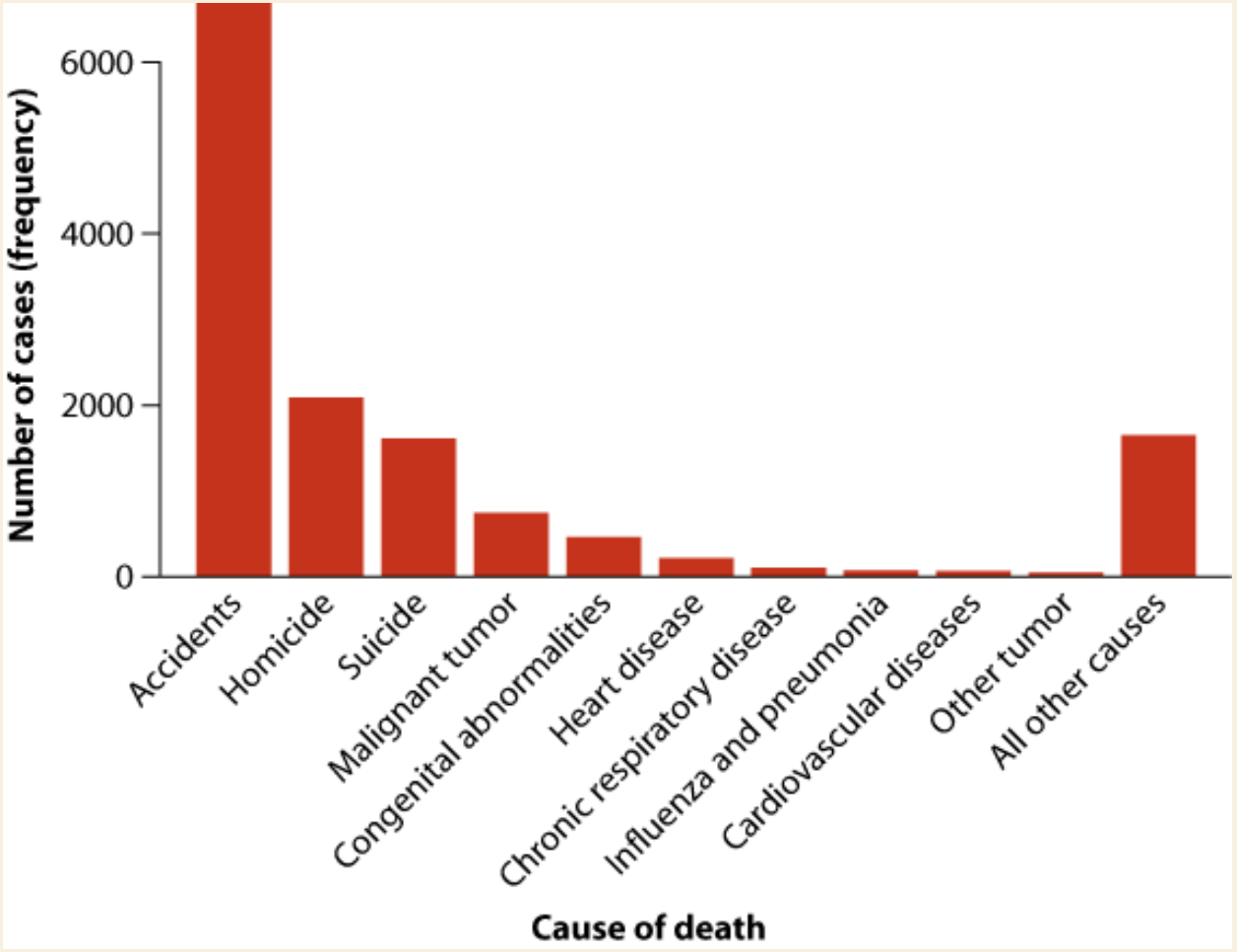
CHECK 5.
TYPES OF DATA

- Time series
- Count/discrete
- Paired
- Continuous
- Categorical
- Ordinal
- Binomial

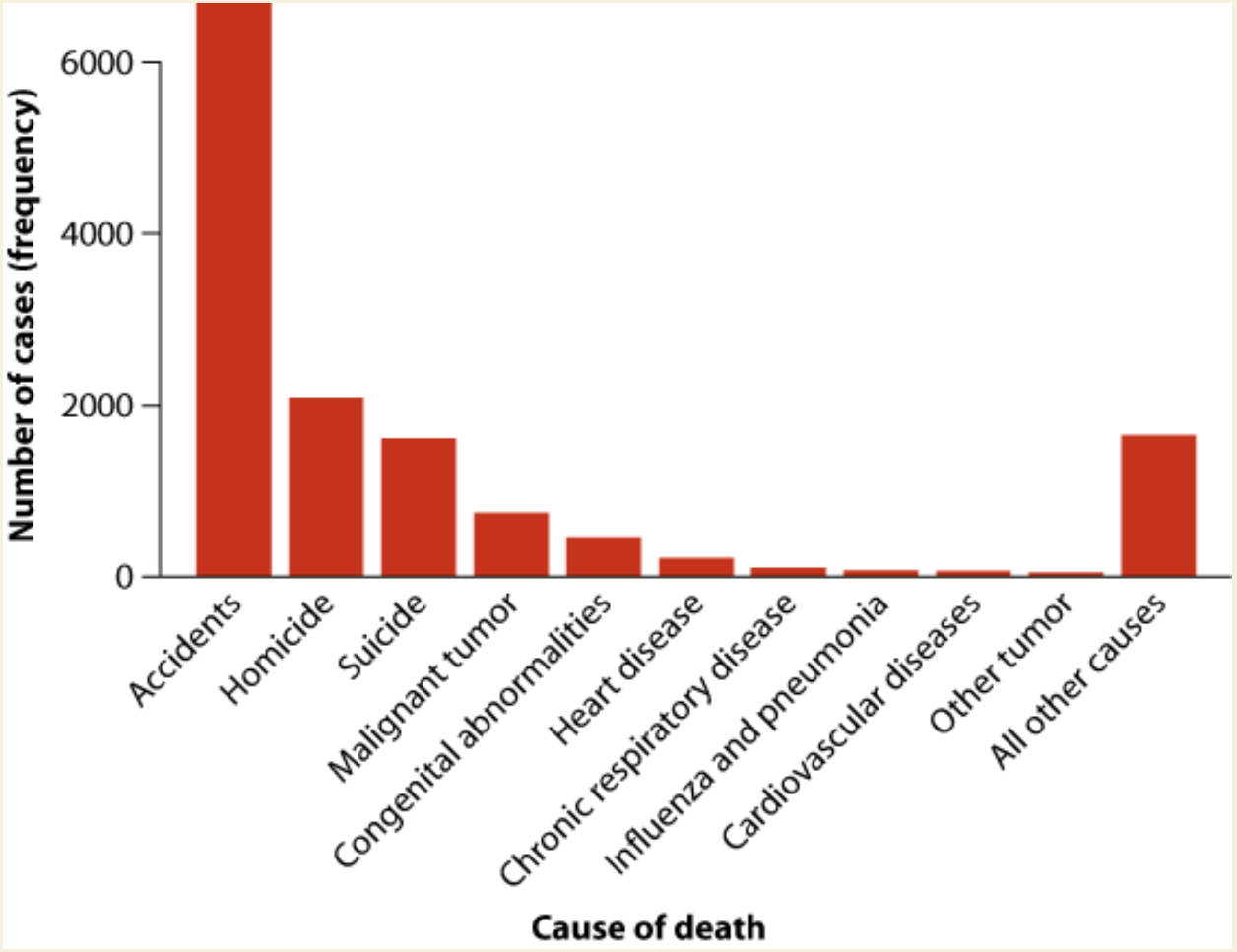
8) On a fishing boat, you record the species of each fish caught.

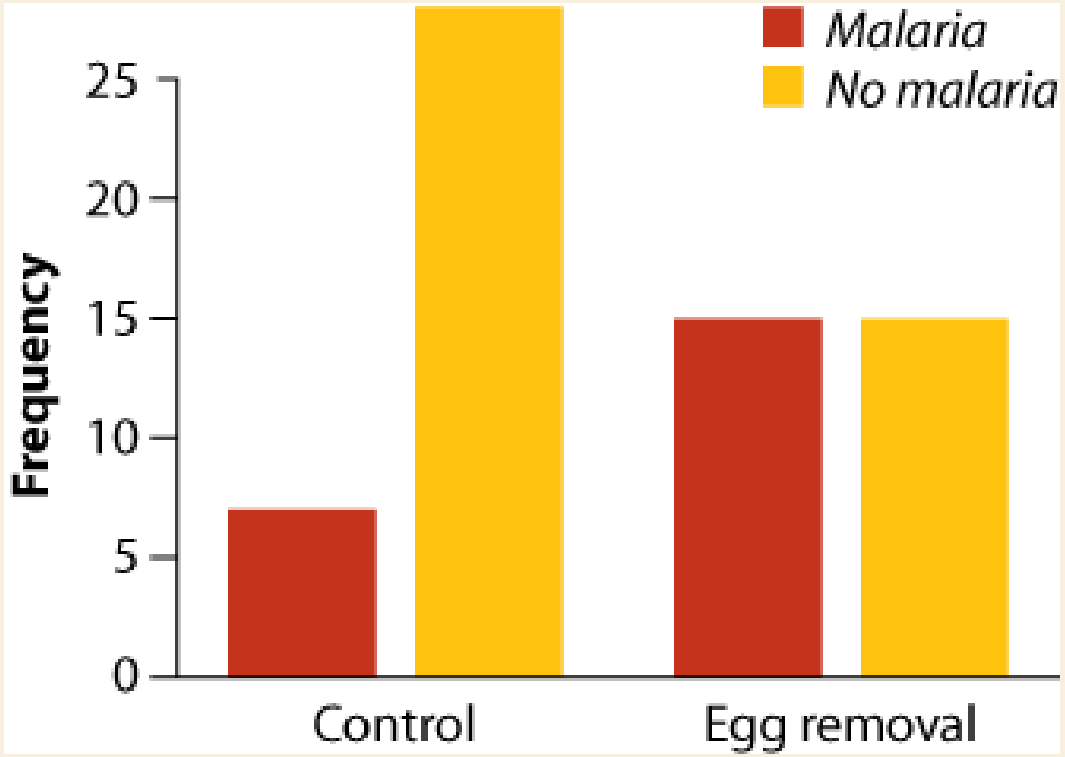
Fish	Species
1	Manta
2	Mero
3	Pompano
…	
200	Pargo

How could you display the data?

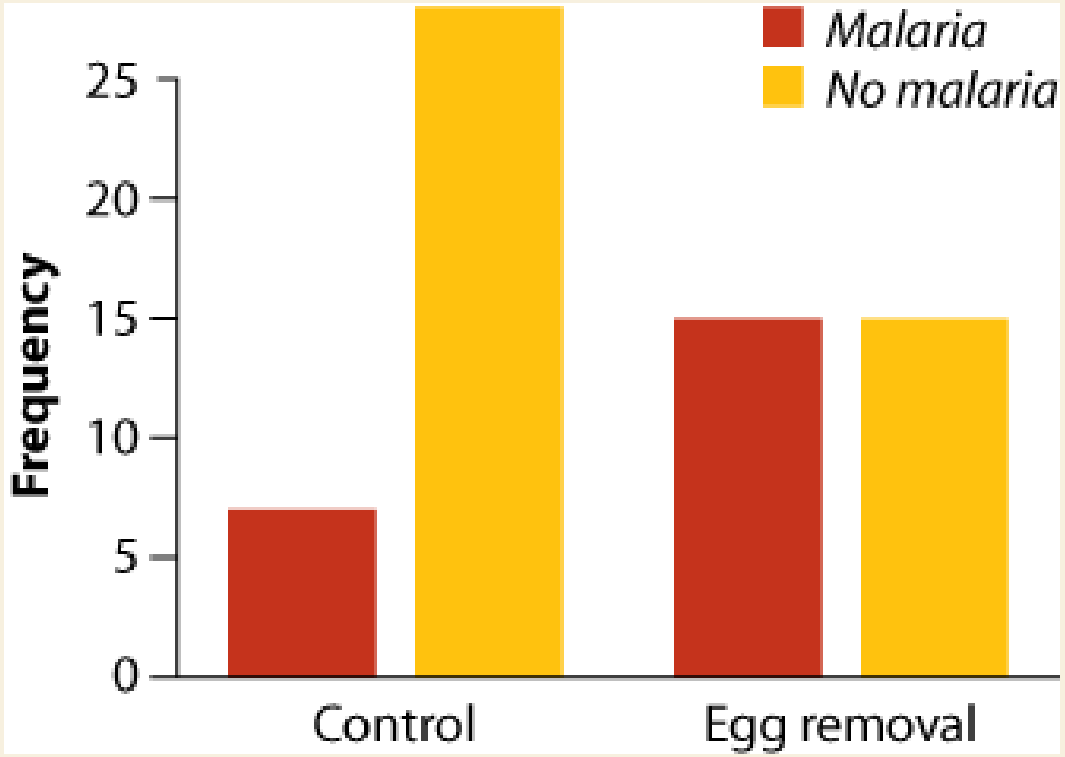

Number of cases (frequency)
6000
4000
2000
0
Accidents
Homicide
Suicide
Malignant tumor
Congenital abnormalities
Heart disease
Chronic respiratory disease
Influenza and pneumonia
Cardiovascular diseases
Other tumor
All other causes
Cause of death

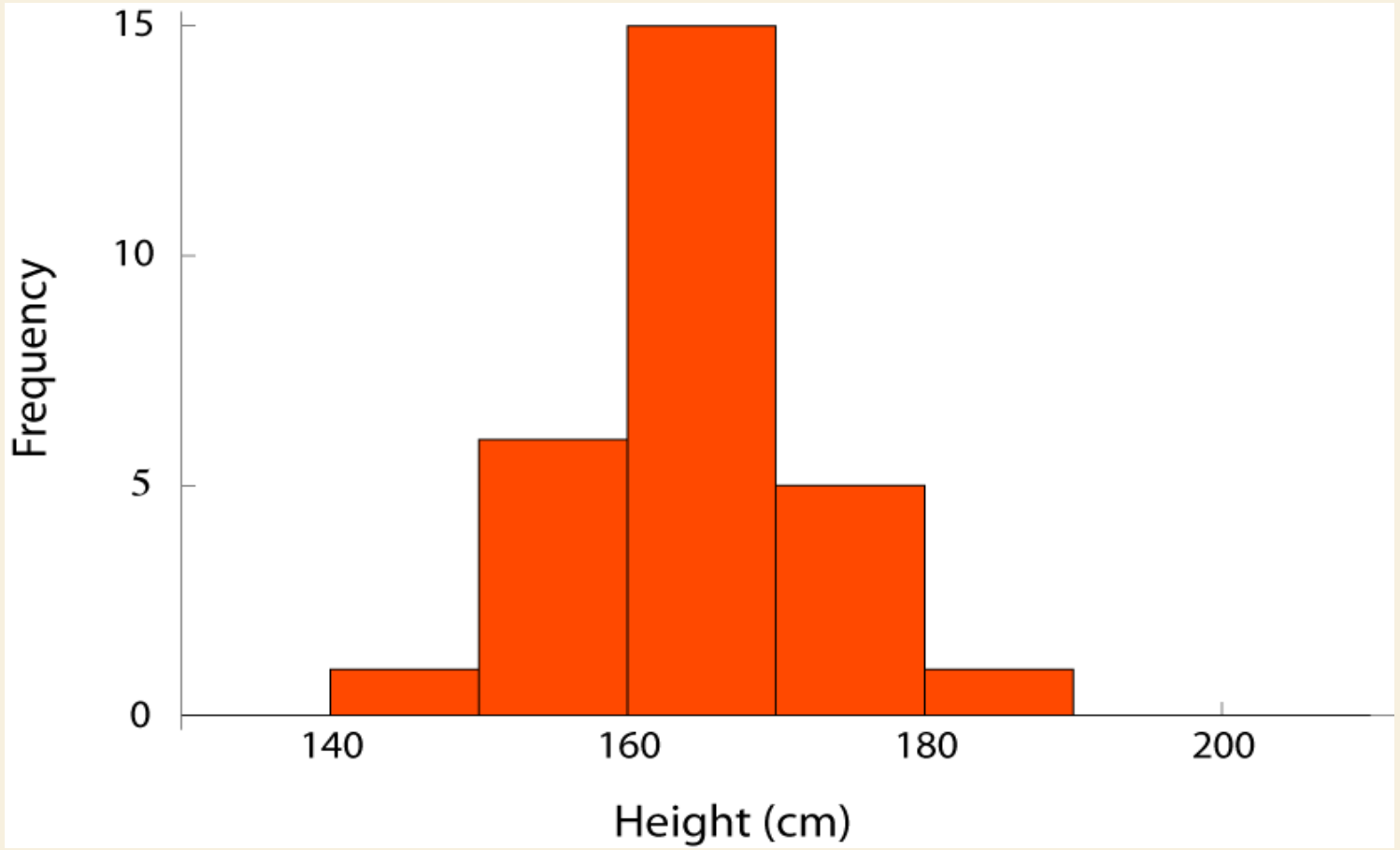
Bar graph


Number of cases (frequency)
6000
4000
2000
0
Accidents
Homicide
Suicide
Malignant tumor
Congenital abnormalities
Heart disease
Chronic respiratory disease
Influenza and pneumonia
Cardiovascular diseases
Other tumor
All other causes
Cause of death

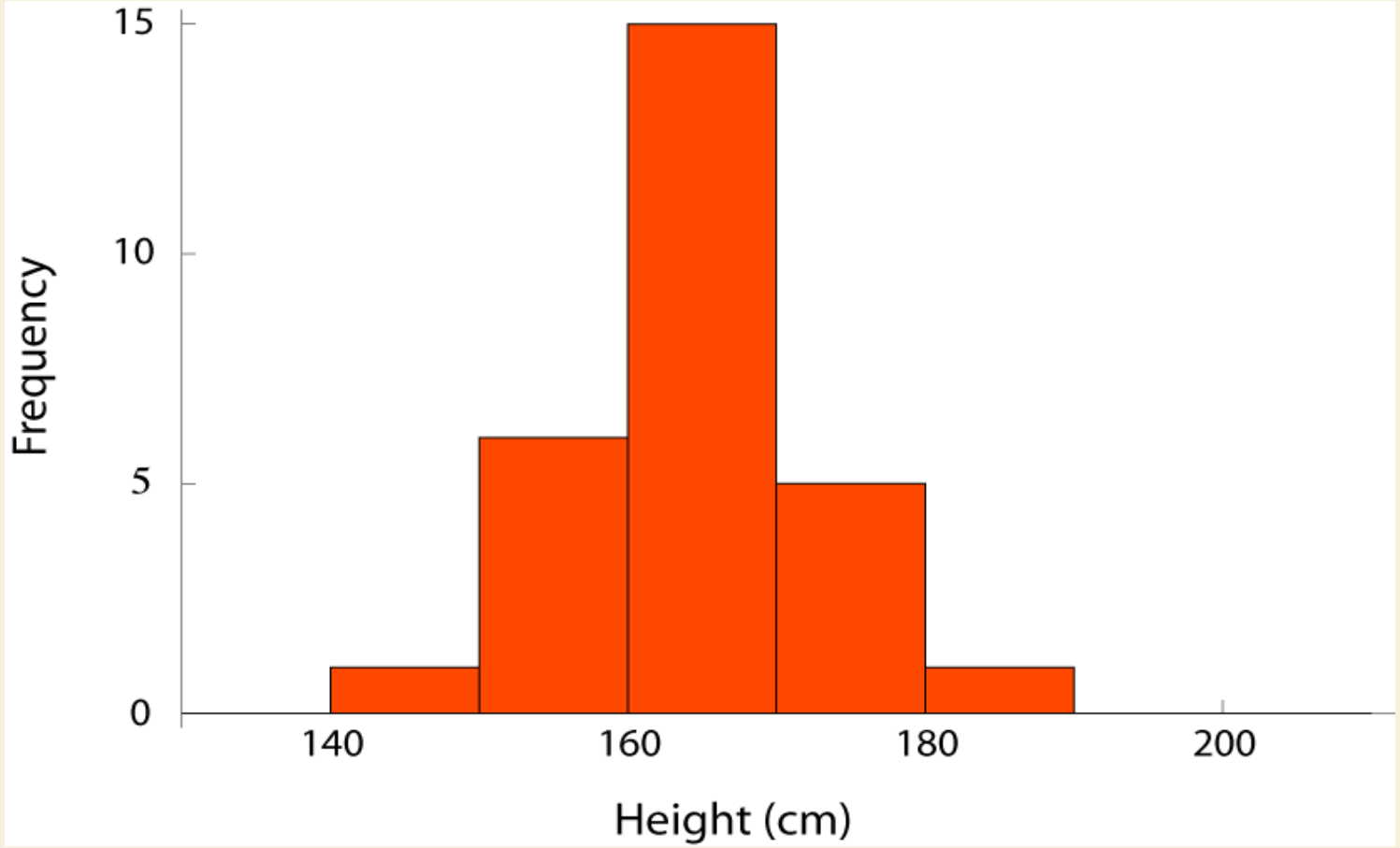

Malaria
No malaria
Frequency
25
20
15
10
5
0
Control
Egg removal

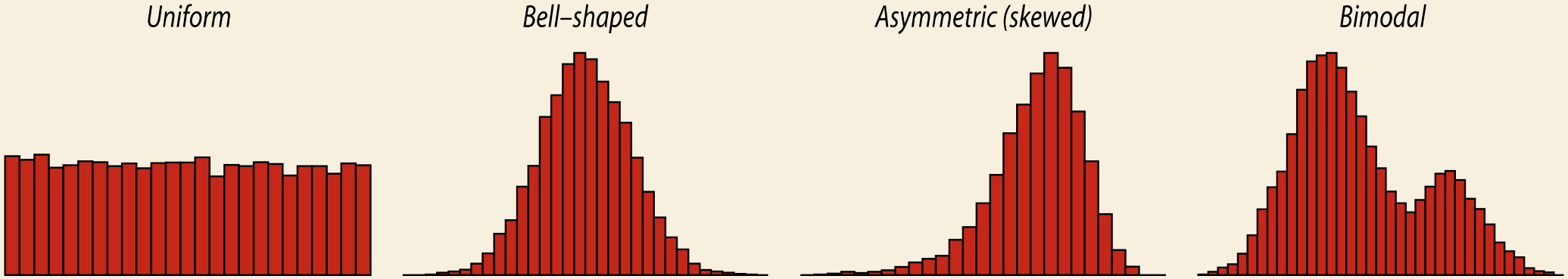
Grouped bar graph



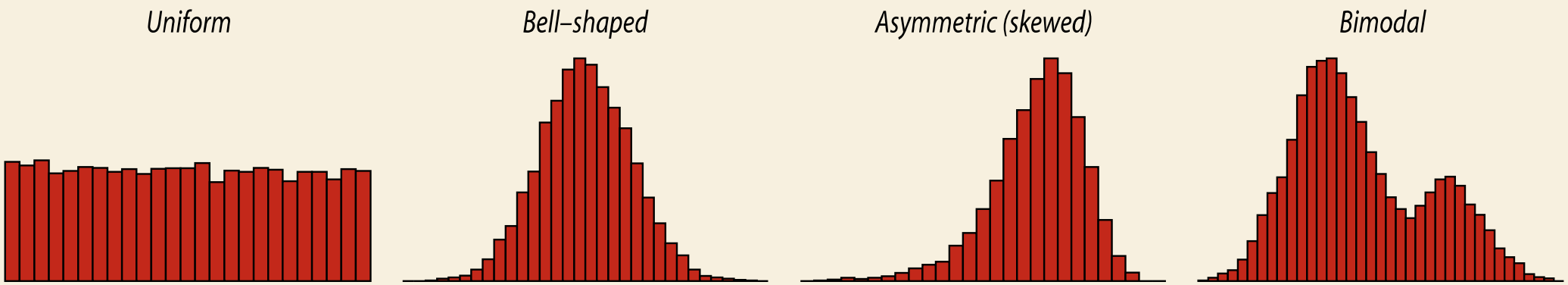

15
10
5
0
Frequency
140
160
180
200
Height (cm)

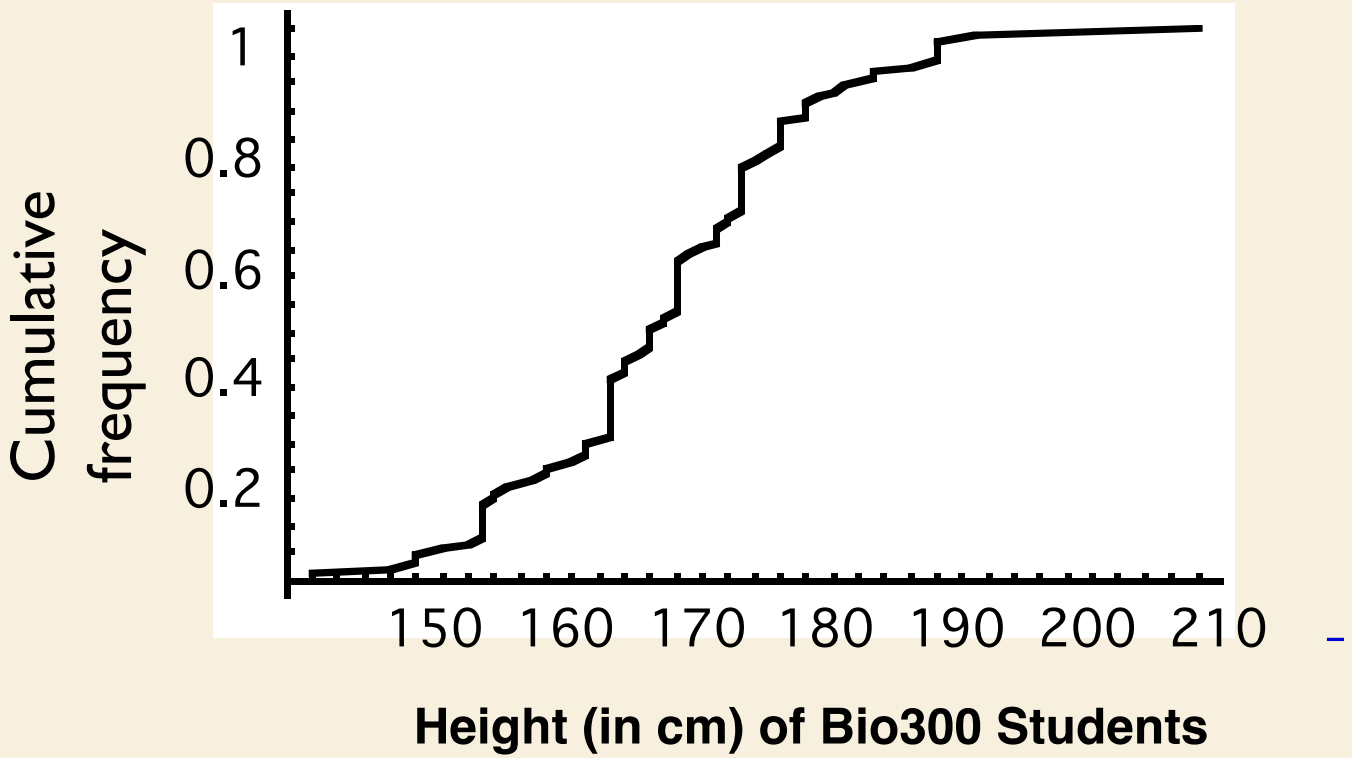
Histogram



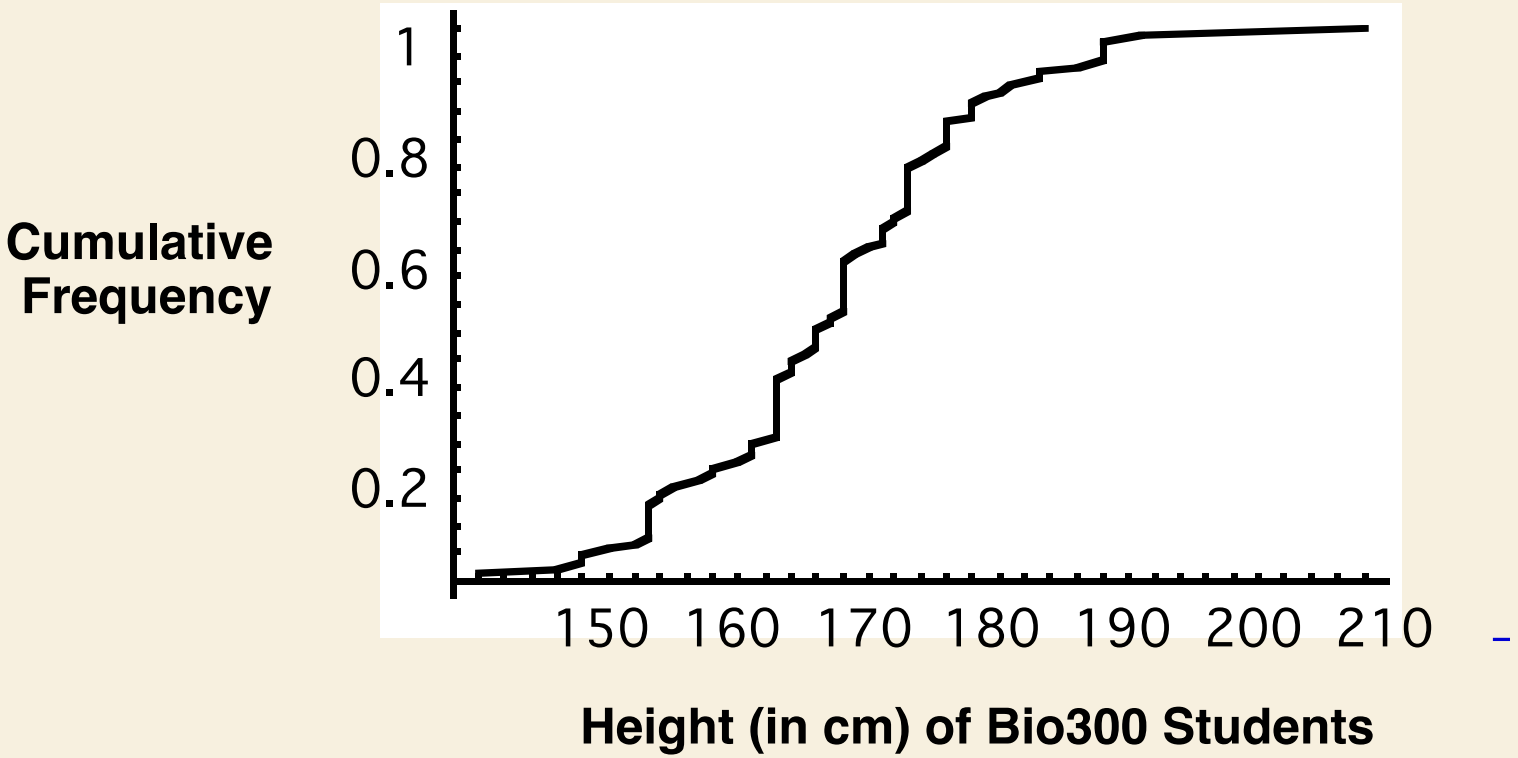

Uniform
Bell–shaped
Asymmetric (skewed)
Bimodal

Histograms



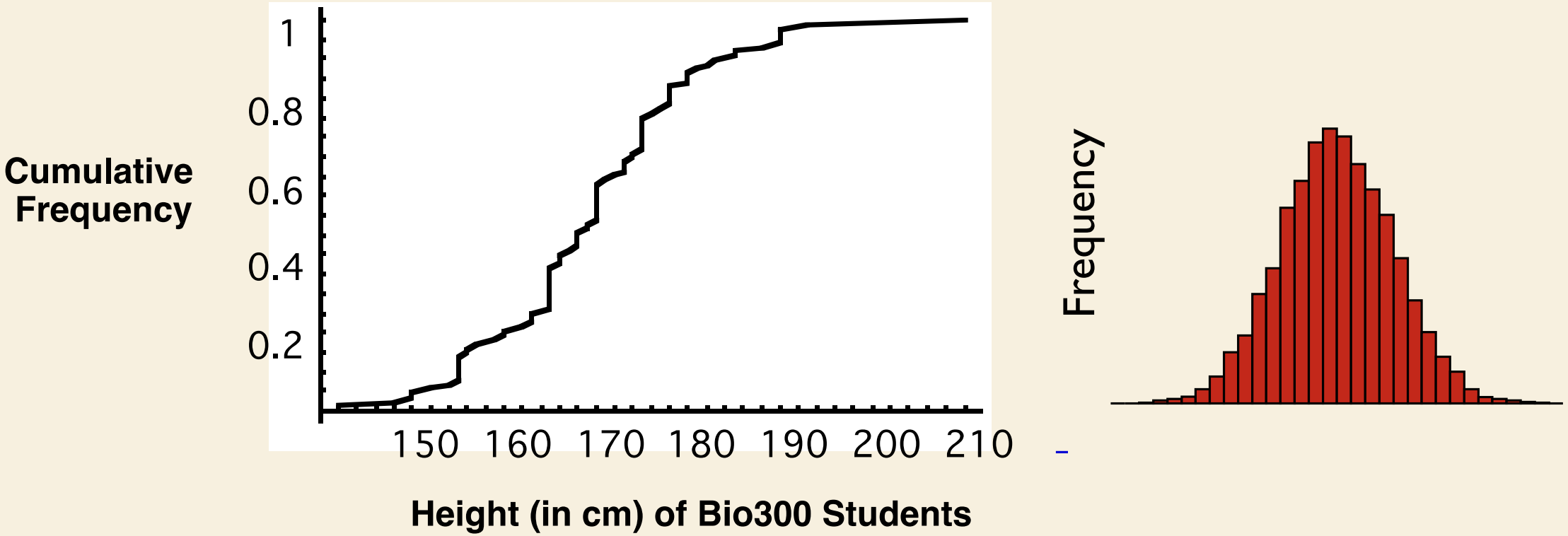

Cumulative frequency
1
0.8
0.6
0.4
0.2
150
160
170
180
190
200
210
Height (in cm) of Bio300 Students

Cumulative Distribution



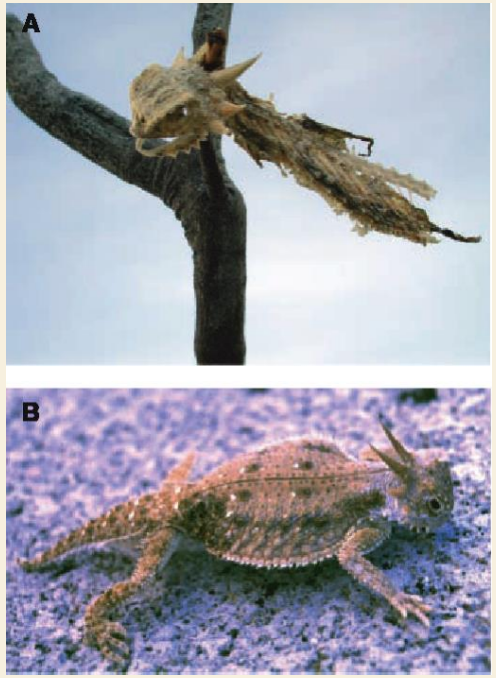
1. What is the **median** height?
2. What percent of students are less than 180 cm? 210 cm?
3. If I am 160 cm tall, what percentile am I?

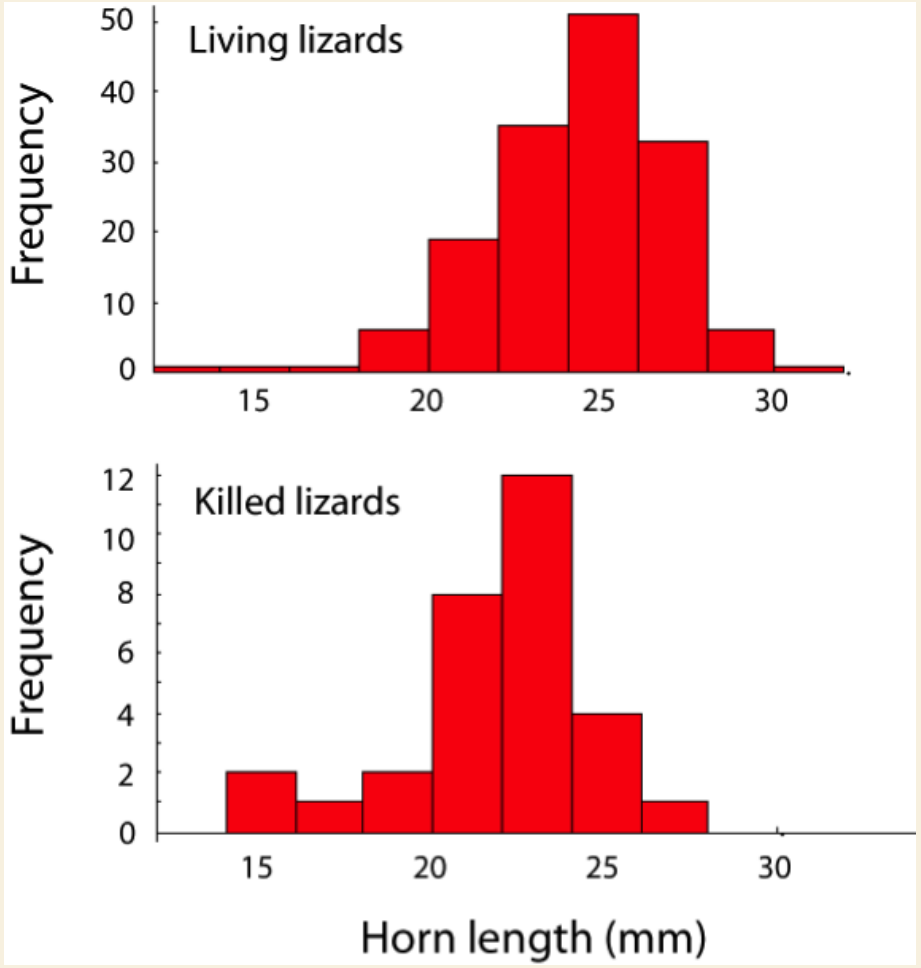
Cumulative Distribution vs. Probability Distribution



The cumulative frequency =
proportion of individuals equal to or less than that value

Multiple histograms

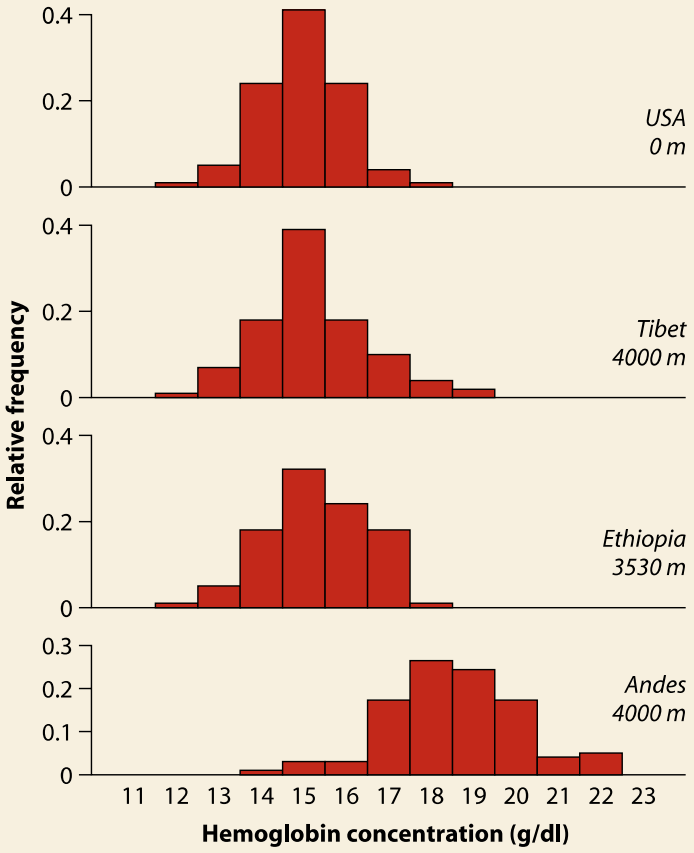


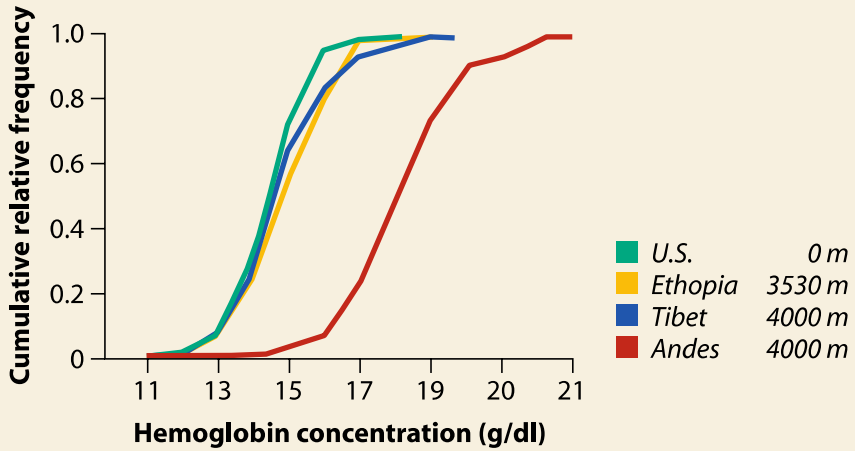


Shrike

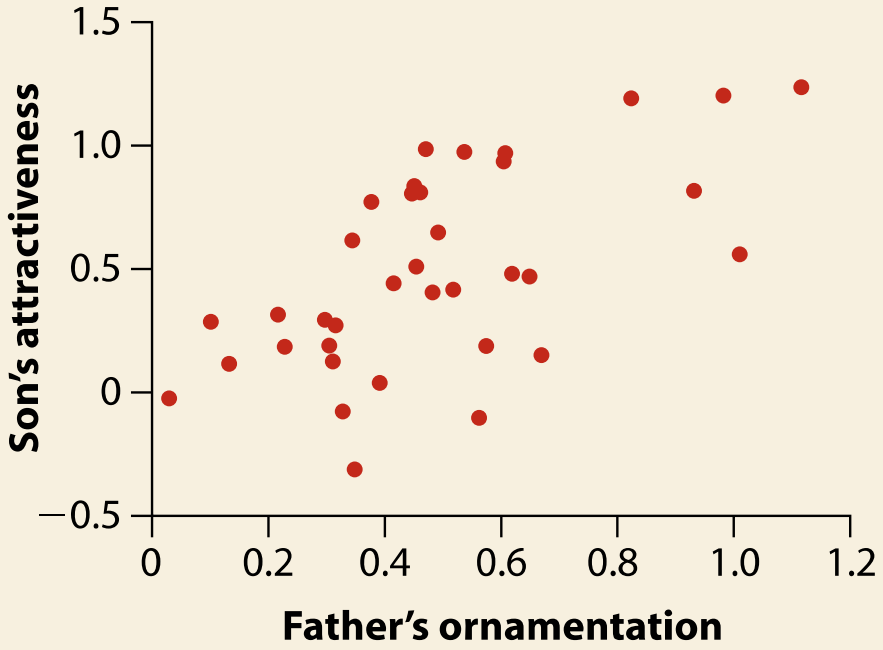
Young, K. V., E. D. Brodie Jr., and E. D. Brodie III. 2004. How the horned lizard got its horns. Science 304:65.

Two ways of comparing 4 distributions



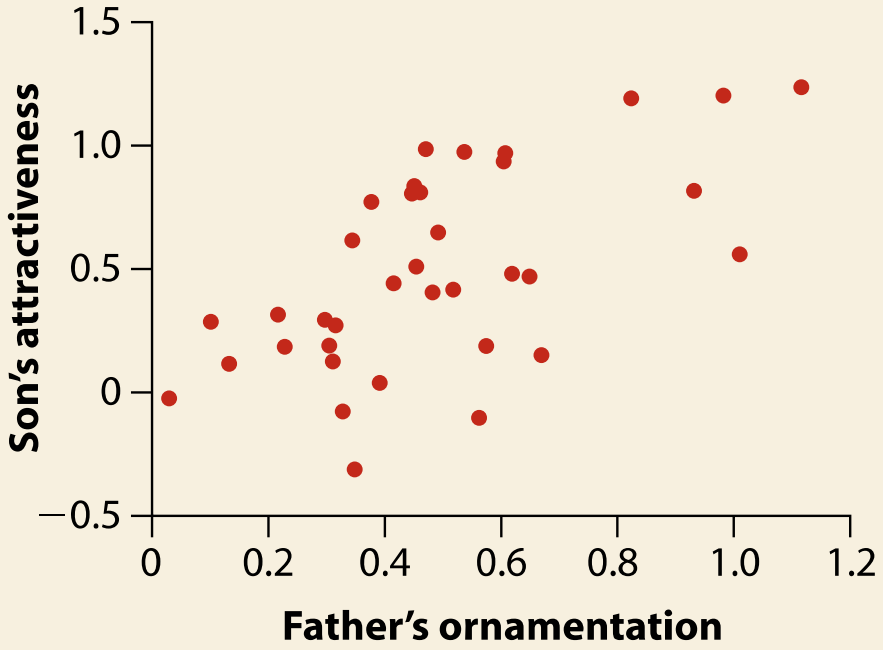


Beall et al. (2002)


Son's attractiveness
1.5
1.0
0.5
0
−0.5
0
0.2
0.4
0.6
0.8
1.0
1.2
Father's ornamentation

Brooks 2000

Scatter plot



Brooks 2000

CHECK 7. DESCRIBING DISTRIBUTIONS

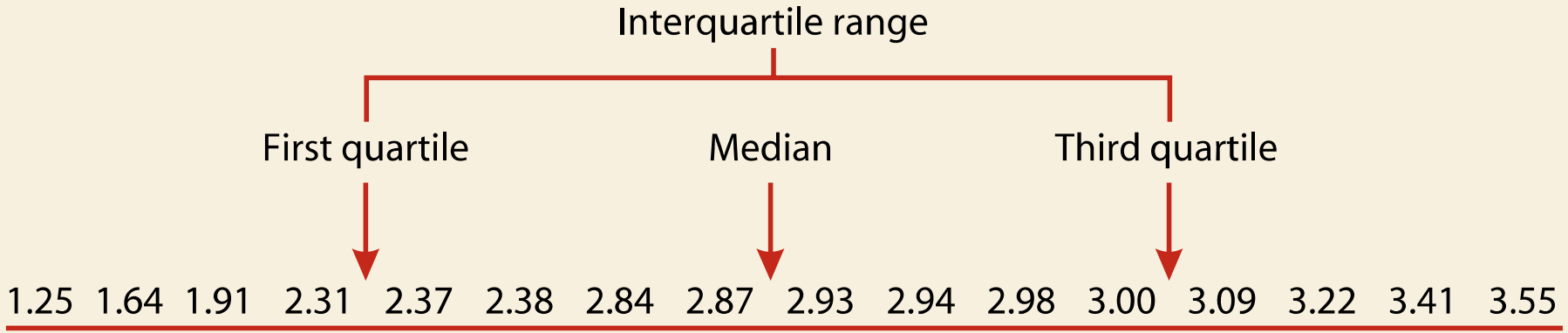
Center

- Mean
- Median
- Mode

Width

- Range
- Standard deviation (SD)
- Variance
- Coefficient of variation (CV)

RANGE


Interquartile range
First quartile
Median
Third quartile
1.25 1.64 1.91 2.31 2.37 2.38 2.84 2.87 2.93 2.94 2.98 3.00 3.09 3.22 3.41 3.55

VARIANCE IN A POPULATION

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

N is the number of individuals in the population.
μ is the true mean of the population.

SAMPLE VARIANCE

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n - 1}$$

n is the sample size

STANDARD DEVIATION (SD)

Positive square root of the variance

σ = true standard deviation

s = sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n - 1}}$$

$$s^2 = 0.70$$

$$s = \sqrt{0.70} = 0.84$$

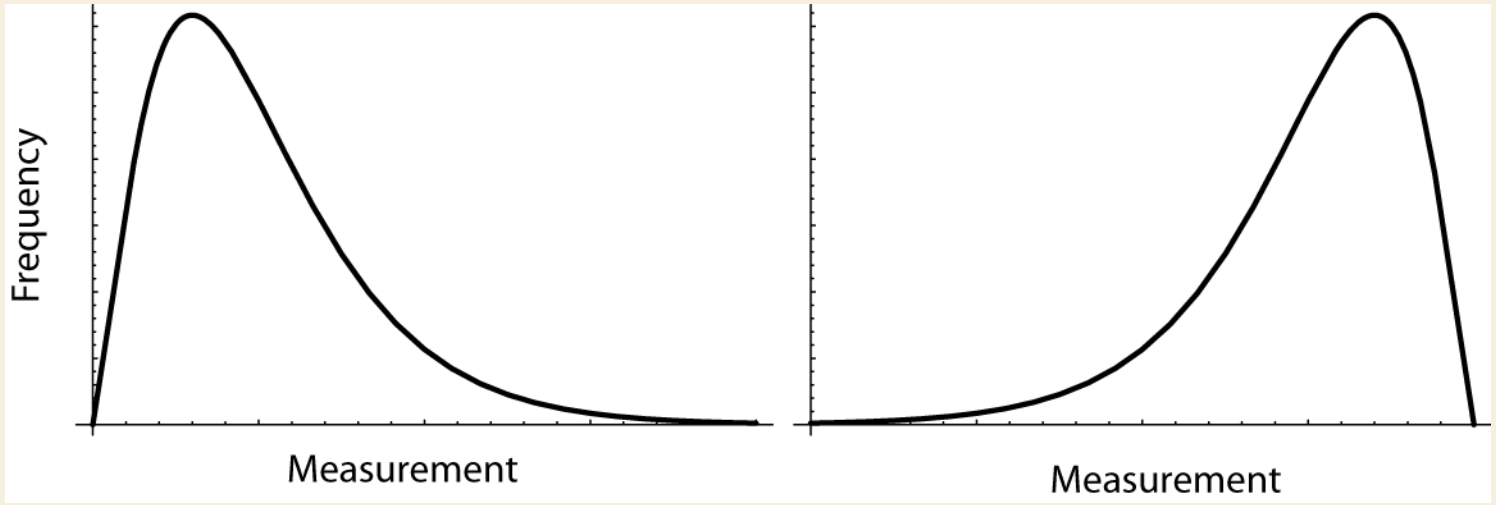
COEFFICIENT OF VARIATION (CV)

The *coefficient of variation* is the standard deviation expressed as a percentage of the mean.

$$CV = 100\% \frac{s}{\bar{Y}}$$

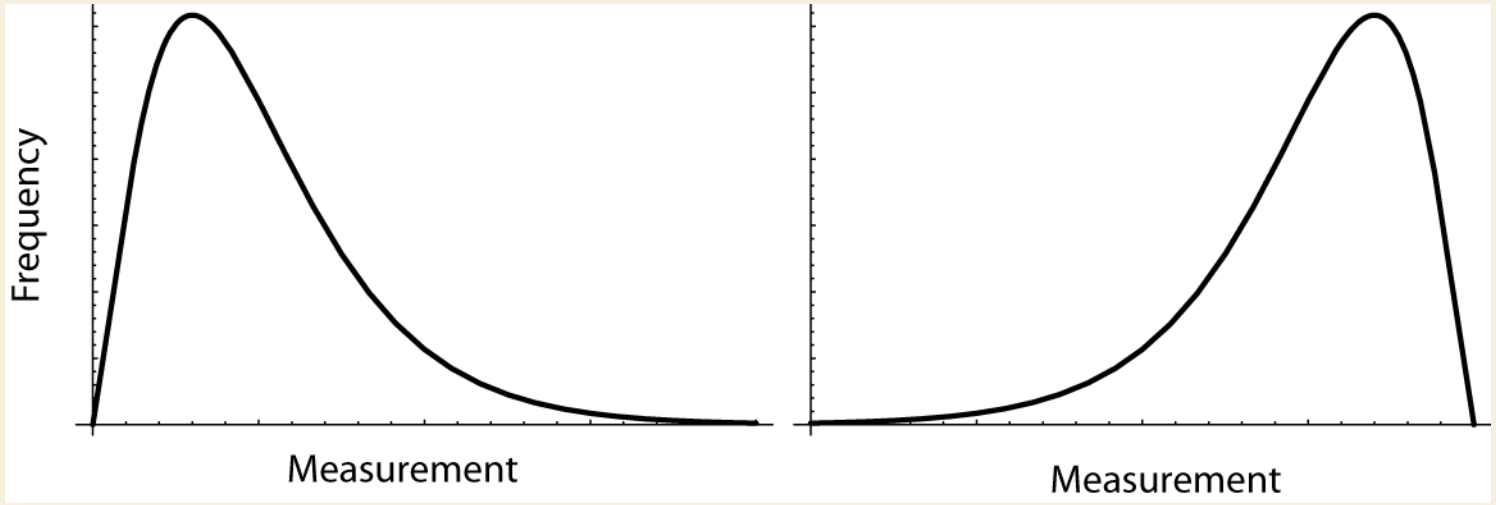
Why calculate CV?

What have we missed by only talking about center and width?



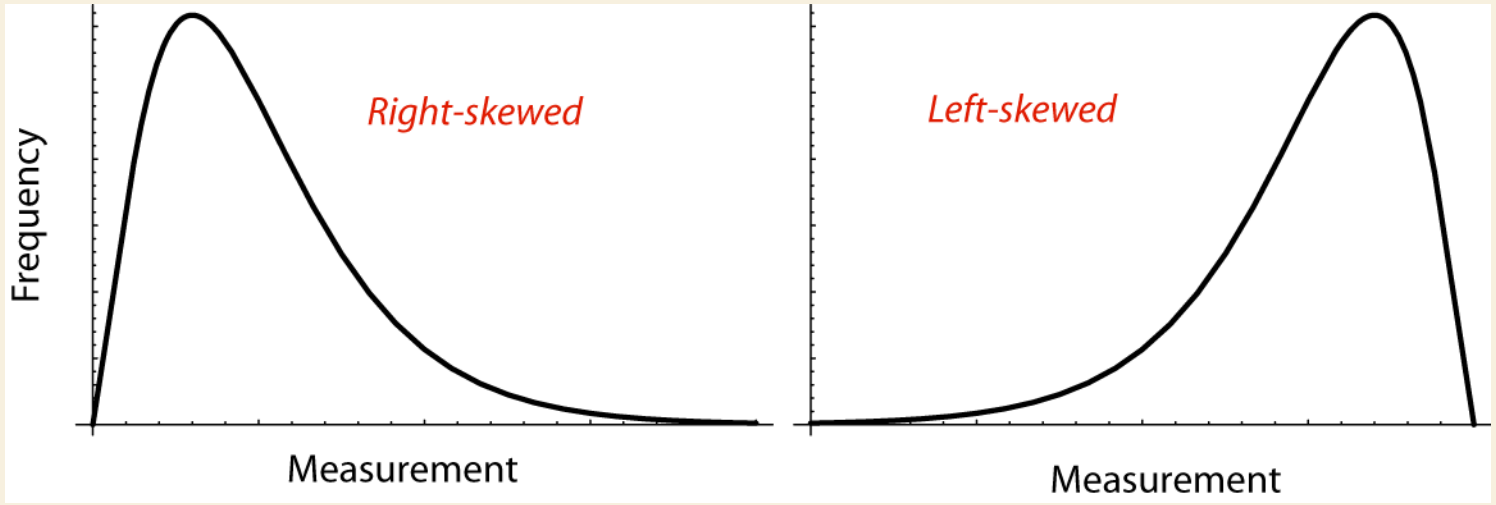
SKEW

- Measure of *asymmetry*
- Skew direction refers to the *pointy tail* of a distribution



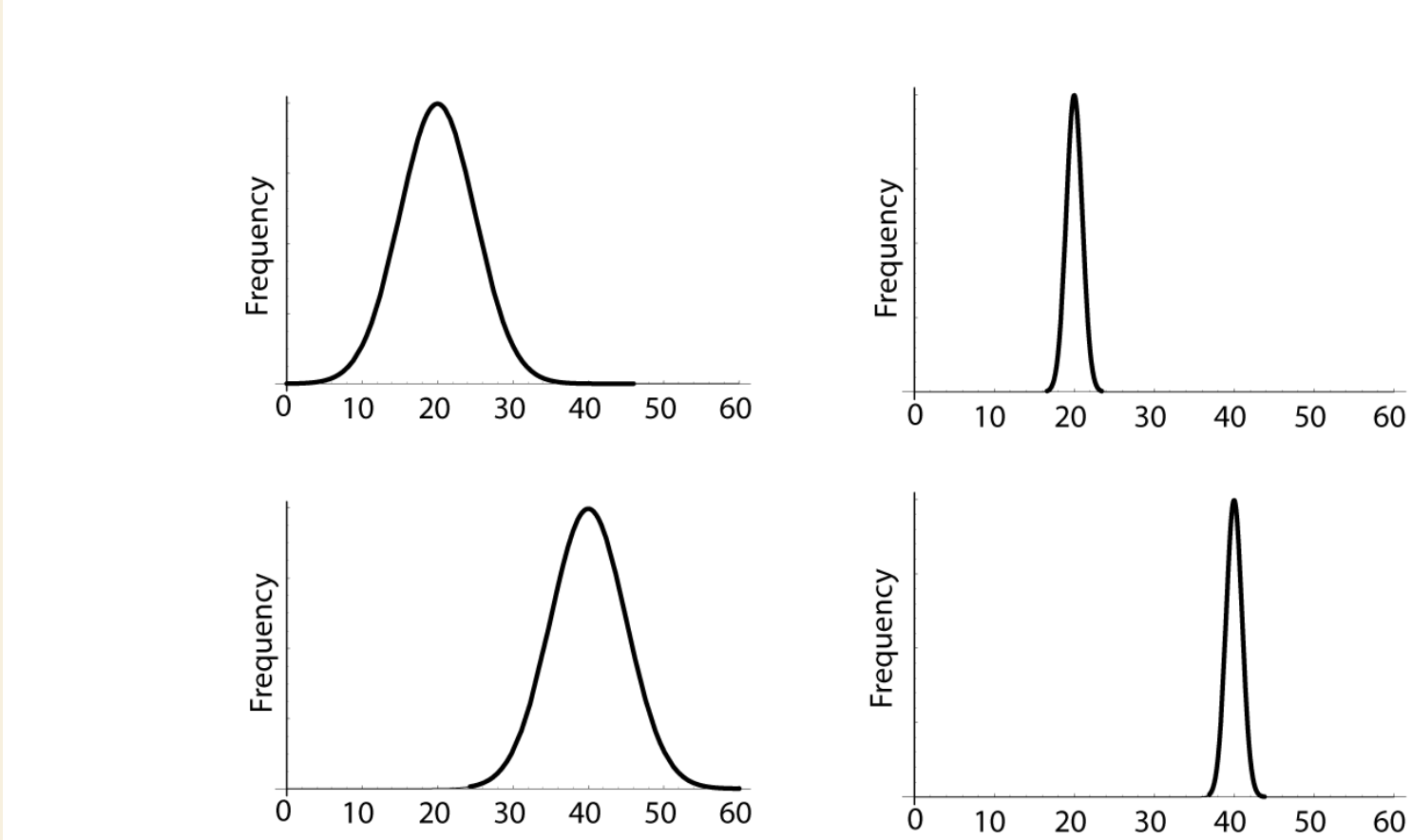
SKEW

- Measure of *asymmetry*
- Skew direction refers to the *pointy tail* of a distribution



CHECK 8.
NORMAL DISTRIBUTION

Describe these normal distributions



CHECK 8.
NORMAL DISTRIBUTION

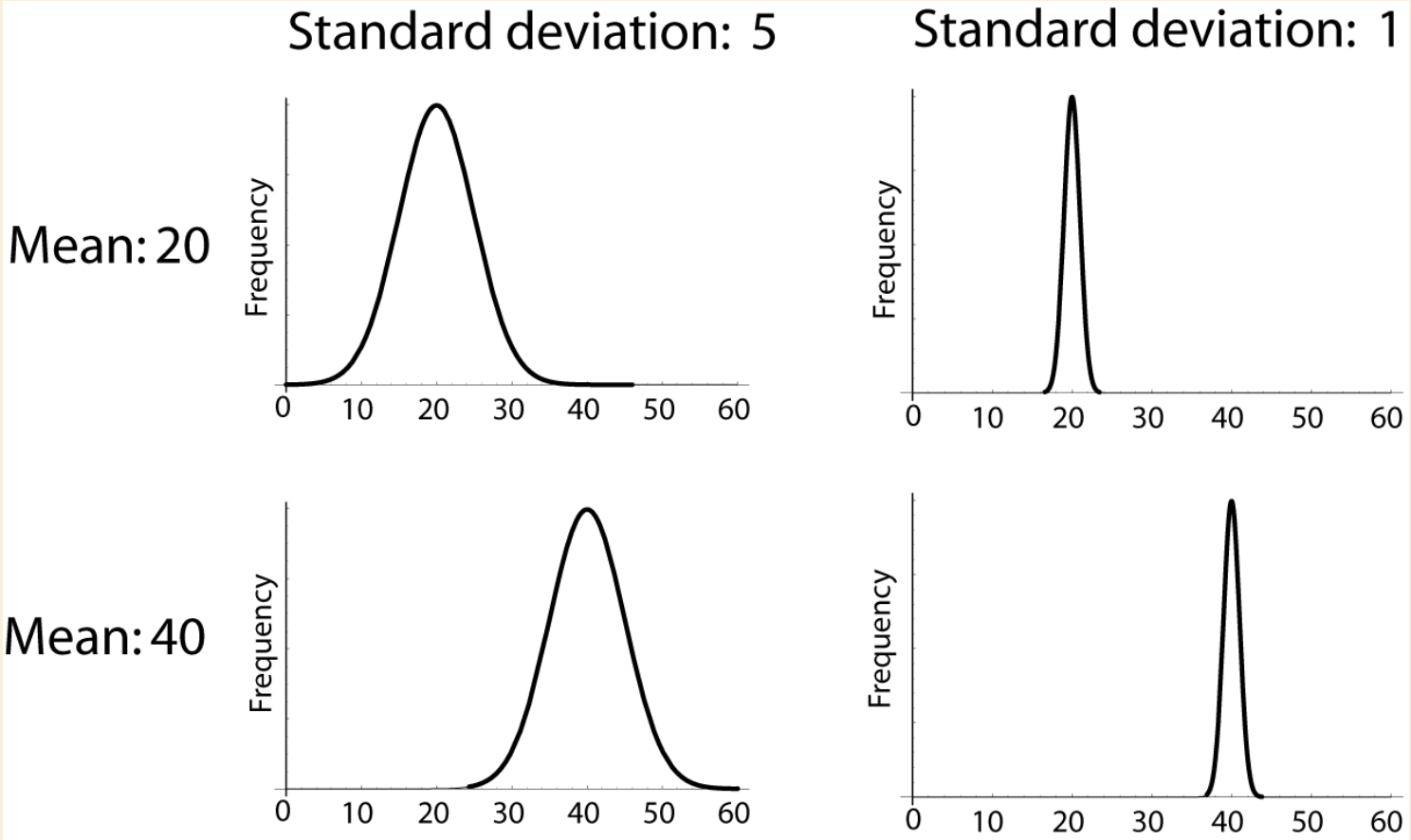
N (mean, sd)

N (20,5)

N (20,1)

N (40,5)

N (40,1)

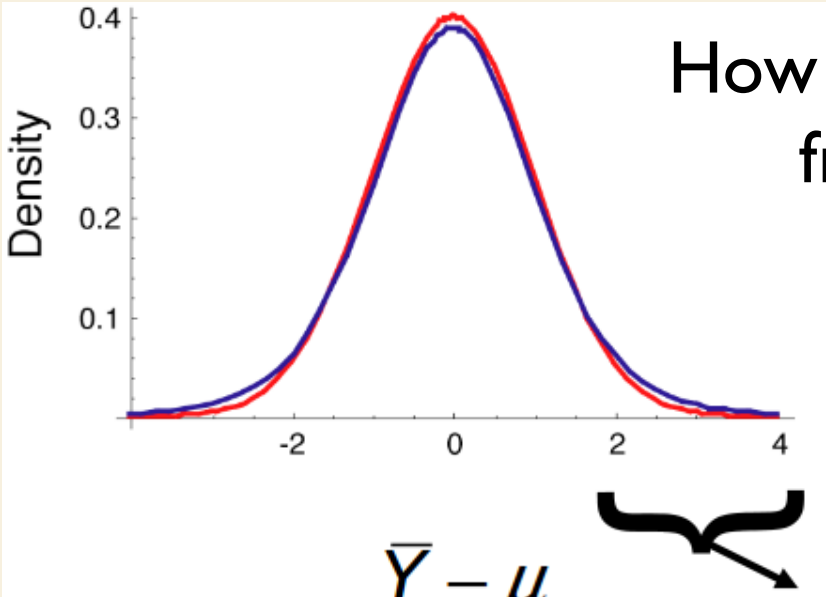


CHECK 9.
T DISTRIBUTION

The *one-sample* t-*test* compares the mean of a random sample from a normal population with the population mean proposed in a null hypothesis.

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

CHECK 9.
T DISTRIBUTION



How is a t-distribution different from a normal distribution?

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

CHECK 9.
T DISTRIBUTION