BIOL 300: Biostatistics

Course web address:
http://www.zoology.ubc.ca/~whitlock/bio300/

**Professor:**
Dr. Michael Whitlock
Professor
Department of Zoology
Office: 216 Biodiversity

e-mail:
whitlock@zoology.ubc.ca

**Textbook**
- Whitlock and Schluter, *The Analysis of Biological Data*

Office hours: Mon. 1:30-3:00
and after class most days
Lab manual

- Available at course web site
- Available at Copyright in the SUB

JMP

- Statistical software for PCs and Macs
- Used in the labs
- Optional

Evaluation

Final 50%
Mid-term 30%
Assignments (homework) 10%
Lab assignments 10%

Lab

- Begins second week of term (September 14-18)
- Biol. Sci. room 2434
Midterm

Mid-October: TBA

Assignments

• Available on course web-page

• Due on Fridays at noon, at your TA’s office

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 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

**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.*
A newer study reports more injuries with longer falls


---

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 sex

A population of starfish

Populations and samples

Random samples of 5 starfish
Populations <-> Parameters; Samples <-> Estimates

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

The 1936 US presidential election

Alf Landon Republican

Franklin Roosevelt Democrat

A biased sample
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
- 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

Number each individual

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