BIOL 300: Fundamentals of Biostatistics

Instructor:
Dr. Darren Irwin

Course web address:

http://www.zoology.ubc.ca/~irwin/BIOL300/

Statistics: likely the most important subject you study at UBC

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

BIOL 300: Fundamentals of Biostatistics

Course web address:

http://www.zoology.ubc.ca/~irwin/BIOL300/

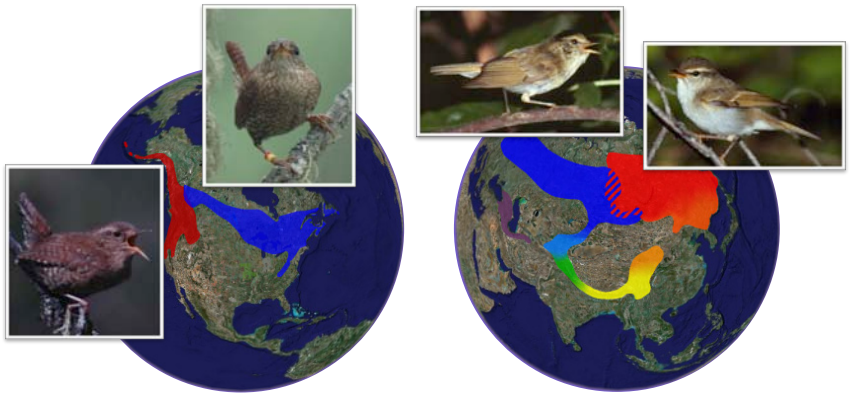
Instructor:

Dr. Darren Irwin
(Associate 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)

Office hours: Fri. 3:30-4:30
(Biodiversity 209)

and after class most days

Please feel free to ask questions during class

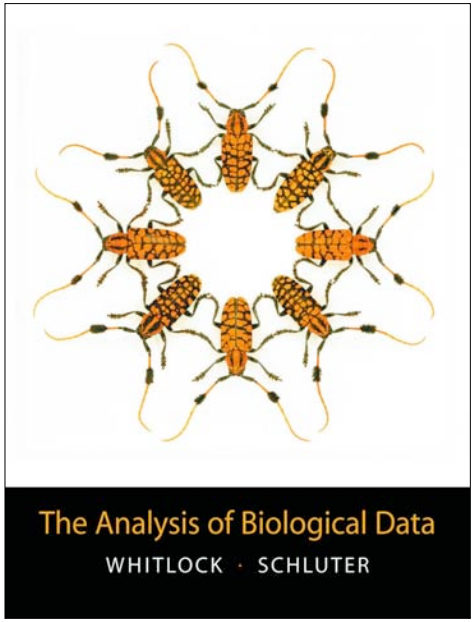
Teaching Assistants

- *Jessica McKenzie*
- *Marc Delepine*
- *Rebecca Kordas*
- *Jocelyn Nelson*
- *Laura Tremblay-Boyer*

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

Textbook

- Whitlock and Schluter (2009) *The Analysis of Biological Data.*



Lab manual

- Available for about $10 at Copiesmart in the UBC Village (near McDonald's)
- Available at course web site

Lab

- Begins second week of term (January 9-13)
- BioSciences room 2434
- Attendance is highly recommended (but technically optional for some labs)
- Great opportunity for learning from TAs, using JMP, and for doing two lab assignments.

JMP

- Statistical software for PCs and Macs
- Used in the labs
- You *might* be interested in buying your own copy (*optional*). Available online: see course website for link.

Evaluation

Homework assignments 15%
Lab assignments 10%
Mid-term 30%
Final 45%

Homework Assignments

- Available on course web-page
- Due on Fridays at noon, at your TA's office
- Intended to help you learn
- First assignment due Jan. 13th

Midterm

February 29, in class

Wait list

- If you are on the wait list, chances are good that you can take the course (but no promises now).
- If you are not registered, try to register for the wait list. If not successful, email me.
- If you do not want to take the course, please de-register yourself (make room for others).

STATISTICS 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 is "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

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."

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

High falls reported to show *lower* injury rates

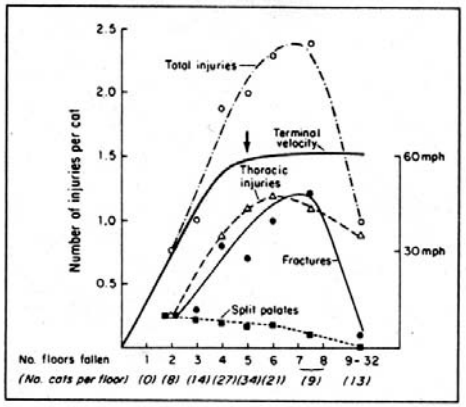


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (—); total number of injuries/cat (○, —·—·—); number of thoracic injuries (pulmonary contusions +pneumothorax)/cat (△, – – –); number of fractures/cat (●, —); number of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

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

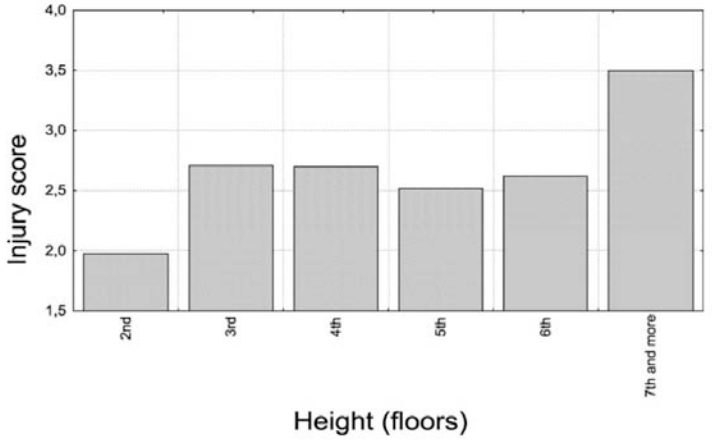


Figure 5 Graph showing the relationship between injury score and height of fall.

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

Limb fractures vs. thoracic trauma

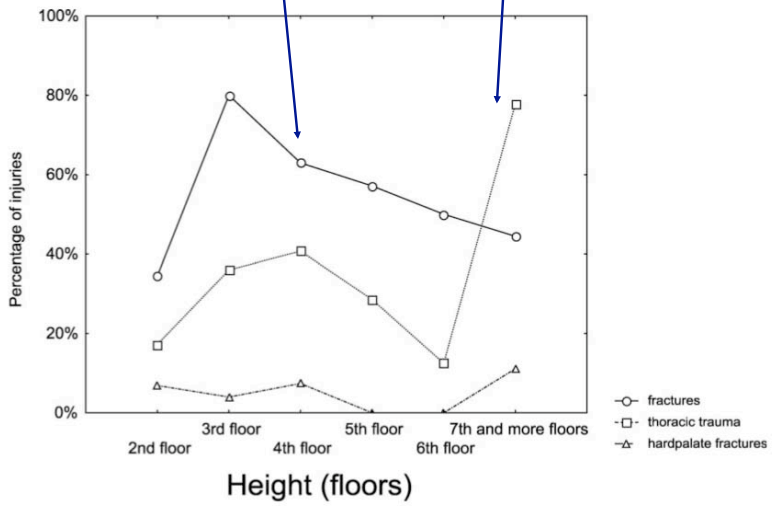


Figure 6 Graph showing the percentage of cats with particular injuries when falling from different heights.

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
- 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 (count half sibs as half)
c) # of cups of coffee consumed in past week
d) Your writing hand (left/right/other)
e) Length of your commute this morning (in minutes)
f) Type of transportation used today (e.g., walk, bike, car, bus)
g) Your weight (indicate lbs or kg)
h) A random number between 0 and 101
i) Your sex (M/F/other)

Read: Chapters 1 & 2

Variables and Data

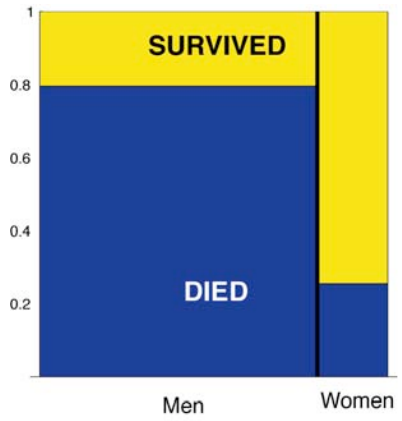
- A **variable** is a characteristic measured on individuals drawn from a population under study.
- **Data** are measurements of one or more variables made on a collection of individuals.

Explanatory and response variables

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

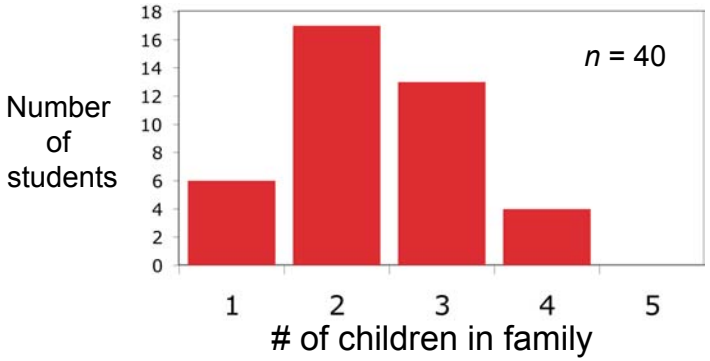
Older terminology:
dependent variable and *independent variable*

Mortality on the *Titanic*, as predicted by sex



Populations and samples

Histogram of family size of a sample of BIOL300 students



average = 2.4

This is higher than national average family size. Why?

You must think carefully about what population is being sampled

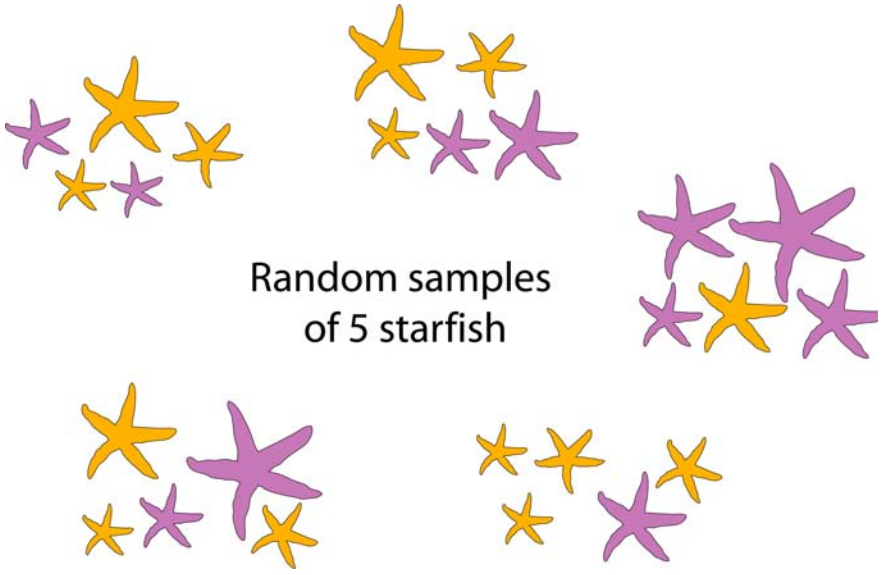
- All cats falling out of windows *vs.* survivors being brought into vets
- Families *vs.* children from families

Populations <-> Parameters; Samples <-> Estimates

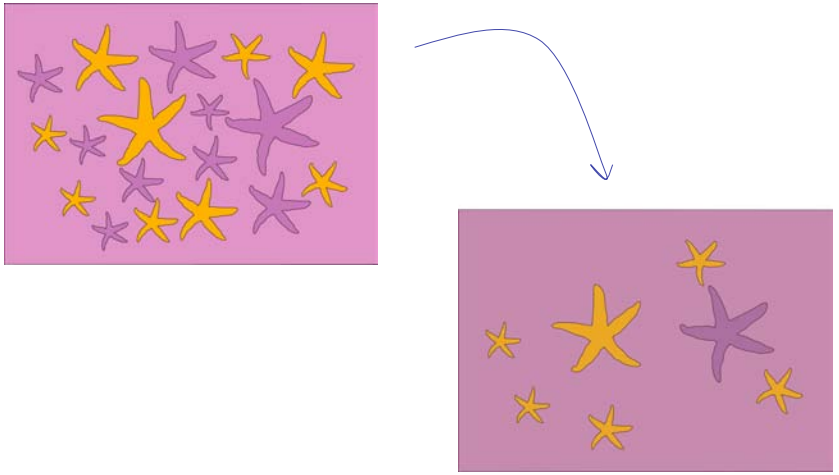
Estimates differ from true population characteristics (parameters) for two reasons:

- Sampling error
- Bias

A population of starfish



A biased sample



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

The 1936 US presidential election

Alf Landon Campaign Poster, 1936

vs.

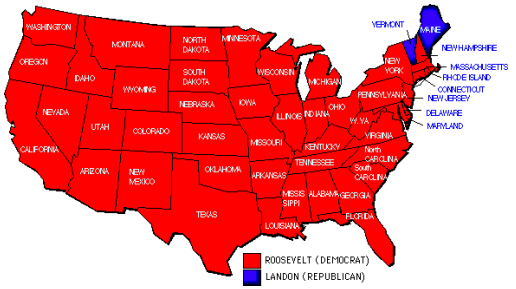
Alf Landon
Republican

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

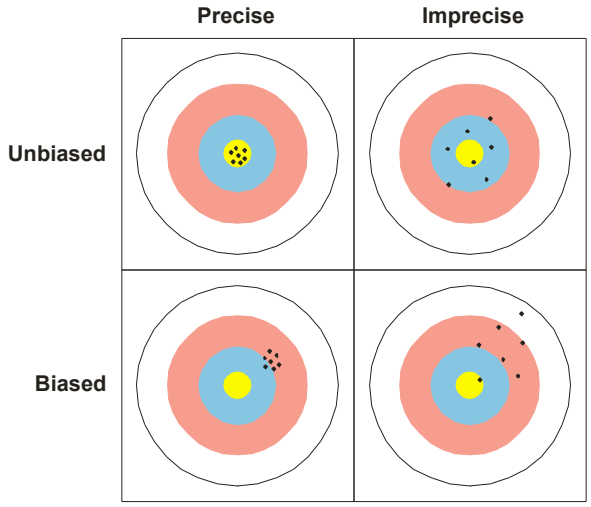
Voting and party preference is correlated with personal wealth

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



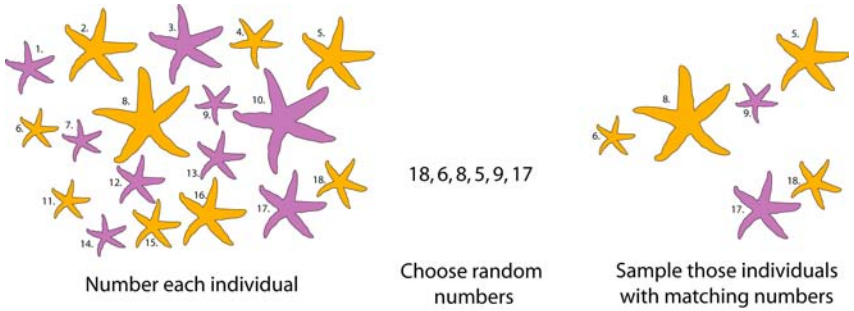
Each point represents an estimate of a parameter.

Properties of a good sample

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

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

One procedure for random sampling



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.

Larger samples on average will have smaller sampling error