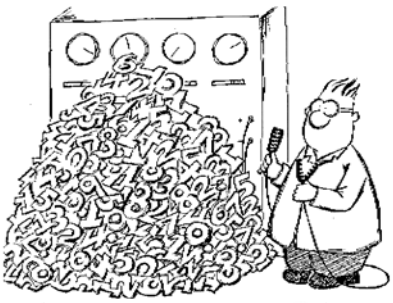
Significance Tests

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.

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point) or at one in a hundred (the 1 percent point). Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

III: One-tailed and two-tailed tests

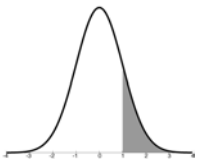
Hypothetical example: items produced by a manufacturer have an average value with respect to some measurement of 100. A sample of 64 from one lot has an average value of 100.25 and an SD equal to 2.

Is there evidence of a problem with that lot?

Test statistic:

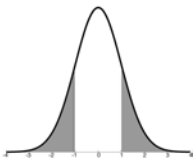
$$z = \frac{100.25 - 100}{.25} = 1$$

one-tailed test



P-value = 16%

two-tailed test



P-value = 32%

One-sided and two-sided alternatives

- In this example, the natural alternative is two-sided, so use a two-tailed test
- In the "healing touch" example, the natural alternative is one sided, so use a one-tailed test

IV: Statistical Significance and Practical Significance

They are not the same thing! For example, in comparing two treatments, the SE of the difference of averages depends on the sample sizes--the larger the sample sizes, the smaller the SE.

So very small differences can be two or more SEs apart if the sample sizes are large. A very small difference may be practically of little importance.

Example

Suppose that math SAT scores in the absence of coaching have a mean of 475 and an SD of 100.

(1) The average of a sample of 100 students who have been coached is 478. Is this significantly different from 475?

SE of average = 100/10 = 10.

Not significantly different.

(2) Suppose a sample of size 10,000 is taken and their average is 478. Is this difference statistically significant?

SE of average = $100/\sqrt{10{,}000} = 1$

(478 - 475)/1 = 3. So P-value is very small.

But is the difference practically significant?

Small differences can be found by large samples and only large differences by small samples.

If the P-value is large, people sometimes say, “the null hypothesis has not been rejected.”

Does that mean that the “null hypothesis is accepted?”

According to Census data, in 1950, 13.4% of the US population lived in the West. In 1990, 21.2% of the population lived in the West. Is the difference practically significant? Is it statistically significant?

V: Could the Results be Due to Chance? The Necessity of a Model for Chance

This is the question that significance testing attempts to answer. In order to do so, the meaning of “due to chance” has to be clarified.

We attempt to do so through a “chance model” (“probability model,” “stochastic model”).

Without a chance model, significance testing is an empty arithmetic exercise.

In this course we have envisioned chance models as draws of tickets from boxes.

The chance model is explicit when there is randomization in sampling or in experimentation.

It is less explicit, but plausible, in contexts in which the Gauss model of measurement error could be used.

Example: The English vocabulary scores of students in a particular high school are higher for those students who study foreign languages than for those who do not. The difference is “significant” as determined by a two sample z-test.

(a) Is the z-test valid?

(b) Can we conclude that studying foreign languages is good for English vocabulary?

Example: The English vocabulary scores of students are higher for those students who study foreign languages than for those who do not. These results were found from a random sample of students in California. The difference is “significant” as determined by a two sample z-test.

(a) Is the z-test valid?

(b) Can we conclude that studying foreign languages is good for English vocabulary?

D3. Two researchers studied the relationship between infant mortality and environmental conditions in Dauphin County, PA. As part of the study they recorded for each baby born there during a six month period, in what season the baby was born, and whether the baby died before reaching one year of age.

	Season of Birth	
	July-Sept	Oct-Dec
Died before 1 year	35	7
Lived 1 year	958	990

Book’s answer: A test of significance is not appropriate here. This is not a probability sample.

Is there a plausible chance model? At a vague intuitive level, chance was at work. What about the following model: There were 1990 total births--993 births in July-Sept and 997 in Oct-Dec. Of these, 42 died during the first year. To say that season of birth is not related to mortality is analogous to a chance model in which 42 tickets are drawn without replacement from 1990 of which 993 are labeled July-Sept and 997 are labeled Oct-Dec. Is the data consistent with this chance model?