	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 $\sum_1^n \left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1} = \frac{(n-1)S_X^2}{\sigma^2}$ $\quad \sum_1^m \left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2_{m-1} = \frac{(m-1)S_Y^2}{\sigma^2}$

- 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 $\sigma_X \neq \sigma_Y$

$\rightarrow$ '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

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

① Divide top & bottom of W by $\sqrt{\sigma_x^2/n+\sigma_y^2/m}$: $W = \frac{\left(\dfrac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}\right)}{\dfrac{\sqrt{S_x^2/n+S_y^2/m}}{\sqrt{\sigma_x^2/n+\sigma_y^2/m}}}$

} $N(0,1)$ by same reasoning

} Ideally, $\frac{S_x^2/n+S_y^2/m}{\sigma_x^2/n+\sigma_y^2/m} = \frac{V}{v}$

not generally true

$$\frac{S_x^2/n + S_y^2/m}{\sigma_x^2/n + \sigma_y^2/m} = \frac{V}{v} \quad \text{where } V \sim \chi^2_v$$

LHS $\rightarrow \sum\left(\frac{x_i - \bar{x}}{\sigma_x}\right)^2 + \sum\left(\frac{y_i - \bar{y}}{\sigma_y}\right)^2$

sums of squared $N(0,1) \rightarrow \chi^2$

Proof(ish):

Rearrange: $\frac{S_x^2}{n} + \frac{S_y^2}{m} = \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right) \cdot \frac{V}{v}$

By equating means and variances of both sides,

$$v = \frac{(\sigma_x^2/n + \sigma_y^2/m)^2}{\sigma_x^4/n^2(n-1) + \sigma_y^4/m^2(m-1)}$$

$\theta = \frac{\sigma_x^2}{\sigma_y^2}$ and divide by σ_y^4

$$= \frac{\left(\frac{1}{n}\cdot\frac{\sigma_x^2}{\sigma_y^2} + \frac{1}{m}\right)^2}{\frac{1}{n^2(n-1)}\left(\frac{\sigma_x^2}{\sigma_y^2}\right)^2 + \frac{1}{m^2(m-1)}} = \frac{\left(\frac{1}{n}\theta + \frac{1}{m}\right)^2}{\frac{1}{n^2(n-1)}\theta^2 + \frac{1}{m^2(m-1)}}$$

multiply by n^2

$$= \frac{\left(\theta + \frac{n}{m}\right)^2}{\frac{1}{n-1}\theta^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2}$$

$\leftarrow$ but don't know θ, plug in $\hat{\theta} = \frac{S_x^2}{S_y^2}$

Note: If we're willing to assume we know σ_x^2 + σ_y^2, can use normal distribution.

$$= \frac{\left(\frac{S_x^2}{S_y^2} + \frac{n}{m}\right)^2}{\frac{1}{n-1}\left(\frac{S_x^2}{S_y^2}\right)^2 + \frac{1}{m-1}\left(\frac{n}{m}\right)^2}$$

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

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2,v} \cdot \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$$

$\leftarrow$ based on W test statistic we just derived

Rejection regions for α-level tests:

2 Inference for a difference in proportions

$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

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

$H_0: p_x = p_y$

$H_1: p_x \neq p_y$

$X \sim Binom(n, p_x)$

$Y \sim Binom(m, p_y)$

$H_0: p_x = p_y$

$H_1: p_x \neq p_y$

2.1 Deriving the GLRT

$$\lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega_1} L(\theta)}$$

$\Omega_0 = \{(p_x, p_y) : 0 \leq p_x = p_y \leq 1\}$

$\Omega_1 = \{(p_x, p_y) : 0 \leq p_x \leq 1, 0 \leq p_y \leq 1, p_x \neq p_y\}$

Since X & Y independent:

$L(p_x, p_y) : \prod_{i,j} f_x(x_i; p_x) f_y(y_j; p_y) = p_x^{x}(1-p_x)^{n-x} p_y^{y}(1-p_y)^{m-y}$

Under H_0: $p_x = p_y = p_0$

$L = p_0^{x+y}(1-p_0)^{n+m-x-y}$

$\ell = (x+y)\ln(p_0) + (n+m-x-y)\ln(1-p_0)$

$\frac{\partial \ell}{p_0} = \frac{(x+y)}{p_0} + \frac{(n+m-x-y)}{1-p_0} \cdot -1 = 0$

$\frac{(x+y)}{p_0} = \frac{(n+m-x-y)}{1-p_0}$

$\frac{1-p_0}{p_0} = \frac{n+m-(x+y)}{(x+y)}$

$1/p_0 - p_0/p_0 = \frac{(n+m)}{(x+y)} - \frac{(x+y)}{(x+y)}$

$1/p_0 = \frac{n+m}{x+y}$

$\hat{p}_0 = \frac{x+y}{n+m}$ ← pooled success proportion

Under H_1: need to maximize p_x & p_y separately

→ simplifies to the 1 sample MLE's

$\hat{p}_x = \frac{x}{n}$ $\hat{p}_y = \frac{y}{m}$

Back to GLRT:

$$\lambda = \frac{\left(\frac{x+y}{n+m}\right)^{x+y}\left(1-\frac{x+y}{n+m}\right)^{n+m-x-y}}{\left(\frac{x}{n}\right)^{x}\left[1-\frac{x}{n}\right]^{n-x}\left(\frac{y}{m}\right)^{y}\left[1-\frac{y}{m}\right]^{m-y}}$$

Intuitive, but ugly. RV's X & Y in numerator, denominator, exponents, etc.

hard to derive λ directly → approximate

2.1.1 Approximation Using the CLT

Under H_0, $\frac{X}{n} - \frac{Y}{m} \,\dot\sim\, N\left(0, \frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right)$

plug in MLE (under H_0) for p to obtain

$$Z = \frac{X/n - Y/m}{\sqrt{\frac{p_p(1-p_p)}{n} + \frac{p_p(1-p_p)}{m}}} \,\dot\sim\, N(0,1) \qquad p_p = \frac{x+y}{n+m}$$

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

$$\left(\frac{X}{n} - \frac{Y}{m}\right) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_y(1-\hat{p}_y)}{m}}$$

→ Probably better, does not assume $p_x = p_y$

Rejection regions for α-level tests:

$$\cdot \sqrt{\frac{\hat{p}_p(1-\hat{p}_p)}{n} + \frac{\hat{p}_p(1-\hat{p}_p)}{m}}$$

$$\cdot \sqrt{\frac{1}{4n} + \frac{1}{4m}}$$

→ most conservative