Theories and numbers

- To determine if a theory is a good one requires collecting and understanding evidence.
  - E.g., Is a theorized effect is actually present?
  - E.g., Exactly what has been observed?

Reasoning with numbers

- Public reasoning and persuasion with numbers often employs them in a highly representative way:
  - a complex state of affairs is boiled down to some number ("state-istic")
- We should always ask:
  - Is that a big, small, worrisome, reassuring, surprising, or intelligible number?
  - It depends on how well we understand what it represents, and on how accurate the number is

Numeracy

- CDF Yearbook: "The number of American children killed each year by guns has doubled since 1950."
- Claim as written in the journal:
  - "Every year since 1950, the number of American children gunned down has doubled."
- CDF: *n* deaths in 1950; therefore 2*n* deaths in 1994
- Journal article: *n* deaths in 1950; therefore *n* x 2^45 deaths in 1995 (n=1; ans=8 trillion in 1995 alone)

Representative numbers

- Interpreting:
  - Percentages
  - Averages
- The crucial questions:
  - Lost information?
  - Misleading suggestion?
  - Intelligibly quantified?

Percentages

- Not (normally) an absolute number
- Meaningfulness depends on the size of the values involved
- Cannot be simply combined with other percentages, without knowing differences in absolute values
    - E.g., 40% of Class 1 got an A grade; 60% of Class 2 got an A grade. We can't average these and conclude that 50% of both classes combined got an A grade.

Psychology and percentages

- Gigerenzer and colleagues have shown that representational format is crucial to reasoning success
- When presented with a reasoning task using percentages (rate of success of mammogram screening, etc.) 10% of participants reasoned correctly
  - when trained, they forgot 1 week later
- When the same information was presented as frequencies in a visual format, 90% reasoned correctly
  - the effect remained for at least 3 months

Joint Waterloo Region Catholic District School Board / St. Jerome's University Lecture

Christian Canada No More?

Bramadat with his son, Max

Paul Bramadat, PhD/University of Winnipeg

According to the 2001 census, the number of Muslims in Canada rose over the previous decade by **129%**, Buddhists by **84%**, Sikhs by **89%** and Hindus by **89%**. Clearly, new waves of immigration have changed the Canadian religious landscape. What does this mean for Canada in the next century?

Author of *The Church on the World's Turf* (2000), Paul Bramadat has written extensively on religion, multiculturalism, and public policy. Associate Professor in the Department of Religious Studies at the University of Winnipeg, he is co-editor (with David Seljak) of ***Religion and Ethnicity in Canada*** (2004).

Friday, November 19, 2004 / 7:30 p.m. / Siegfried Hall / St. Jerome's University

➔ **Book Launch:** Come celebrate the launch of ***Religion and Ethnicity in Canada***, edited by Paul Bramadat and David Seljak (Director of the St. Jerome's Centre for Catholic Experience).

Immigration example

- 'According to the 2001 census, the number of Muslims in Canada rose over the previous decade by 129%, Buddhists by 84%, Sikhs by 89%, and Hindus by 89%. Clearly, new waves of immigration have changed the Canadian religious landscape.'

Canadian wealth distribution

Stats Canada: http://www4.hrsdc.gc.ca/.3ndic.1t.4r@-eng.jsp?iid=22

- Highest quintile wealth increase, absolute terms
  - 1997: $94,500 2007: $126,700
- Lowest quintile increase, absolute terms
  - 1997: $12,400 2007: $13,900
- Percentage change:
  - Highest +34% Lowest +12%
- Absolute change
  - Highest $32,000 Lowest $1500
- In absolute terms, the highest income increased 21 times that of lowest (less obvious looking at % change)

Percent up vs down

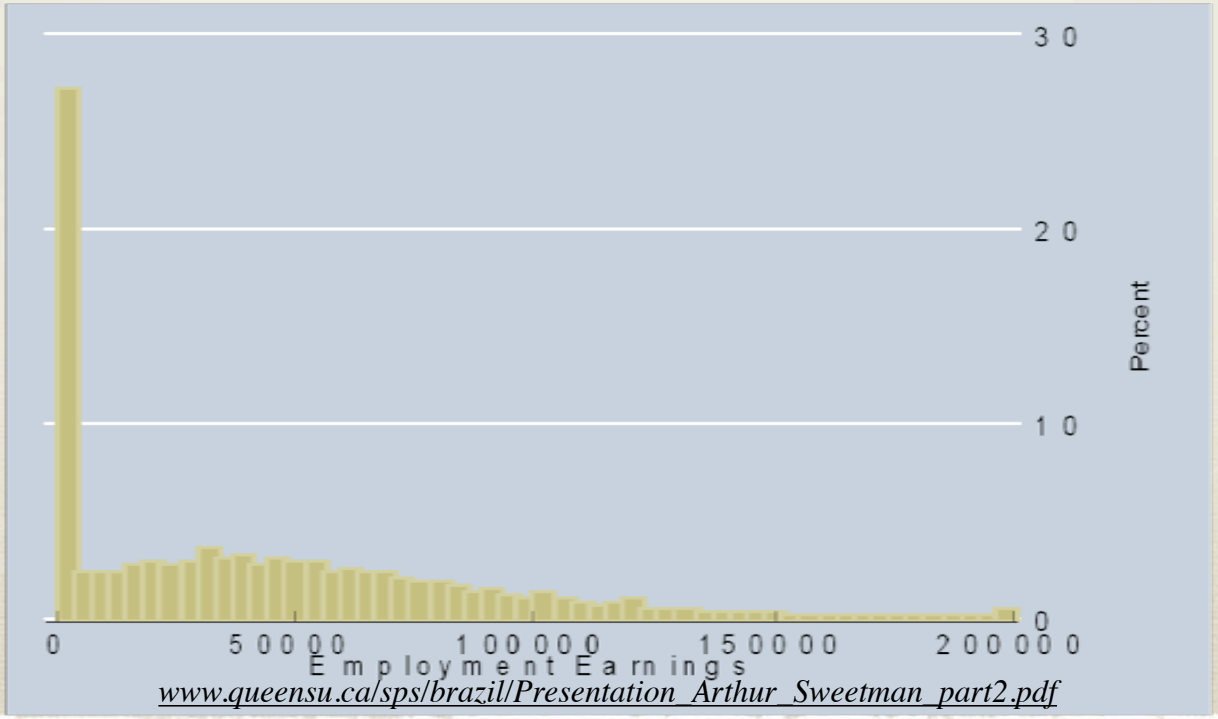
- $1000 investment
- "Markets decreased by 20%"
  - $800
- "Markets increased by 20%"
  - $960
- Need a 25% increase to 'break even'
- Same effect in the opposite order as well

Averages

- Averages are representative (like percentages)
  - a single number that represents something about a set of data
- The good (convenience):
  - distills complex information to a single number
  - permits simple comparisons of data sets

Averages

- The bad (loss of information)
  - The size of the data set isn't transparent
  - The nature or distribution of the data set isn't transparent



www.queensu.ca/sps/brazil/Presentation_Arthur_Sweetman_part2.pdf

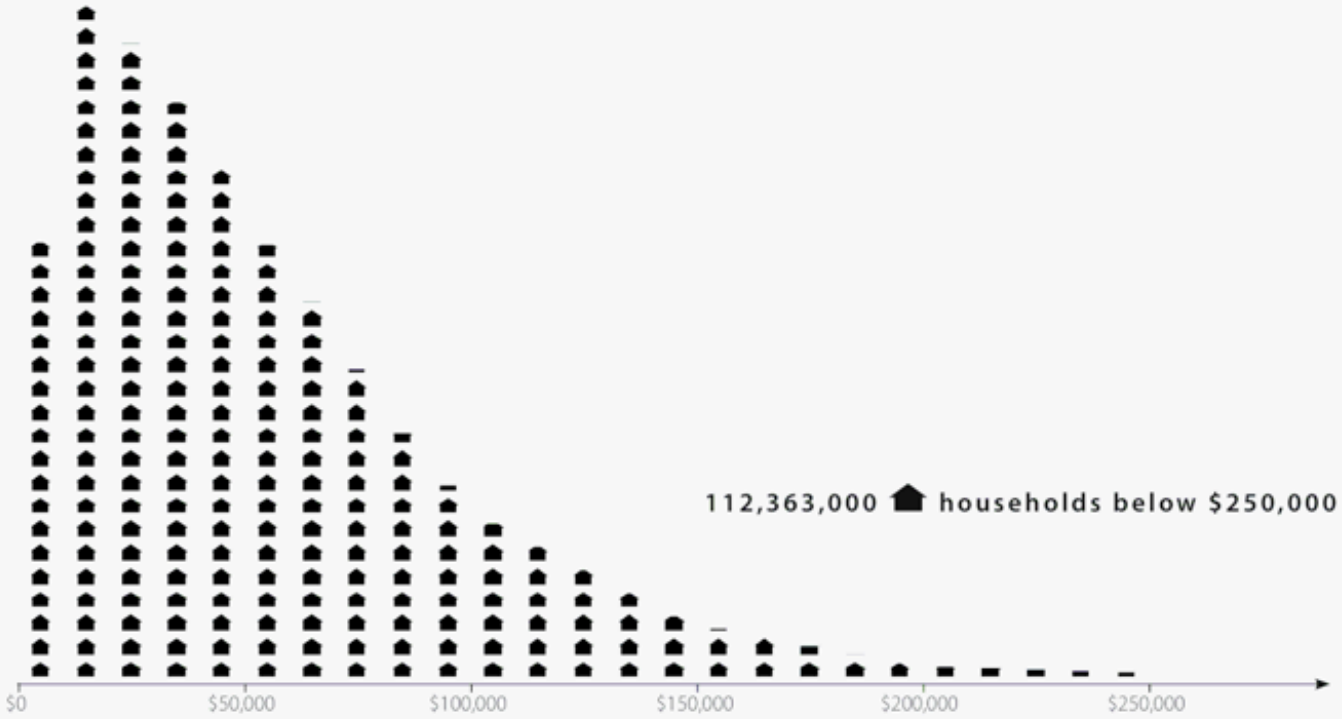
Visualizing Economics
Making the *Invisible Hand* Visible

Visit www.visualizingeconomics.com
to view more examples

2005 United States

Income Distribution (Bottom 98%)

Each ⌂ equals 500,000 households



Averages

- Two related concepts:
  - Mean (average)
  - Median

Mean

- The mean: An arithmetically calculated average, representing the sum of the values of a sample divided by the number of elements in the sample.
- Usually ‘average’ means the arithmetical mean.
    - US mean income (for bottom 98%): $63K

Median

- The element in the set for which half the elements have a greater value and half have a lesser value
  - US median income (for bottom 98%): $46K

Equivalence of averages

- The following pairs of data sets have the same mean:
  - {0, 25, 75, 100}, {50, 50, 50, 50}
- As grades in a seminar over two years, important differences are lost in simply noting a constant average, or median
  - "The class did fine. The average was around 70%."
- Sometimes the full distribution is most informative

Pitfalls of statistical data

- The underlying statistics used to generate averages and percent
- ages have to be gathered properly.
- Must be careful of:
  - Broad trends
  - Sampling issues (comparison class)

Broad trends

- All statistical trends that are uniform (i.e. increasing or decreasing) over a long period of time are correlated
- However, these 'monotonic' trends are not good evidence for a *causal* correlation
- There must be an independent reason to think the two factors are related (or an up & down relation).
- Sometimes you can undermine such claims by noticing that one of the trends was in place before the other.

Broad trends

- Examples of broad trends
  - decline of arable land and decline of puddle ducks in Maryland
  - increase in average height and increase in air travel
  - decline of the cheetah population, increased use of cell phones
- What about:
  - Increase of IQ scores, increased complexity of common visual displays
    - Need independent (testable) reasons to think this might be a relevant correlation

Sampling

- Issues with averaging are compounded when we are only taking a sample from some larger set of data
  - Determining the average height of Canadians involves taking a (relatively small) sample of Canadians and determining their average height
  - Is average is representative?
  - Is the data we sample representative?
- How do we get a representative sample? Alternatively, why should we wonder whether someone else's claims about an average are based on a representative sample?

Sampling

- Bad sampling can result from:
  - having a biased selection technique
    - Comparison class
    - You have to know what the statistic has been gathered with respect to (i.e., know the sampling procedure) in order to evaluate it.
  - getting unlucky

Sampling bias

- Any means of gathering data that tends towards an unrepresentative sample
  - Using a university's alumni donations address list for a survey on past student satisfaction
  - An email survey measuring people's level of comfort with technology
  - A Sunday morning phone survey about church-going
  - A swinger's website for relationship statistics
  - Voluntary responses in general

Sampling bias

- Media:
  - ‘unscientific’ polls
  - screening calls (it’s a debate after all)
  - sports reporting (x-game losing/winning streak)
- Red flags:
  - Small samples (esp. of 1)
  - Arbitrary cutoffs: “Last 17 years…”, “Since 1998…”

Sampling bias

- Examples of comparison class problems
  - The full moon increases accident rates
    - But statistics are gathered only on weekends
- Compare
  - The average Canadian has one testicle
  - The average Canadian male has one testicle.

Sampling bias

- Dropout error
  - Long term studies tend to have 'dropouts' (i.e., those that don't complete the study).
  - This error is the error of not including those who leave in the final data.
    - Any 'course' that makes claims regarding those who finished the course may be relying on 'filtering factors' to get those who won't succeed out of the course early on (or they may make the course so long, or boring, that most people stop going). Then those who actually finish are more likely to succeed.
  - Specific kind of comparison class error

Sampling bias

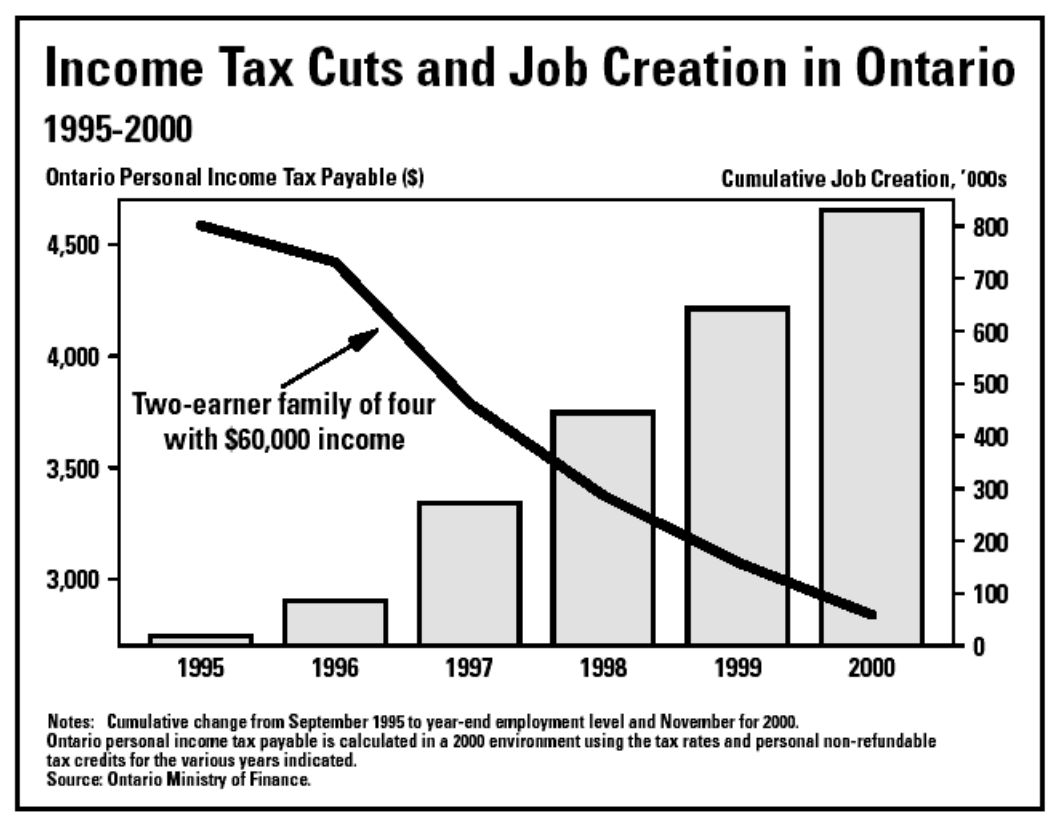
- Dropout error examples
  - 'Innerchange Freedom Initiative' helps reduce recidivism
    - But only 'graduates' (not the starting cohort) of the program are counted for final comparison
  - Dieting is notorious:
    - "Few high-quality studies have assessed weight loss programs. Many of the existing studies present the best-case scenario because they do not account for people who drop out of the program. Of the programs the authors evaluated, Weight Watchers had the strongest studies to support it. The best study found that participants lost 5% of their initial body weight (about 10 pounds) in 6 months and kept off 3% (about 5 pounds) at 2 years. The authors found no published high-quality studies of Jenny Craig or L A Weight Loss" (Tsai et al, 2005, Annals of Internal Medicine)

Comparison class

- The 'right statistic' examples
  - E.g., Five hundred new professors will be hired in the next year ...or... The number of professors will increase by 30% in 2003.
    - But, the increase in enrolment is expected to more than double (double cohort), so this will still make the quality of education decline.
  - Administration: "Arts has hired 13 new faculty this year! Don't tell me I don't have a strong commitment to Arts."
    - But 15 faculty members retired.
  - Amartya Sen has pointed out that despite a hugely lower 'daily income' in many parts of the world, the general population in those countries has a higher standard of living than blacks living in the US.

The right statistic

- "Tax cuts have paid real dividends in terms of a strong economy and jobs in Ontario." ... or not!



Other sampling problems

- Even without a biased sampling technique, we might just get unlucky. Surveying height, we might happen to pick a set of people who are all taller than average, or shorter than average.
- How do we rule out being unlucky in this way?
  - By taking the largest sample we can afford to take.
  - By qualifying our confidence in our conclusions, according to the likelihood of getting unlucky with a sample of the size we chose.

Other sampling problems

- Quality of statistics concerns:
  - US Department of Justice:
    - "There are ... a number of reasons why people do not report crimes: mistrust of police or the criminal justice system, a belief that the police would not act, a feeling that the incident was not important enough, fear of reprisal or embarrassment, or a belief that the incident was a private or personal matter" http://www.ojp.usdoj.gov/bjs/pub/ascii/tcsadnci.txt
  - European Commissioner Poul Nielson
    - "Development statistics are powerful because they document poverty. The repression of knowledge is the strongest weapon in the hands of the privileged in their resistance against social change. And for the poor, knowing the reality is essential for mobilizing awareness...Good statistics are hard to get. They cannot be imported when local capacity is lacking."

Significance

- When we draw inferences from a set of data, we can only be confident in the conclusion to some degree
- *Statistical significance:*
  - a measure of the confidence we are entitled to have in our probabilistic conclusion
  - A function of how precise a conclusion we want

Significance

- Determination of correlation is relative to a *null hypothesis*
  - null hypothesis: observed correlation is accidental
- Significance is measured by a p-value
  - usually .01 or .05 (99% or 95% chance of data being non-random)
- Still need to figure out *why* it's non-random (cause? common cause? other confound?)

Significance and error

- Confidence is cheap. We can always be 100% confident that the probability of some outcome is somewhere between 0 and 1 inclusive -- at the price of imprecision.
- The more precise we want our conclusion, the more data we need in order to have high confidence in it.

Significance and error

- So when we are told the result of some sample, we need to know both:
  - the margin of error (or confidence interval) – that is, how precise the conclusion is
  - and the degree of significance (the p-value)
- This why polls report having, for example, "a 3% margin of error 19 times out of 20"
- If we conducted the same poll repeatedly, we'd have .95 (19/20) probability of getting a result within 3% (on either side) of the reported value

Significance and error

- So what does it mean if a poll reports a 3% difference in the popularity of two political candidates, when it has a +/-3% margin of error at 95% confidence?
  - The difference is at the boundary of the margin of error
  - This doesn't mean that the difference is nothing
  - It does mean that we can't be 95% confident in that difference
- Could convert our 95% confidence into 99% confidence
  - either have to increase the margin of error,
  - or get more data

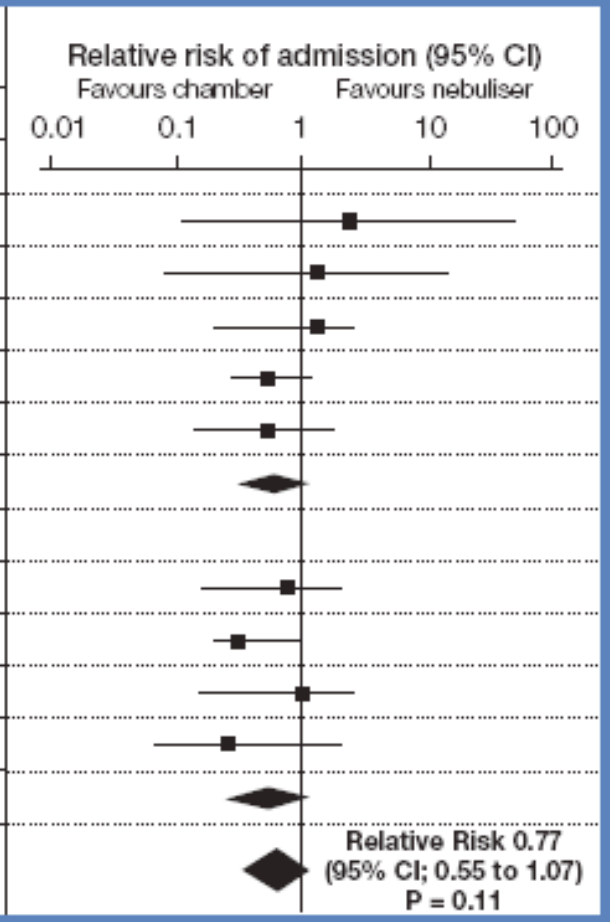
Medical example

- Asthma
- If CI crosses null, not significant
- p-value: Not significant
- But, spacers usually as good as Neb.

Spacers versus nebulisers

Spacers versus nebulisers — relative risk of admission to hospital in acute asthma

Study	Spacer n/N	Nebuliser n/N
Adults		
Colacone	1/40	0/40
Idris	1/15	1/20
Rodrigo	5/49	4/48
Rodriguez	14/36	17/33
Turner	4/27	5/26
Sub-total	**25/176**	**27/176**
Children		
Chou	4/71	5/81
Leversha	10/30	18/30
Ploin	3/31	3/32
Williams	2/42	2/18
Sub-total	**19/183**	**28/170**
Total	**44/359**	**55/346**



n = number in study arm admitted; N = total number in study arm

http://www.nntonline.net/ebm/newsletter/2005/July/P_values_and_confidence_intervals.html

Iraq mortality study

- The Lancet 2004: The number of deaths additional to what would have occurred without the invasion: 98,000.
  - .95 confidence interval was 8,000-194,000.
  - Widely reported as demonstrating a sloppy technique
  - But at .90 CI the study shows a lower bound of approximately 40,000 additional deaths.
    - Most confident of values at the centre of the CI
    - The confidence tails off, but does not abruptly cease, as we consider outlying values of the CI

Terminology

- There may be an ambiguity in the term 'significant'
- News stories often include phrases like, "Scientists reported that eating _____ had a significant effect in lowering the risks of ____".
- But this could mean: A *very tiny* difference in probability was detected, but it was detected 19 times in 20.

Significance

- If the experiment is run 100 times, we should expect 5 of the outcomes to be different from our current result 'R'
- The existence of 5 studies indicating not-R is exactly what is predicted by (i) the truth of R and (ii) the performance of 100 studies testing R to .95 significance
- So citing 3 or 4 *properly conducted* studies supporting R does not mean that there is good evidence for R
- You need to know how many studies were performed altogether!

Consequences

- This sets the standards for careful debate or belief-formation on contested questions quite high
- Note: if results defying expectations are 'more interesting' and if 'interesting' results are more likely to be reported in the media...
- ...then the 1 dissenting study in 20 is likely to be disproportionately reported in the media
  - e.g., MMR & autism

Summary

- A set of data permits you to be
  - confident, to a degree
  - precise, to a degree
- Understanding a statistical claim requires knowing both degrees
  - Using fixed standards of significance (p-values) is the most common way of simplifying the interpretation of a statistical claim, but says nothing of effect size

Summary

- p-value test (e.g., .01, .05, etc.)
  - The chance that an observed result is 'bad luck'
  - Passing means the result is 'significant'
- Confidence interval (CI)
  - An interval around the mean that you're X% certain that the mean is within
- If the X% confidence interval overlaps the null hypothesis (i.e. that the result is 'bad luck'), it will fail that p-value test (1-X%).
  - Suppose a chance result is 50%. If I perform an experiment where the number of positive results is 64/100 (64%) and the 95% CI is +-15%, then the experiment fails the .05 p-value test.

Statistical fallacies

- We've already seen:
  - Seeing patterns (Hot hand, V1 bombing)
  - Regression fallacy (Sports Illustrated jinx)
  - Missing data fallacies (Adopting & conception)
  - Base-rate errors (GRE score and grad school)

More statistical fallacies

- Gambler's fallacy
  - Mistaking the likelihood of the next event for the likelihood of a larger set of events (I'm due!)
- Simpson's paradox
  - The surprising result that the rate for an aggregate statistic is very different from the rates for the sub-groups making up the aggregate

Simpson's paradox

	SUCCESSES	ATTEMPTS	% SUCCESSES
Batter A (lefties)	40	100	**.400**
Batter A (righties)	10	40	**.250**

	SUCCESSES	ATTEMPTS	% SUCCESSES
Batter B (lefties)	39	100	**.390**
Batter B (righties)	1	5	**.200**

A's batting is better than B's for both lefties and righties.
Does it follow that A's is a better batter than batter B?

	SUCCESSES	ATTEMPTS	% SUCCESSES
Batter A	50	140	**.357**
Batter B	40	105	**.366**

Question

- Give an example of events that are statistically independent, say why they are independent. Given an example of events that are statistically dependent, say why they are dependent.