

Lecture 21

Inference on Two Population Means

STAT 8010 Statistical Methods I

October 9, 2019

Whitney Huang
Clemson University

- We often interested in comparing two groups, for example:
 - Does a particular pesticide increase the yield of corn per acre?
 - Do men and women in the same occupation have different salaries?
- The common ingredient in these questions: They can be answered by conducting statistical inferences of two populations using two (independent) samples

- Parameters:

- Population means: μ_1, μ_2
- Population standard deviations: σ_1, σ_2

- Statistics:

- Sample means: \bar{X}_1, \bar{X}_2
- Sample standard deviations: s_1, s_2
- Sample sizes: n_1, n_2

- Point estimate: $\bar{X}_1 - \bar{X}_2$
- Interval estimate: Need to figure the sampling distribution of $\bar{X}_1 - \bar{X}_2 \Rightarrow \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} \approx N(0, 1)$ by CLT. However, since we estimate $\sigma_{\bar{X}_1 - \bar{X}_2}$ using $s_{\bar{X}_1 - \bar{X}_2}$. We have $\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} \approx t(df)$
- Hypothesis Testing:
 - Upper-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 > 0$
 - Lower-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 < 0$
 - Two-tailed test: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 \neq 0$

If we are willing to **assume** $\sigma_1 = \sigma_2$, then we can “pool” these two (independent) samples together to estimate the common σ using s_p :

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

We can then derive the (estimated) standard error of $\bar{X}_1 - \bar{X}_2$:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Therefore, we obtain the $(1 - \alpha) \times 100\%$ CI for $\mu_1 - \mu_2$:

$$\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{point estimate}} \pm \underbrace{t(\alpha/2, n_1 + n_2 - 2)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{\text{margin of error}}$$

Confidence Intervals for $\mu_1 - \mu_2$: What if $\sigma_1 \neq \sigma_2$?

- We will use s_1^2, s_2^2 as the estimates for σ_1^2 and σ_2^2 to obtain the standard error:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- The formula for the degrees of freedom is somewhat complicated:

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

- We can then construct the $(1 - \alpha) \times 100\%$ CI for $\mu_1 - \mu_2$:

$$\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{point estimate}} \pm \underbrace{t(\alpha/2, \text{df calculated from above}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{margin of error}}$$

To Pool ($\sigma_1 = \sigma_2$) or Not to Pool ($\sigma_1 \neq \sigma_2$)?

We could perform the following test:

- $H_0 : \sigma_1^2/\sigma_2^2 = 1$ vs. $\sigma_1^2/\sigma_2^2 \neq 1$
- Test statistic: $F^* = s_1^2/s_2^2$. Under H_0 , $F^* \sim F(n_1 - 1, n_2 - 1)$
- For a given α , we reject H_0 if the P-value $< \alpha$ (or $F_{obs} > F_\alpha(n_1, n_2)$)
- If we fail to reject H_0 , then we will use s_p as an estimate for σ and we have $s_{\bar{X}_1 - \bar{X}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Otherwise, we use

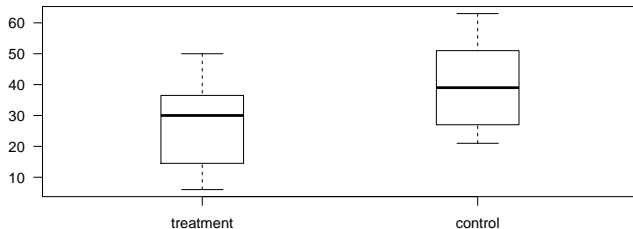
$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age and health was randomly divided into two groups. Twelve of the lambs were injected with the drug (treatment group) and the remaining twelve were left untreated (control group). After a 6-month period, the worm counts were recorded:

Treatment	18	43	28	50	16	32	13	35	38	33	6	7
Control	40	54	26	63	21	37	39	23	48	58	28	39

Plot the Two Samples



- $n_1 = n_2 = 12 \Rightarrow$ sample size is perhaps not large enough for CLT to work. But the boxplots suggest the distributions are symmetric with no outliers
- The untreated lambs (control group) appear to have higher average worm counts than the treated lambs (treatment group). But do we have enough evidence ?

Example Cont'd

```
> apply(dat, 2, mean)
treatment  control
 26.58333   39.66667
> apply(dat, 2, sd)
treatment  control
 14.36193   13.85859
> var.test(treatment, control)
```

F test to compare two variances

```
data: treatment and control
F = 1.074, num df = 11, denom df = 11, p-value = 0.9079
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3091686 3.7306092
sample estimates:
ratio of variances
      1.073959
```

We fail to reject $\sigma_1 = \sigma_2 = \sigma$. Therefore we will use s_p , the pooled standard deviation, as an estimate for σ

- Place a 95% confidence interval on $\mu_1 - \mu_2$ to assess the size of the difference in the two population means
- Test whether the mean number of tapeworms in the stomachs of the treated lambs is less than the mean for untreated lambs. Use an $\alpha = 0.05$ test

Another Example

A Simple random sample with sample size 37 is taken and are subjected to a treatment ($\bar{X}_1 = 19.45, s_1 = 4.3$). A Simple random sample with sample size 31 is taken and given a placebo ($\bar{X}_2 = 18.2, s_2 = 2.2$). At the 10% level can we say that the means are different between the two groups?

In this lecture, we learned

- Point/Interval estimate for $\mu_1 - \mu_2$
- Test if $\sigma_1^2 = \sigma_2^2$
- Hypothesis Testing for $\mu_1 - \mu_2$

In next lecture we will learn

- Tests with matched samples
- Analysis of Variance (ANOVA)