10.2: The Two-Sample t Test and Confidence Interval

Taylor

University of Virginia

Motivation

Most of the time population variances are unknown. If we use sample variances, we can use CLT justification to show that z-score like test statistics are approximately normally distributed for large data. However, this section comes in handy when at least one of the sample sizes is smaller, but we can still assume all the data is randomly sampled from normal distributions.

Assumptions for Approximate Inference

1. $X_1, \ldots, X_m \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$
2. $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$
3. The Xs and Ys are independent

We can check (1) and (2) by looking at the two histograms or by looking at two normal probability plots (we didn't learn about the latter, though).

Theorem

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

APPROXIMATELY follows a t-distribution with degrees of freedom

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

1. Proof is a quiz question. Follow page 500 for help.
2. if ν is a decimal, round down.

Our two sample confidence approximate confidence interval is then:

$$\bar{x} - \bar{y} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Hypothesis Test

And our two sample approximate hypothesis test is:

1. $H_0 : \mu_1 - \mu_2 = \Delta_0$
2. $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$
3. If $H_a : \mu_1 - \mu_2 > \Delta_0$, then reject when $t > t_{\alpha,\nu}$
4. If $H_a : \mu_1 - \mu_2 < \Delta_0$, then reject when $t < -t_{\alpha,\nu}$
5. If $H_a : \mu_1 - \mu_2 \neq \Delta_0$, then reject when $t > t_{\alpha/2,\nu}$ or $t < -t_{\alpha/2,\nu}$

Assumptions for Exact Inference

There is another method for two sample t-distribution inference, though. The book's section on this is called "Pooled t Procedures." We stick with our original assumptions:

1. $X_1, \ldots, X_m \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$
2. $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$
3. The Xs and Ys are independent

but then we ALSO assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Derivation

We know how $\frac{(m-1)S_1^2}{\sigma^2}$ and $\frac{(n-1)S_2^2}{\sigma^2}$ are two independent χ^2 rvs. So adding them together gives us another χ^2 rv

$$\frac{(m-1)S_1^2}{\sigma^2} + \frac{(n-1)S_2^2}{\sigma^2} = \frac{(m+n-2)S_p^2}{\sigma^2}$$

where $S_p^2 = \frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2}$

Derivation

So in this case we get an exact t distribution

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma^2(1/m+1/n)}} \div \sqrt{\frac{(m+n-2)S_p^2}{\sigma^2}\frac{1}{m+n-2}} = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

So this expression follows a t distribution with $n+m-2$ degrees of freedom.