12: TWO-SAMPLE INFERENCE

Larsen & Marx 9.2, 9.4

Prof Amanda Luby

Today, we're going to continue our exploration of inference for a few different settings beyond inference for the mean or proportion of a population. Specifically, we're going to derive the (approximate) sampling distributions for a *difference in means* and a *difference in proportions*. We'll see that even in simple settings where we're able to make "nice" assumptions, deriving exact test statistics quickly becomes unwieldy.

1 Inference for a difference in means

One of the most common settings for inference is comparing the means for two groups. For example, if we split a random sample of patients into a *treatment* and a *placebo* group in a clinical trial, do we obtain different amounts of improvement? We could also be interested in measuring differences between existing subgroups within a population, like those who grew up within a 50 mile radius of a superfund site compared to those who did not.

1.1 Assuming $\sigma_X = \sigma_Y$

Two-sample t statistic

Let $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ and let $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$, and let all X_i's and Y_j's be independent. Let S_X^2 and S_Y^2 be the corresponding *sample* variances, and let s_p^2 be the *pooled* variance, where

$$S_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} = \frac{\sum(X_i - \bar{X})^2) + \sum(Y_i - \bar{Y})^2}{n+m-2}$$

Then,

Proof:

Proof (cont):

Form for a $(1-\alpha)\%$ confidence interval:

Rejection regions for α-level tests:

1.2 Assuming $\sigma_X \neq \sigma_Y$

Welch's 2-sample t statistic

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

has an approximate T_ν distribution, where

$$\nu = \frac{(\frac{S_X^2}{S_Y^2} + \frac{n}{m})^2}{\frac{1}{n-1}(\frac{S_X^2}{S_Y^2})^2 + \frac{1}{m-1}(\frac{n}{m})^2}$$, rounded to the nearest integer

$H_0 : p_x = p_y$

Proof(ish):

Form for a $(1 - \alpha)\%$ confidence interval:

Rejection regions for α-level tests:

2 Inference for a difference in proportions

Suppose that m Bernoulli trials have resulted in X successes, and suppose n Bernoulli trials have resulted in Y successes; where all trials are independent. A common test is:

$H_0 : p_x = p_y$

$H_1 : p_x \neq p_y$

2.1 Deriving the GLRT

2.1.1 Approximation Using the CLT

Form for a $(1-\alpha)\%$ confidence interval:

Rejection regions for α-level tests: