Inference for Two Means

Estimating a difference in means

Download the section 17 .Rmd handout to `STAT240/lecture/sect16-two-means`.

Download the file `TIM.txt` to `STAT240/data`.

Continue working with the Boston Marathon data.

What is the change in the average time for an 18-34 year old female runner from 2010 to 2011?

Assume each group (2010 and 2011) was taken from a different population. Let's look at the data.

Formally,

$$X_{1,i} \sim D_1(\mu_1,\ \sigma_1)$$
$$X_{2,i} \sim D_2(\mu_2,\ \sigma_2)$$

and populations 1 and 2 are independent.

We allow the distributions to differ completely. All that matters is $\bar{X}_1$ and $\bar{X}_2$ are approximately normal.

point estimate ± critical value × standard error

- For $\mu_1 - \mu_2$, the point estimate is $\bar{x}_1 - \bar{x}_2$
- The standard error combines σ_1 and σ_2

The standard error is

$$\sqrt{se_1^2 + se_2^2} \;=\; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This gives sampling distribution

$$\bar{X}_1 - \bar{X}_2 \quad \dot{\sim} \quad N\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

This suggests a normal critical value. However, we need to estimate the se.

$$\hat{se}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Again, we use a T distribution for our CI and hypothesis test.

The T is approximate, and the df is messy.

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

This is the **Welch** procedure. The df are "penalized" according to s_1 and s_2.

The formula for the CI is

$$\bar{X}_1 - \bar{X}_2 \ \pm \ t_{\alpha/2,w}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We are 95% confident that the true difference is within (-1.416, 1.961). `t.test` confirms this.

Build a 90% CI for the difference in mean time between female and male 50-54 year old runners in 2010.

- Find a se term using both groups
- Calculate approximate df for critical value
- Build and interpret your interval
- Check with `t.test`.

Are athletes faster in 2010?

Is the change in the average time for an 18-34 year old female runner from 2010 to 2011 less than 0?

$$H_0 : \mu_1 - \mu_2 \geq 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 < 0$$

This test will be performed similarly to the one-sample T test. In general, a T statistic is

$$\frac{\text{point estimate} - \text{value under } H_0}{\text{estimated standard error}}$$

$$\frac{\bar{X}_1 - \bar{X}_2 - 0}{\hat{se}(\bar{X}_1 - \bar{X}_2)}$$

Our model is

$$X_{1,i} \sim D_1(\mu_1,\ \sigma_1)$$
$$X_{2,i} \sim D_2(\mu_2,\ \sigma_2)$$

and $H_0 : \mu_1 - \mu_2 \geq 0$.

Our observed test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If H_0 is true, T has a T distribution with w degrees of freedom. We observe a test stat of $t_{obs} = 0.316$.

For hypotheses

$$H_0 : \mu_1 - \mu_2 \geq 0 \quad \text{versus} \quad H_A : \mu_1 - \mu_2 < 0$$

we have p-value

$$P(T_w < 0.316) = 0.624$$

and fail to reject H_0.

P-value calculation:

- If the alternative is $H_A : \mu < \mu_0$, the p-value is the area *below* the test statistic.
- If the alternative is $H_A : \mu > \mu_0$, the p-value is the area *above* the test statistic.
- If the alternative is $H_A : \mu \neq \mu_0$, the p-value is $2\times$ the area *outside* of the test statistic.

Sometimes, two samples of data are **paired**.

Do children grow taller from age 13 to 14? It doesn't make sense to take independent samples of 13 and 14 year olds.

Instead, measure the same individuals at 13 and 14.

Five individuals had their heights measured at age 13 and 14.

Height at 13	44.1	59.0	65.9	58.7	49.3
Height at 14	46.3	60.5	68.2	59.4	50.6

The two samples are *not independent*, so we can't use one of the previous methods.

Instead, study the *differences*.

Height at 13	44.1	59.0	65.9	58.7	49.3
Height at 14	46.3	60.5	68.2	59.4	50.6
Difference	2.2	1.5	2.3	0.7	1.3

We are back in a one-sample setting! Parameter:

$$\mu_{diff} \quad \text{instead of} \quad \mu_1 - \mu_2$$

Assume the differences are independent draws from a population:

$$\mathit{diff}_i \;\sim\; D(\mu, \sigma)$$

We then have hypotheses

$$H_0 : \mu_{\mathit{diff}} \leq 0 \quad \text{versus} \quad H_A : \mu_{\mathit{diff}} > 0$$

We use the usual one-sample T test statistic

$$T = \frac{\bar{diff} - \mu}{S/\sqrt{n}}$$

which follows a T with $n - 1$ df if H_0 is true.

Let's complete the test with $\alpha = 0.05$.

We have a large positive test statistic and a small p-value, so we reject H_0.

This is confirmed in `t.test`.

If we had done a two independent sample test, we get the opposite result!