# Section 6.3 - Use and Abuse of Tests

Statistics 104

Autumn 2004

# Choosing a Significance Level

Choosing a significance level usually only makes sense if you want to make a decision, e.g.

Should this drug be allowed to go to market

You want to act like $H_0$ is true of like $H_0$ is not true.

Most of the time, there is no sharp border between statistical "significance" and "insignificance".

Do you want to treat a $p$-value $= 0.049$ differently from a $p$-value $= 0.051$. (or 0.04999 vs 0.05001)

Also the amount of evidence that you want to use to make a decision will vary with the problem of interest.

What is the cost if you make the wrong decision? What is the benefit if you make the correct decision?

"Extraordinary claims require extraordinary evidence" (Carl Sagan, 1934 - 1996)

Would you want to use the same $\alpha$ for positive claims for ESP, therapeutic touch, or cold fusion as for radiation being harmful, a new beta-blocker will help lower blood pressure, or is Ohio State University running a crooked football program.

My $\alpha$ level for a particular problem may be different than your $\alpha$ level.

# What Statistical Significance Doesn't Mean

- Statistical Significance $\neq$ Practical Significance

  Big sample sizes make it possible to identify small deviations from $H_0$

  $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$

  $n = 1000000$; $\bar{x} = 0.01$; $\sigma = 1$

  $z = 10$; $p$-value $\approx 0$

  99% CI $= 0.01 \pm 2.576 \times 0.001 = 0.01 \pm 0.002567$

  Can have strongly statistically significant results with small effects.

- Rejecting $H_0$ doesn't mean its not true.

  You will reject $H_0$ $100\alpha\%$ of the time when $H_0$ is actually true. (Assuming that you always do $\alpha$ level tests.)

  The cutoffs for hypothesis tests are based on the sampling distribution of the test statistic assuming $H_0$ is true. The least likely outcomes totally to $100\alpha\%$ end up in the rejection region.

# Don't Ignore Lack of Significance

- Not rejecting $H_0$ doesn't mean its actually true.

  You many have a small sample and not have enough information to show the difference.

  Suppose you want to test $H_0 : \mu = 0$ but the truth is actually $\mu = 1$. Also suppose that the sample size is small so that you get a 95% CI of the form $\bar{x} \pm 5$ with your sampling procedure.

  A lot of the time, 0 will be in the confidence intervals if you did repeated studies (around 93%). You would only correctly reject $H_0$ about 7% of the time.

  In a larger study, you may just get unlucky and observe an unlikely event.

| | Reject $H_0$ | Don't Reject $H_0$ |
|---|---|---|
| $H_0$ true | ✗(Prob $= \alpha$) | ✓ |
| $H_0$ false | ✓ | ✗ |

The probability of not rejection $H_0$ when you should depends on what the true parameter value is.

In almost all problems, any data set can be observed under both the Null and Alternative hypotheses, though with different likelihoods. So it usually is impossible to avoid both types of errors.

- Results with no statistical significance can be interesting

  People tend to look for small $p$-values. They may not report studies with "insignificant" results.

  Information from these studies can be useful for future researchers (parameter estimates and CIs).

  Should further studies be done. CIs from past studies suggest that an "interesting" result is somewhat plausible.

If so, what sample size should be used – the estimate of $\sigma$ can be used to be used to give a new sample size.

Information from small studies can be combined to give more precise information. These techniques are known as meta analysis. However for valid inference, all similar trials performed are required.

Because of these problems, some groups are moving away from fixed level testing (American Psychological Association). Instead they are requiring $p$-values, parameter estimates, and standard errors (or CIs) to be reported in articles published in their journals.

There is also a movement to get all clinical trials to be reported to a registry, regardless of funding sources. The movement for this registry has to do with good science and the belief that some researchers try to hide bad results.

# Statistical Inference is not Valid for All Sets of Data

Biased samples give biased confidence intervals and biased tests.

Underlying tests and confidence intervals is randomization. Without it, the probability results associated with these tests and CIs may not hold

Garbage In $\longrightarrow$ Garbage Out

# Beware of Searching for Significance

Example: Factors influencing infection rates associated with liver transplants

| | | | |
|---|---|---|---|
| Age | Drugs | Surgery time | Blood transfused |
| Donor type | Bilirubin level | Complications | Creatinine level |
| Stool sample 1 | Stool sample 2 | Stool sample 3 | Stool sample 4 |

Suppose I performed a 5% test examining each of these factors, the chance that at least one of them will be found significant, even if none of them actually influence infection rates, will be greater than 5%.

In fact, it could be as high as 60%. (Usually it will be less.)

If you did 100 tests at 5%, you should expect around 5 to be significant due to chance, even if $H_0$ is true in each case.

You can't mindlessly apply significance test.

It is possible to adjust the analysis to take account of the multiple comparisons. We will see examples of it later, for example when we get to Analysis of Variance (ANOVA).

In addition, doing exploration like this can be useful. Just don't take the $p$-values for each test at face value.

You can use results of the data exploration to design future studies where you can focus on the interesting factors and design valid tests to examine their significance.