Mathematical Biostatistics Bootcamp: Lecture 10,

Brian Caffo

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 13, 2012

Table of contents

Independent group t confidence intervals

- Suppose that we want to compare the mean blood pressure between two groups in a randomized trial; those who received the treatment to those who received a placebo
- We cannot use the paired t test because the groups are independent and may have different sample sizes
- We now present methods for comparing independent groups

Notation

- Let $X_1, \ldots, X_{n_x}$ be iid $N(\mu_x, \sigma^2)$
- Let $Y_1, \ldots, Y_{n_y}$ be iid $N(\mu_y, \sigma^2)$
- Let $\bar X$, $\bar Y$, S_x, S_y be the means and standard deviations
- Using the fact that linear combinations of normals are again normal, we know that $\bar Y - \bar X$ is also normal with mean $\mu_y - \mu_x$ and variance $\sigma^2(\frac{1}{n_x} + \frac{1}{n_y})$
- The pooled variance estimator

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$$

is a good estimator of σ^2

Note

- The pooled estimator is a mixture of the group variances, placing greater weight on whichever has a larger sample size
- If the sample sizes are the same the pooled variance estimate is the average of the group variances
- The pooled estimator is unbiased

$$
\begin{aligned}
E[S_p^2] &= \frac{(n_x - 1)E[S_x^2] + (n_y - 1)E[S_y^2]}{n_x + n_y - 2} \\
&= \frac{(n_x - 1)\sigma^2 + (n_y - 1)\sigma^2}{n_x + n_y - 2}
\end{aligned}
$$

- The pooled variance estimate is independent of $\bar{Y} - \bar{X}$ since S_x is independent of $\bar{X}$ and S_y is independent of $\bar{Y}$ and the groups are independent

Result

- The sum of two independent Chi-squared random variables is Chi-squared with degrees of freedom equal to the sum of the degrees of freedom of the summands
- Therefore

$$
\begin{aligned}
(n_x + n_y - 2)S_p^2/\sigma^2 &= (n_x - 1)S_x^2/\sigma^2 + (n_y - 1)S_y^2/\sigma^2 \\
&= \chi^2_{n_x - 1} + \chi^2_{n_y - 1} \\
&= \chi^2_{n_x + n_y - 2}
\end{aligned}
$$

Putting this all together

- The statistic

$$\frac{\frac{\bar Y - \bar X - (\mu_y - \mu_x)}{\sigma\left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}}}{\sqrt{\frac{(n_x + n_y - 2)S_p^2}{(n_x + n_y - 2)\sigma^2}}} = \frac{\bar Y - \bar X - (\mu_y - \mu_x)}{S_p\left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}}$$

  is a standard normal divided by the square root of an independent Chi-squared divided by its degrees of freedom
- Therefore this statistic follows Gosset's t distribution with $n_x + n_y - 2$ degrees of freedom
- Notice the form is (estimator - true value) / SE

Confidence interval

- Therefore a $(1-\alpha) \times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar Y - \bar X \pm t_{n_x + n_y - 2, 1 - \alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

- Remember this interval is assuming a constant variance across the two groups
- If there is some doubt, assume a different variance per group, which we will discuss later

Likelihood method

- Exactly as before,

$$\frac{\bar Y - \bar X}{S_p\left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}}$$

  follows a non-central t distribution with non-centrality parameter $\frac{\mu_y - \mu_x}{\sigma\left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}}$
- Therefore, we can use this statistic to create a likelihood for $(\mu_y - \mu_x)/\sigma$, a standardized measure of the change in group means

Example

Unequal variances

- Note that under unequal variances

$$\bar Y - \bar X \sim N\left(\mu_y - \mu_x, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

- The statistic

$$\frac{\bar Y - \bar X - (\mu_y - \mu_x)}{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)^{1/2}}$$

approximately follows Gosset's t distribution with degrees of freedom equal to

$$\frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y - 1)}$$