BIOL 300: Biostatistics

Course web address:

http://www.zoology.ubc.ca/~whitlock/bio300/

Canvas: https://canvas.ubc.ca/courses/39708

Prof. Michael Whitlock
Department of Zoology
Office: 216 Biodiversity
whitlock@zoology.ubc.ca

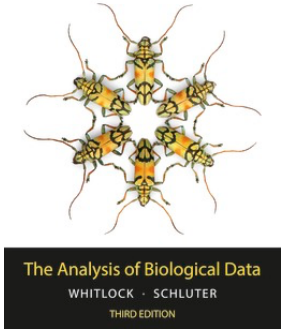
Dr. Matt Pennell
Dept. of Zoology
pennell@zoology.ubc.ca

Office hours: Mon. 1:30-2:30

and after class most days

Textbook

Whitlock and Schluter,
The Analysis of Biological Data



Lab manual

Available at course web site

R

Statistical software for PCs and Macs

Available for free download:

http://www.zoology.ubc.ca/~whitlock/bio300/labs/downloadingR.html

Evaluation

Final 25%
Mid-term 30%
Project final 25%
Assignments (homework) 10%
Lab assignments 10%

Lab

Begins second week of term (September 14-18)

~~Biosci 2004~~

Virtual labs

Midterm

Mid-October:

Class vote: Oct. 22 or 29

Assignments

Available on canvas

Due on Fridays at noon

Lab assignments

Due in canvas the following week

Statistics course pairings

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

These are paired with BIOL 300, but *do not count* as biology courses.

Introduction to statistics

Statistics are "a quantitative technology for empirical science; it is a logic and methodology for the measurement of *uncertainty* and for an examination of that uncertainty."

The key word here is "uncertainty." Statistics become necessary when observations are variable.

Goals of statistics

- Estimate the values of important parameters
- Test hypotheses about those parameters

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."

High falls show *lower* injury rates

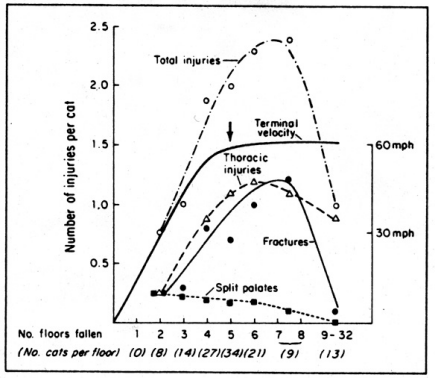


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (——); total number of injuries/cat (○, —·—·—); number of thoracic injuries (pulmonary contusions +pneumothorax)/cat (△, – – –); number of fractures/cat (●, ——); number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

Or not…

A *sample of convenience* is a collection of individuals that happen to be available at the time.

Read: Chapters 1 & 2

A newer study reports more injuries with longer falls

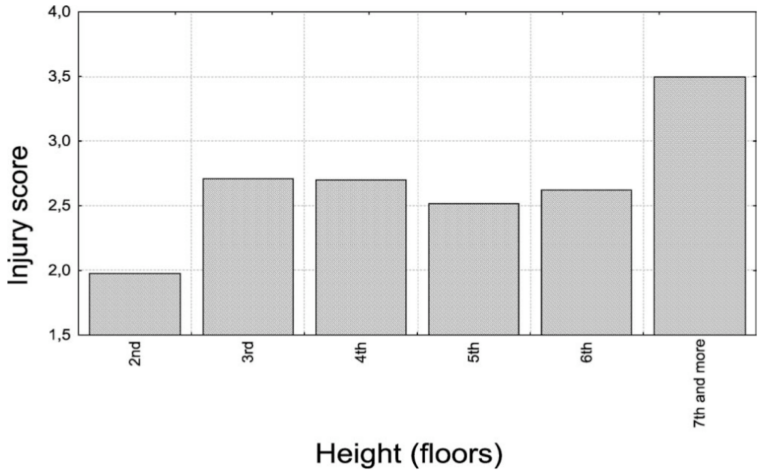


Figure 5 Graph showing the relationship between injury score and height of fall.

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

Variable

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

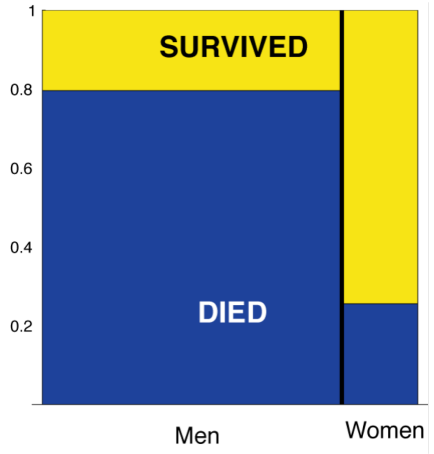
Data are measurements of one or more variables made on a collection of individuals.

Explanatory and response variables

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

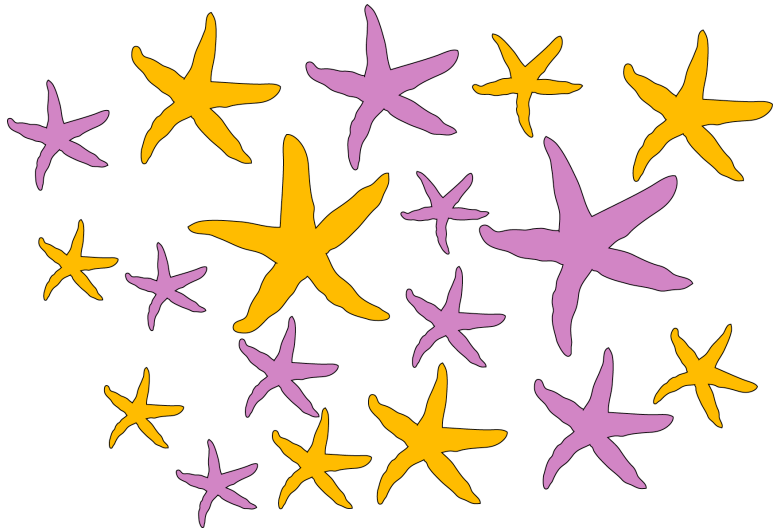
Older terminology:
dependent variable and *independent variable*

Mortality on the *Titanic*, as predicted by gender



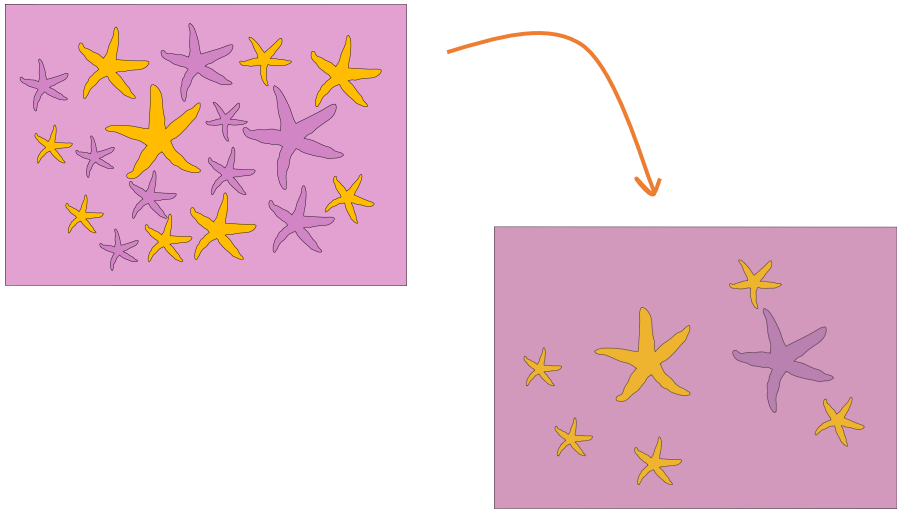
Populations and samples

A population of starfish



Populations <-> Parameters;
Samples <-> Estimates

A biased sample



Bias is a systematic discrepancy between estimates and the true population characteristic.

The 1936 US presidential election

Alf Landon Campaign Poster, 1936

Alf Landon
Republican

vs.

Franklin Roosevelt
Democrat

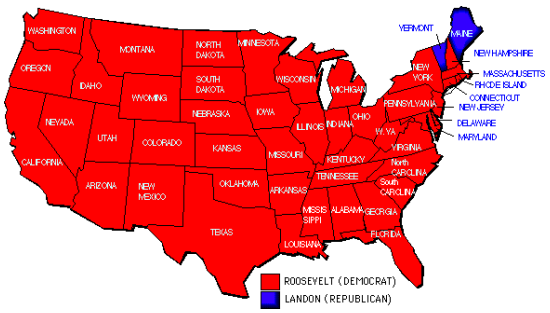
1936 *Literary Digest* Poll

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

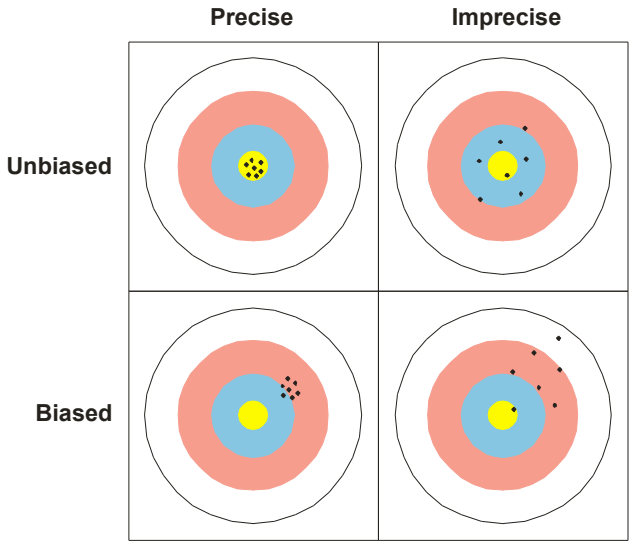
Voting and party preference is correlated with personal wealth

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.

Goals of estimation

- Accuracy (on average gets the correct answer)
- Precision (gives a similar answer repeatedly)

Properties of a good sample

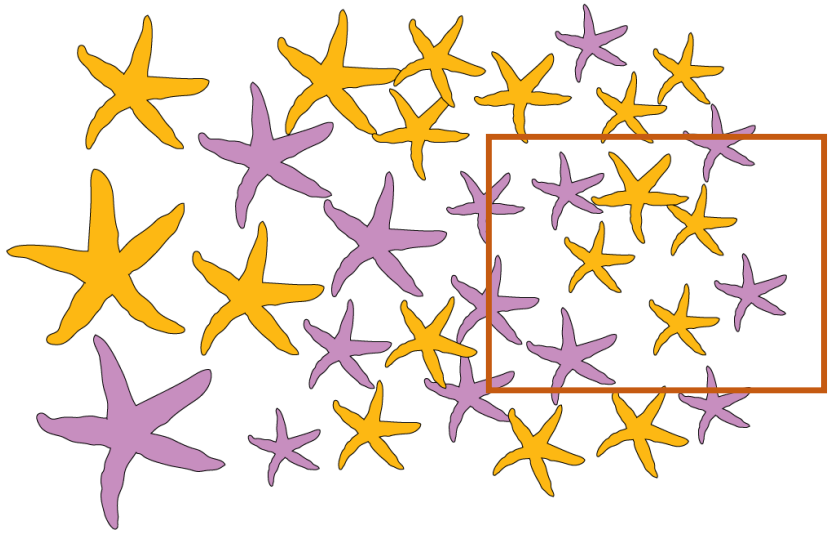
- Independent selection of individuals
- Random selection of individuals
- Sufficiently large

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

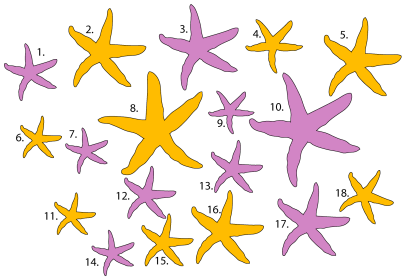
Independent sampling

The chance of an individual being included in the sample does not depend on who else is sampled.

Non-independent sample



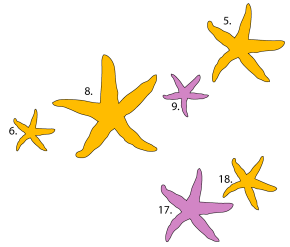
One procedure for random sampling



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.

Larger samples on average will have smaller sampling error.

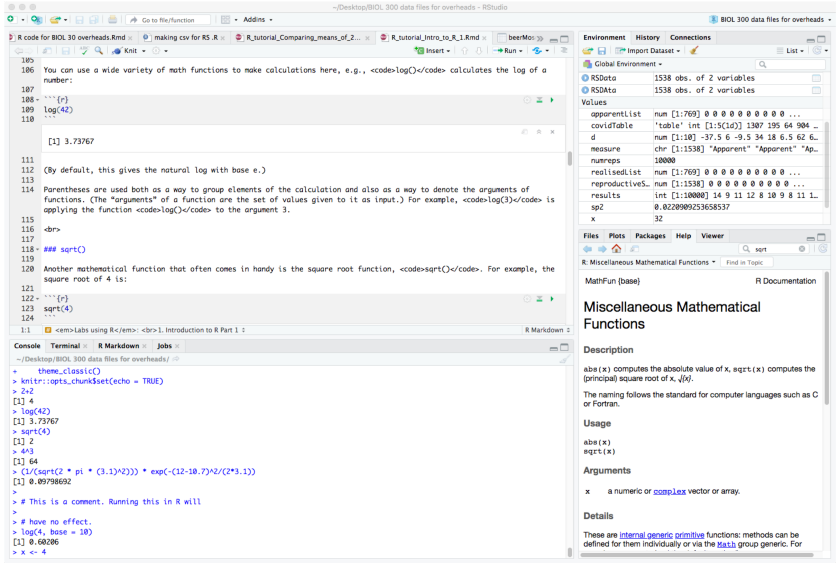
Sampling error

The difference between the estimate and average value of the estimate

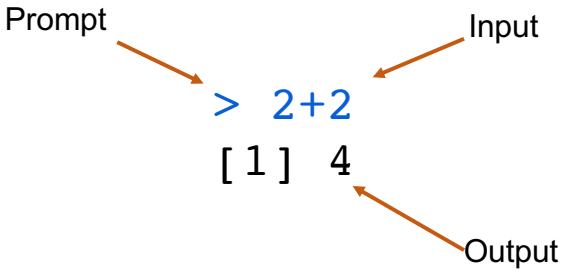
Introduction to R

R is a free, open-source statistical language, widely used by scientists worldwide.

RStudio



A simple command in R



Basic arithmetic

```
> log(42)
[1] 3.73767
> sqrt(4)
[1] 2
> 4^3
[1] 64
```

Functions can have multiple inputs

```
> log(4, base = 10)
[1] 0.60206
```

You can define variables

```
> x <- 4
> x + 3
[1] 7
```

Data frames

```
> titanicData <- read.csv("DataForLabs/titanic.csv")

> head(titanicData)

  passenger_class                                      name     age    embarked
1             1st                  Allen,MissElisabethWalton 29.0000 Southampton
2             1st                   Allison,MissHelenLoraine  2.0000 Southampton
3             1st             Allison,MrHudsonJoshuaCreighton 30.0000 Southampton
4             1st Allison,MrsHudsonJ.C.(BessieWaldoDaniels) 25.0000 Southampton
5             1st                  Allison,MasterHudsonTrevor  0.9167 Southampton
6             1st                            Anderson,MrHarry 47.0000 Southampton
```

Each row is an individual;
each column is a variable describing each individual.

Calculations on data

```
> mean(titanicData$age, na.rm=TRUE)
[1] 31.19418
```

Calculations on data

```
> t.test(titanicData$age ~ titanicData$survive)

      Welch Two Sample t-test

data:  titanicData$age by titanicData$survive
t = 1.9947, df = 570.96, p-value = 0.04655
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 0.03644633 4.71184176
sample estimates:
 mean in group no mean in group yes
         32.24811          29.87396
```