Lauren, Natalie F , Jeffry, Valeria, Naomi

Stat 145, Fri 29-Oct-2021 -- Fri 29-Oct-2021

Biostatistics

Spring 2021

Friday, October 29th 2021

Wk 9, Fr

Topic:: Inference on two means

Read:: Lock5 6.10-6.13

Ch.6 theme

Redoing CI and Hyp. Tests

now equipped w/ formulas for SE

t-distributions appeared

- as a substitute for normal dists
- only in context of inference on a mean

main focus: μ — estimate by $\bar{x}$

sidelight: needed σ $\left(SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \right)$

estimate by s

Two-Sample t Inference

Thomas Scofield

October 29, 2021

Recall this fact:

Theorem: Suppose X and Y are independent variables, and both are normally distributed, with $X \sim \text{Norm}(\mu_X, \sigma_X)$ and $Y \sim \text{Norm}(\mu_Y, \sigma_Y)$. Then their difference $X - Y$ also has a normal distribution, with $(X - Y) \sim \text{Norm}(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$.

Two-sample t context

Imagine you have two groups/populations in mind, and you take *independent* samples, one of size n_1 from Group 1, and one of size n_2 from Group 2. The variable you measure is quantitative, so you can talk about

variable that has mean, sd

- μ_1, μ_2, means for the two populations
- σ_1, σ_2, standard deviations for the two populations
- $\overline{x}_1$, $\overline{x}_2$, means for the two samples
- s_1, s_2, standard deviations for the two samples

Note that

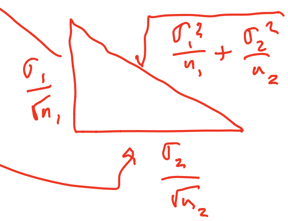
- $\overline{x}_1$, $\overline{x}_2$ should be independent, since the samples are.
- If either $n_1 \geq 30$, or if Population 1's values are reasonably symmetric, bell-shaped, then

$$\overline{x}_1 \sim \text{Norm}\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right).$$

CLT says this

- Likewise, if either $n_2 \geq 30$, or if Population 2's values are reasonably symmetric, bell-shaped, then

$$\overline{x}_2 \sim \text{Norm}\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right).$$



Under these conditions, the theorem tells us

$$\overline{x}_1 - \overline{x}_2 \sim \text{Norm}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1}{n_1^2} + \frac{\sigma_2}{n_2^2}}\right).$$

This is a statement about the sampling distribution for $\overline{x}_1 - \overline{x}_2$—that (under conditions) it is approximately normal. Thus, the *spread* of that sampling distribution is rightly called the **standard error** of $\overline{x}_1 - \overline{x}_2$:

$$\boxed{\text{SE}_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^{2}}{n_1^{\cancel{2}}} + \frac{\sigma_2^{2}}{n_2^{\cancel{2}}}}.}$$

Confidence Intervals for $\mu_1 - \mu_2$

Again, the overarching process is the centered interval approach:

$$(\text{point estimate}) \pm (\text{critical value})(\mathrm{SE}_{\overline{x}_1 - \overline{x}_2}).$$

As we will almost never know σ_1, σ_2, we will estimate this standard error using the approximation

$$\mathrm{SE}_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

As before, the need to estimate another (and non-central) parameter forces us to employ t-distributions. Thus, the line above looks like

$$(\overline{x}_1 - \overline{x}_2) \pm (t^*)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Choosing degrees of freedom

If we specify 95% as the confidence level, then we must choose the best t-distribution so as to have a corresponding success rate of 95%. It is not known how to do so so that we *always* obtain the desired success rate. There are several strategies:

1. **Satterthwaite formula**:

   aka Satterthwaite - Welch

   aka Welch approximation

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

probably the best option

less appealing than next option

2. **Conservative estimate**:

   Taught in Lock 5

$$df = \underbrace{\min(n_1, n_2)}_{\text{smaller of the two sample sizes}} - 1.$$

then subtract 1

Example data:

1. Case: summary data is all we know

   Means are for number of beetle larvae per stem in oat crop

Group	n	x-bar	s
Control	13	3.47	1.21
Malathion	14	1.36	0.52

unstandardized test statistic $\bar{x}_1 - \bar{x}_2 = 3.47 - 1.36 = 2.11$

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1.21^2}{13} + \frac{0.52^2}{14}} \doteq 0.36323$$

   Construct a 95% CI for difference mu_C - mu_M

$$t^* = \text{qt}(0.975,\ \underline{df = 12})$$

   Test hypothesis that mu_C-mu_M = 0 vs. one-sided alternative

2. CaffeineTaps data

1.(continued) A 95% CI

$$(\text{point est./unstandardized test stat}) \pm (t^*)(SE_{\bar{x}_1 - \bar{x}_2})$$

$$2.11 \pm \underline{2.1788}\,(0.36323)$$

results in CI $(1.319,\ 2.901)$.

2. favstats(Taps ~ Group, data = CaffeineTaps)

gives $\bar{x}_c = 248.3$, $s_c = 2.2136$, $n_c = 10$

$\bar{x}_p = 244.8$, $s_p = 2.3944$, $n_p = 10$

point est: $\bar{x}_c - \bar{x}_p = 248.3 - 244.8 = 3.5$

est. SE $= \sqrt{\frac{2.2136^2}{10} + \frac{2.3944^2}{10}} = 1.0312$

for 96% conf. $t^* = \text{qt}(.98,\ df = 9) = 2.3984$

96% CI

$$3.5 \pm (2.3984)(1.0312) \quad \text{or} \quad (1.0268,\ 5.973)$$

All-in-one-step calculations provided by t.test( ) <u>but</u> you have to have access to the raw data. So, it's usable for problem 2, but not for Problem 1.

```
t.test( Taps ~ Group, data = CaffeineTaps, conf.level = 0.96 )
```