04/03/2023

Chapter 9 - Inferences from Two Samples

Consider two sets of responses (measurements) that are random samples from two populations. The two samples may be independent from one another or "matched pairs".

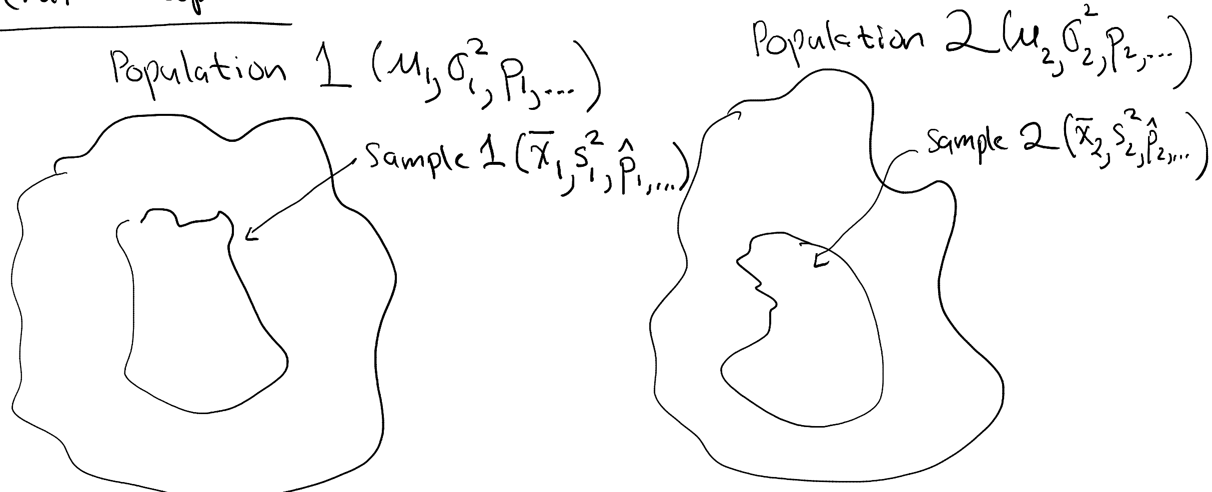
Ex: Compare two drugs given for high blood pressure.

Drug A
Drug B } compare the treatment means in terms of blood pressure reduction.

Ex: Living at home (born between 1981 and 1996)

Missouri
Illinois } proportion of millennials living at home with parents

General Concept:



Two sample sizes may be different. i.e., n_1 may not equal n_2.

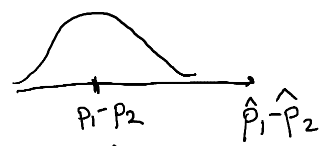
We wish to compare μ_1 and μ_2 or p_1 and p_2.

Sample from Population 1 Sample from Population 2

	Sample from Population 1	Sample from Population 2
Sample size	n_1	n_2
Sample mean	$\bar{x}_1$	$\bar{x}_2$
Sample variance	s_1^2	s_2^2
sample proportion	$\hat{p}_1$	$\hat{p}_2$

A "large-sample" CI for $p_1 - p_2$

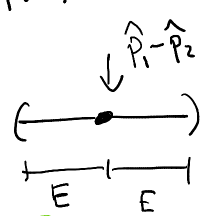
If two random samples are independent and $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$, and $n_2(1-\hat{p}_2) \geq 5$, then the CLT for $\hat{p}_1 - \hat{p}_2$ results in



$$\hat{p}_1 - \hat{p}_2 \overset{\text{approx}}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

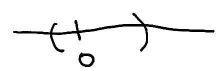
An approximate $100(1-\alpha)\%$ CI for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm E$$



or

$$\left[(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$



$p_1 - p_2 = 0$ when $p_1 = p_2$

Ex:

	Missouri Millennials	Illinois Millennials
sample size	$300 = n_1$	$400 = n_2$
Number of "successes"	81	84
sample proportion	$\frac{81}{300} = 0.27 = \hat{p}_1$	$\frac{84}{400} = 0.21 = \hat{p}_2$

300 400

Assumptions? Two independent $\overset{\text{random}}{\wedge}$ samples

Check: $n_1\hat{p}_1 \geq 5$ ✓ $n_2\hat{p}_2 \geq 5$ ✓

$n_1(1-\hat{p}_1) \geq 5$ ✓ $n_2(1-\hat{p}_2) \geq 5$ ✓

An approximate 95% CI for $p_1 - p_2$ is

$$(0.27 - 0.21) \pm 1.96\sqrt{\frac{(0.27)(1-0.27)}{300} + \frac{(0.21)(1-0.21)}{400}}$$

$$0.06 \pm 0.064$$

or $(-0.004,\ 0.124)$

Likely values of $p_1 - p_2$ are between -0.004 and 0.124 based on the two samples collected.

Q: Is it possible, based on data, that $p_1 - p_2 = 0$?

A: Yes

$H_0: p_1 = p_2 \quad (p_1 - p_2 = 0)$

$H_1: p_1 \neq p_2 \quad (p_1 - p_2 \neq 0)$

$\alpha = 0.05$

Fail to reject H_0.

Sample evidence does not substantiate the claim that the proportion of millennials living at home is different in Missouri and Illinois.

CI for $\mu_1 - \mu_2$ Based on Normal Populations or

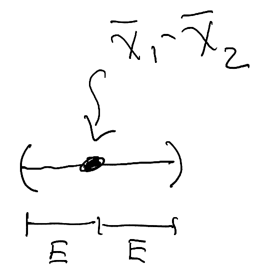
CI for $\mu_1 - \mu_2$ Based on Normal Populations or the CLT for $\bar{x}_1 - \bar{x}_2$

Two independent random samples.

Assume (1) both populations are normally distributed or (2) $n_1 \geq 30$ and $n_2 \geq 30$.

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

$$\left[(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$



$$(\bar{x}_1 - \bar{x}_2) \pm E$$

Remarks: 1. Textbook suggests a "conservative" df formula, which is

$$df = \text{minimum}\{n_1 - 1, n_2 - 1\}$$

2. See textbook (or software) for an alternative formula.

Ex: (see handout)

	Treatment	n	Sample mean (mm Hg)	Sample standard deviation (mm Hg)
Group 1	Biofeedback	40	$\bar{x}_1 = 10.4$	$s_1 = 12.5$
Group 2	Control	50	$\bar{x}_2 = 4.0$	$s_2 = 13$

$n_1 = 40$, $\bar{x}_1 = 10.4$, $s_1 = 12.5$

$n_2 = 50, \quad \bar{x}_2 = 4.0, \quad s_2 = 13$

$t_{\frac{0.05}{2}} = t_{0.025} = 2.023$ since $df = 40 - 1 = 39$

Assumptions?

A 95% CI for $\mu_1 - \mu_2$ is

$$(10.4 - 4.0) \pm 2.023\sqrt{\frac{(12.5)^2}{40} + \frac{(13)^2}{50}}$$

$$6.4 \pm 5.46$$

or $(0.94 \text{ mm Hg}, \; 11.86 \text{ mm Hg})$

Likely values of $\mu_1 - \mu_2$ are from 0.94 mm Hg to 11.86 mm Hg.

Q: Based on samples, does μ_1 differ from μ_2?

A: Yes, because zero is not a likely value of $\mu_1 - \mu_2$.

$H_0: \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$

$H_1: \mu_1 \neq \mu_2 \quad (\mu_1 - \mu_2 \neq 0)$

$\alpha = 0.05$ (equivalent to using 95% CI)

Reject H_0.

Ex: Backfat thickness of pigs.

Does it appear that the mean backfat thickness of pigs on Diet 1 exceeds

Does it appear that the mean backfat thickness of pigs on Diet 1 exceeds that of pigs of Diet 2?

	n	sample mean (cm)	Sample standard deviation (cm)
Diet 1	10	3.49	0.40
Diet 2	8	3.18	0.37

Two Independent Sample t-test for $\mu_1 - \mu_2$

04/05/2023

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$ (right tail test)

$\alpha = 0.05$

Aside:

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 > 0$

test Statistic

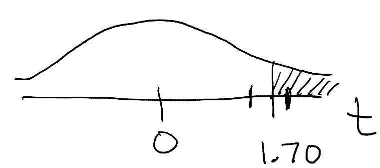
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Assumptions?

observed test statistic

$$t = \frac{(3.49 - 3.18) - 0}{\sqrt{\frac{(0.40)^2}{10} + \frac{(0.37)^2}{8}}} = 1.70$$

$df = 8 - 1 = 7$



One can show that

$$0.05 < \text{P-value} = P(t \geq 1.70) < 0.10$$

P-value $> \alpha$ so fail to reject Ho.

Sample evidence does not support the claim that the mean backfat thickness of pigs on Diet 1 exceeds that of pigs on Diet 2.