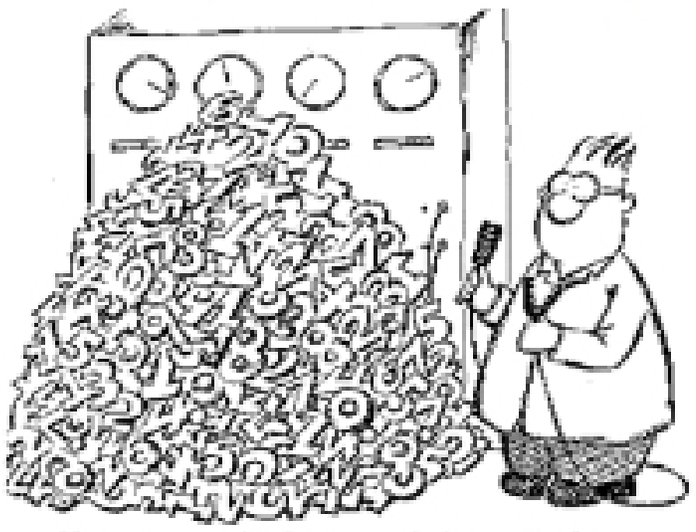
Significance Tests

Uses and Abuses

I: The Dangers of Data Snooping



If you torture the data enough, it may confess

Testing Many Hypotheses

Psychiatrists examined a sample of schizophrenics and a sample of non-schizoids. Measured 77 variables including religion, family background...

- Did 77 hypothesis tests of the null hypothesis of no difference
- Found 2 of 77 had P-value < .05

If you tested null hypotheses 77 times and they were all true, how many times would you expect the P-value to be less than .05?

$$.05 \times 77 = 3.85$$

What is the chance that at least one would have a P-value < .05 if all tests were independent?

$$1 - .95^{77} = 1 - .0193 = .98$$

Example: Screening for Carcinogens

Specially bred rats are used to screen for possible carcinogens. Half of a group are given a normal diet and half are given the test chemical. Cancer rates in the two groups are compared using the two sample z-test.

Investigators look at cancer rates in about 25 organs. Suppose that the P-value for liver is less than 1%. What can you conclude?

Since so many hypotheses have been tested, the P-value isn't "really" 1%.

If only one true null hypothesis is tested, the chance of such a P-value is 1%.

If 25 hypotheses are tested, we expect

$$25 \times .01 = .25$$

of them to have a P-value of 1%

This is a case of multiple hypothesis tests, or data-snooping, or a fishing expedition: we have to be careful in interpreting results.

Surely, cancer-screening is a good thing. What can be done in situations in which many hypotheses are tested?

The soundest strategy is to replicate the study, testing particularly for those things found significant in the first study.

Another strategy is to split the data randomly in half, data-snoop on one half and test rigorously on the other half.

Results should make scientific sense.

II: The Interpretation of P-Values

Less than .05: "statistically significant"

Less than .01: "highly significant"

It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occcur more than once in twenty trials."

R. A. Fisher

Ronald A. Fisher (1890-1962)

Fisher was a student at a time when there was still controversy about Darwin's theories and when Mendel's work on genes had just been rediscovered. Fisher made important discoveries in statistics (eg. maximum likelihood), genetics, selection and (genetic) dominance. It could be said that he invented a large part of modern statistics.