Introduction

Geog 301a

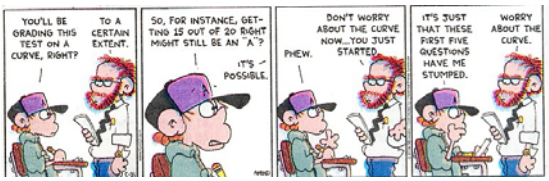
Why have this course?

- Increase your options
- Increase your understanding

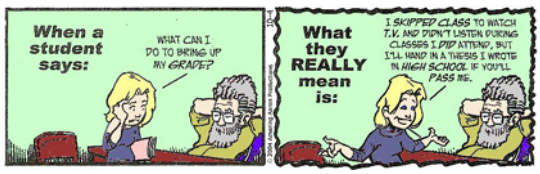
Course website

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

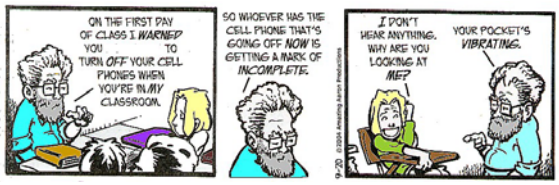
Grading



Don't be like this



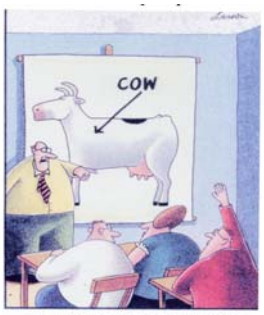
No cell phones



Don't wait for help

- Old statisticians never die, they just become nonsignificant

There are no stupid questions



Birthday Paradox

- Chance is 50% with 23 people
- The reason this is so surprising is because we are used to comparing our particular birthdays with others. For example, if you meet someone randomly and ask him what his birthday is, the chance of the two of you having the same birthday is only 1/365 (0.27%). In other words, the probability of any two individuals having the same birthday is extremely low. Even if you ask 20 people, the probability is still low -- less than 5%. So we feel like it is very rare to meet anyone with the same birthday as our own.
- When you put 20 people in a room, however, the thing that changes is the fact that each of the 20 people is now asking each of the other 19 people about their birthdays. Each individual person only has a small (less than 5%) chance of success, but each person is trying it 19 times. That increases the probability dramatically.
- If you want to calculate the exact probability, one way to look at it is like this. Let's say you have a big wall calendar with all 365 days on it. You walk in and put a big X on your birthday. The next person who walks in has only a 364 possible open days available, so the probability of the two dates not colliding is 364/365. The next person has only 363 open days, so the probability of not colliding is 363/365. If you multiply the probabilities for all 20 people not colliding, then you get:
- **364/365 * 363/365 * ... 365-20+1/365 = Chances of no collisions That's the probability of no collisions, so the probability of collisions is 1 minus that number. The next time you are with a group of 30 people, try it!**

When statistics are valuable

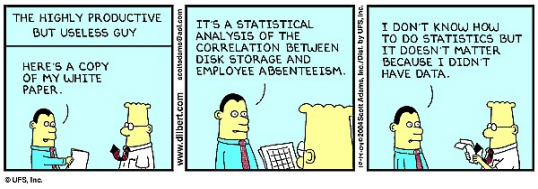
- Can only give answers if the data collection and the data collected allow such answers
- User is aware the statistics is just another strategy for finding, patterns in the data
- Statistics are based on certain assumptions If those assumptions are not true the technique can still be applied but significance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

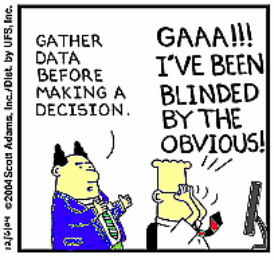
- Why this matters
  - When we sample, we are really drawing a single sample from all possible samples (i.e., from the sampling distribution)
  - We know the real shape of the sampling distribution
  - For any observed statistic, we can infer things about population parameter
  - This resolves some of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

Example: Aleve

- Fevered reaction to a recent warning from the National Institutes of health that the over-the-counter painkiller Aleve might cause heart attacks may be overblown, some medical experts say.
- Yesterday in an interview, Dr. Breitner said that evidence of naproxen side effects in the randomized study was 'not really' statistically significant. Only by combining several side effects and 'splicing and dicing' the data was it possible to detect an effect, he says, and then only for nonfatal side effects. … 'So there is no inference you can draw', say Dr. Breitner.

- From Wall Street Journal, Dec 23, 2004, pg B1, 'Some Scientists Say Aleve's Dangers May be Overblown

- *What's Statistics* video
- Web address for viewing video: http://www.learner.org/resources/series65.html#
- You need to sign up to get access but it is free



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

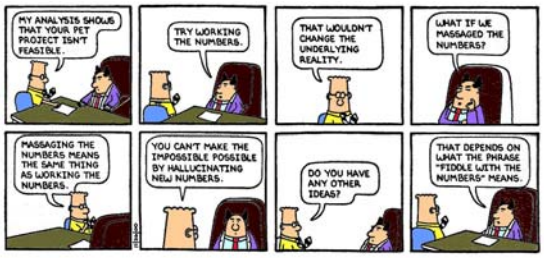
Missing data

- **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?



Why not use Excel to do Stats?

- Many statistical procedures are not available
  - Spearman's and Kendall's rank correlation coefficients
  - 2-way ANOVA with unequal sample sizes (unbalanced data)
  - p-values for two-way ANOVA
  - Nonparametric tests, including rank-sum and Kruskal-Wallis

Why not use Excel to do Stats?

- Excel uses naive algorithms that are vulnerable to rounding and truncation errors and may produce very inaccurate results in extreme cases
- Excel doesn't do regression properly if there is a high degree of multicollinearity
- Excel is unreliable when relying on standard deviation calculations (e.g. t-tests) where there are large numbers with low variation

- Routines for handling missing data may be incorrect (pre 2000 version)
- Ranks of tied data are computed incorrectly
- **Friends Don't Let Friends Use Excel for Statistics!**

Excel useful for

- viewing your data in graphs to detect errors, unusual values, trends and patterns
  - **Caution!** the graphs aren't really of publishable quality
- summarizing data with means and standard deviations