BIOL 300: Fundamentals of Biostatistics

Instructor: Darren Irwin
(Professor, Dept. of Zoology)

Statistics: possibly the most important subject you study at UBC

- Statistics is about how we can use information to infer something about **Truth**, while taking into account **Uncertainty**.
- Applicable in all fields.
- Vital for scientists, especially biologists (and medical professionals).
- Understanding of statistical principles is important for everyone.
  - Making decisions (e.g. medical / safety / environmental / purchasing)
  - Interpreting news reports, voting, etc.

My three goals for you

- As scientists, know how to design studies and do statistical analysis on your own data.
- Be able to evaluate whether other people have done statistics correctly.
- Become skilled at statistical thinking.

Statistical thinking *will one day be as necessary for efficient citizenship as the ability to read and write.*

-H.G. Wells (paraphrased)

BIOL 300: Fundamentals of Biostatistics

Course web site:

On UBC Canvas

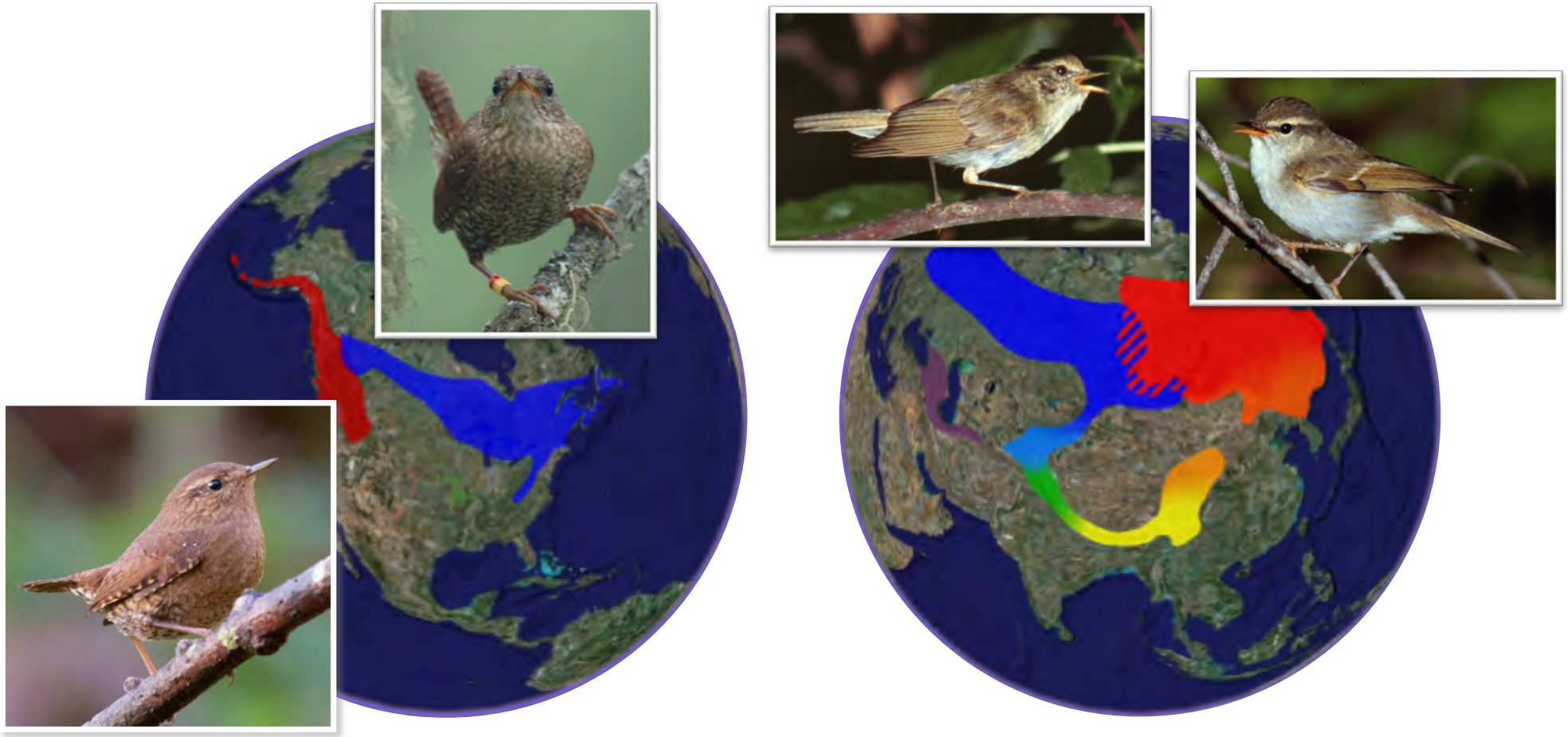
Instructor:

Darren Irwin
(Professor, Dept. of Zoology)

Office: 209 Biodiversity
(Beaty Biodiversity Research Centre)

e-mail: irwin@zoology.ubc.ca

Speciation in birds: lots of statistics!



Genes, plumage, body shape, habitat, migration

Also: population trends (for conservation)

My BIOL300 office hours: Mondays 10-11:30am
(Biodiversity 209)

Please feel free to come by to talk about BIOL300 material or anything else.

Also, available for short questions right after class most days (in hallway outside of lecture room).

Teaching Assistants
(all graduate students in biology)

- *Sydne Guevara Rozo*
- *Nicola Love*
- *Rashika Ranasinghe*
- *Ilan Rubin*
- *Seth Watt*

Please: Respect the TA's; Respect each other.

Respect each other

*Please **do***:

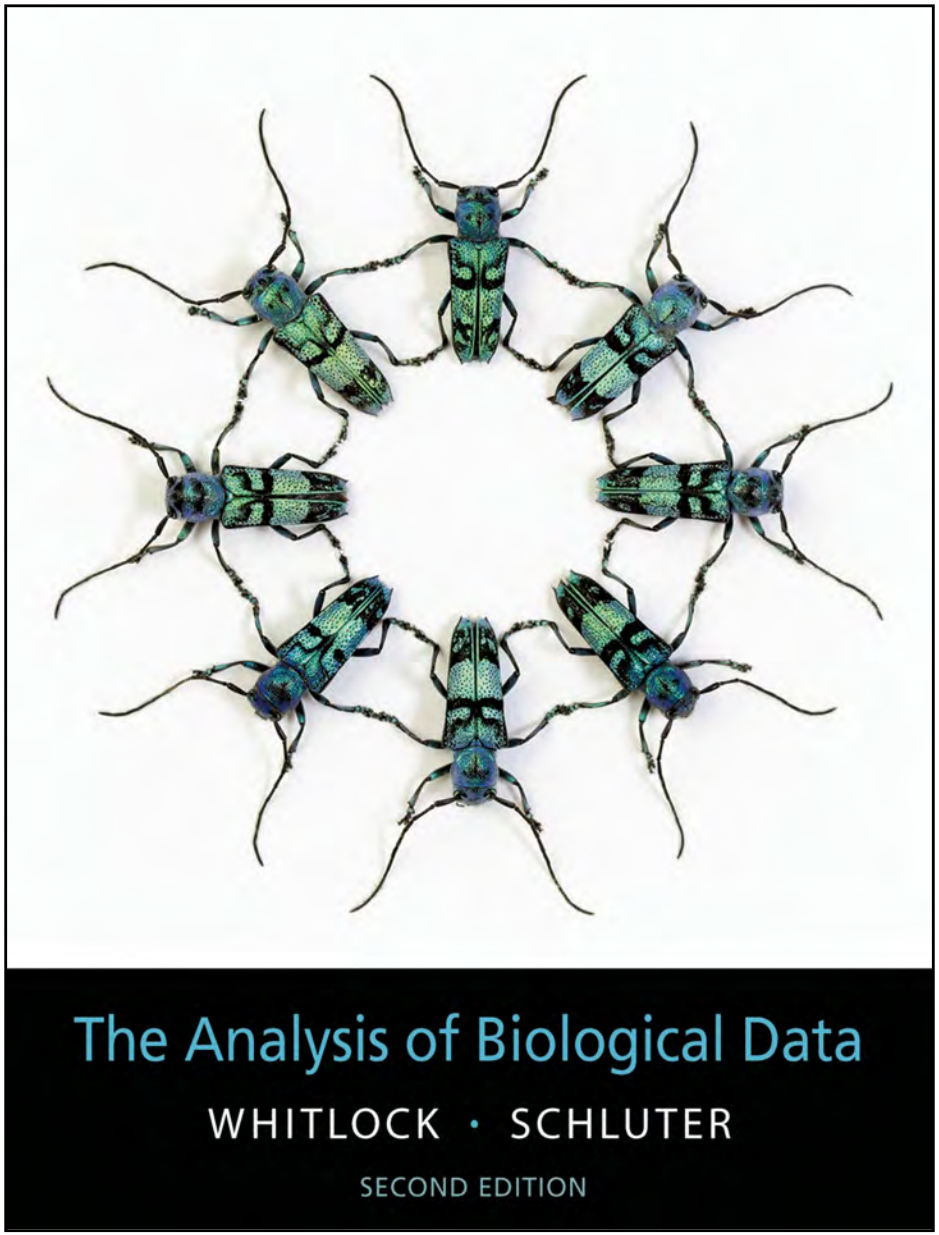
- Come to class ready to think about statistics.
- Participate in class, & in an appropriate way
  - Ask / answer questions during class, and also allow others to do so. (leave some questions for after class)

*Please **don't***:

- Distract your fellow students.
  - (e.g., texting, checking social media, fighting monsters, having conversations)

Textbook

- Whitlock and Schluter, *The Analysis of Biological Data*, ***2nd Edition***.



Lab manual

- Available at the course web site (on Canvas)

Lab sections

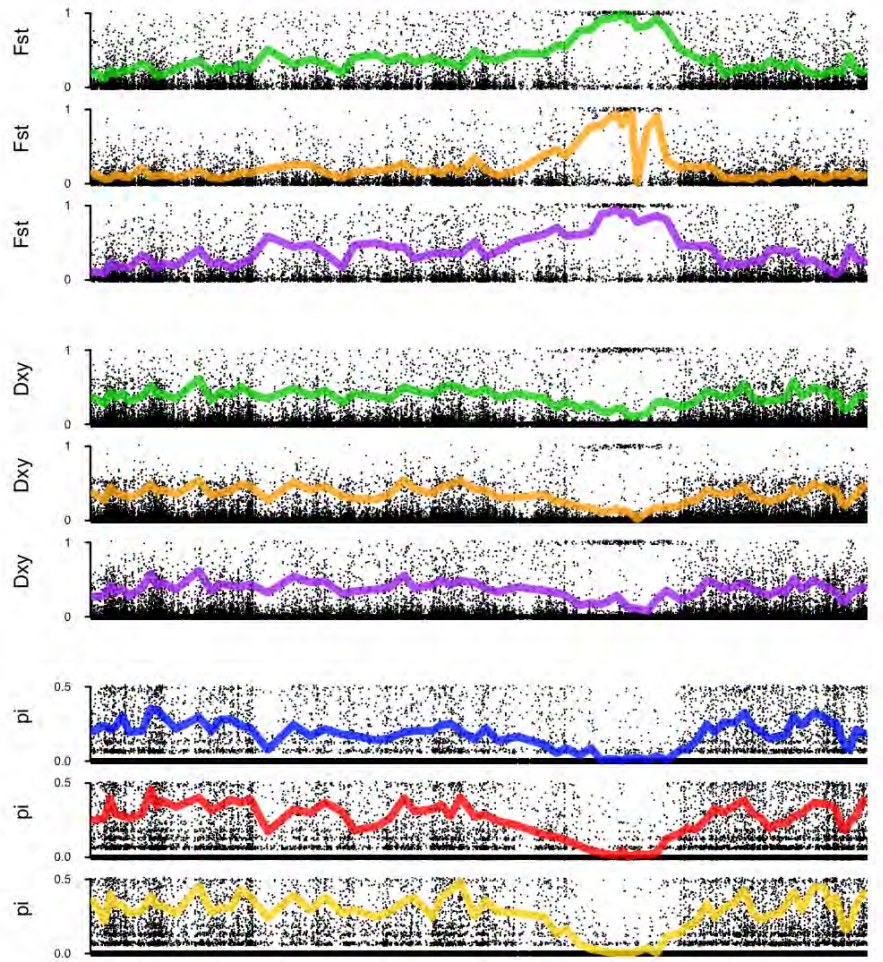
- Begin second week of term (January 13-17), in BioSciences room 2004
- Attendance is **strongly recommended** (and **required for some labs**)
- Great for learning from TAs, using R, and for doing lab assignments.

- “R is a language and environment for statistical computing and graphics”
- Used widely by scientists
- Used in the BIOL 300 labs
- Available for free download to your own computer (or use the lab computers)
  - To do so, follow links from the course website

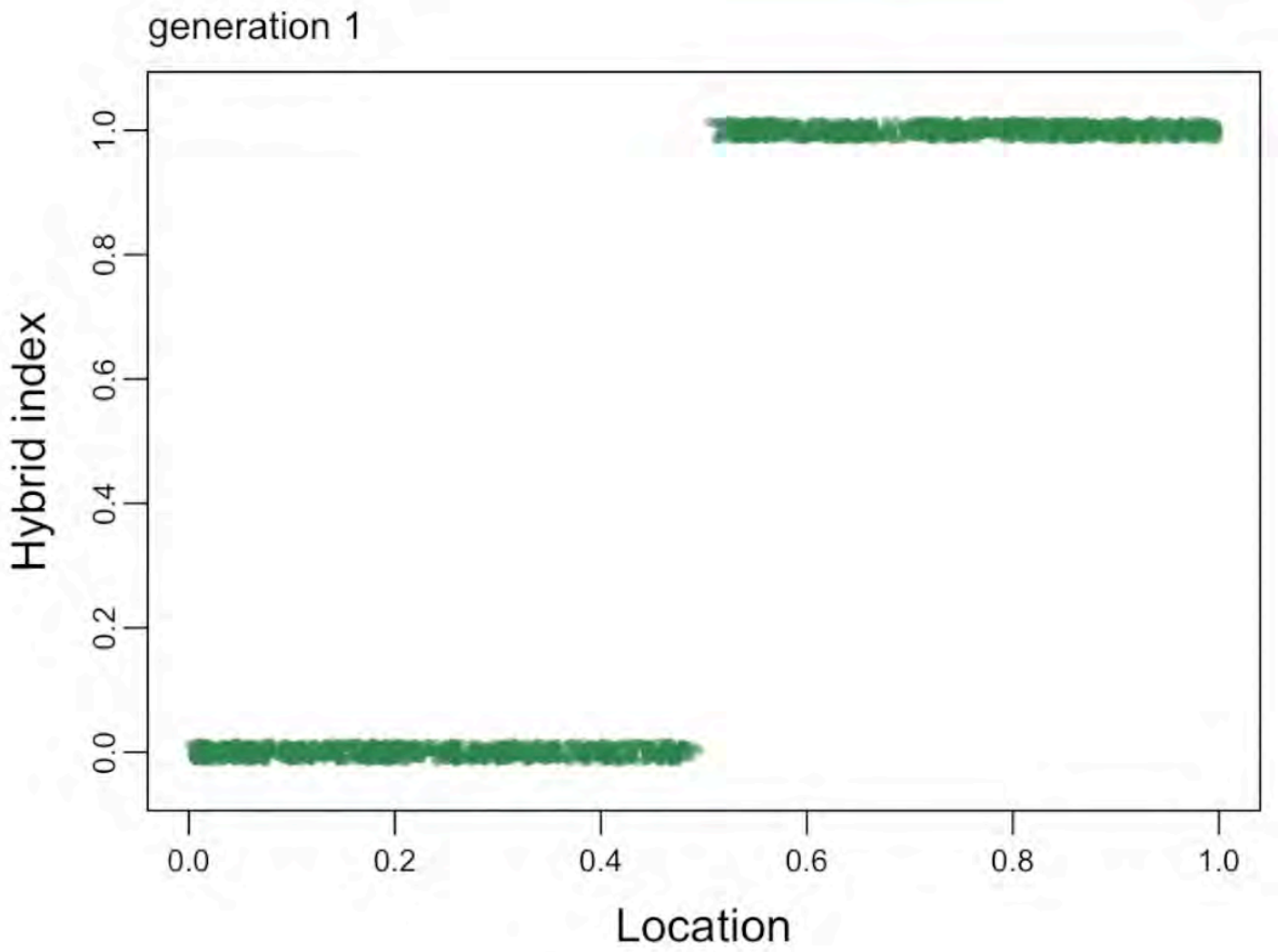
R can do simple things, and very complex things too

```
> 1 + 2
[1] 3
```

```
> x <- c(1, 3, 7)
> mean(x)
[1] 3.666667
```



An example of a simulation written in R: hybridization between two species



Homework Assignments

- Intended to help you learn
- Most questions are in textbook (numbers will be listed on website); some from lab material.
- Assigned each Friday.
- Due following Friday exactly at noon (12pm), in your TA's box (at entrance to Stats Lab: BioSci 2004). Feel free to turn in early (even days early), but not later than 12pm Friday.
- First assignment due Jan. 17th:
  - Chapter 1, problems 14, 15, 19, 21
  - Chapter 2, problems 20, 22a-d, 23, 24

Evaluation

Homework assignments 10%

Lab assignment(s) 10%

Mid-term 30%

Final 50%

Policy on academic honesty:

Your performance on the exams, homework, and assignments is expected to reflect your own work. Copying another's work or allowing your work to be copied can result in suspension or expulsion from UBC.

Midterm

March 2nd, in class

Wait list

- If you are on the wait list, we'll have to see how much room opens up in the next week.
- If you are not registered, try to register for the wait list. If you have questions about the registration process, contact the Biology Program Office in person or email Tammy Tromba: tromba@zoology.ubc.ca.
- If you do not want to take the course, please de-register yourself (make room for others).

Other UBC Statistics Courses

- Credit given for only one of BIOL 300, FRST 231, PSYC 218, 366, STAT 200

These other courses are paired with BIOL 300, but *do not count* as Biology courses (usually, Biology majors should take BIOL 300).

Check with your academic advisor for details.

Origin of “Statistics”

In the 1700’s, the term “*Statistics*” was used to describe the collection and analysis of demographic and economic data by states (i.e., governments).

The term was gradually applied to any sort of data from a population.

In biology, the need for analyzing data from variable populations led to great advances in statistical methods.

Introduction to Statistics

Statistics is the study of methods to describe and measure aspects of nature from samples. It provides tools to quantify the **uncertainty** of these measures, allowing us to determine their likely magnitude of departure from the **truth**.

Statistics become necessary when you have limited information (just a **sample**), but want to infer something about reality more generally (i.e., about a **population**).

Goals of statistics

- Estimate the values of important parameters

- Test hypotheses about those parameters

Parameter: a characteristic of a population.

Statistics is also about good scientific practice

Feline High-Rise Syndrome (FHRS)

The injuries associated with a cat falling out of a window.

"The diagnosis of high-rise syndrome is not difficult. Typically, the cat is found outdoors, several stories below, and a nearby window or patio door is open."

Two veterinarians decided to examine data from cats brought in to their clinics in New York City . . .

http://www.petplace.com/cats/highrise-syndrome-in-cats/page1.aspx

High falls reported to show *lower* injury rates

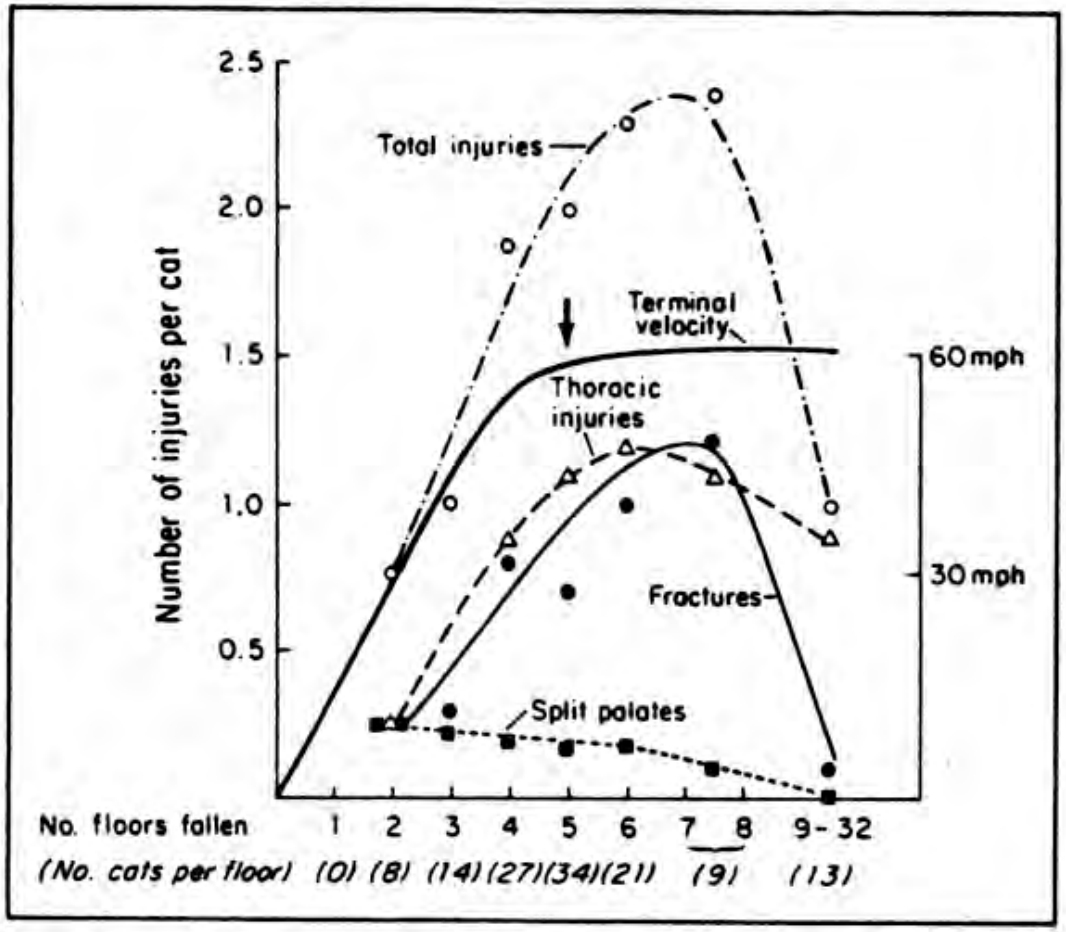


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (——); total number of injuries/cat (○, –·–·–·); number of thoracic injuries (pulmonary contusions +pneumothorax)/cat (△, – – –); number of fractures/cat (●, ——); number of split palates/cat (■, – – – – –).

Whitney and Mehloff, *Journal of the American Veterinary Medicine Association*, 1987

Why?

1. Cats have high surface-to-volume ratios
2. Cats have excellent vestibular systems
3. Cats reach terminal velocity quickly, relax, and therefore absorb impact better
4. Cats land on their limbs and absorb shock through soft tissue

Jared Diamond, *Nature* 1988

Why?

1. Cats have high surface-to-volume ratios
2. Cats have excellent vestibular systems
3. Cats reach terminal velocity quickly, relax, and therefore absorb impact better
4. Cats land on their limbs and absorb shock through soft tissue

Jared Diamond, *Nature* 1988

Or not…

Sample of convenience:
a collection of individuals that happen to be available at the time.

A newer study reports more injuries with longer falls

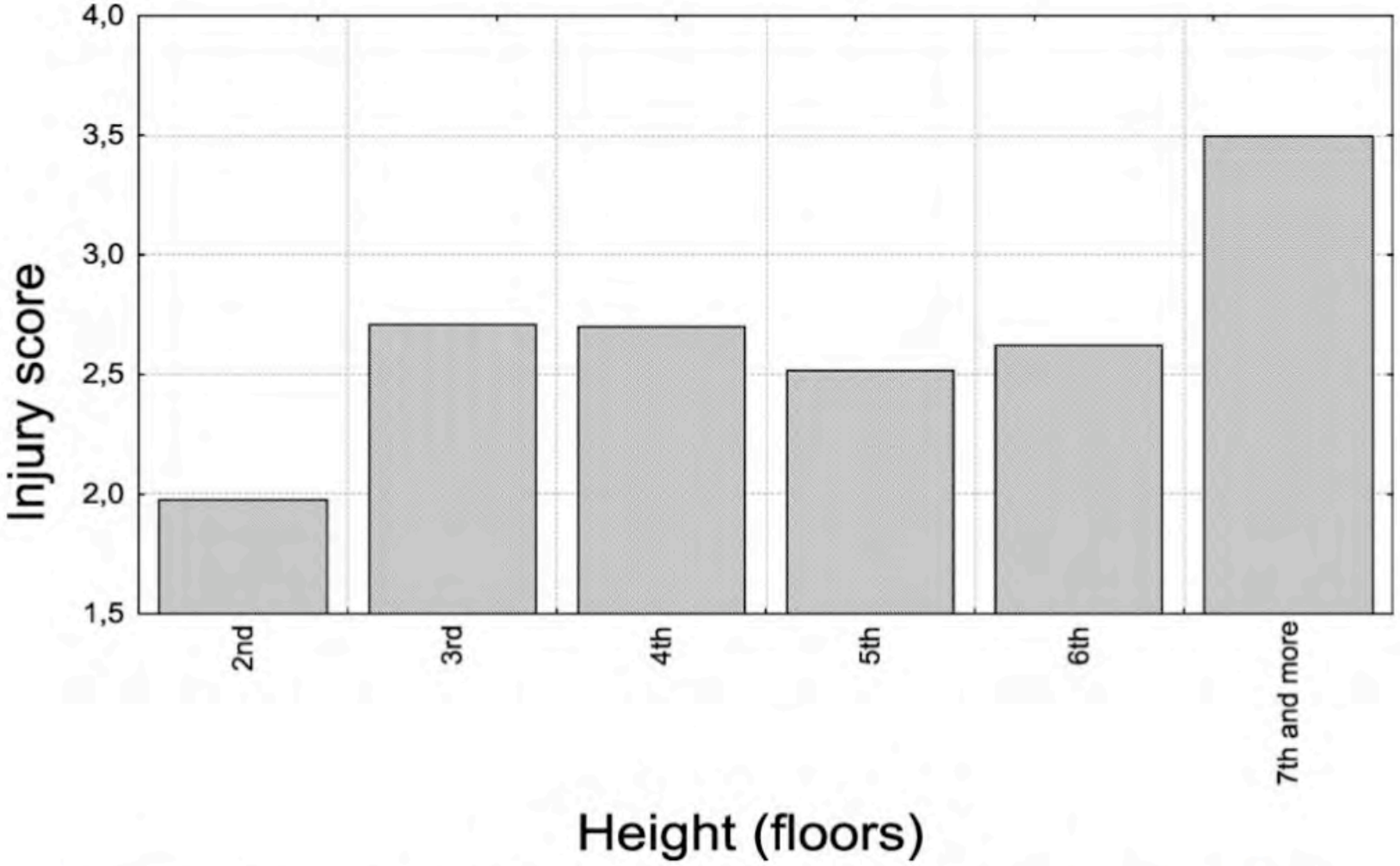


Figure 5 Graph showing the relationship between injury score and height of fall.

Vnuk et al. 2004. Feline high-rise syndrome: 119 cases (1998-2001). *J. Fel. Med. Surg.* 6:305-312.

FHRS illustrates importance of:

- Unbiased sample
- Large sample size
- Replication of studies
- Careful choice of variables measured
  - Are they really what you want to know?
- Careful interpretation of data

Let's collect some data . . .

On an index card, please write (all anonymous and optional):

a) Your height (indicate inches or cm)
b) Number of siblings you have (include half sibs)
c) # of cups of coffee or tea consumed today
d) Your favorite color
e) Length of your commute this morning (in minutes)
f) Type of transportation used today (e.g., walk, bike, car, bus)
g) A random integer from 1 to 5

Read:

Chapters 1 and 2

(If you don't yet have textbook, ch1-3 are available as PDFs on website.)

In future weeks, read each chapter around the time I cover it in lecture, (and/or when I assign homework from it).

Variables and Data

- A **variable** is a characteristic measured on individuals drawn from a population under study.

- **Data** are measurements of one or more variables made on a collection of individuals.

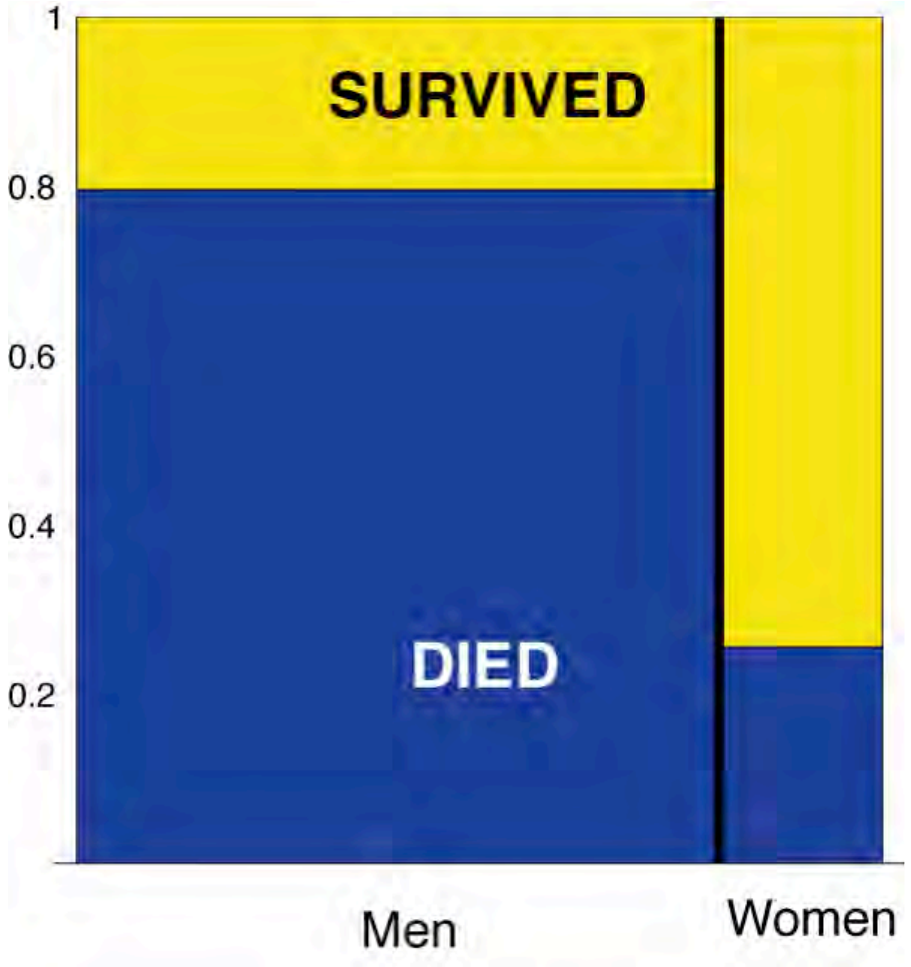
Explanatory and response variables

We try to predict or explain a **response variable** from an **explanatory variable.**

Older terminology:

dependent variable and *independent variable*

Mortality on the *Titanic*, as predicted by sex



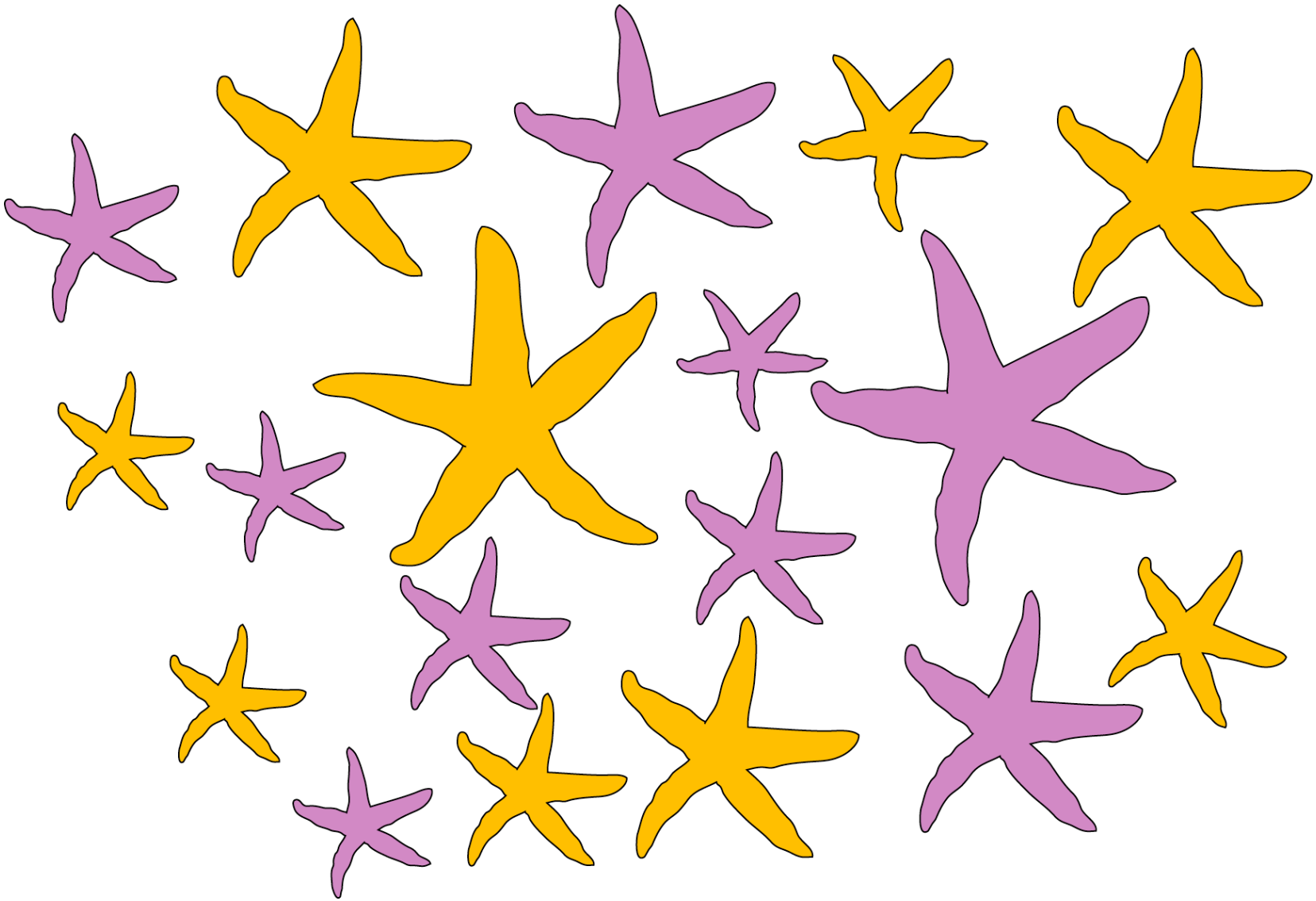
Populations and samples

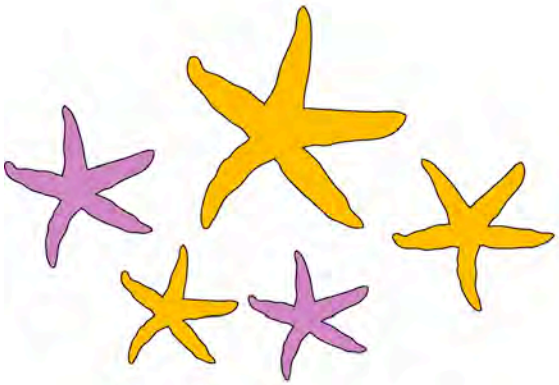
Populations <-> Parameters;
Samples <-> Estimates

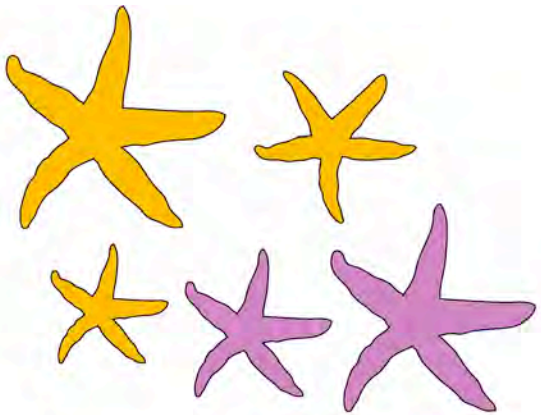
Estimates almost always differ from Parameters, for a variety of reasons . . .

Pisaster sea stars

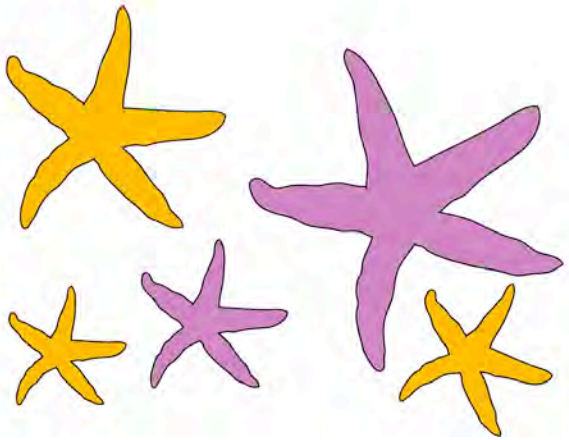
A population of starfish



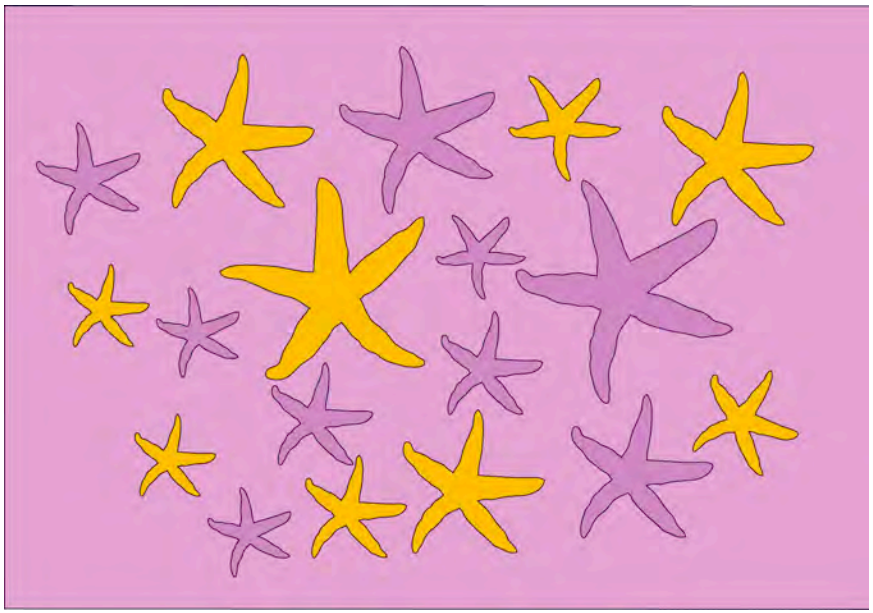


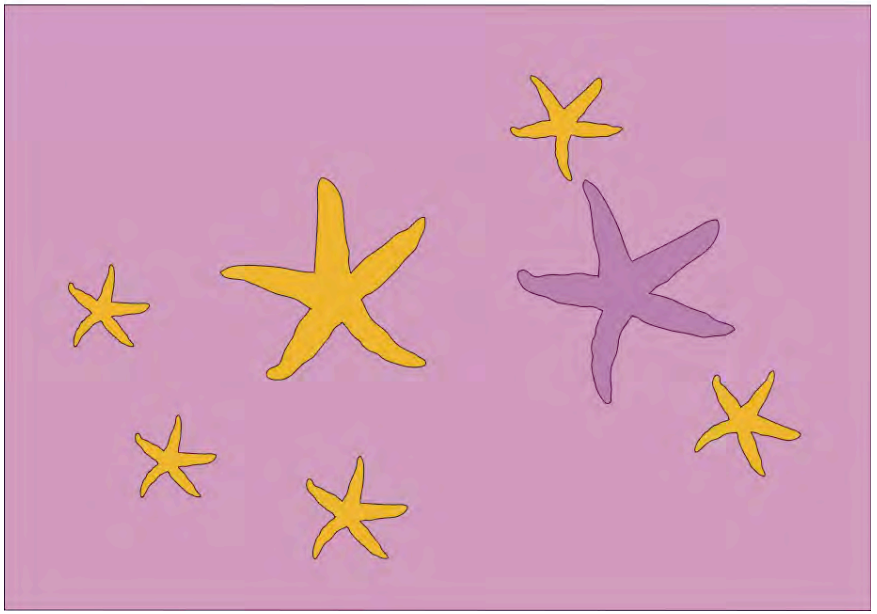


Random samples
of 5 starfish



A biased sample





Bias refers to the tendency of a measurement process to over- or under-estimate the value of a true population characteristic.

If estimates tend to differ from a parameter in a certain direction, the estimate is *biased*.

The 1936 US presidential election

Alf Landon Campaign Poster, 1936

vs.

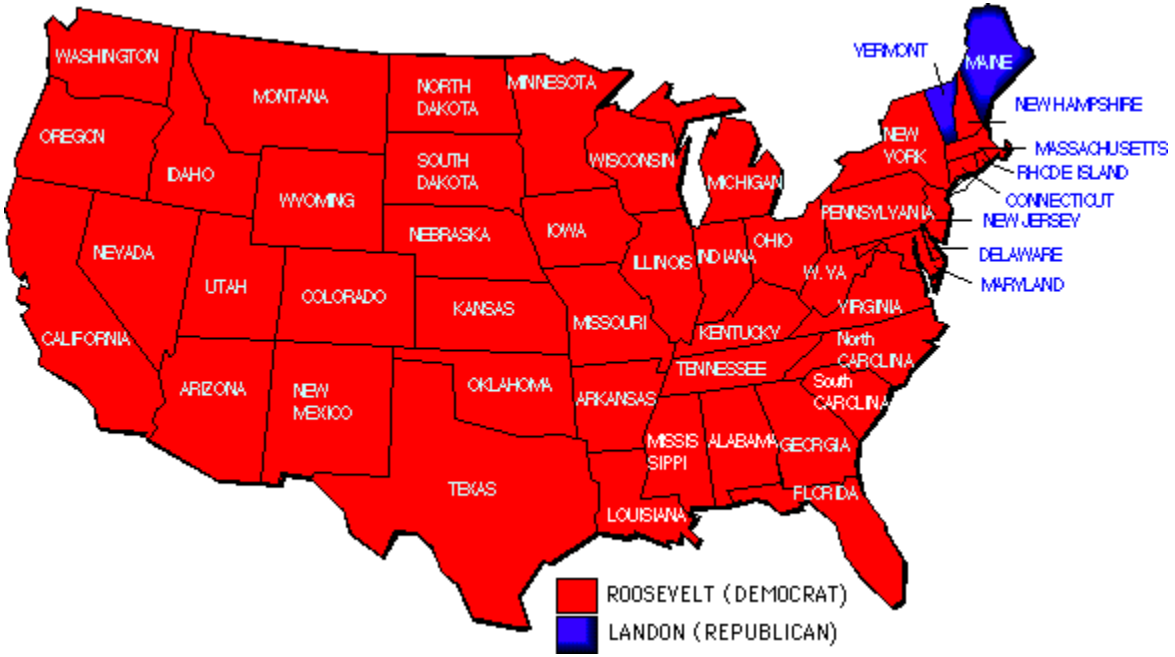
Alf Landon
Republican

Franklin Roosevelt
Democrat

1936 *Literary Digest* Poll

- 2.4 million respondents
- Based on questionnaires mailed to 10 million people, chosen from telephone books and club lists
- Predicted Landon wins: Landen 57% over Roosevelt 43%

1936 election results



Roosevelt won with 62% of the vote

What went wrong?

Subjects given the questionnaire were chosen from telephone books and clubs, biasing the respondents to be those with greater wealth

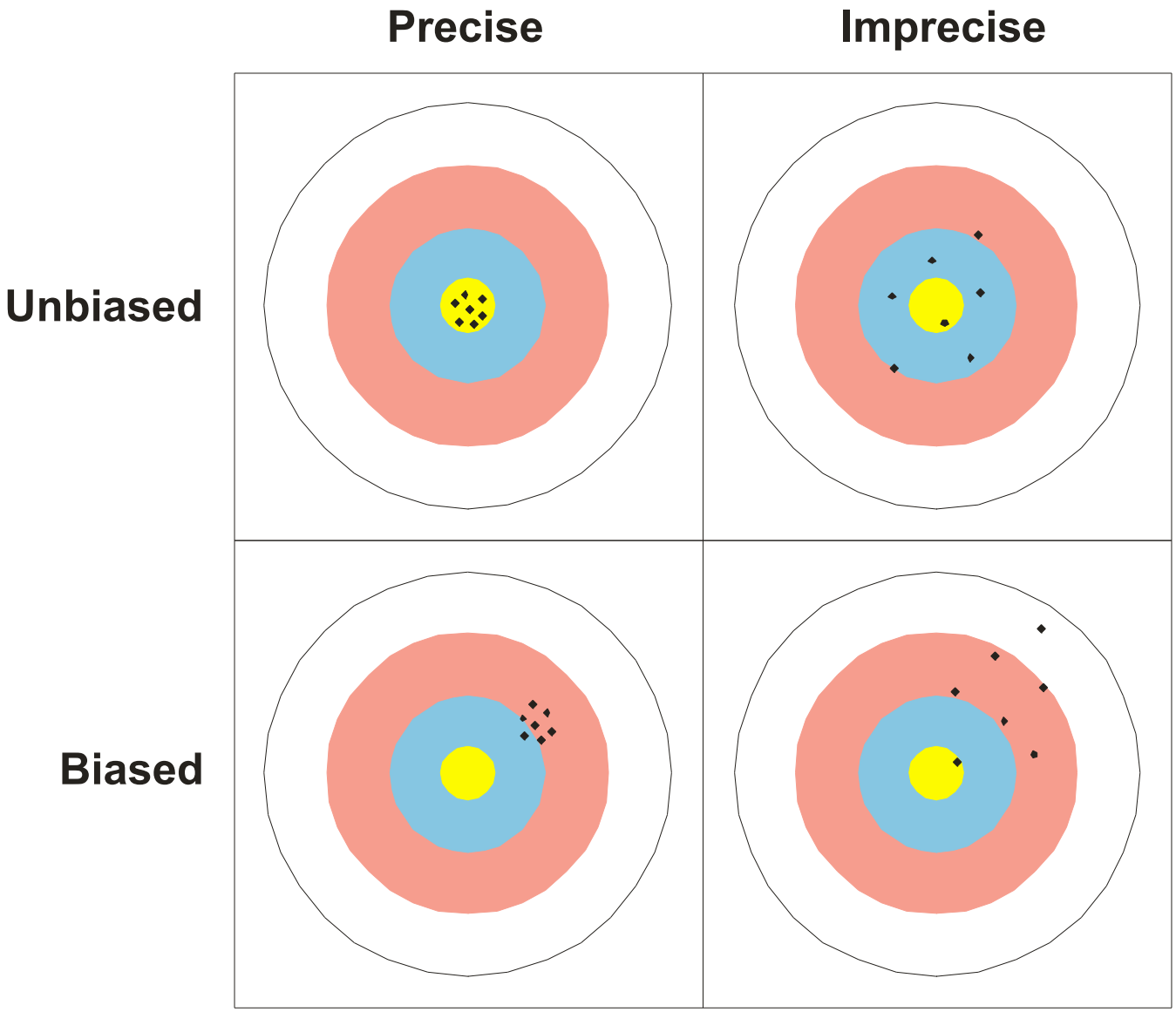
Voting and party preference is correlated with personal wealth

Volunteer bias

Volunteers for a study are likely to be different, on average, from the population

For example:

- Volunteers for sex studies are more likely to be open about sex
- Volunteers for medical studies may be sicker than the general population



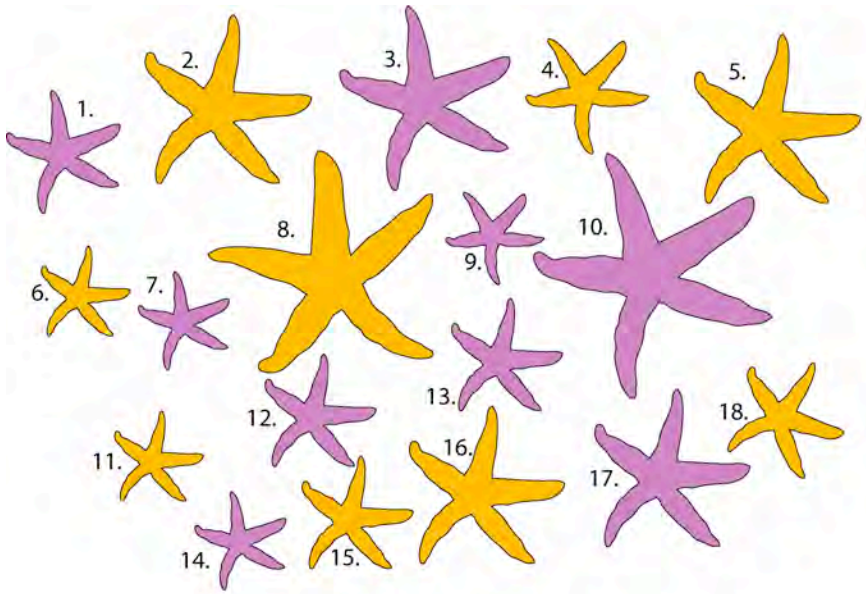
Each point represents an estimate of a parameter.

Properties of a good sample

- Random selection of individuals (each individual has equal probability of being selected)
- Independent selection of individuals
- Sufficiently large

In a *random sample*, each member of a population has an equal and independent probability of being selected.

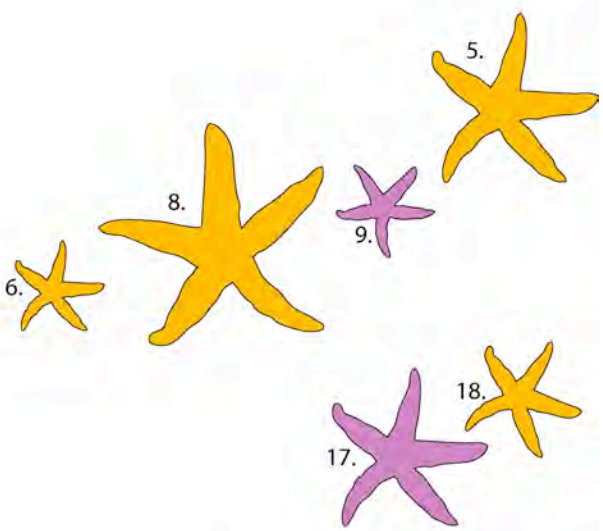
One procedure for random sampling



Number each individual

18, 6, 8, 5, 9, 17

Choose random numbers



Sample those individuals with matching numbers

Population parameters are *constants* whereas estimates are *random variables*, changing from one random sample to the next from the same population.

Sampling error

- The chance difference between an estimate and the population parameter being estimated.

  (note that sampling bias is not included here)

The good news:

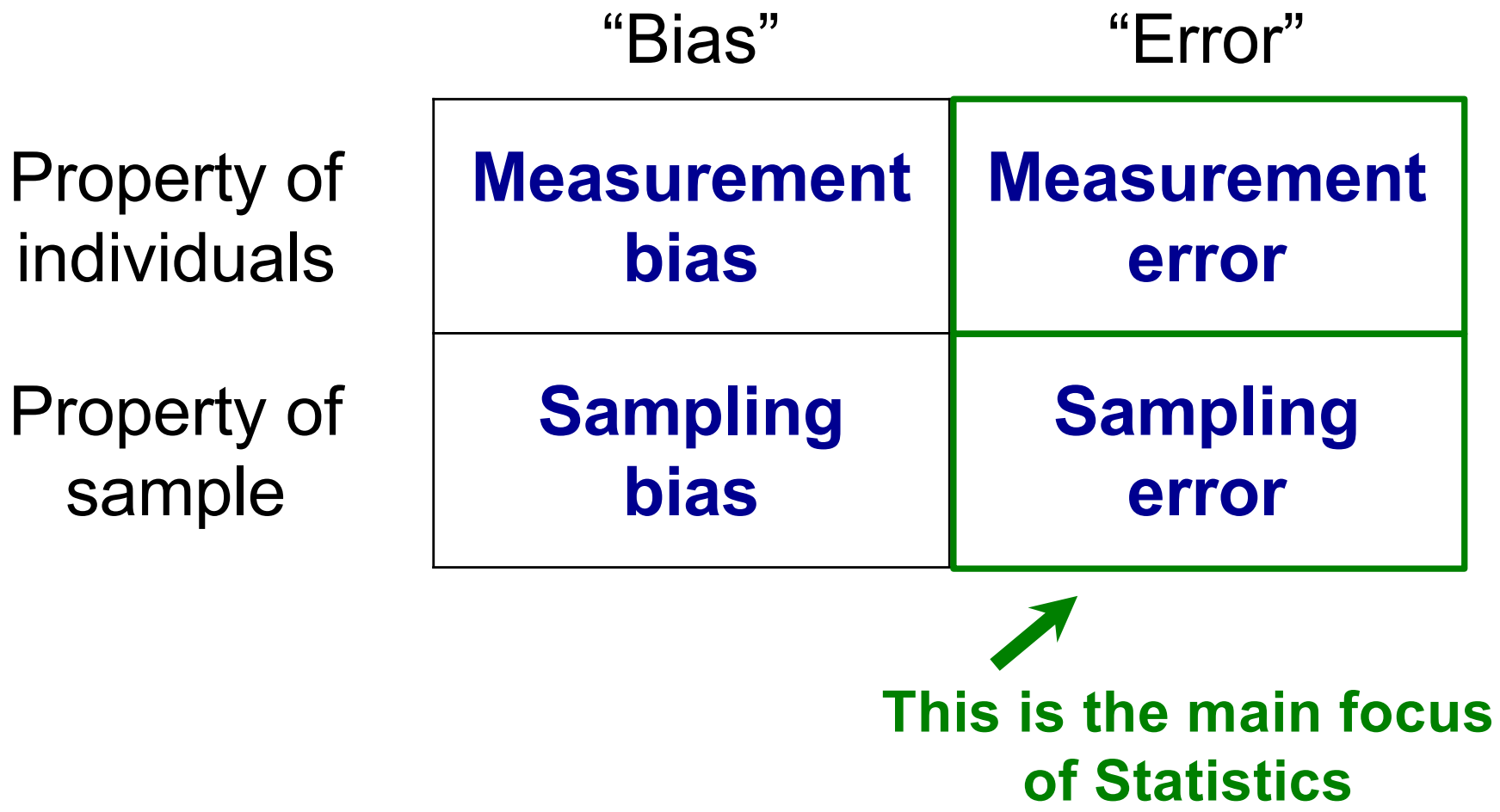
We can estimate the magnitude of sampling error using properties of the sample.

Bias vs. error

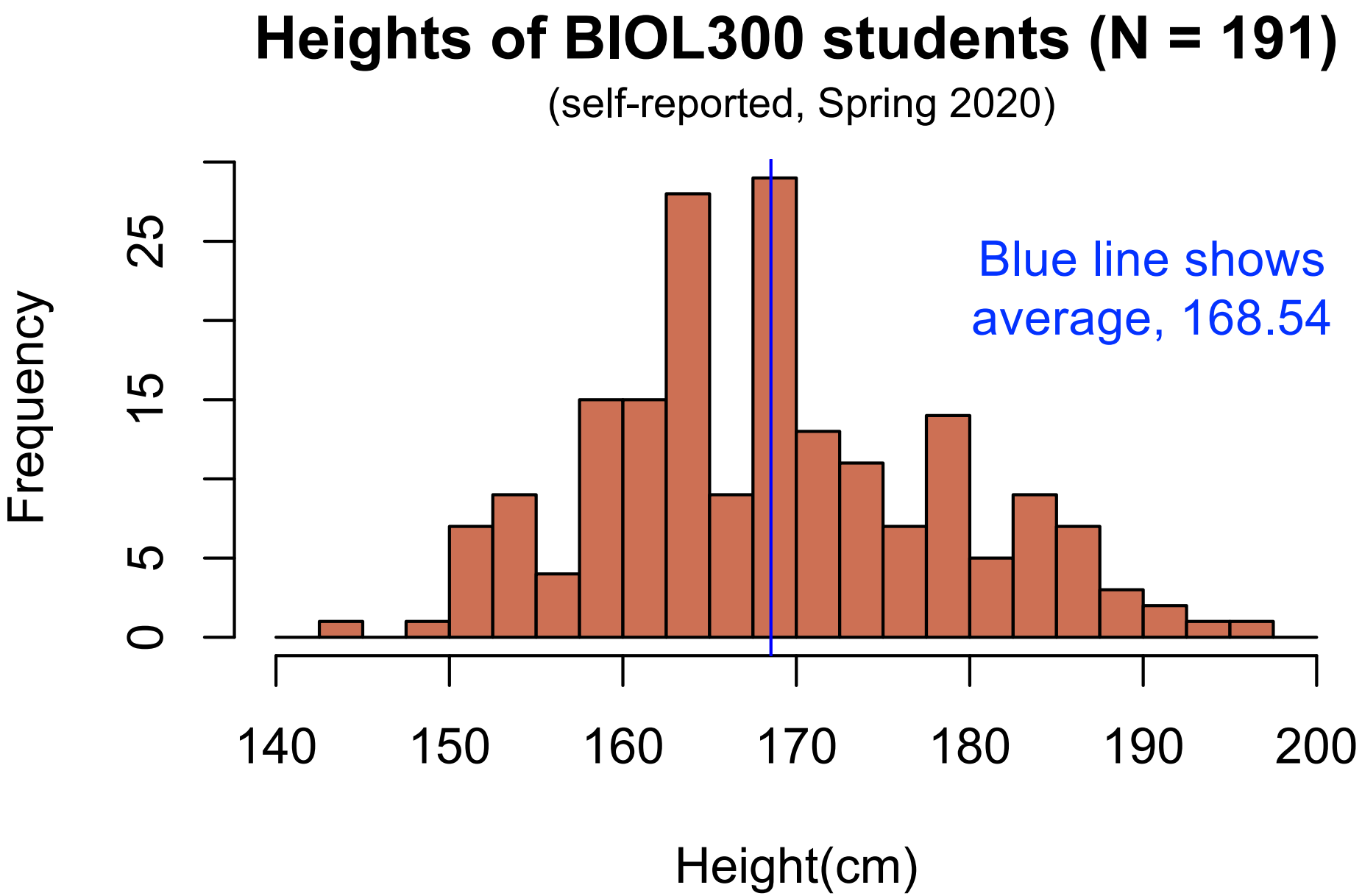
Bias is a systematic discrepancy (tending in a certain direction) between an estimate and the true population characteristic.

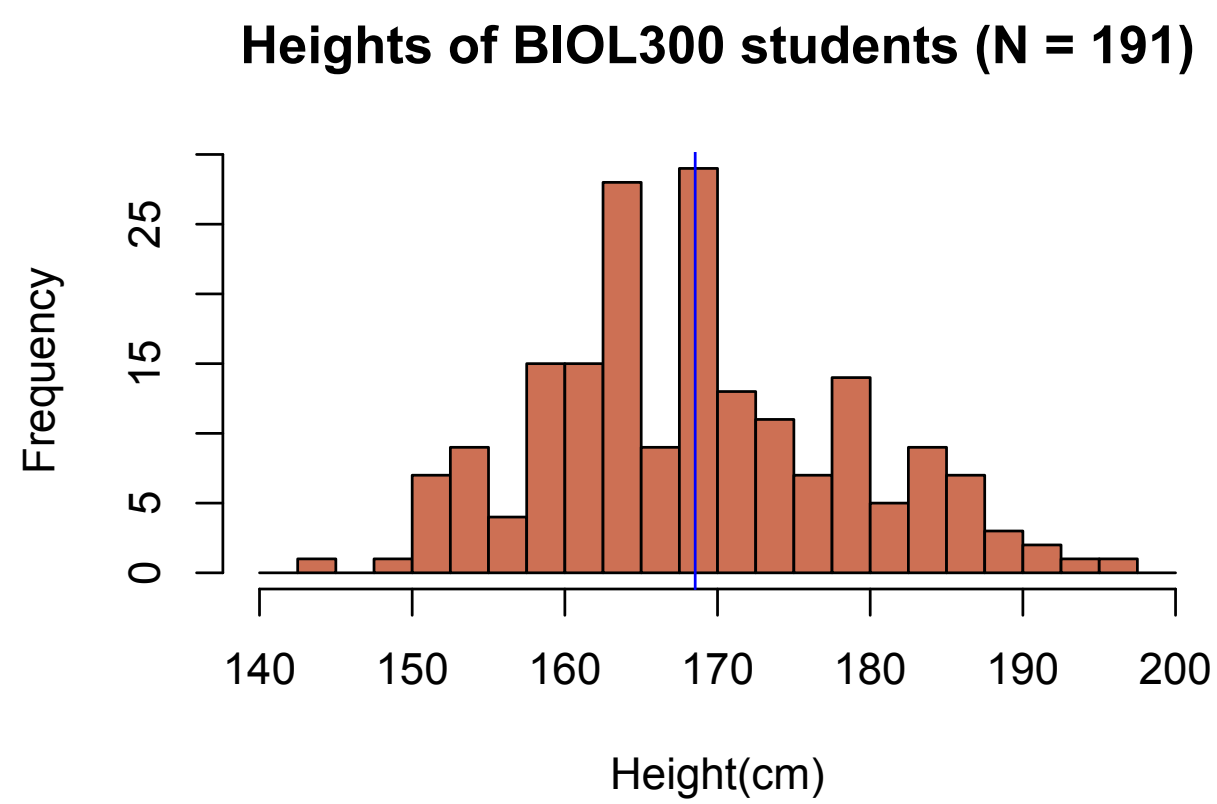
Error is a random difference (not tending in any direction) between an estimate and the true population characteristic.

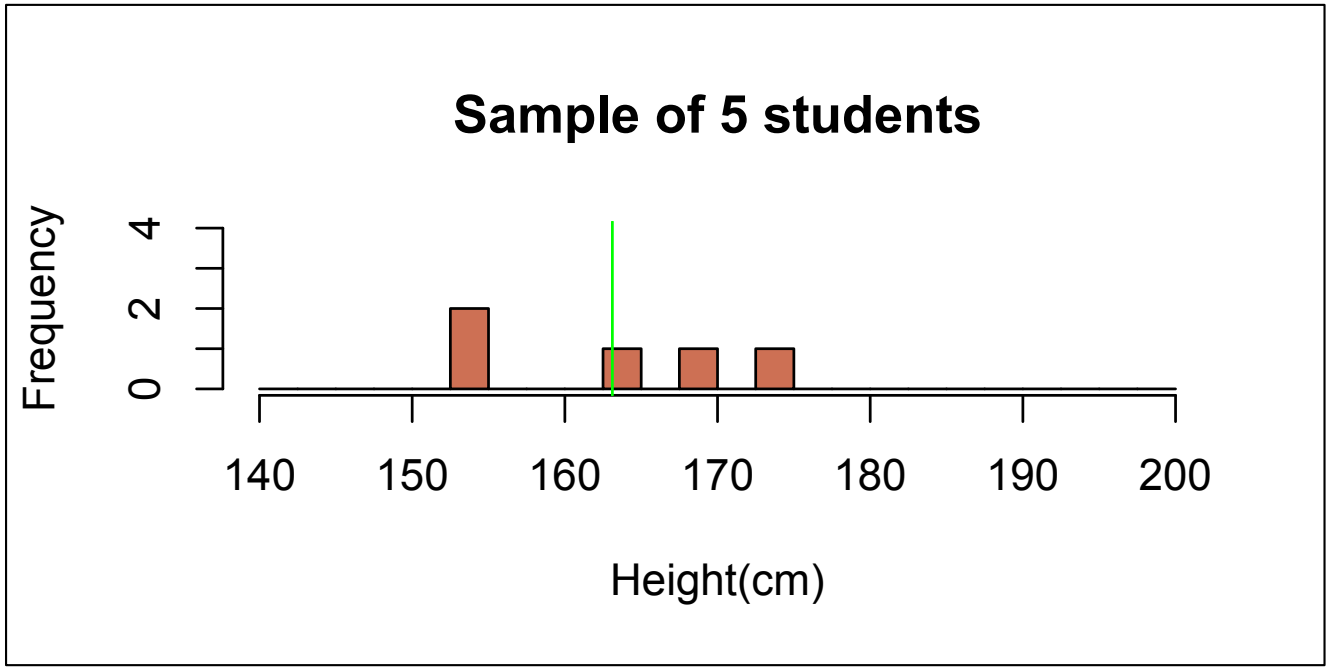
Estimates (from samples) differ from **Parameters** (of populations) for four reasons:

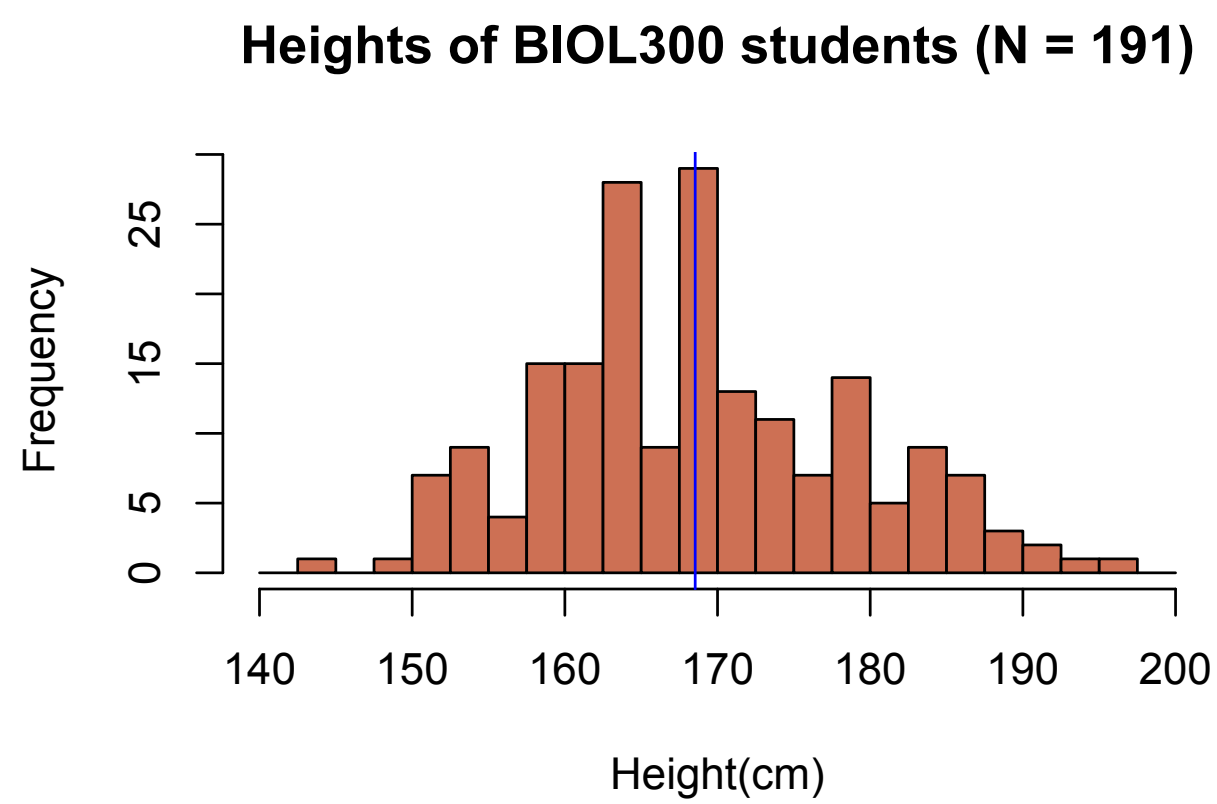


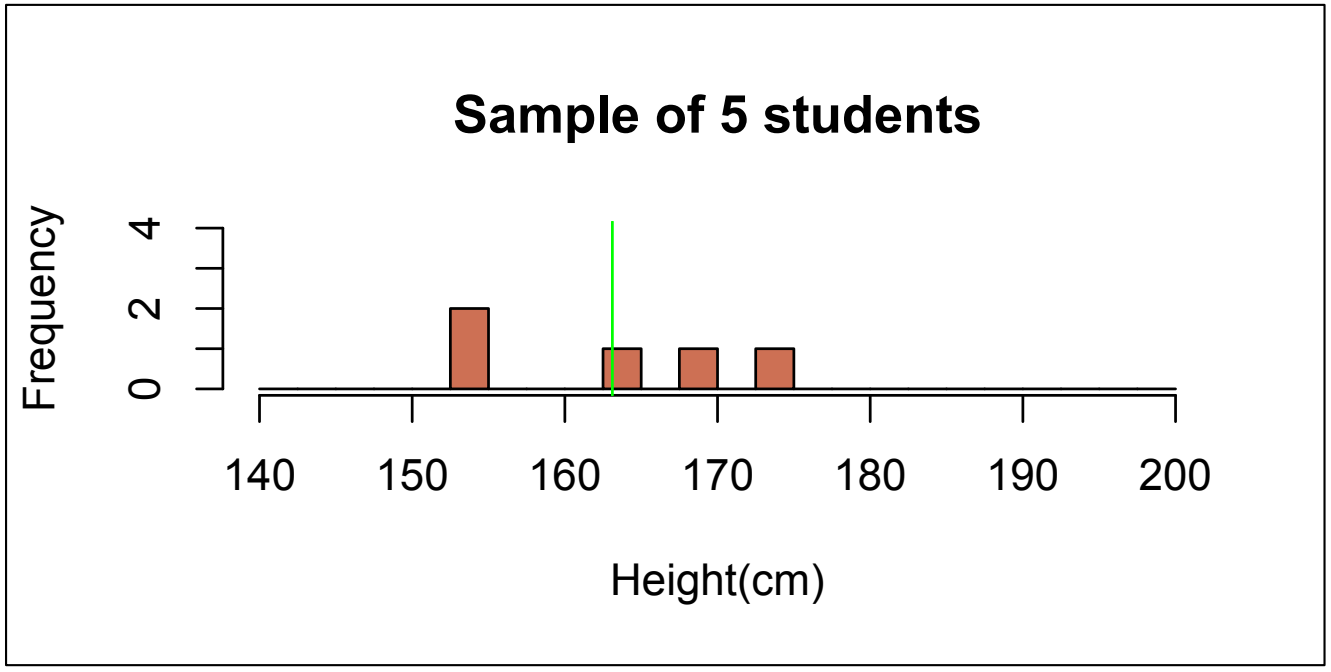
Larger samples on average will have smaller sampling error

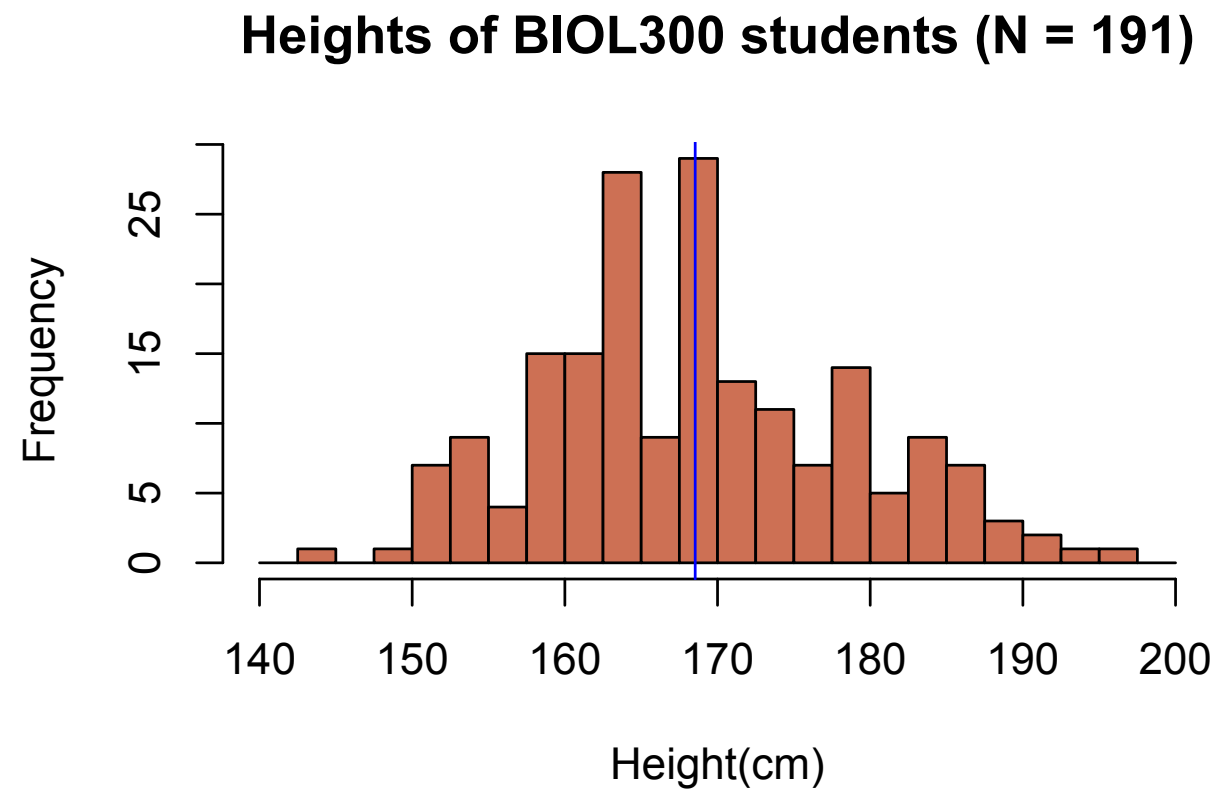

Heights of BIOL300 students (N = 191)
(self-reported, Spring 2020)
Frequency
0
5
15
25
Blue line shows
average, 168.54
140
150
160
170
180
190
200
Height(cm)

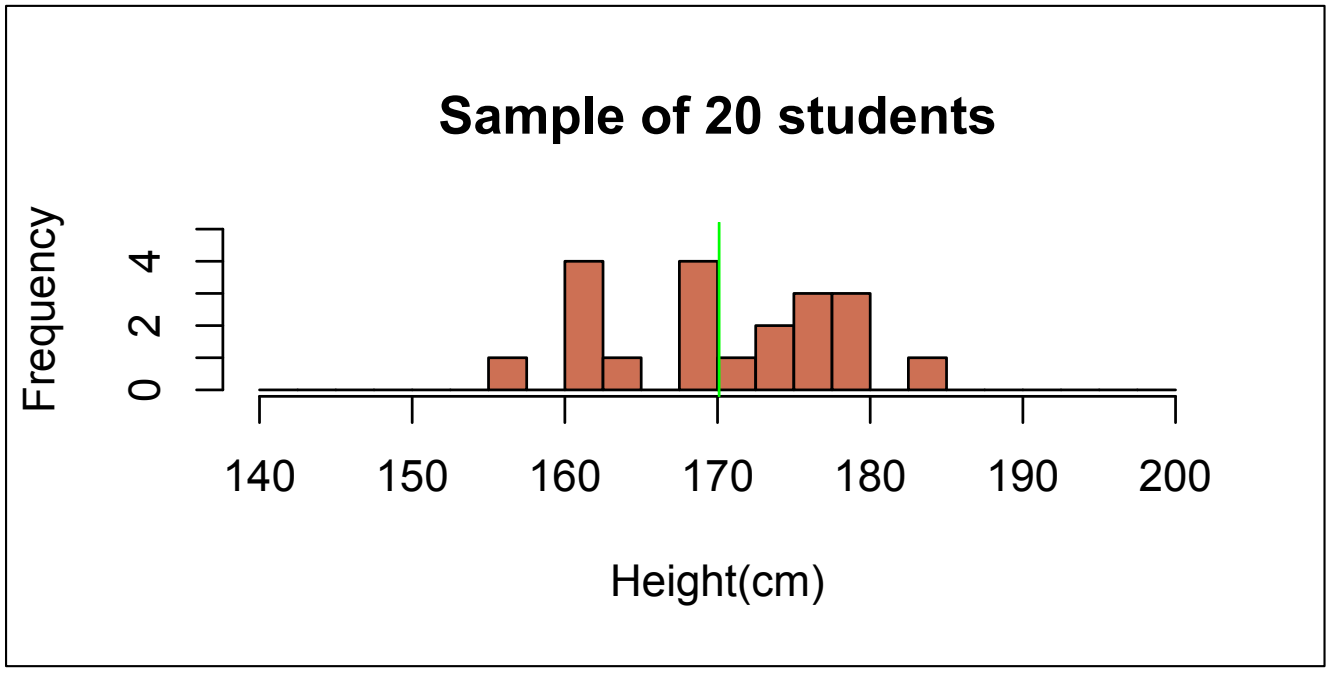

Heights of BIOL300 students (N = 191)
Frequency
0
5
15
25
140
150
160
170
180
190
200
Height(cm)

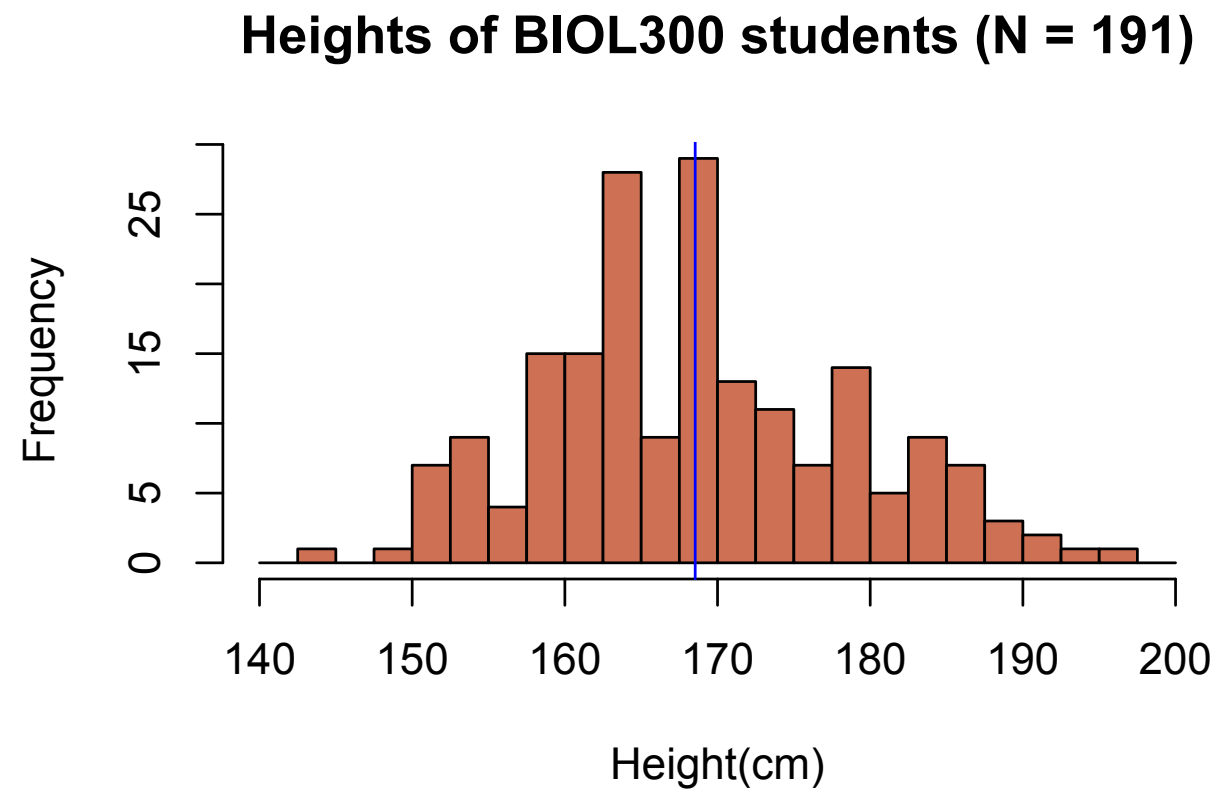

Sample of 5 students
Frequency
0
2
4
140
150
160
170
180
190
200
Height(cm)

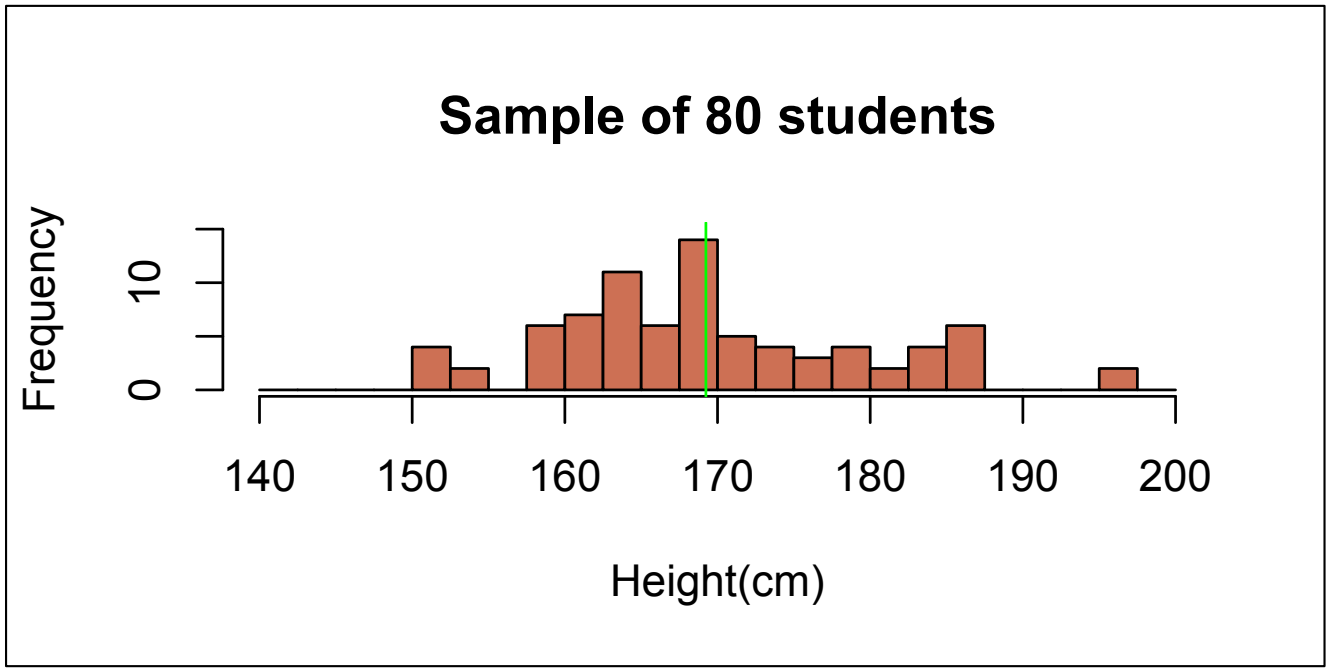

Heights of BIOL300 students (N = 191)
Frequency
0
5
15
25
140
150
160
170
180
190
200
Height(cm)


Sample of 5 students
Frequency
0
2
4
140
150
160
170
180
190
200
Height(cm)


Heights of BIOL300 students (N = 191)
Frequency
0
5
15
25
140
150
160
170
180
190
200
Height(cm)


Sample of 20 students
Frequency
0
2
4
140
150
160
170
180
190
200
Height(cm)

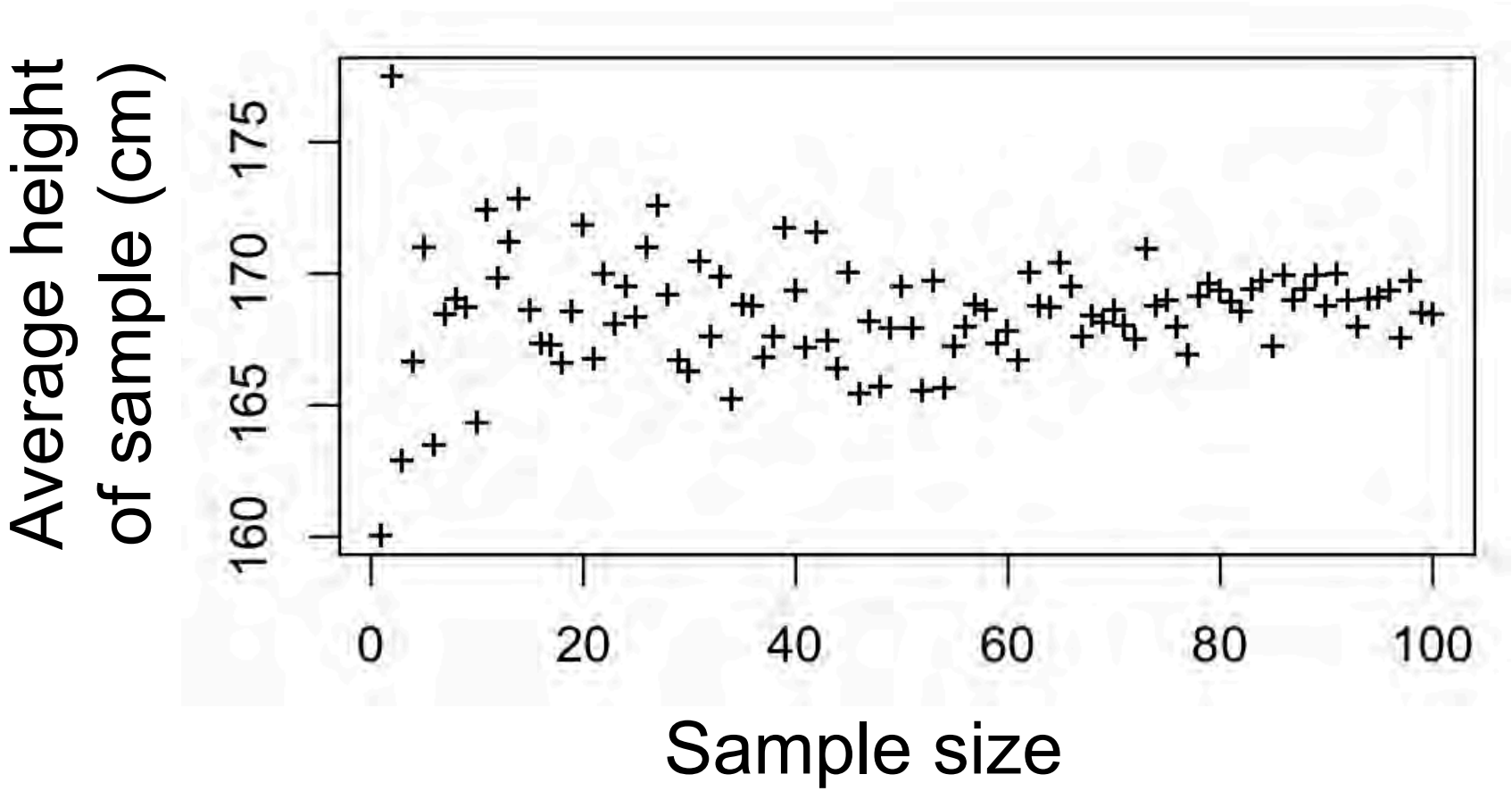

Heights of BIOL300 students (N = 191)
Frequency
0
5
15
25
140
150
160
170
180
190
200
Height(cm)


Sample of 80 students
Frequency
0
10
140
150
160
170
180
190
200
Height(cm)

Larger samples *tend* to have less sampling error

Sample size vs. mean height

(2020 Spring; each cross is a single sample average)



Good samples are a foundation of good science

- When thinking about any study, one should always ask “How were the data collected?”

Some recent poll-related headlines

- **Poll on best prime ministers of the 20th century suggests regional divide** https://vancouversun.com/pmn/news-pmn/canada-news-pmn/poll-on-best-prime-ministers-of-the-20th-century-suggests-regional-divide/wcm/ff8e3e98-3ad6-43ee-a514-532e9ccd5a63
  - **"The Leger online poll conducted the week of Nov. 11 surveyed 2,295 Canadians but cannot be assigned a margin of error because polls from internet panels are not random samples."** (I could not find the underlying methodology or data at either the Leger or ACS sites)

- **British Columbians more supportive of Trans Mountain pipeline: poll** https://bc.ctvnews.ca/british-columbians-more-supportive-of-trans-mountain-pipeline-poll-1.4746283
  - Some info from Research Co.: https://researchco.ca/2019/12/18/pipelines-british-columbia/

Be a careful consumer of information.

You must think carefully about what population is being sampled

- All cats falling out of windows *vs.* survivors being brought into vets
- American voters *vs.* Americans who have telephones who respond to survey through mail
- British Columbians *vs.* British Columbians willing to fill out online survey