Keele University

Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

Keele University

Hypothesis Testing

Keele University

Aim

- Introduce Hypothesis testing framework
  - Explaining problems
- Introduce concept of Type 1 and Type 2 error and power
- Assessing required size of samples
- Addressing multiple hypothesis tests

Hypothesis testing

- Compare two or more groups of objects
  - With data collected on each object
  - With respect to some metric
    - Usually the mean sometimes the variance
  - In order to decide whether the groups differ with respect to the metric
    - Is the difference "substantial" by some criterion?
- Done within context of experiment or quasi-experiment

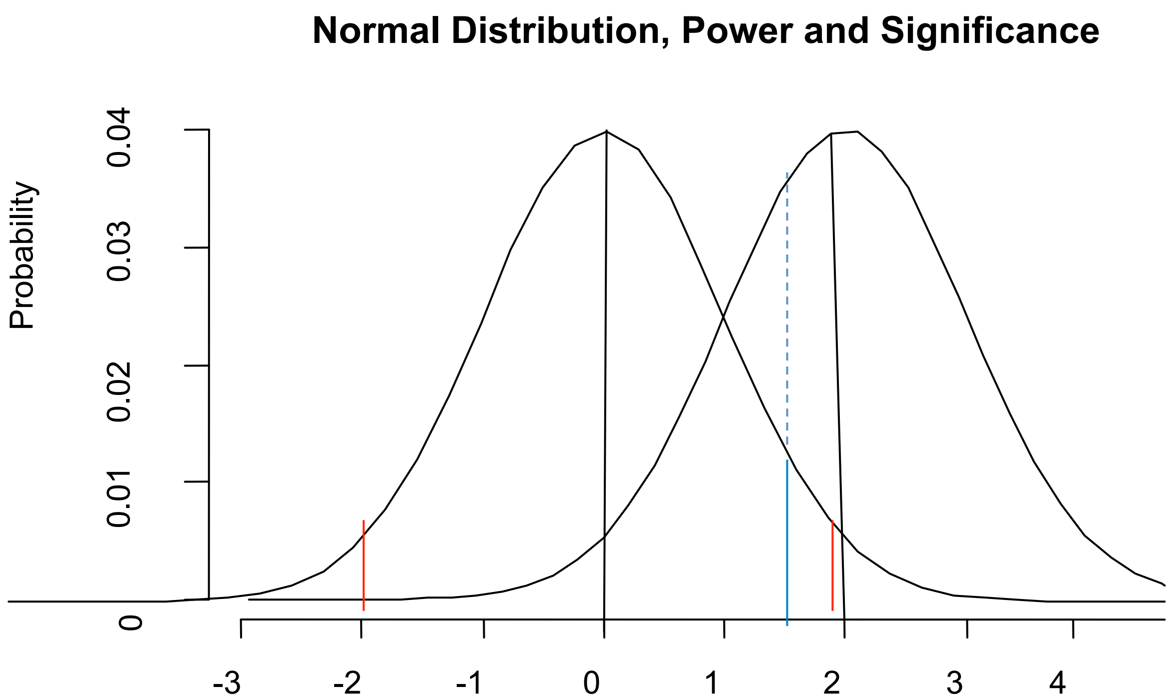
Decision making framework

- Hypothesis that groups are the same
  - Referred to as Null hypothesis (H0)
  - Estimate of metric of interest obtained from group1 is the same, within sampling error, as the estimate from group 2
    - H0 : $\theta_1 = \theta_2$
- Hypothesis that groups are different
  - Referred to as Alternative Hypothesis (H1)
  - One-sided Hypothesis
    - H1 : $\theta_1 >$ (or $<$) θ_2
  - Two-sided Hypothesis
    - H1 : $\theta_1 \neq \theta_2$
  - Difference matters!
    - One sided $\alpha=0.05$ significance $\theta_2 > \theta_1$,critical value $z=1.65$
    - Two-sided $\alpha=0.05$ significance, critical value $|z|=1.96$

Errors and Power

- Type I error
  - Null hypothesis true but rejected
    - Probability of incorrectly rejecting null hypothesis
  - "Controlled" by selected alpha level
- Type 2 error
  - Null hypothesis wrong but not rejected
    - Probability of incorrectly failing to reject null hypothesis
  - Alternative is true but it is rejected
  - Referred to as beta (β)
- Power of a test
  - Probability of correctly rejecting null hypothesis
    - $(1-\beta)$

Comparing Two Distributions



Keele University

Power and sample size

- Important to have reasonable power
  - Advice is β~0.2, power=0.8
- Power is determined by
  - Sample size
  - Alpha level
  - Mean Difference
  - Variance
- Mean difference and variance combined into
  - Effect size =Mean difference/ Standard deviation

Example

- Two theoretical distributions had
  - Mean Difference= 2
  - Variance = 1
  - Alpha level =0.05
  - One-sided test
- From unit normal distribution
  - Value of z corresponding to alpha=1.645
  - Corresponds to z on H1 curve=2-1.645=0.335
    - If alternative distribution re-centered on 0
  - Beta is area of Normal curve to left of -0.355
    - =0.3726
  - Power = 0.6274
- For “real” power analysis, we need to consider a sample

R package

- Package=pwr
- Library(pwr)
- Handles all main situations
  - t-test, ANOVA, correlation, chi-squared etc.
- pwr.t.test(n= ,d= ,sig.level= ,power= ,type= , alternative=)
- alternative is “**two-sided**”, “less”, “greater”
- type=“**two.sample**”, “one.sample”, “paired”
- Estimate missing value of n or power
- If d unknown, choose based on best guess
  - Small effect d=0.2, Medium d=0.5, Large d=0.8

Example

- d=0.5
- alpha= 0.05
- Two-sided, two-sample t-test
- pwr.t.test(d=0.5,sig.level=0.05,power=.8)
  - Requires n=64 entities in each group
  - How many if d=0.8?
  - What power if n=15 in each group?
- Power analysis only tractable in simple cases

Effectiveness of tests

- Statisticians use simulation studies to assess effectiveness of tests
  - Extract a sample from each of two of theoretical populations
  - Perform test for the sample for specific alpha level
  - Record outcome test (i.e. reject or accept H0)
  - Repeat for many different pairs of samples
- When the two samples are from an identical distribution
  - The proportion of reject outcomes should ~ α
- When samples are from different distributions
  - The proportion of rejects estimates the power i.e. $(1-\beta)$
- Used to
  - Assess impact of deviations from Normality
  - Assess relative effectiveness of alternative tests

Hypothesis Testing Problems

- Level of significance is arbitrary
  - Why use 0.05, 0.01 rather than 0.025?
- Significance is not the same as importance
  - Recall $var(\overline{x}_1 - \overline{x}_2) = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$
    - Variance of difference between means decreases as n_1 and n_2 increase
    - Any small difference is importance with large enough sample sizes
- Do enough tests and you'll find something significant
  - With 10 tests probability of one or more by chance
    - $1-[(1-.05)^{10}]=0.4013$

Compromise position

- Report
  - Confidence limits not just p-values
  - Effect size not just “t” or “z” values
    - Effect size removes reliance on sample size

$$d = \frac{(\bar{x} - \mu)}{s}$$

- Adjust significance level depending on number of tests

Adjusting p-values

- Bonerroni
  - Set new value p= α/n, for n=# tests
  - Very conservative
- Rom's "sequentially retentive" method
  - Most effective in a study of 5 alternative methods
  - Tables for alpha 0.05 & 0.01, and n=1 to 10
  - Order the p values for set of tests in descending order i.e. largest p value first
  - Set k=1, if $p_{[k]}<d_k$ from table reject all null hypotheses
  - Otherwise accept null hypothesis H01 and put k=k+1
  - Continue until all hypotheses are accepted or rejected

Hochberg's method

- Hochberg's method similar to Rom's and is simpler when many tests
  - Let p1,…,pC be the α probabilities from C tests
  - Order the p-values in descending order
    - $p[1] \geq p[2] \ldots . \geq p[C]$
  - Put k=1
    - Reject all hypotheses if $p[k] \leq \alpha/k$ (i.e. α) & exit
      - Otherwise fail to reject hypothesis 1 and continue
  - Increment k by 1. If $p[k] \leq \alpha/k$ stop and reject all remaining hypotheses
  - If $p[k] > \alpha/k$ keep hypothesis k , repeat previous step

Example of ROM's method

ROM's Table

Number of tests	alpha= 0.05	alpha= 0.01
1	0.05	0.01
2	0.025	0.005
3	0.0169	0.00334
4	0.0127	0.00251
5	0.0102	0.0021
6	0.00851	0.00167
7	0.00730	0.00143
8	0.00639	0.00126
9	0.00568	0.00112
10	0.00511	0.00101

Example results

Number of Tests	p-values	p-values ordered
1	0.006	0.054
2	0.025	0.049
3	0.033	0.033
4	0.054	0.025
5	0.049	0.010
6	0.010	0.006

Conclusions

- Hypothesis testing has philosophical problems
- However, it is advisable to be pragmatic
  - The purpose is to be honest
  - And to be seen to be honest
- The most important things are
  - Be careful about multiple tests
  - Try to ensure adequate power
    - As many independent observations as possible