Some pointers on statistics for your senior thesis

Senior Seminar
February 4, 2015
Matt Palmer & Spahr Webb

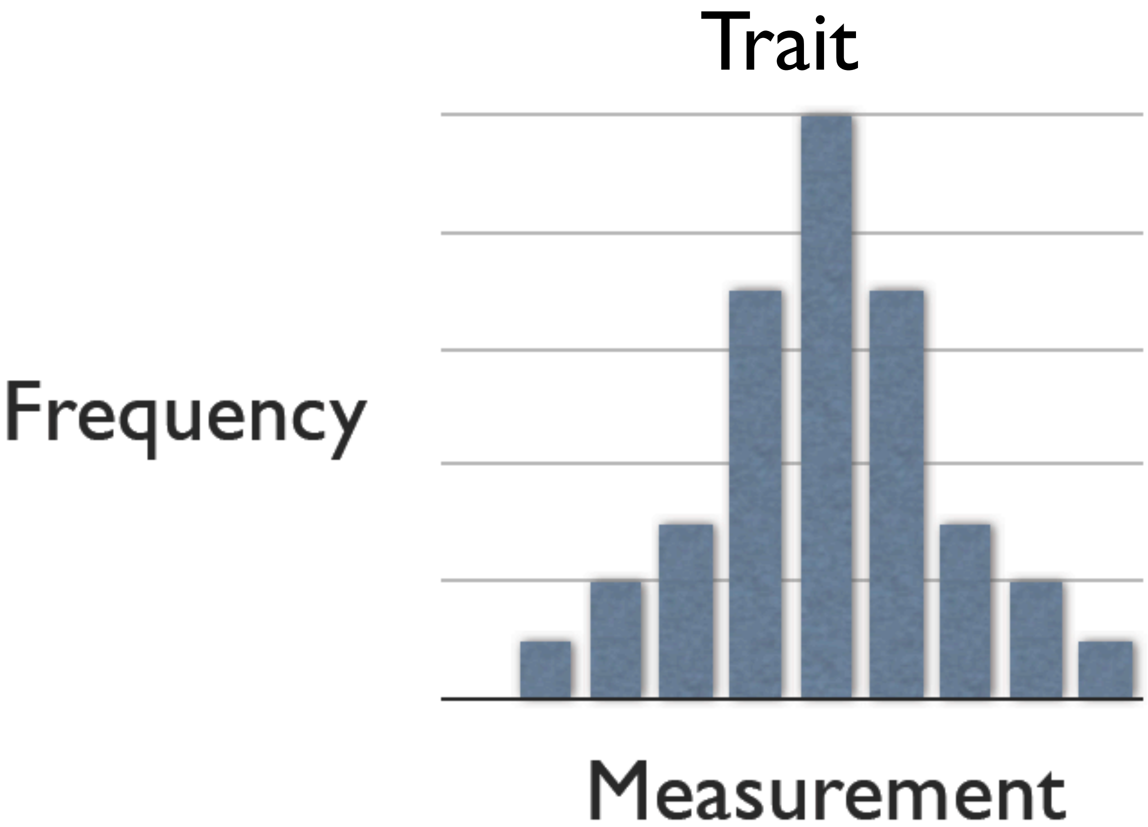
General Advice

- Look at the methods used in related work (other papers, especially those by your mentor and their colleagues)
- Talk to your mentor (or their lab managers, graduate students, postdocs, techs, etc.)
- Use the statistical consulting service: consult@stat.columbia.edu
- Read a (portion of a) book!

Author’s Responsibility

- Know what your statistical methods do!
- Be aware of the assumptions and limitations of your statistical tests
- Report all the proper results
- Understand what your results mean. (Plot and look at your data, do the data make sense?)

Variation



This will usually be a Normal Distribution, but not always.

Reporting variation

- Every measure which summarizes a distribution (e.g., a mean) should include some measure of spread (e.g., a standard deviation)

- A graph without error bars is incomplete and potentially misleading!

(Need to show whether differences between data are significant)

Hypothesis testing

- Comparing two or more hypotheses in light of the data
- Scientists generally make a null hypothesis of no effect - any variation in the data is just random
- We reject the null when the data deviate strongly from random. This lends support to the hypothesis that some phenomenon is responsible for part of the variation

Probability

$$P = \frac{\text{number of outcomes}}{\text{number of trials}}$$

So P=0.05, means you expect an outcome one time in 20 trials, by chance.

Binomial Distribution

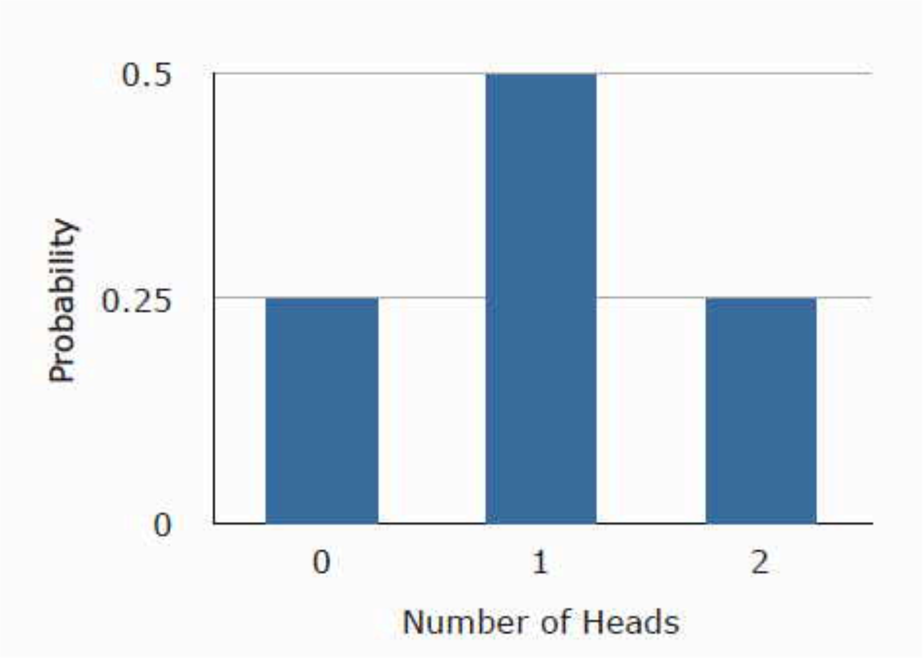


Figure 1. Probabilities of 0, 1, and 2 heads.

Data from experiments with a discrete number of outcomes- flipping coins, which choice food item an animal makes, and so on, are often described by a binomial distribution.

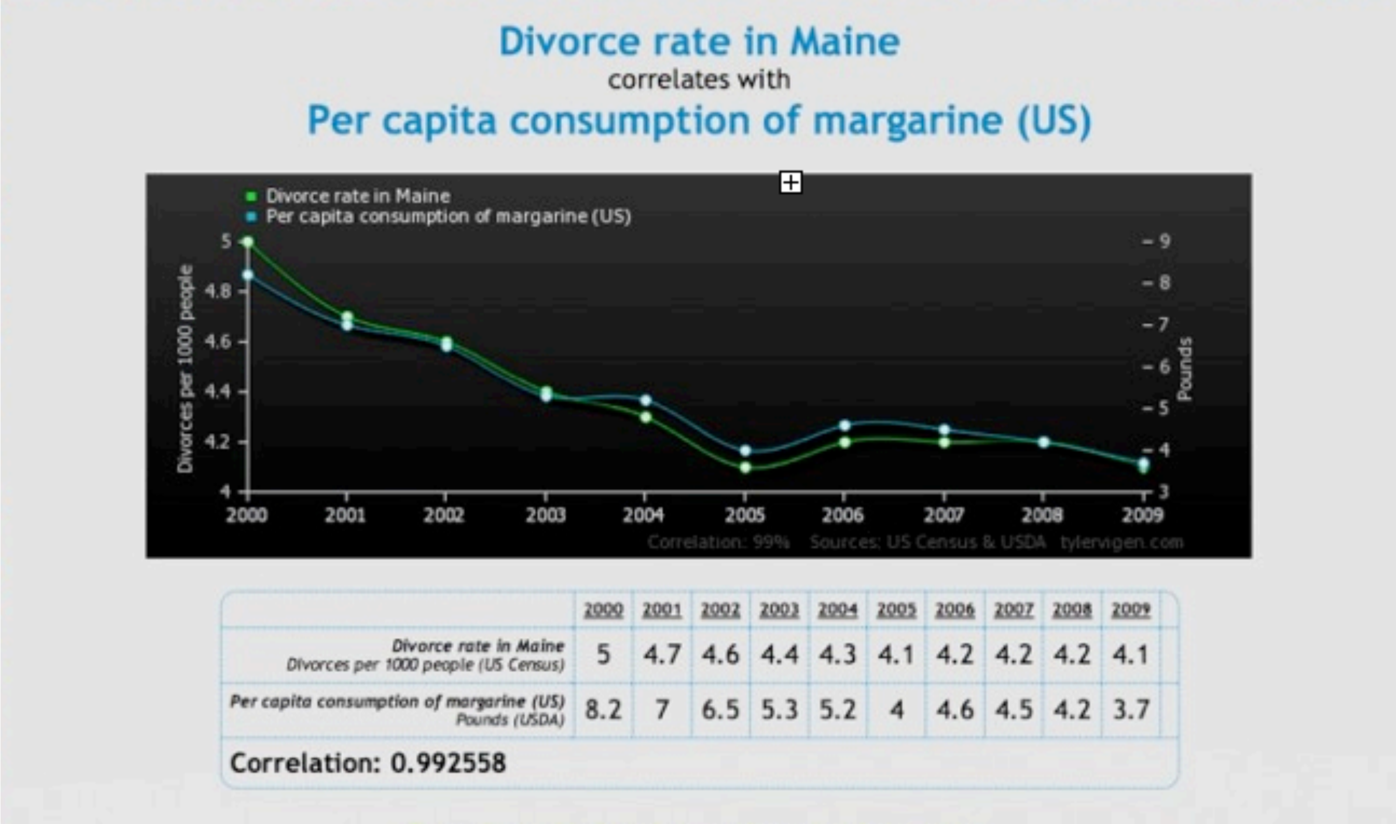
Other kinds of data

Proportional data (0 to 1).
Count data (0, 1, 2, 3...)
Poisson (things like times between random events)
Zero inflated (Poisson, except lots of non-events data like insurance claims).

Many kinds of data are not described by a Normal distribution or Student t- distribution (ANOVA).

For each of these there is likely a set of statistical tests that are appropriate.

Correlation



Divorce rate in Maine

correlates with

Per capita consumption of margarine (US)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Be careful.
Small number of “degrees of freedom”
leads to false apparent correlations.

Other Resources

(1) Barnard College Empirical Reasoning Lab:
http://erl.barnard.edu/
(located in the Barnard Library)

(2) CU - Digital Social Science Center
(Lehman Social Science Library)
http://library.columbia.edu/locations/dssc.html

(3) CU Dept. of Statistics
- they offer statistics consulting
http://stat.columbia.edu/consulting-information/

(4) Applied Statistic Center
http://applied.stat.columbia.edu/
(see their consulting tab)

Playroom time
On Tuesdays from 2:15pm-5:15pm, people from the Applied Statistics Center are available in the Playroom (IAB 707). You can stop in during that time with questions and we will direct you to the right place.

For software and basic data problems try the two library options (1) and (2) first.
They also offer GIS advice.