STAT 515 fa 2023 Lec 16 slides

Two-sample testing

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Think about comparing two populations:

- Compare μ_1 with μ_2 by comparing $\bar{X}_1$ and $\bar{X}_2$.
- Compare p_1 with p_2 by comparing $\hat{p}_1$ and $\hat{p}_2$.

1. Inference about $\mu_1 - \mu_2$
   - Sampling distribution of difference in sample means
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$
2. Inference about $p_1 - p_2$

Consider two random samples:

$$X_{11}, \dots, X_{1n_1} \text{ a rs from a pop. with mean } \mu_1 \text{ and variance } \sigma_1^2$$
$$X_{21}, \dots, X_{2n_2} \text{ a rs from a pop. with mean } \mu_2 \text{ and variance } \sigma_2^2$$

Goals:

1. Build confidence intervals for $\mu_1 - \mu_2$
2. Test null and alternate hypotheses of the form

$$\begin{array}{lclcl} H_0\colon \mu_1 - \mu_2 \geq \delta_0 & \text{or} & H_0\colon \mu_1 - \mu_2 = \delta_0 & \text{or} & H_0\colon \mu_1 - \mu_2 \leq \delta_0 \\ H_1\colon \mu_1 - \mu_2 < \delta_0 & & H_1\colon \mu_1 - \mu_2 \neq \delta_0 & & H_1\colon \mu_1 - \mu_2 > \delta_0. \end{array}$$

In most situations we have $\delta_0 = 0$.

Exercise: Write down the null and alternate hypotheses for the following:

1. Do honors grads earn more in first post-grad year than non-honors grads?
2. Do PhD holders earn at least twice as much as Bachelor's degree holders?
3. Does a fertilizer increase crop yields?

Sampling distribution of difference in sample means

1. If both populations are Normal

$$\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

2. Otherwise, as long as $\sigma_1^2 < \infty$ and $\sigma_2^2 < \infty$,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \text{ behaves more and more like } Z \sim \text{Normal}(0, 1)$$

for larger and larger n_1 and n_2.

1. Inference about $\mu_1 - \mu_2$
   - Sampling distribution of difference in sample means
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$
2. Inference about $p_1 - p_2$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2_{\text{common}}$, then we can estimate σ^2_{common} with

$$S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Sampling distribution of difference in sample means ($\sigma_1^2 = \sigma_2^2$)

1. If both populations are Normally distributed, then

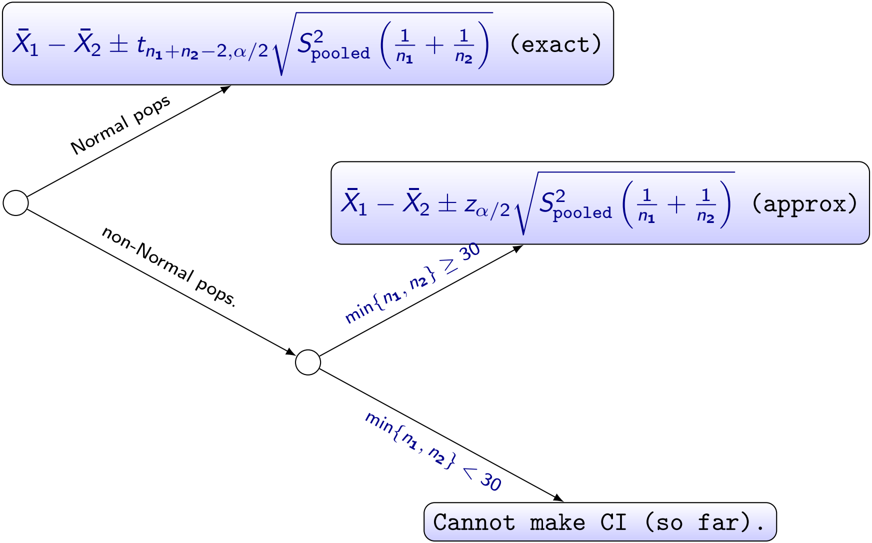
$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}.$$

2. Otherwise

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ behaves more and more like } Z \sim \text{Normal}(0, 1)$$

for larger and larger n_1 and n_2.

Formulas for $(1-\alpha)100\%$ confidence intervals for $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$.



$$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S^2_{\texttt{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \texttt{(exact)}$$

Normal pops

non-Normal pops.

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{S^2_{\texttt{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \texttt{(approx)}$$

$\min\{n_1, n_2\} \geq 30$

$\min\{n_1, n_2\} < 30$

`Cannot make CI (so far).`

Exercise: We wish to know whether drug tablets produced at two different sites have the same average concentration of the drug. Download .Rdata file (Ex 6.92 in [2]).

Assuming Normality and $\sigma_1^2 = \sigma_2^2$, build a 95% confidence interval for the difference in means.

Let $X_{k1}, \dots, X_{kn_k} \overset{\text{ind}}{\sim} \text{Normal}(\mu_k, \sigma_k^2)$, $k = 1, 2$.

Tests about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$

For some null value δ_0, define the test statistic

$$T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{S_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Then the following tests have $P(\text{Type I error}) \leq \alpha$.

$H_0\colon \mu_1 - \mu_2 \geq \delta_0$ $H_1\colon \mu_1 - \mu_2 < \delta_0$	$H_0\colon \mu_1 - \mu_2 = \delta_0$ $H_1\colon \mu_1 - \mu_2 \neq \delta_0$	$H_0\colon \mu_1 - \mu_2 \leq \delta_0$ $H_1\colon \mu_1 - \mu_2 > \delta_0$
Reject H_0 if $T_{\text{test}} < -t_{n_1+n_2-2,\alpha}$	Reject H_0 if $\lvert T_{\text{test}}\rvert > t_{n_1+n_2-2,\alpha/2}$	Reject H_0 if $T_{\text{test}} > t_{n_1+n_2-2,\alpha}$
$p\text{-val} = P(T < T_{\text{test}})$	$p\text{-val} = 2 \cdot P(T > \lvert T_{\text{test}}\rvert)$	$p\text{-val} = P(T > T_{\text{test}})$

For computing the p-values, let $T \sim t_{n_1+n_2-2}$. **Draw pictures.**

Exercise: We wish to know whether drug tablets produced at two different sites have the same average concentration of the drug. Download .Rdata file (Ex 6.92 in [2]).

1. What are the null and alternate hypotheses?
2. Assuming Normality and $\sigma_1^2 = \sigma_2^2$, find the p-value.

1. Inference about $\mu_1 - \mu_2$
   - Sampling distribution of difference in sample means
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$
2. Inference about $p_1 - p_2$

Sampling distribution of difference in sample means ($\sigma_1^2 \neq \sigma_2^2$)

1. If both populations are Normally distributed, then

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{approx}}{\sim} t_{\nu^*}.$$

2. Otherwise

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ behaves more and more like } Z \sim \text{Normal}(0,1)$$

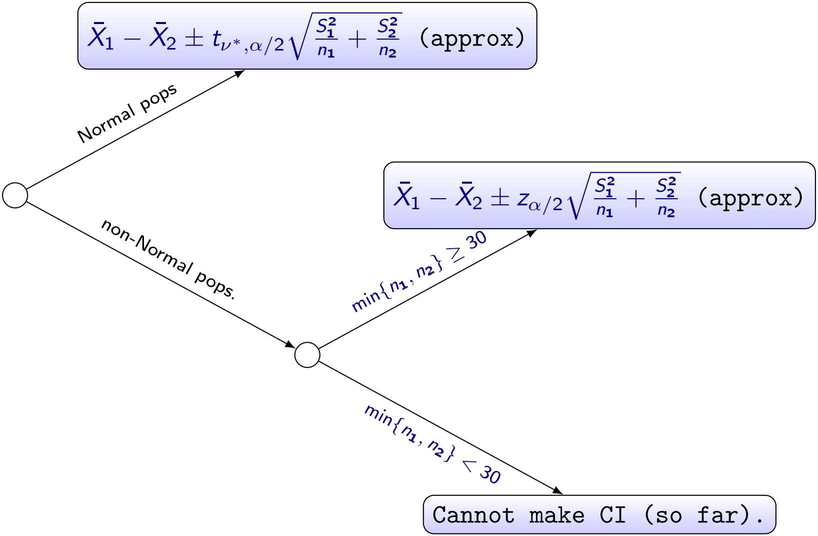
for larger and larger n_1 and n_2.

In the above

$$\nu^* = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left[\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right]^{-1}$$

😭

Formulas for $(1-\alpha)100\%$ confidence intervals for $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$.



$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu^*,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \texttt{(approx)}$$

Normal pops

non-Normal pops.

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \texttt{(approx)}$$

$\min\{n_1, n_2\} \geq 30$

$\min\{n_1, n_2\} < 30$

Cannot make CI (so far).

Let $X_{k1}, \dots, X_{kn_k} \overset{\text{ind}}{\sim} \text{Normal}(\mu_k, \sigma_k^2)$, $k = 1, 2$.

Tests about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$

For some null value δ_0, define the test statistic

$$T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2 - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Then the following tests have $P(\text{Type I error}) \leq \alpha$.

$H_0\colon \mu_1 - \mu_2 \geq \delta_0$ $H_1\colon \mu_1 - \mu_2 < \delta_0$	$H_0\colon \mu_1 - \mu_2 = \delta_0$ $H_1\colon \mu_1 - \mu_2 \neq \delta_0$	$H_0\colon \mu_1 - \mu_2 \leq \delta_0$ $H_1\colon \mu_1 - \mu_2 > \delta_0$
Reject H_0 if $T_{\text{test}} < -t_{\nu^*,\alpha}$	Reject H_0 if $\lvert T_{\text{test}} \rvert > t_{\nu^*,\alpha/2}$	Reject H_0 if $T_{\text{test}} > t_{\nu^*,\alpha}$
$p\text{-val} = P(T < T_{\text{test}})$	$p\text{-val} = 2 \cdot P(T > \lvert T_{\text{test}} \rvert)$	$p\text{-val} = P(T > T_{\text{test}})$

For computing the p-values, let $T \sim t_{\nu^*}$. **Draw pictures**.

Exercise: We wish to know whether drug tablets produced at two different sites have the same average concentration of the drug. Download .Rdata file (Ex 6.92 in [2]).

Assuming Normality and $\sigma_1^2 \neq \sigma_2^2$:

1. Find the p-value for testing

$$H_0\colon \mu_1 - \mu_2 = 0 \text{ versus } H_1\colon \mu_1 - \mu_2 \neq 0.$$

2. Build a 95% confidence interval for the difference in means.

Save trouble by using the `t.test()` function in R. 😂

1. For $\sigma_1^2 = \sigma_2^2$, we can use

```
t.test(x1,x2,var.equal=TRUE,alternative="two.sided")
```

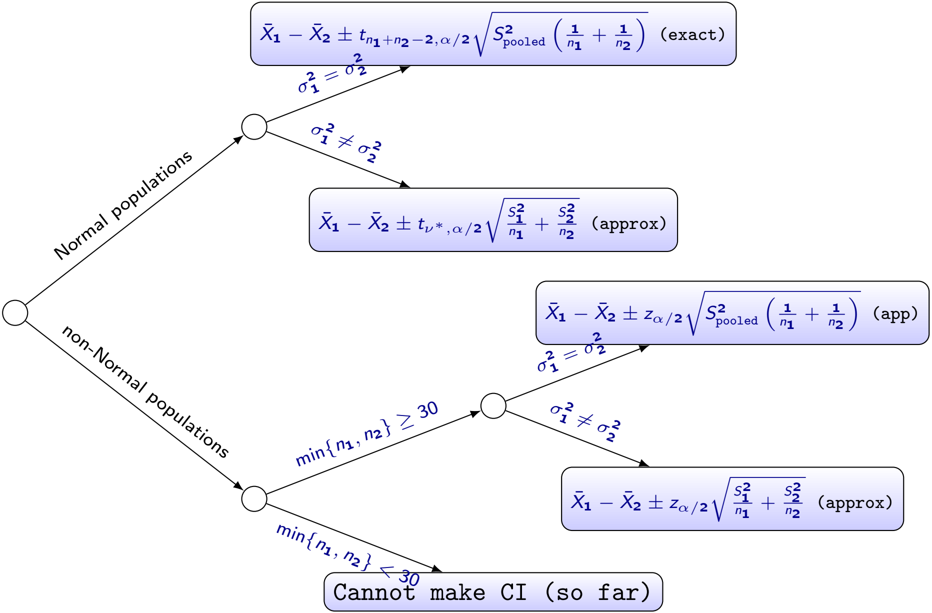
2. For $\sigma_1^2 \neq \sigma_2^2$, we can use

```
t.test(x1,x2,var.equal=FALSE,alternative="two.sided")
```

We can change the `alternative` argument to test other sets of hypotheses.

Run `?t.test` to read the documentation.

Exercise: Replicate results for drug data using the `t.test()` function.



Normal populations

$\sigma_1^2 = \sigma_2^2$

$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ (exact)

$\sigma_1^2 \neq \sigma_2^2$

$\bar{X}_1 - \bar{X}_2 \pm t_{\nu^*,\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ (approx)

non-Normal populations

$\min\{n_1, n_2\} \geq 30$

$\sigma_1^2 = \sigma_2^2$

$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ (app)

$\sigma_1^2 \neq \sigma_2^2$

$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ (approx)

$\min\{n_1, n_2\} < 30$

Cannot make CI (so far)

Exercise: It is of interest to know whether drug tablets produced at two different sites have the same average concentration of the drug. [Download .Rdata file].

1. What are the null and alternate hypotheses?
2. Assuming Normality and $\sigma_1^2 = \sigma_2^2$, find the p-value.
3. Assuming Normality and $\sigma_1^2 \neq \sigma_2^2$, find the p-value.

1. Inference about $\mu_1 - \mu_2$
   - Sampling distribution of difference in sample means
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 = \sigma_2^2$
   - Inference about $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$

2. Inference about $p_1 - p_2$

Exercise: Write down the null and alternate hypotheses for the following:

1. Do same number of honors and non-honors students pursue grad school?
2. Does a vaccine reduce the probability of getting an infection?
3. Do rural and urban voters differ in their preferences for a candidate?

Let $X_{k1}, \dots, X_{kn_k} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k)$, $k = 1, 2$, and let $\hat{p}_1 = \bar{X}_1$, $\hat{p}_2 = \bar{X}_2$.

Sampling distribution of difference in sample proportions

We have

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \text{ behaves more and more like } Z \sim \text{Normal}(0,1)$$

for larger and larger n_1 and n_2.

Rule of thumb: Need $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Confidence interval for difference in proportions

An approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

provided $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Exercise: It is reported that among the 319 adult first class passengers aboard the Titanic, 197 survived, while among the 627 adult third class passengers, 151 survived. The data are taken from [1].

Build a 95% confidence interval for the difference in the "true" proportions as a way of assessing whether the probability of surviving was affected by class.

Tests about $p_1 - p_2$

Define the test statistic

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

Then for n_1, n_2 large, the following tests have (approx) $P(\text{Type I error}) \leq \alpha$.

$H_0\colon p_1 - p_2 \geq 0$ $H_1\colon p_1 - p_2 < 0$	$H_0\colon p_1 - p_2 = 0$ $H_1\colon p_1 - p_2 \neq 0$	$H_0\colon p_1 - p_2 \leq 0$ $H_1\colon p_1 - p_2 > 0$
Reject H_0 if $Z_{\text{test}} < -z_\alpha$	Reject H_0 if $\lvert Z_{\text{test}}\rvert > z_{\alpha/2}$	Reject H_0 if $T_{\text{test}} > z_\alpha$
p-val $= P(Z < Z_{\text{test}})$	p-val $= 2 \cdot P(Z > \lvert Z_{\text{test}}\rvert)$	p-val $= P(Z > Z_{\text{test}})$

In the above $\hat{p}_0 = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$.

Exercise: Suppose that in random samples of size 1000 of 15-17 yr-olds and 25-35 yr-olds, 6% and 3%, respectively, were found to have used JUUL in the last month. You wish to know if the proportion is higher in the younger age group. This exercise is based on some summary statistics given in [3].

1. Give the hypotheses of interest.
2. What is our conclusion at the $\alpha = 0.01$ significance level?

Robert J MacG Dawson.
The “unusual episode” data revisited.
Journal of Statistics Education, 3(3), 1995.

J.T. McClave and T.T. Sincich.
Statistics.
Pearson Education, 2016.

Donna M Vallone, Morgane Bennett, Haijun Xiao, Lindsay Pitzer, and Elizabeth C Hair.
Prevalence and correlates of juul use among a national sample of youth and young adults.
Tobacco Control, 2018.