Lecture 23: Comparing Means

Dewayne E Perry
ENS 623
perry@ece.utexas.edu

Preliminaries

- $p = \alpha = \text{significance}\ \mathit{level}$
- Significance *test* = effect size x study size
- *p* determined from the specific significance *test*
- The higher *test* value, the lower the *p* value
  - the greater the *significance* of the results
  - the lower the *significance level*
- *Explanatory Variable* = independent variable
- *Response variable* = dependent variable

t Tests, *r* and *d*

- Most common situation: comparing means
  - experimental vs control
  - one treatment vs another
  - etc
- Most common method: *t* test
  - no differences in the mean, or equivalently
  - no relationship between independent and dependent variables
- *t* test made up of two components
  - the size of the effect
  - the size of the study
- *t* and *r* relationship
  - *r* indexes the size of the effect
  - *df* = number of pairs of scores less two ($N - 2$)
  - relationship defined previously

t Tests, *r* and *d*

- when change the size-effect index, also change the study size index
- Alternative for *t*
  - index it by the standardized difference between the group means *(M1 - M2)/S*
  - $t = (M_1 - M_2 / S) \times (1 / \sqrt{(1/n_1)/(1/n_2)})$
  - when interested in the effect size in the sample
- *Cohen's d*
  - effect size in the population, not sample
  - $d = (M_1 - M_2) / \sigma$

Maximizing *t*

- *t* can be maximized in 3 ways
  - driving the means further apart
    - strong treatments
  - decreasing S and σ , the variability within groups
    - maximizing standardization of our procedures
    - using fairly homogeneous samples that are substantially correlated with the independent variable
  - increasing the effective size of the study
    - as sample size increases, *t* increases
    - make n1 and n2 as nearly equal as possible - unequal group sizes effectively loses subjects
    - for any give effect size, *t* would be about the same if we had 99 and 1 as it would if we had 2 in each group

Interpreting *t*

- Generally like large values of *t*
  - ie, events unlikely to occur if the null hypothesis were true
- 2 ways of thinking about H0 and the *t* test
  - the means do not differ in the populations from which we have randomly sampled the subject
    - treatments A and B have the same mean benefit score
  - there is no relationship between the *X* and *Y*, the independent and dependent variables
    - correlation between the treatment condition and the benefit score is 0
- Generally think of *t* as a single test of statistical significance, but . . .

Interpreting *t*

- Might better think of it as a family of tests of significance
  - different distribution of t values for each possible value of $n1 + n2 - 2 = df$
  - 2 extremes:
    - $df = 1$ lower center and higher at the tails
    - df = infinity a normal distribution
- Most *t* distributions resemble std normal
  - all are symmetrical, centered at 0, having tails that do not touch down
- for a given level of significance *p*, the *t* value required is smaller as *df* increases
  - for $p = .001$,
    - $t = 318$ if $df = 1$
    - $t = 3.55$ if $df = 20$
  - remember, *t* = effect size x study size

Interpreting *t*

- If null hypothesis were true
  - Means do not differ, no effect, effect size index $r = 0$
  - Most likely *t* would be 0
  - We could still get a non-zero *t* value by chance
    - With *df* = 8, *t* could be 1.40 10% of the time by chance
- Have to decide at what significance level we decide against H0
  - standard level is $p < .05$
  - often considered dichotomous decision
  - But what is the difference between $p = .05$ or $p = .06$
  - can always reach any significance level merely by adding more subjects

Computing *t*

- Have seen a number of differing ways to compute *t*
- Most generally useful: M_1 and M_2 are the means, n_1 and n_2 are the number of sampling units, and S^2 is the pooled estimate of the population variance

$$S^2 = \frac{\sum\left(X_1 - \overline{X}_1\right)^2 + \sum\left(X_2 - \overline{X}_2\right)^2}{n_1 + n_2 - 2}$$

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S^2}}$$

Computing *t*

- Example - 2 groups
  - (2, 3, 4, 5) and (1, 2, 1, 2)
  - Sums: 14 and 6
  - Means: 3.5 and 1.5
  - *n* : 4 and 4
  - With 6 *df*, a *M* of 2.83 is significant at the .02 level

$$S^2 = \frac{5+1}{4+4-2}$$

$$t = \frac{3.5-1.5}{\sqrt{\left(\frac{1}{4}+\frac{1}{4}\right)1.00}} = \frac{2}{\sqrt{.5}} = 2.83$$

Non-Independent Samples

- Previous: independent samples
  - Eg, 4 girls, 4 boys
  - Independent if all from different families
  - Non-independent if come from say 4 families
    - Might well be family correlations
    - Eg, boy and girl from same family
- For correlated (or repeated measurements, or matched pairs)
  - Perform calculations on the differences between the n1 and n2 scores *(D = X1 - X2)*
- Treating the previous example as matched pairs
  - Obtain a *t* of 3.46 with *df* of 3 (4 - 1) not 6 and a *p* value of about .02 (the same as if they were independent groups)
    - The larger *t* is offset by a smaller *df*

Non-Independent Samples

- The relationships of *t* for correlated observations to size of effect and size of study are analogous to *t* for independent observations – but for *n* matched pairs

$$S_D^2 = \frac{\sum(D-\overline{D})^2}{n-1} = \frac{4}{3} = 1.333$$

$$t = \frac{\overline{D}}{\sqrt{\left(\frac{1}{n}\right)S_D^2}} = \frac{2}{\sqrt{\left(\frac{1}{4}\right)1.333}} = 3.46$$

Non-Independent Samples

- Two ways to consider how we get *r* for the size of the treatment or group effect for matched pairs tests
  - Via analysis of variance
  - Sums of squares *SS* and an error term

$$r = \sqrt{\frac{SS_{groups}}{SS_{groups} + SS_{error}}} = \sqrt{\frac{8}{8+2}} = \sqrt{.82} = .894$$

t Test Assumptions

- May make incorrect inferences if assumptions not met
- Basic assumptions often summarized by "errors are *IID normal*"
  - Independently and identically distributed in a normal distribution
- *IID normal* translates in to three assumptions
  - The errors are independent - *independence*
    - If observations strongly correlated, might get larger *t* than should
    - Eg, people in a group might be come too much alike –
      - ✓ then need to calculate for groups not individual samples
      - ✓ reduction in *df*
    - See Kenny and Judd 86 for valuable discussion

t Test Assumptions

- The errors are identically distributed – *homogeneity of variance*
  - *t* more accurate if variances of the populations are more nearly equal
  - Only if the two populations variances and sample sizes are very different will this likely lead to serious consequences
  - Can transform the data to make it closer
    - ✓ Square root, logs, reciprocals etc
    - ✓ Goal: make the transformed data most nearly homogeneous in its variances
- The errors are normally distributed – *normality*
  - If not too skewed or not to bi-modal, there seems to be little cause for concern

→ These same assumptions underlie the non-parametric *F* test as well

Nonparametric Procedures

- Sometimes have to make fewer assumptions – referred to as
  - Non-parametric statistics
  - Distribution free statistics
  - Or more generally, *sturdy* statisitics
- Do make the independence assumption
  - But not homogeneity of variance or crude normality assumptions
- Ordinarily equivalent to parametric procedures applied to transformed data
  - Some times cannot make the transformations
- Sometimes samples are too small

F Test

- Comparing 2 or more means – test for
  - No difference between the means, or
  - No relationship between the membership in any particular group and score on the response (dependent) variable
- Like the *t* test, a test of significance
  - Significance test = effect size + study size