Statistics and Imaging

Jon Clayden <j.clayden@ucl.ac.uk>

DIBS Teaching Seminar, 23 Mar 2015
“Statistics is a subject that many medics find easy, but most statisticians find difficult”

— Stephen Senn (attrib.)
**Purposes**

- Summarising data, describing features such as **central tendency** and dispersion

- Making inferences about the **population** that a given **sample** was drawn from
Hypothesis testing

• A null hypothesis is a default position (no effect, no difference, no relationship, etc.)

• This is set against an alternative hypothesis, generally the opposite of the null

• A hypothesis test estimates the probability, \( p \), of observing data at least as extreme as the sample, under the assumption that the null is true

• If this \( p \)-value is less than a threshold, \( \alpha \), usually 0.05, then the null is rejected and treated as false

• 5\% of rejections are therefore expected to be false positives

• The rate at which the null hypothesis is correctly rejected is the power

• NB: Failing to reject the null hypothesis does not constitute strong evidence in support of it
The \textit{t-test}

- A test for a difference in means ...
- ... which may be of a particular sign (\textit{one-tailed}) or either sign (\textit{two-tailed}) ...
- ... either between two groups of observations (\textit{two sample}), or one group and a fixed value, often zero (\textit{one sample}) ...
- ... which is valid under the assumptions that the groups are approximately \textit{normally-distributed}, \textit{independently sampled} and (for some implementations) have \textit{equal population variance}
Anatomy of a test

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \]

\[ \nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left( \frac{s_1^2}{n_1} \right)^2 \left( \frac{1}{n_1-1} \right) + \left( \frac{s_2^2}{n_2} \right)^2 \left( \frac{1}{n_2-1} \right)} \]
In R

> t.test(a, b)

Welch Two Sample t-test

data:  a and b
t = -2.6492, df = 197.232, p-value = 0.008722
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.63820792 -0.09351402
sample estimates:
 mean of x  mean of y
-0.1366332  0.2292278

> se2.a <- var(a) / length(a)
> se2.b <- var(b) / length(b)
> t <- (mean(a) - mean(b)) / sqrt(se2.a + se2.b)
> t

[1] -2.6492

> df <- (se2.a + se2.b)^2 / ((se2.a^2)/ (length(a)-1) + (se2.b^2)/(length(b)-1))
> df

[1] 197.2316

> pt(t, df) * 2

[1] 0.00872208
Effect of sample size

Mean of 1000 $p$-values at each $n$
Other common hypothesis tests

- $t$-test for significant correlation coefficient
- $t$-test for significant regression coefficient
- $F$-test for difference between multiple means
- $F$-test for model comparison
- Nonparametric equivalents, e.g. signed-rank test
- Robustness to violations of assumptions varies
Issues with significance tests

- Arbitrary $p$-value threshold
- Significance vs effect size, especially with many observations
- Publication bias: non-significant results are rarely published
- Choice of null hypothesis can be controversial
- Ignores any prior information
- Probability of data (obtained) vs probability that hypothesis is correct (often desired)
The big-picture problem

Unlikely results
How a small proportion of false positives can prove very misleading

1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.

2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.

3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.

Source: The Economist

The Economist, 19th October 2013
Multiple comparisons

See R’s `p.adjust` function for $p$-value adjustments
The picture in imaging

- Hypothesis tests may be performed on a variety of scales
- Worth carefully considering the appropriate scale for the research question
- Dimensionality reduction can be helpful
- Mass univariate testing (e.g. voxelwise) produces a major multiple comparisons issue
SPM

Savitz et al., Sci Reports, 2012
Beyond hypothesis tests

- Models of data as outcomes
- Parameter estimates, confidences intervals, etc.
- Model comparison via likelihood, information theory approaches
- Clustering
- Predictive power, e.g. ROC analysis
- Measures of uncertainty via resampling methods
- Bayesian inference: prior and posterior distributions
Some advice

- Plan ahead
- Be clear what you really want to know
- Use R
- Visualise and understand your data
- Save scripts
- Keep statistical tests to a minimum
- Be aware of sources of bias
- Use available resources at ICH and beyond