Required for next meeting: January 9, 1996

- For short project 4, a list of between 2 and 4 sources from which you'll start looking for information on your essay.

Another example of Simpson's paradox

This is the graduate school admissions data from Berkeley that I mentioned last week. In fall quarter, 1973, there were 8,442 men who applied for admission to graduate school, and 4,321 women. About 44% of the men and 35% of the women were admitted. Since admissions are made separately for each major, the admissions data were broken down to find out which majors were discriminating against women. Major by major, there did not seem to be any bias against women. The data for the six largest majors, accounting for over one third of the total number of applicants, is given below:

	Men		Women	
Major	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7
Total	2691	44	1835	30

Over 50% of the men applied to the first two majors, which were relatively easy to get into. Over 90% of the women applied to the last four majors, which were much harder to get into. The confounding variable is the major applied to; it is hidden in a comparison of the overall admission rates.

Reference Freeman, D., Pisani, R., Purves, R. and Adhikiri, A. (1991) *Statistics* (2nd edition). WW Norton.

Technical note: testing hypotheses, Part II

There are three main components to testing hypotheses:

- Formulate the question
- Choose a method to answer the question
- Compute the answer

The question

Are the data I'm looking at consistent with the null hypothesis?

The first component is formulating a null hypothesis, first conceptually, and then mathematically. This aspect was discussed in the first technical note on testing hypotheses (November 21).

For example, in the hot hand article, in the section "Cold Facts from the NBA", Tversky and Gilovich used as their null hypothesis "there is no streak shooting". This was formulated mathematically as: "for each player (on the Philadelphia 76ers, in the 1980-81 season), field goal attempts follow the probability law of coin-tossing[1], i.e. successive attempts at field goals are independent, and the probability of a hit on each attempt is constant". (See, in particular, p.19,column 1 "assuming independent shots with a constant hit rate".)

In the cheating papers, the null hypothesis was "the suspect (or pair of suspects) did not cheat". I haven't been able to come up with a concise summary of the mathematical model used to describe this. It seems to be different for each case of cheating, and less precisely described than in the case of streak shooting.

The method

A summary number, called a test statistic, is computed, to measure the agreement of the observed data with what would be expected under the null hypothesis.

In the cold facts section of the hot hand article, there were several test statistics used (but not explained): the serial correlation, the Wald-Wolfowitz runs statistic, and the chi-square statistic. These are all fairly specialized, but the least specialized is the chi-square statistic. It is a single number determined from the table of "observed" and "expected" numbers of "low", "moderate", and "high" performance sequences. The authors constructed such a table for each player, but didn't include any of the tables in the paper. A summary table for all 9 players taken together is given, and is also on the handout of November 21.

In the cheating examples, the test statistic was usually a so-called z score, based on the number of matching wrongs.[2] In Peeper-Sein, they considered all possible pairs among examinees with similar final scores, and counted the number of matching wrongs among all these possible pairs. Then they computed the mean and standard deviation of these counts, which turned out to be 12.81 and 3.13, respectively. Peeper and Sein had 53 matching wrongs. In the first C&S case, they considered all pairs with S and the other examinees, and again counted the number of matching wrongs. The mean and standard deviation for these counts were 4.75 and 1.65, respectively. S and C had 14 matching wrongs.

[1] The probability of heads was not assumed to be 1/2, but was estimated from each player's total season hit rate.

[2] Several of the cheating examples did not include tests of hypotheses, but presented the information graphically instead. The Peeper-Sein case used graphical display (Figure 1) and a formal test; the Francis Foecke case didn't seem to use any statistical analysis; the Irish case used a graphical display (Figure 1). The two copying cases used z scores.

The answer

The answer is usually given as a "p-value", also called a "significance level", or else is given as the phrase "statistically (in)significant".

For example, in the hot hand articles, we read:

- "these [serial correlations] are not significantly different from 0"
- "a series of Wald-Wolfowitz runs tests reveal that the observed number of runs ... does not depart from chance expectation"
- "none of the nine chi-square [test statistics] approached statistical significance".

In the cheating articles, we read (p.24, col.3) "the odds of obtaining a value this high or higher by chance alone are less than 1 in 1,000,000,000" (10^{-9}, or 1 in a billion). For the copying example (p.13, col.2), "the z-statistic of 5.6 gave a p-value of less than 10^{-8}" (1 in 100 million).

The p-value quantifies whether or not the data are consistent with the null hypothesis on a probability scale: it is the probability of the test statistic taking a value as or more extreme than the data value, if the null hypothesis is true. If the p-value is very small, then we've observed something that is extremely unlikely under the null hypothesis. The conclusion is then that the data are not consistent with the null hypothesis. is not very small, then the data *are* consistent with the null hypothesis. A commonly used cut-point is 0.05 (1 in 20): p-values that are larger than 0.05 are often not reported, but the result "not statistically significant" is reported instead.

Does this all make sense

At each of the three stages there are choices to be made that may or may not be reasonable.

Formulating the null hypothesis mathematically is often very difficult, and is at best an approximation to reality. The criticism of the hot hand article given by Hooke (in the second article) turns on the choice of the null hypothesis. He seems to be suggesting that the null hypothesis should be "there is streak shooting", and see if the data is consistent with this.

There are many ways to summarize data, and choosing a test statistic, a single number summary, is also a choice that can be made well or badly. Since large values of the test statistic will be used as evidence that the data do not support the null hypothesis, the test statistic should ideally take large values when the null hypothesis is not true.[3] There is a large body of statistical theory directed towards "good" choices of test statistics for various situations, but when explanations other than the null hypothesis are not well understood or formulated, the theory doesn't

[3]The explanation, or model, for the data if the null hypothesis is not true is called the alternative hypothesis.

help much. Other published criticisms of the hot hand article turned on the choice of test statistics: the authors argued that none of the three test statistics considered would tend to be very large under plausible models for streak shooting.

The final outcome of all this juggling is the phrase that gets the most attention: "statistically significant", or "not statistically significant", or in extreme cases "the probability that this data arose by chance alone is less than 1 in ...". Most people don't realize just how much juggling went on before this number was arrived at. And, of course, using a p-value of 0.05 to determine whether a result gets the label "statistically significant" is completely arbitrary, although very well established.

In the Globe and Mail this week

- "Tests support psychic phenomena" December 2. A new analysis of psychic research by a professor of statistics at the University of California, Davis, has uncovered statistical evidence for psychic phenomena. A panel of experts was contracted to evaluate the results of a secret CIA program called Stargate. This made national TV news in the US. The details are available on the web http://www.stat.ucdavis.edu/users/utts/.

- "Ozone hole' growth termed unprecedented" December 2. The ozone hole grew at an unprecedented rate in 1996, according to the World Meteorological Organization.

- "Civic leaders face some tough choices" November 30 (A4). Article about how municipalities will have less money but more flexibility. Includes an innovative version of the pie chart – I guess you'd call it a donut chart.