Introduction

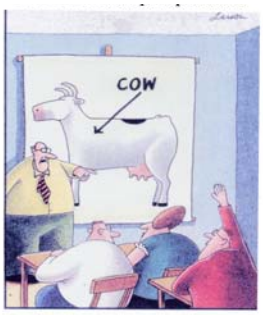
Geog 301a

Course website

- https://instruct.uwo.ca/geog/301
- Password: geog30104
- Username: student301

- Old statisticians never die, they just become nonsignificant

There are no stupid questions



"Yes ... I believe there's a question in the back."

When statistics are valuable

- Can only give answers if the data collection and the data collected allow such answers
- User is aware the statistics is just another strategy for finding, patterns in the data
- Statistics are based on certain assumptions If those assumptions are not true the technique can still be applied but significance tests must treated with caution

When statistics are valuable

- User is aware that techniques are mathematical models. Reality in all its complexity cannot be modeled in a useful way. Complex models may imitate reality but they will be equally complex and therefore not useful. Summarizing data in a complex way is not a step forward.
- Data exploration needs to be done before any higher level modeling

Users can attack complex retail problems with canned applications for correlation analysis, t-tests, analysis of variance, chi-squared tests, factor analysis and least-squares regression and be satisfied that state-of-the-art sophistication has been applied to the problem. But the ease with which these canned techniques are implemented also presents a danger. Poorly applied, these methods can backfire, but in extremely subtle ways of which few are even cognizant.

From: Gross, Bryan, 2000, The Retail Model Maze, *Business Geographics*, June, pg. 24

Statistical inference

- Why this matters
  - When we sample, we are really drawing a single sample from all possible samples (i.e., from the sampling distribution)
  - We know the real shape of the sampling distribution
  - For any observed statistic, we can infer things about population parameter
  - This resolves some of the fundamental uncertainty about empirical social science

- *What's Statistics* video
- Web address for viewing video: http://www.learner.org/resources/series65.html#

How does a Statistics test work?

- Statistics test analyses the data (numbers) submitted (by the researcher) to calculate the chances of obtaining a difference when there is none i.e. probability of obtaining a spurious difference.
- It does not indicate
  - whether your design is right or wrong
  - whether the type of data is correct or wrong
  - the magnitude of the difference
  - whether the difference will be practically useful
- All it can point out is whether the obtained difference between two groups is REAL or FALSE

What does a Statistics test infer?

- Statistics test → Data → p value
- When $p<0.05$, it shows that the chances of obtaining a false difference is less than 5% (1 in 20) [$p<0.01$ – 1 in 100; $p<0.001$ – 1 in 1000]
- Since we consider 5% p is small, we conclude that the difference between groups is TRUE
- Truth is something which is most likely to be true and 100% certainty is impossible.

Mechanics of hypothesis testing

- Statement of null hypothesis
  - Null hypothesis of theoretical interest
  - Vast majority of times, researchers hope to disprove null hypothesis
    - Null hypothesis: smoking lots of cigarettes does not cause cancer
    - Having a highly developed economy does not make a country likely to be more democratic
    - Paul Martin has 50% approval rating

Mechanics of hypothesis testing

- Select sampling distribution and choose alpha (define critical region)
  - May choose any p (or α, it's the same thing) we want; 0.05 is standard in literature

Statistical significance

- Statistical significance: a statistic is statistically significant at the X% level if we are X% confident that the result is not due to chance

p-values

- All of the following are equivalent statements:
  - The statistic is significant (at traditional level)
  - We can rule out the null hypothesis with 95% confidence
  - The *p-value* is less than .05
  - The 95% confidence interval does not include my null hypothesis
  - We can state with 95% confidence that the result was not due to random sampling variability

Data

Missing Data

Missing data

- conventional methods of dealing with missing data
- **listwise deletion** - observations are deleted is there are any missing values
  - 2 advantages
    - a) can be used in any statistical analysis
    - b) no special computational methods are required

Missing data

- **pairwise deletion** - (also known as available case analysis)
  - each pairwise case with existing values are utilized
  - May cause seriously biased results if data is data is not randomly missing

Missing data

- **dummy variable adjustment**
- code a new variable that takes on a value of 0 if the independent variable of interest has a value, a value of 1 if it is missing
- include this new variable in the model
- this approach also produces biased results

Missing data

- **imputation**
- basic idea is to substitute some reasonable guess for the missing value
  - simplest is to use the mean of the variable but produces serious bias and should be avoided
  - there are 2 better methods for missing data estimation
    - maximum likelihood
    - multiple imputation
      - but both are beyond the scope of this course

Missing data

- bottom line is:
- use listwise if you don't lose too many cases
- otherwise use pairwise but realize the estimates will be biased in your analysis

Massage the numbers?