# Inference for difference in sample means

ST551 Lecture 19

Charlotte Wickham

2017-11-01

# From last time

**Setting**: two **independent** samples

$Y_1, \ldots, n$ i.i.d from population with c.d.f $F_Y$, and
$X_1, \ldots, m$ 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 $\sigma_Y^2$, $\sigma_X^2$.

# 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 $\sigma^2$.
- Populations variances are unknown and not assumed equal.

# Equal variances

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

pooled variance

$$s_p^2 = \underbrace{\hat{\sigma}^2}_{\text{estimate of } \sigma^2} = \frac{\overbrace{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}^{\text{obs. from first pop.}} + \overbrace{\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2}^{\text{obs from second pop}}}{(n-1)+(m-1)}$$

each obs. contributing equally to the estimate

$$s_p^2 = \boxed{\frac{(n-1)\underbrace{s_Y^2}_{\text{sample variance } Y\text{'s}} + (m-1)\underbrace{s_X^2}_{\text{sample variance for } X\text{'s}}}{n+m-2}}$$

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 $\underbrace{n+m-2}$ degrees of freedom.

Also leads to CI of form: for $\mu_Y - \mu_X$

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

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 $\underbrace{E(s_p^2/n + s_p^2/m)}_{\text{estimate of } Var(\bar{Y}-\bar{X})}$ to $\underbrace{Var(\overline{Y} - \overline{X})}_{\text{true variance}}$

$$\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}$$

# 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 | $\sigma_X^2$ | n | $\sigma_Y^2$ | 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 $\boxed{\frac{\sigma_Y^2}{n} + \frac{\sigma_X^2}{m}}$?

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

Plugging into Z-stat:

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

$$\frac{Z}{\chi/d} \quad Z \sim N(0,1) \quad X \sim \text{Chisq}(d$$

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

# Welch-Satterthwaite

Could compare

$t(\delta_0)$ to $N(0, 1)$

Slightly better than just using a Normal approximation.

Compare to $t$ with $v$ degrees of freedom, where

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

might not be an integer

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

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

compare to $t_v$

call this "Welch's t-test"

· Procedure

1. Look at data to determine procedure

$\left( \frac{s_y}{s_x} > C \right)$ Two sample variance test

Welch's t-test

Two-sample equal variance

2. Do the test, report test results

Consequence : doesn't stated performance