Biostatistics

Lecture 7

Hypothesis Testing: Two-Sample Inference

Study Design Issues

- Instead of comparing sample statistics against ‘known’ statistics, we are comparing statistics from two samples
- Two types of studies are important here
  - Longitudinal study
  - Case-control study

Longitudinal study

- Same ‘cohort’ of people followed over time
- Once in the cohort, always in the cohort
- No new members to the cohort after recruitment is closed
- Can be a two visit design (paired) or a multiple visit design (repeated measures)
- Research questions deal with changes over time
  - May be related to initiation of treatment
- Samples are correlated = Data are not independent
- Controls for inherent factors that may not be measured

Cross-sectional study

- A sample of people ‘seen’ only once
- Sampling may be done according to some classification criteria to answer questions similar to those from a longitudinal study, but with different groups of people seen only once
- Samples are independent = Data are independent
- Less expensive to see people only once without worrying about tracking, scheduling repeat appointments, etc.
- However, cannot truly assess changes over time or as a result of a treatment

Paired T-Test

- Test to deal with two observations on the same individuals
  - Before vs. after treatment
  - Before vs. after some biological milestone
- Approach is to calculate differences between two measurements for each individual and then test the difference against zero
  - $d_i = X_{i1} - X_{i2}$
  - Test statistic $t = \bar{d} / \left(s_d / \sqrt{n} \right)$
  - And test against the t-distribution with d.f. = n-1 and associated p-value

Confidence interval for Mean of Paired Differences

- We know that mean of sample differences d_i is normally distributed with standard deviation $s_d/\sqrt{n}$ – the standard error
- The two-sided 100%x(1-α) confidence interval for the mean of d_i is given by

$$\bar{d} \pm t_{n-1,1-\alpha/2} s_d / \sqrt{n}$$

Two-sample t-test for Independent Samples with Equal Variances

- Assume now that samples are independent samples from a cross-sectional study
- May be a subset of a longitudinal study taken at one time point, say, at baseline
- Test statistic $t = (\bar{x}_1 - \bar{x}_2) / s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

- We can use a pooled estimate of the variance as

$$s = \sqrt{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right] / (n_1 + n_2 - 1)}$$

Confidence interval for Comparison of Means from Two Independent Samples with Equal Variances

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2} s / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Testing for Equality of Two Variances

- How do we test for equal variances for the t-test
- We calculate the ratio of the variances
  - $F = s_1^2 / s_2^2$
- Test that against the F distribution with n_1 numerator and n_2 denominator degrees of freedom
  - Also called the F test – a major part of regression analysis
  - Two-sided tests, so we reject for small and large values of the F statistic
- Because it is two-sided, does not matter which variance is in the numerator vs. the denominator

Two-sample t-test for Independent Samples with Unequal Variances

- Essentially the same statistic, but with a different standard error of the difference
- Test statistic

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Distribution of this is difficult to determine exactly (i.e., difficult to determine exact d.f. for t-distribution, so use Satterthwaite approximation for d.f.

Confidence interval for Comparison of Means from Two Independent Samples with Unequal Variances

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2} / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Strategy for Testing Equality of Means

- Test equality of variances
- If not equal, use t-test for unequal variances
- If equal, use t-test for equal variances

Outliers

- Extreme values can have a major effect on both means and standard deviations as well as measures of relationship, such as correlation
- How to identify
  - Visual inspection
    - Stem and leaf plots
    - Other graphical devices
    - Listings of 5 largest and smallest values
  - ESD procedure
    - Not used much in practice
    - Other similar procedures exist

Handling outliers

- Try to find cause
  - Data entry error, transcription error, etc.
- Exclude possible outliers if values are not possible
- Run analyses with outlier(s) in and out of data set
- ‘Winsorize’ the data so that extreme values are assigned a large, but possible, value
  - Usually done for only a few values

Sample Size Estimation for Comparison of Two Means

- Assuming equal sample sizes in the two groups, the sample size in each group is estimated by

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2}$$

  - n is the sample size in each group to have power of 1-β of finding a significant difference based on a two-sided α level significance test, assuming the absolute value of the difference between the two means is Δ
- Slightly modified if we know that the two groups may not be equal size with $n_2 = kn_1$

Power Estimation for Comparison of Two Means

- For a specific comparison, the power is approximated by

$$Power = \Phi\left[-z_{1-\alpha/2} + \frac{\sqrt{n_1}\Delta}{\sqrt{\sigma_1^2 + \sigma_2^2 / k}} \right]$$

Sample-size estimation for longitudinal studies

- The major issue with longitudinal studies is the correlation between measurements, which affects the variance of the difference
- If measurements are independent, the variance of the difference is simply the sum of the variances
- If the measurements are correlated, the variance of the difference is

$$\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

Sample-size estimation for longitudinal studies

- Thus, sample size is estimated by

$$n = \frac{2\sigma_d^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

- and power is estimated by

$$Power = \Phi\left[-z_{1-\alpha/2} + \frac{\sqrt{n}\Delta}{\sigma_d \sqrt{n}} \right]$$