Inference for difference in sample means

ST551 Lecture 19

Charlotte Wickham

2017-11-03

From last time

Setting: two **independent** samples

$Y_1, \ldots, Y_n$ i.i.d from population with c.d.f F_Y, and
$X_1, \ldots, X_m$ i.i.d from population with c.d.f F_X

Parameter: Difference in population means $\mu_Y - \mu_X$

Properties of sampling distribution for $\overline{Y} - \overline{X}$, lead to Z-test and associated intervals:

$$Z(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{\sigma_Y^2/n + \sigma_X^2/m}}$$

With known population variances σ_Y^2, σ_X^2.

When variances aren't known

Like in one-sample Z-test, we proceed by substituting in good estimates for the variances, then alter reference distibutions accordingly.

Two scenarios:

- Populations variances are unknown but assumed equal, $\sigma^2 = \sigma_Y^2 = \sigma_X^2$. Both samples give information about σ^2.
- Populations variances are unknown and not assumed equal.

Equal variances

Need to use both samples to estimate $\sigma^2 = \sigma_Y^2 = \sigma_X^2$

$$s_p^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 + \sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2}{(n-1)+(m-1)}$$
$$= \frac{(n-1)s_Y^2 + (m-1)s_X^2}{n+m-2}$$

where s_Y^2 and s_X^2 are the samples variances for the Y_i and X_i respectively.

Intuition: weighted average of sample variances, so that larger sample should contribute more in the average.

Plugging in to Z-stat

Hypothesis: $H_0 : \mu_Y - \mu_X = \delta_0$

Assumption: $\sigma_Y^2 = \sigma_X^2$

Leads to test statistic:

$$t(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{s_p^2/n + s_p^2/m}} = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{(\overline{Y} - \overline{X}) - \delta_0}{s_p\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Leads to equal variance t-test

Compare $t(\delta_0)$$ to a t-distribution with $n + m - 2$ degrees of freedom.

Also leads to CI of form:

$$(\overline{Y} - \overline{X}) \pm t_{(n+m-2),1-\alpha/2}\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$$

This distribution is **exact** if the populations are Normal.

Assymptotically exact otherwise.

For large sample sizes, it doesn't make much difference $t_{m+n-2} \to z$ as $n + m - 2 \to \infty$

Equal variance assumption: What can go wrong?

Compare $E(s_p^2/n + s_p^2/m)$ to $Var(\overline{Y} - \overline{X})$

Equal variance assumption: What can go wrong?

$$\text{Actual} = Var(\overline{Y} - \overline{X}) = \frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}$$

$$\text{Estimated} = E(\widehat{Var}(\overline{Y} - \overline{X})) \approx \frac{\sigma_Y^2}{m} + \frac{\sigma_X^2}{n}$$

m	σ_X^2	n	σ_Y^2	Actual	Estimated
10	1	50	4	0.18	0.42
10	9	50	1	0.92	0.28

Equal variance assumption: Consequences

The expected value of the estimated variance is:

- Larger than it should be when the smaller sample comes from the population with the smaller variance.
  - Test statistic will be closer to zero than it should be, and rejection rates will be smaller.
- Smaller than it should be when the smaller sample comes from the population with the larger variance.
  - Test statistic will have a larger absolute value than it should, and rejection rates will be larger.

If we don't assume equal variance?

What's the best estimate of $\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}$?

$$\frac{s_Y^2}{n} + \frac{s_X^2}{m}$$

Plugging into Z-stat:

$$t(\delta_0) = \frac{(\overline{Y} - \overline{X}) - \delta_0}{\sqrt{s_Y^2/n + s_X^2/m}}$$

Reference distribution? Even when populations are Normal, this test statistic doesn't have exactly a t-distribution.

Welch-Satterthwaite

Slightly better than just using a Normal approximation.

Compare to t with v degrees of freedom, where

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

Somewhere between $\min(m-1, n-1)$ and $m+n-2$