STAT22000 Autumn 2013 Lecture 23

Yibi Huang

November 20, 2013

7.2 Comparing Two Means

Outline

- Two-sample z statistic
- Two-samples t procedures
- Two-sample t tests and confidence intervals ($\sigma_1 \neq \sigma_2$)
- Pooled two-sample t tests and confidence intervals ($\sigma_1 = \sigma_2$)
- Robustness

Two Sample Problems (1)

- E.g., is the air more polluted in Chicago than in LA?
- E.g., are the midterm scores of the students who prefer to sit in the front of the class higher than the scores of those who prefer to sit in the back?
- E.g., are smokers suffering less from depression than non-smokers?
- E.g., are the response in the treatment group different from that in the control group?

Two Sample Problems (2)

- The goal is to compare the means (of some quantity) μ_1 and μ_2 of the two populations.
  Suppose the SDs of the two populations are respectively σ_1 and σ_2.
  To compare μ_1 and μ_2, we will take a simple random sample from each of the two populations.

  SRS of size n_1 from population 1 : $X_{1,1}, X_{1,2}, \ldots X_{1,n_1}$
  SRS of size n_2 from population 2 : $X_{2,1}, X_{2,2}, \ldots\ldots, X_{2,n_2}$
- The responses in each group are **independent** of those in the other group
- Unlike the matched pairs design, there is **no matching** of the observations in the two samples and the two samples may be of different sizes

Two Sample Problems (3)

- The first sample has the sample mean
$$\overline{X}_1 = \frac{1}{n_1}(X_{1,1} + X_{1,2} + \cdots + X_{1,n_1}) \quad \text{which estimates } \mu_1.$$
- The second sample has the sample mean
$$\overline{X}_2 = \frac{1}{n_2}(X_{2,1} + X_{2,2} + \cdots + X_{2,n_2}) \quad \text{which estimates } \mu_2.$$
- Therefore $\overline{X}_1 - \overline{X}_2$ estimates $\mu_1 - \mu_2$.

How close is $\overline{X}_1 - \overline{X}_2$ to $\mu_1 - \mu_2$?
What is the **sampling distribution** of $\overline{X}_1 - \overline{X}_2$?

Two Sample Problems (3)

Recall that
$$E(\overline{X}_1) = \mu_1, \quad \mathrm{Var}(\overline{X}_1) = \sigma_1^2/n_1$$
$$E(\overline{X}_2) = \mu_2, \quad \mathrm{Var}(\overline{X}_2) = \sigma_2^2/n_2.$$
Observe $\overline{X}_1 - \overline{X}_2$ is an **unbiased estimate** of $\mu_1 - \mu_2$ because
$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2.$$
Furthermore, since the two samples are independent, $\overline{X}_1$ and $\overline{X}_2$ are independent, we have
$$\mathrm{Var}(\overline{X}_1 - \overline{X}_2) = \mathrm{Var}(\overline{X}_1) + \mathrm{Var}(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
Thus the **standard deviation** of $\overline{X}_1 - \overline{X}_2$ is
$$\mathrm{SD}(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Two-Sample z-Statistic When σ_1, σ_2 Are Known

If the standard deviations σ_1, σ_2 are *known*, the two-sample z-statistic for the difference $\mu_1 - \mu_2$ is

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Then, Z is approximately $N(0, 1)$ if

- both populations are normal, or
- both sample sizes n_1 and n_2 are large (CLT)

Two Sample Problem with Known σs: CIs and Tests

The $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$(\overline{X}_1 - \overline{X}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $z^* = z_{\alpha/2}$ is the $\alpha/2$ critical value of the standard normal distribution.

To test the hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2$ or equivalently $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$, we use

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \text{ under } H_0$$

The p-value is calculated as before.

Two-Sample t-Statistic When σ_1, σ_2 Are Unknown

Of course, σ_1^2 and σ_2^2 are often unknown. Thus we substitute them by the sample variances s_1^2 and s_2^2.

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \quad \text{where} \quad \begin{aligned} s_1^2 &= \frac{\sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_1)^2}{n_1 - 1} \\ s_2^2 &= \frac{\sum_{i=1}^{n_2}(X_{2,i} - \overline{X}_2)^2}{n_2 - 1} \end{aligned}$$

- Unfortunately, the two-sample t-statistic does NOT have a t-distribution
- Fortunately, it can be approximated by a t-distribution with a certain degrees of freedom.

See the next slide for the approximation

Approximate Distribution of the Two-Sample t-Statistic

The two-sample t-statistic has an **approximate t_k distribution**.
For the degrees of freedom k we have two formulas:

1. software formula:

$$k = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}, \qquad \begin{aligned} w_1 &= s_1^2/n_1, \\ w_2 &= s_2^2/n_2. \end{aligned}$$

2. simple formula: $\boxed{k = \min(n_1 - 1, n_2 - 1)}$

Comparison of the two formulas:

- The software approximation is more accurate. It gives larger degrees of freedom and yields shorter CIs and smaller P-value
- The simple formula is more conservative. I.e., it yields wider CIs and larger P-values than the actual P-value
- For this course, it is fine to **just using the simple formula**.

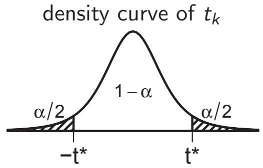
Confidence Intervals for $\mu_1 - \mu_2$

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is given by

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

where $t^* = t_{k,\alpha/2}$ is the $\alpha/2$ critical value of the t distribution with k degrees of freedom.

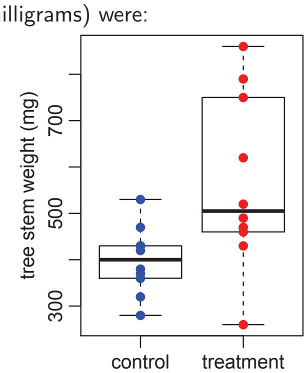
density curve of t_k



Example: Nitrogen Effect on Tree Growth

20 northern red oak seedlings
Half received no nitrogen
All grown in same type of soil in same greenhouse
After 140 days, stem weights (in milligrams) were:

Control no nitrogen		Treatment nitrogen	
320	430	260	750
530	360	430	790
280	420	470	860
370	380	490	620
470	430	520	460
mean = 399		mean = 565	
SD = 186.74		SD = 72.79	
$n_C = 10$		$n_T = 10$	



Example: CI for the Nitrogen Effect on Tree Growth

The two samples have 10 observations each. Thus the degrees of freedom is the smaller one of $10-1$ and $10-1$, which is 9.

From Table D, we can find the critical value for 95% confidence level is $t^* = t_{9,0.025} = 2.262$.

So the 95% CI for $\mu_T - \mu_C$ (treatment mean - control mean) is

$$\overline{X}_T - \overline{X}_C \pm t^* \sqrt{\frac{s_T^2}{n_1} + \frac{s_C^2}{n_2}} = 565 - 399 \pm 2.262 \sqrt{\frac{(186.74)^2}{10} + \frac{(72.79)^2}{10}}$$
$$\approx 166 \pm 143.4 = (22.6, 309.4)$$

Since 0 (zero) is NOT inside the CI, it appears that there **is** a difference in the population mean stem weights of the treatment and control groups.

We conclude that Nitrogen has an effect on stem weight.

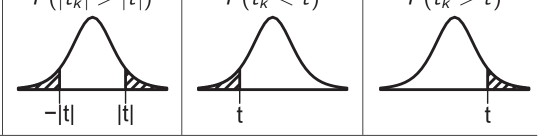
Hypothesis Tests for $\mu_1 - \mu_2$

To test the null hypothesis H_0: $\mu_1 - \mu_2 = \delta_0$, the two-sample t-statistic is

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

which has an approximate t_k-distribution, where the degrees of freedom is $k = \min(n_1 - 1, n_2 - 1)$, and the P-value is computed as follows depending on the alternative hypothesis H_a.

	Two-Sided	Lower One-Sided	Upper One-Sided
H_1	$\mu_1 - \mu_2 \neq \delta_0$	$\mu_1 - \mu_2 < \delta_0$	$\mu_1 - \mu_2 > \delta_0$
P-value	$P(\lvert t_k \rvert > \lvert t \rvert)$	$P(t_k < t)$	$P(t_k > t)$



The bell curves above is the t-curve with k degrees of freedom.

Example: Test for the Nitrogen Effect on Tree Growth

For testing $H_0 : \mu_T - \mu_C = 0$ v.s. $H_a : \mu_T - \mu_C \neq 0$, the t-statistic is

$$t = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{s_T^2/n_T + s_C^2/n_C}} = \frac{565 - 399}{\sqrt{\frac{(186.74)^2}{10} + \frac{(72.79)^2}{10}}} = \frac{166}{63.38} \approx 2.619.$$

The degrees of freedom is $10 - 1 = 9$.
From Table D, we see that $P(t_9 > 2.619)$ is between 0.01 and 0.02. So the two-sided P-value is between 0.02 and 0.04.

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781

The difference is significant at 5% level.
We conclude that Nitrogen has an effect on stem weight.

What if $\sigma_1 = \sigma_2$?

So far we have assumed that $\sigma_1 \neq \sigma_2$. What if we have reason to believe $\sigma_1 = \sigma_2 = \sigma$ albeit σ is unknown?

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, both s_1^2 and s_2^2 are unbiased estimates of σ^2. We can combine s_1^2 and s_2^2 to get a better estimate for σ^2, which is the so-called **pooled samples variances**

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

Observe that s_p^2 is a weighted average of s_1^2 and s_2^2, and it gives more weights to the sample with larger size.

Moreover, as $s^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$, we can see that

$$s_p^2 = \frac{\sum_i (X_{1,i} - \overline{X}_1)^2 + \sum_i (X_{2,i} - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

is simply an "average" of the *combined sum of squares*, though we divide by $n_1 + n_2 - 2$ but not $n_1 + n_2$.

The Pooled Two-Sample t-Statistic (When $\sigma_1 = \sigma_2$)

The two-sample t-statistic then becomes

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is specifically called **the pooled two-sample t-statistic**.

- It has an **exact** t-distribution with $n_1 + n_2 - 2$ **degrees of freedom** when the two populations are normal.
- It is approximately $t_{(n_1+n_2-2)}$ as long as the sample size n_1, n_2 is not too small.
- The degrees of freedom, $n_1 + n_2 - 2$ is greater the degrees of freedom given by the software formula or the simple formula when $\sigma_1 \neq \sigma_2$

Two Sample Problems w/ Equal but Unknown σs

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$(\overline{X}_1 - \overline{X}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t^* = t_{k,\alpha/2}$ is the $\alpha/2$ critical value of the $t_{(n_1+n_2-2)}$ distribution

To test the hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, we use

$$t = \frac{\overline{X}_1 - \overline{X}_2 - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \qquad \text{under } H_0$$

Tree Growth Example Revisit: Assuming $\sigma_1 = \sigma_2$

If assuming $\sigma_1 = \sigma_2$, the pooled SD is

$$s_p = \sqrt{\frac{(10-1)(186.74)^2 + (10-1)(72.79)^2}{10+10-2}} \approx 141.72$$

The degrees of freedom is $n_T + n_C - 2 = 10 + 10 - 2 = 18$.
From Table D, we can find the critical value for 95% confidence level is $t^* = t_{18,0.025} = 2.101$.

So the 95% CI for $\mu_T - \mu_C$ (treatment mean - control mean) is

$$\overline{X}_T - \overline{X}_C \pm t^* s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}} = 565 - 399 \pm 2.101 \times 141.72 \times \sqrt{\frac{1}{10} + \frac{1}{10}}$$
$$\approx 166 \pm 133.2 = (32.8, 299.2)$$

Observe the CI become shorter. As the degrees of freedom k increases, the critical value $t_{k,0.025}$ decreases.

Tree Growth Example Revisit: Assuming $\sigma_1 = \sigma_2$

For testing $H_0 : \mu_T - \mu_C = 0$ v.s. $H_a : \mu_T - \mu_C \neq 0$, assuming $\sigma_1 = \sigma_2$ the pooled t-statistic is

$$t = \frac{\overline{X}_T - \overline{X}_C}{s_p\sqrt{1/n_T + 1/n_C}} = \frac{565 - 399}{141.72\sqrt{1/10 + 1/10}} = \frac{166}{63.38} \approx 2.619.$$

The degrees of freedom is $n_T + n_C - 2 = 10 + 10 - 2 = 18$.

From Table D, we see $P(t_{18} > 2.619)$ is between 0.005 and 0.01. So the two-sided P-value is between 0.01 and 0.02.

	Upper-tail probability p											
df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	.0005
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922

The pooled t-test makes the P-value smaller and the result more significant.

Robustness of Two-Sample t-Procedures (1)

Strictly speaking, unless the two samples are both drawn from normal distributions, neither

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

nor

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution.

Nonetheless, the actual distributions of the two-sample t-statistics are well approximated by t-distributions, even when the populations are not normal, as long as the sample sizes are not too small.

This is the so-called **robustness** of the two-sample t-procedures.

Robustness of Two-Sample t-Procedures (2)

- Given a fixed sum of the sample sizes $n = n_1 + n_2$ the t-approximation works the best when the sample sizes are equal $n_1 = n_2$
  - In planning a two-sample study, choose equal sample sizes if you can
- The t-approximation is generally good if $n_1 + n_2$ is not too small (say, ≥ 15), the data are not strongly skewed, and there are no outliers.
  - Check the back-to-back stemplots or side-by-side boxplots of the data
- With $n_1 + n_2$ sufficiently large (say $n_1 + n_2 \geq 40$), the approximation is good even when the data are clearly skewed.

One-Sample or Two-Sample or Matched-Pairs? (1)

For each of the following scenario, determine whether it is a one-sample, two-sample, or a matched-paired problem.

- Comparing vitamin content of bread immediately after baking vs. 3 days later (the same loaves are used on day one and 3 days later).
  - matched-pairs
- Comparing vitamin content of bread immediately after baking vs. 3 days later (tests made on independent loaves).
  - two-sample
- Is blood pressure altered by use of an oral contraceptive? Comparing a group of women not using an oral contraceptive with a group taking it.
  - two-sample

One-Sample or Two-Sample or Matched-Pairs? (2)

- Average fuel efficiency for 2005 vehicles is 21 miles per gallon. Is average fuel efficiency higher in the new generation "green vehicles"?
  - one-sample
- Review insurance records for dollar amount paid after fire damage in houses equipped with a fire extinguisher vs. houses without one. Was there a difference in the average dollar amount paid?
  - two-sample