STAT 22000 Lecture Slides
Analysis of Two-Sample Data

Yibi Huang
Department of Statistics
University of Chicago

Coverage

This set of slides covers Section 5.3 in the 3rd edition of *OpenIntro Statistics* (or Section 7.3 in the 4th edition).

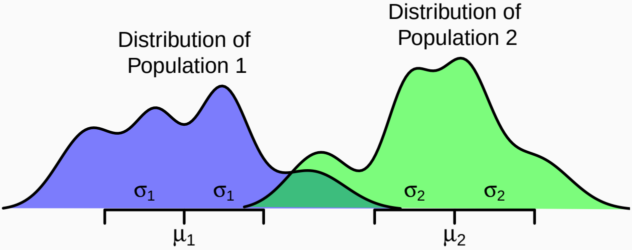
Two Sample Problems (1)

- E.g., is the air more polluted in Chicago than in LA?
- 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)

Population 1

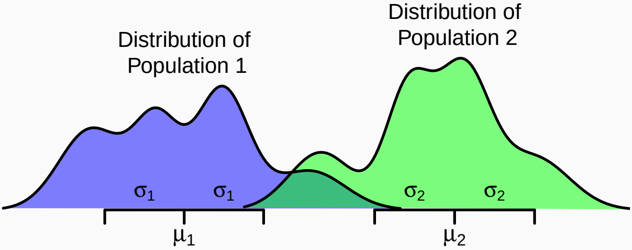
Population 2



Population distributions may NOT be normal or of the same shape for large samples.

Population 1

Population 2

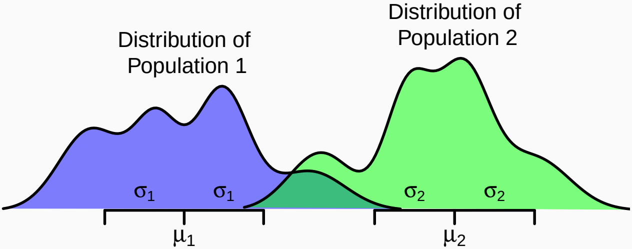


Population SDs σ_1 and σ_2 may not be equal.

Two Sample Problems (2)

Population 1

Population 2

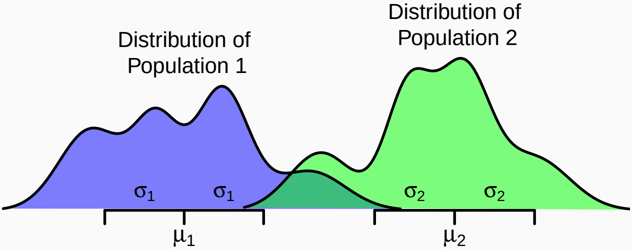


Goal: inference about difference of population means $\mu_1 - \mu_2$.

Two Sample Problems (2)

Population 1

Population 2



Data may come from experiments or observational studies.

Two Sample Data

For an Observational Study:

$$\begin{array}{ll} \text{Population 1} & \longrightarrow \text{random sample } X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} \\ \text{Population 2} & \longrightarrow \text{random sample } X_{2,1}, X_{2,2}, \ldots\ldots\ldots, X_{2,n_2} \end{array}$$

For a Randomized Experiment:

$$\begin{array}{ll} \text{Treatment 1} & \longrightarrow \text{observations } X_{1,1}, X_{1,2}, \ldots, X_{1,n_1} \\ \text{Treatment 2 (Control)} & \longrightarrow \text{observations } X_{2,1}, X_{2,2}, \ldots\ldots\ldots, X_{2,n_2} \end{array}$$

- The responses in each group are **independent** of those in the other group

Two Sample Problems (3)

A natural estimate of $\mu_1 - \mu_2$ is the difference of the two sample means $\overline{X}_1 - \overline{X}_2$.

How close is $\overline{X}_1 - \overline{X}_2$ to $\mu_1 - \mu_2$?

Two Sample Problems (4)

Recall that

$$E(\overline{X}_1) = \mu_1, \quad \mathrm{V}(\overline{X}_1) = \sigma_1^2/n_1$$
$$E(\overline{X}_2) = \mu_2, \quad \mathrm{V}(\overline{X}_2) = \sigma_2^2/n_2.$$

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

$$\mathrm{V}(\overline{X}_1 - \overline{X}_2) = \mathrm{V}(\overline{X}_1) + \mathrm{V}(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Thus the **standard error** 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 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.

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

- 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:
$$k = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}, \quad \text{where} \quad \begin{array}{l} w_1 = s_1^2/n_1, \\ w_2 = s_2^2/n_2. \end{array}$$
2. simple formula: $\boxed{k = \min(n_1 - 1, n_2 - 1)}$

Comparison of the two formulas:

- The software formula is more accurate. It gives larger d.f. and yields shorter CIs and smaller p-value
- The simple formula is conservative. I.e., it yields wider CIs and larger p-values than the actual p-value
- In STAT 220, it is fine to **just use the simple formula**.

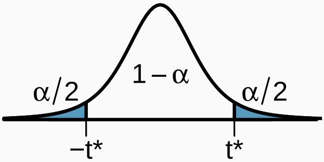
Confidence Intervals for $\mu_1 - \mu_2$

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^* is the value of the t distribution with k degrees of freedom such that

density curve of t_k



which can be found in R using the `qt()` command.

```
qt(alpha/2, df=k, lower.tail=F)
```

Example: Nitrogen Effect on Tree Growth

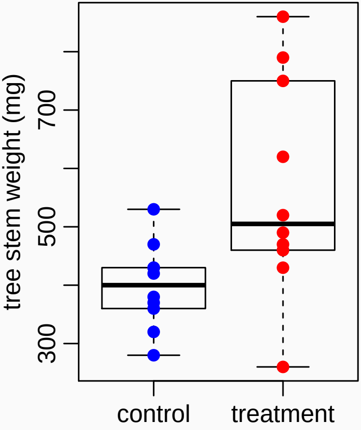
20 northern red oak seedlings

half received nitrogen, and half didn't.

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 = 72.79		SD = 186.74	
$n_C = 10$		$n_T = 10$	



Example: CI for the Nitrogen Effect on Tree Growth

The df is $\min(10-1, 10-1) = 9$. The critical value $t^* \approx 2.26$ for 95% CI can be found in R as follows

```
> qt(0.05/2, df=9, lower.tail=F)
[1] 2.262157
```

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

$$\overline{X}_T - \overline{X}_C \pm t^* \sqrt{\frac{s_T^2}{n_1} + \frac{s_C^2}{n_2}} = 565 - 399 \pm 2.26 \sqrt{\frac{(186.74)^2}{10} + \frac{(72.79)^2}{10}}$$

$$\approx 166 \pm 143.4 = (22.6, 309.4)$$

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.

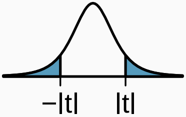
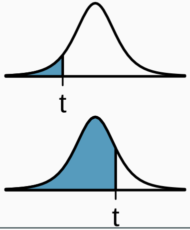
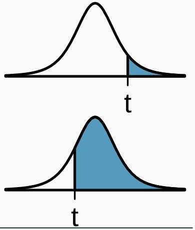
We conclude that Nitrogen has an effect on stem weight.

Hypothesis Tests for $\mu_1 - \mu_2$

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

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

where the df is $k = \min(n_1 - 1, n_2 - 1)$, or the one given by the software formula, and the p-value is computed as follows depending on H_a.

H_a	$\mu_1 - \mu_2 \neq \delta_0$	$\mu_1 - \mu_2 < \delta_0$	$\mu_1 - \mu_2 > \delta_0$
p-value			

The bell curve 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

$$t = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{s_T^2/n_T + s_C^2/n_C}} = \frac{565 - 399}{\sqrt{\frac{(186.74)^2}{10} + \frac{(72.79)^2}{10}}} = \frac{166}{63.38} \approx 2.62.$$

The degrees of freedom is $10 - 1 = 9$. The two-sided P-value can be found in R to be ≈ 0.0278.

```
> 2*pt(2.62, df=9, lower.tail=F)
[1] 0.02781094
```

The difference is significant at 5% level.

We conclude that Nitrogen has an effect on stem weight.

Two-Sample Tests/CIs in R

```
> ctrl = c(320,430,530,360,280,420,370,380,470,430)
> trt = c(260,750,430,790,470,860,490,620,520,460)
```

By default, the R command `t.test` does NOT assume $\sigma_1 = \sigma_2$.

```
> t.test(ctrl, trt)

Welch Two Sample t-test

data:  ctrl and trt
t = -2.6191, df = 11.673, p-value = 0.02286
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -304.52438  -27.47562
sample estimates:
mean of x mean of y
      399       565
```

Note the df $= 11.673$ given above is based on the software formula, which is more accurate than the simple formula.

Robustness of Two-Sample *t*-Procedures (1)

Even when the populations are not normal, the two-sample statistics

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

can be well-approximated by t-distributions, 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)

- The t-approximation is generally good if $n_1 + n_2$ is not too small (both ≥ 15), the data are not strongly skewed, and there are no outliers.
  - Check histograms or side-by-side boxplots of the data
- With $n_1 + n_2$ sufficiently large (say both ≥ 30), the approximation is good even when the data are clearly skewed.
- 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

Analysis of Two Sample Data When $\sigma_1 = \sigma_2$ (Optional Topic)

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 sample 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 *squared deviations from the corresponding means*, 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**.

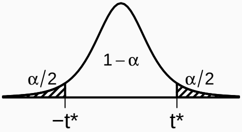
- It has an **exact** t-distribution with $n_1 + n_2 - 2$ **degrees of freedom** when the two populations are normal.
- It is approximately $t_{(n_1+n_2-2)}$ as long as the sample size n_1, n_2 is not too small.
- The degrees of freedom, $n_1 + n_2 - 2$ is greater the degrees of freedom given by the 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 where t^* is the value of the t distribution with $n_1 + n_2 - 2$

degrees of freedom such that  .

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 $n_T + n_C - 2 = 10 + 10 - 2 = 18$. The critical value $t^* \approx 2.101$ for 95% CI can be found in R as follows

```
> qt(0.05/2, df=18, lower.tail=F)
[1] 2.100922
```

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^* 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 df is $n_T + n_C - 2 = 10 + 10 - 2 = 18$.

The two-sided P-value can be found in R to be

```
> 2*pt(2.619, df=18, lower.tail=F)
[1] 0.01738994
```

The pooled t-test gives smaller p-value and the result appears more significant.

Two-Sample Tests/CIs in R

One can force σ_1, σ_2 to be equal by the argument `var.equal = T`.

```
> t.test(ctrl, trt, var.equal = T)

Two Sample t-test

data:  ctrl and trt
t = -2.6191, df = 18, p-value = 0.01739
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -299.15788  -32.84212
sample estimates:
mean of x mean of y
      399       565
```

Which Two-Sample Tests/CIs to Use?

We have introduced two different two-sample tests/CIs:

- the one assuming $\sigma_1 = \sigma_2$ used the **pooled SD**.
- the one w/o assuming $\sigma_1 = \sigma_2$ is called **Welch's method**.

Though in many cases, the two methods agree in the conclusion, but they can provide different answers when:

- the sample SDs are very different, and
- the sizes of the groups are also very different

So which method should I use?

- When σ_1 and σ_2 are indeed equal, the method based on pooled SD is more powerful
- However, it is usually hard to check whether $\sigma_1 = \sigma_2$. So it's safer to use Welch's method.