	Mon	Wed	Fri
This week	Notes 12, Lab 06	Notes 12, Notes 13 HW 9 due Quiz 2 corrections due	Notes 13 + review
next week	Quiz 03! HW 10 due (completion)	asynch. activity. zoom OH during class time Lab 06 due	BREAK
after Thanksgiving	Start regression	Project proposal due (NO HW)	

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.

Idea: $\bar{X} \sim N(\mu_x, \sigma_x^2/n)$, $\bar{Y} \sim N(\mu_y, \sigma_y^2/m)$ by CLT. Interested in $\mu_x - \mu_y$ → Depending on what we're willing to assume about our σ_x^2, σ_y^2, we obtain different test statistics

1.1 Assuming $\sigma_X = \sigma_Y$

Two-sample t statistic

Let $X_1, ..., X_n \sim N(\mu_X, \sigma^2)$ and let $Y_1, ..., Y_m \sim N(\mu_Y, \sigma^2)$, 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

weighted average of S_x^2 + S_y^2

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

Then,

$$T_{n+m-2} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{1/n + 1/m}}$$ has a T_{n+m-2} distribution

Proof: Idea: $T_v = \frac{Z}{\sqrt{V/v}}$ where $Z \sim N(0,1)$ $V \sim \chi^2_v$

① Divide top & bottom by $\sigma\sqrt{1/n + 1/m}$

$$T = \frac{\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{1/n + 1/m}}}{\sqrt{S_p^2/\sigma^2}}$$

② $\bar{X} \sim N(\mu_x, \frac{\sigma^2}{n})$ $\bar{Y} \sim N(\mu_y, \frac{\sigma^2}{m})$ → $\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ → top is a $N(0,1)$

$$T = \frac{\dfrac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sigma\sqrt{1/n+1/m}}}{\sqrt{S_p^2/\sigma^2}} \quad \begin{matrix} \} N(0,1) \\ \} \sqrt{V/v} \end{matrix}$$

Proof (cont):

③ $\sum_1^n \left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} = \frac{(n-1)S_x^2}{\sigma^2}$ $\qquad \sum_1^m \left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2_{m-1} = \frac{(m-1)S_y^2}{\sigma^2}$

- X_i's & Y_i's are independent, so $\frac{(n-1)S_x^2}{\sigma^2}$ & $\frac{(m-1)S_y^2}{\sigma^2}$ are also indep.
- sum of χ^2 RV's, we obtain a χ^2

$$\frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

④ can write S_p^2/σ^2 as $\left(\frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2}\right) \cdot \frac{1}{n+m-2} = \frac{\chi^2_{n+m-2}}{n+m-2}$

$\rightarrow T_{n+m-2} = \frac{Z}{\sqrt{V/n+m-2}}$ so T_{n+m-2} has a T_{n+m-2} df.

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

for $\mu_x-\mu_y$: $(\bar{X}-\bar{Y}) \pm t_{\alpha/2,n+m-2} \cdot S_p\sqrt{1/n+1/m}$

Rejection regions for α-level tests: $H_0: \mu_x=\mu_y$ let $t = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/n+1/m}}$

$H_1: \mu_x > \mu_y$

reject if $t \geq t_{\alpha,n+m-2}$

$H_1: \mu_x < \mu_y$

$t \leq t_{\alpha,n+m-2}$

$H_1: \mu_x \neq \mu_y$

$t \leq -t_{\alpha/2,n+m-2}$

$t \geq t_{\alpha/2,n+m-2}$

1.2 Assuming $\sigma_X \neq \sigma_Y$ → 'Best guess': S_x S_y

$\frac{\text{sample stat} - \text{pop param}}{SE(\text{sample stat})}$

Welch's 2-sample t statistic

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

Ideally, $= \frac{Z}{\sqrt{V/v}}$

but in general don't know df of $\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}$:(

has an approximate T_ν distribution, where

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

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: