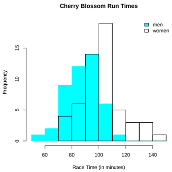
Lecture 31: t Distribution for Difference of Two Means

Chapter 5.4

Question for Today

In Chapter 5.2 we asked: Did men ($n_m = 45$) run faster than women ($n_w = 55$) in the Cherry Blossom Race?



What can we say about $\mu_1 - \mu_2$ when n_1 and n_2 are both small?

Components

Similarly to one-sample t-tests, now we use the two sample t-test:

1. The point estimate of $\mu_1 - \mu_2$ is $\overline{x}_1 - \overline{x}_2$
2. The standard error of the sampling distribution

$$SE_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3. Confidence intervals using t^*_{df}
4. Hypothesis tests using t-statistic

Components

But what degrees of freedom df do we use?

The true formula for degrees of freedom is

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

Rather, for this class, use the smaller of n_1 and n_2 minus 1 i.e.

$$\min(n_1, n_2) - 1$$

Conditions of Two Sample t-Test

- Both samples meet the conditions for using the t distribution
  - Sample observations are nearly normal
  - Sample observations are independent within their respective populations
- The two samples are independent

Pooled Standard Deviation Estimate

Say however, you suspect both populations have similar true population standard deviations $\sigma_1 = \sigma_2 = \sigma$.

If so, we can leverage this fact to make the t distribution approach slightly more precise.

Pooled Standard Deviation Estimate

The pooled standard deviation estimate is

$$s_{pooled}^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

So use s_{pooled}^2 instead of s_1^2 and s_2^2 in SE:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} = \sqrt{s_{pooled}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Pooled Standard Deviation Estimate

You can think of s_{pooled}^2 as being very close to a weighted average of the two sample standard deviations:

$$\begin{aligned} s_{pooled}^2 &= s_1^2 \times \frac{n_1 - 1}{n_1 + n_2 - 2} + s_2^2 \times \frac{n_2 - 1}{n_1 + n_2 - 2} \\ \text{close to} &\approx s_1^2 \times \frac{n_1}{n_1 + n_2} + s_2^2 \times \frac{n_2}{n_1 + n_2} \end{aligned}$$

The -1 and -2 are degrees of freedom corrections.

Pooled Standard Deviation Estimate

Benefits: If σ's are equal, we have more precise model of the sampling distribution of $\overline{x}_1 - \overline{x}_2$

Caveats: Only pool when background research/intuition indicates the population σ_1 and σ_2 of the two groups are nearly equal.