

Comparing two Populations Parameters

LEARNING OBJECTIVES

At the end of this section, you should be able to do the following:

- Discuss the logic behind and demonstrate the techniques for using **independent samples** to test hypotheses and develop interval estimates for the difference between **two population means**.
- Develop confidence interval estimates and conduct hypothesis tests for the difference between two population means for **paired samples**.
- Carry out hypothesis tests and establish interval estimates, using sample data, for the difference between **two population proportions**.

CASE STUDY

SOCIAL MEDIA MARKETING

- YouTube advertising is more effective than Facebook advertising as a method of attracting new subscribers to HBO GO.

OR

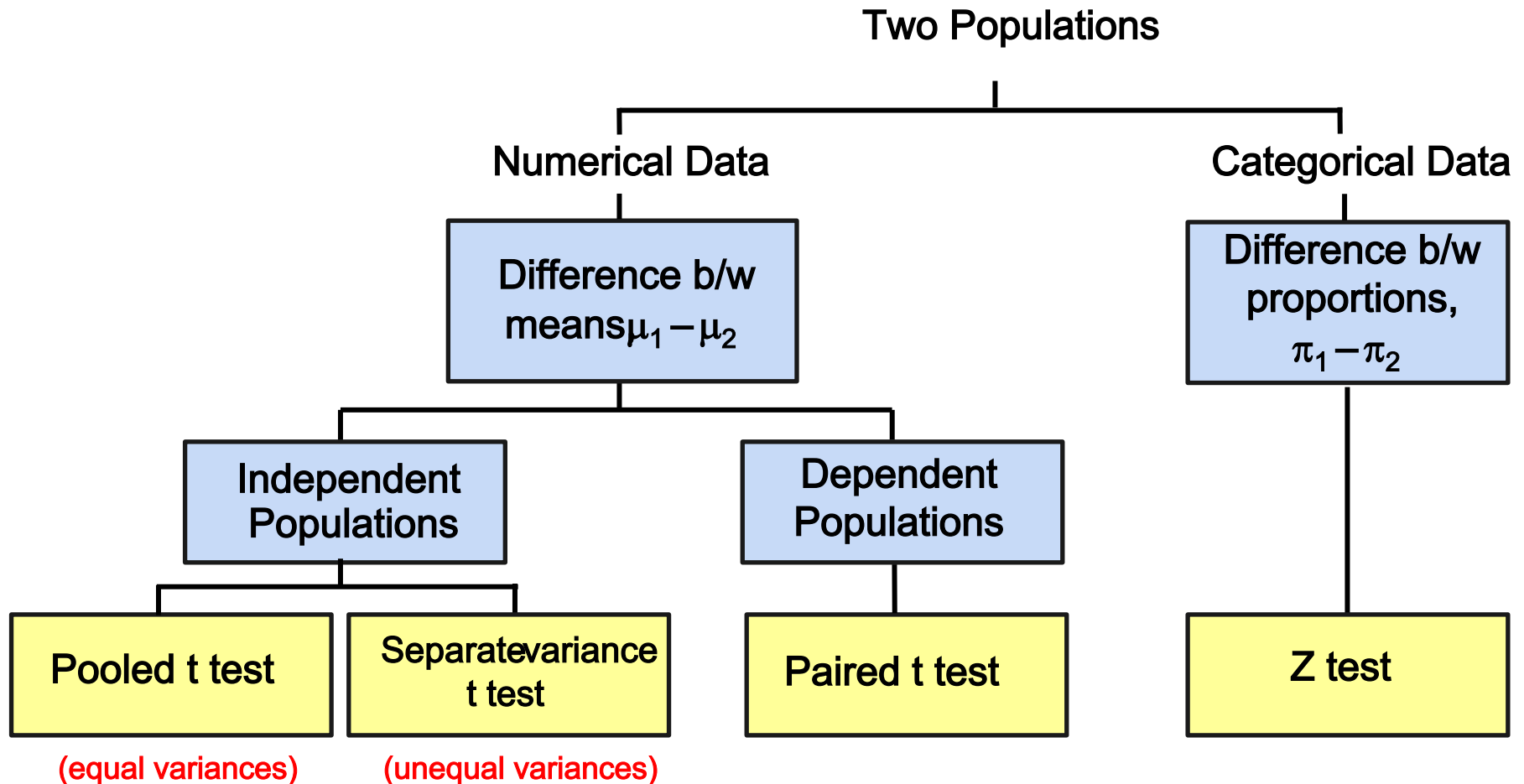
- The mean number of advertisements it takes for a YouTube user to become a paying HBO GO subscriber is less than that of a Facebook user.



COMPARING TO SAMPLES

- Up until now we have only considered situations involving **one sample**.
- There are many situations where we are interested in **comparing two samples**.
- To do this, we need to extend our earlier techniques.

INFERENCEAL TECHNIQUES FOR TWO SAMPLES



DIFFERENCE BETWEEN TWO MEANS: INDEPENDENT SAMPLES


Independent Samples

- Samples selected from two or more populations in such a way that the occurrence of values in one sample has no influence on the probability of the occurrence of values in the other sample(s).
- For **equal variances**, we use the **pooled t-test**
- For **unequal variances**, use the **separate variance t-test**
- There are **formal ways** to test for equal variances. However, a simple check (such as comparing sample variances or standard deviations) will suffice.

A REVIEW OF HYPOTHESIS TESTING PROCESS

Step 1: Set up H_0 and H_1 (in words and symbols)

Step 2: Decide on the direction of our test?
(Two tail, Lower tail, Upper tail)




Critical z/t is determined differently

Step 3: Decide on α , the level of significance and determine critical values of z or t .

Step 4: Decision Rule (using critical values of z or t .)

Step 5: Sample (Perform relevant calculations)



Sample z or t statistics are calculated differently

Step 6: Conclusion (About H_0 and H_1 and in Practical Terms)

HYPOTHESIS TESTING FOR TWO POPULATIONS MEANS DIFFERENCE

Assumptions:

- The populations are normally distributed
- The populations have equal variances
- The samples are independent
- Hypothesis testing should be done using t-distribution.

WHAT ARE THE HYPOTHESES IN THE CASE STUDY?

- The mean number of advertisements it takes for a YouTube user to become a paying HBO GO subscriber is less than that of a Facebook user.

H_0 : The true mean number of ads required for YouTube user to become a paying HBO GO subscriber is **greater than or equal to** that of a Facebook user.

$$H_0: \mu_{1=\text{YouTube}} \geq \mu_{2=\text{Facebook}}$$

OR

$$\mu_{1=\text{YouTube}} - \mu_{2=\text{Facebook}} \geq 0$$

H_1 : The true mean number of ads required for a YouTube user to become a paying HBO GO subscriber is **less than** that of a Facebook user.

$$H_1: \mu_{1=\text{YouTube}} < \mu_{2=\text{Facebook}}$$

OR

$$\mu_{1=\text{YouTube}} - \mu_{2=\text{Facebook}} < 0$$

HYPOTHESIS TESTING (TWO POPULATION MEANS)

- Lower tail test
- Assume 5% significance.
- $t_{\text{critical}} = -1.645$ ($df = n_1 + n_2 - 2$)
- If $t_{\text{statistic}} < t_{\text{critical}}$, then reject H_0 . Otherwise DO NOT reject H_0 .

	YouTube	Facebook
\bar{x}	3.5	4.9
s	0.5	0.3
n	200	150

- $$t_{\text{statistic}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

NOTE

One sample t -statistic

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Two sample (Independent)
Pooled t -statistic

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

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

HYPOTHESIS TESTING - CONT'D

(TWO POPULATION MEANS)

- $$t_{\text{statistic}} = \frac{(3.5 - 4.9) - (0)}{\sqrt{0.18 \left(\frac{1}{200} + \frac{1}{150} \right)}} = -30.42$$

	YouTube	Facebook
\bar{x}	3.5	4.9
s	0.5	0.3
n	200	150

- The sample statistic of $t = -30.89 < \text{critical value of } -1.645$ and therefore lies in the rejection region. We reject H_0 .
- At 5 percent significance, there is sufficient evidence to conclude that the true average number of ads required to turn a YouTube user into a paying HBO GO subscriber is less than that of the Facebook users.

TWO INDEPENDENT SAMPLES – UNEQUAL VARIANCES

Independent Samples
(unequal variances)
t-statistic

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

Degrees of Freedom
(corresponding to critical *t*-value)

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

ESTIMATING DIFFERENCE (TWO POPULATION MEANS)

- Step 1: Define the population parameter of interest and select independent samples from the two populations
- Step 2: Specify the desired confidence level
- Step 3: Compute the point estimate ($\bar{x}_1 - \bar{x}_2$)
- Step 4: Determine the standard error of the sampling distribution
- Step 5: Calculate the degrees of freedom and determine the critical value, t, from the t-distribution table
- Step 6: Develop the confidence interval estimate

CONFIDENCE INTERVAL (ESTIMATING MEANS DIFFERENCE)

- Assume 95% confidence

- Point Estimate:

$$\bar{x}_{1=\text{YouTube}} - \bar{x}_{2=\text{Facebook}} = 3.5 - 4.9 = -1.4$$

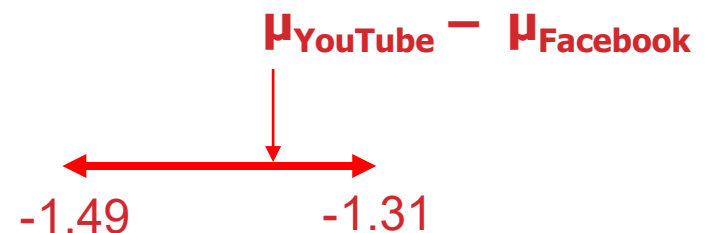
- Standard Error:

$$SE = 0.0460$$

- Critical t-value $(CI = 95\%, df = n_1 + n_2 - 2) = 1.966$.
- Margin of Error = Critical Value \times SE = 0.09
- Confidence Intervals

- We are 95% confident that the true average number of ads required to turn a YouTube user to a paid HBO GO customer is between 1.31 to 1.49 less than that of Facebook users.

	YouTube	Facebook
\bar{x}	3.5	4.9
s	0.5	0.3
n	200	150



CASE STUDY

SOCIAL MEDIA MARKETING

- YouTube advertising is more effective than Facebook advertising as a method of attracting new subscribers to HBO GO.

OR

- The proportion of YouTube users who became a paying HBO GO subscriber is greater than that of Facebook users.



WHAT ARE THE HYPOTHESES IN THE CASE STUDY?

- The proportion of YouTube users who became a paying HBO GO subscriber is greater than that of Facebook users.

H_0 : The true proportion of YouTube users who become a paying HBO GO subscriber is **less than or equal to** that of Facebook users.

$$H_0: \pi_{1=\text{YouTube}} \leq \pi_{2=\text{Facebook}} \quad \text{OR} \quad \pi_{1=\text{YouTube}} - \pi_{2=\text{Facebook}} \leq 0$$

H_1 : The true proportion of YouTube users who become a paying HBO GO subscriber **is greater than** that of Facebook users.

$$H_1: \pi_{1=\text{YouTube}} > \pi_{2=\text{Facebook}} \quad \text{OR} \quad \pi_{1=\text{YouTube}} - \pi_{2=\text{Facebook}} > 0$$

HYPOTHESIS TESTING (TWO POPULATION PROPORTIONS)

- Upper tail test
- Assume 5% significance.
- $z_{\text{critical}} = + 1.644$
- If $z_{\text{statistic}} > z_{\text{critical}}$, then reject H_0 . Otherwise DO NOT reject H_0 .

	YouTube	Facebook
x	143	109
n	200	150

- $$z_{\text{statistic}} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Pooled estimator of overall proportion

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

HYPOTHESIS TESTING - CONT'D

(TWO POPULATION PROPORTIONS)

- $$z_{\text{statistic}} = \frac{(-0.0117) - (0)}{\sqrt{0.72(1-0.72)\left(\frac{1}{200} + \frac{1}{150}\right)}}$$

= -0.240

	YouTube	Facebook
x	143	109
n	200	150

- The sample statistic of $z = -0.240 < \text{critical value of } +1.644$ and therefore lies in the non-rejection region. We DO NOT reject H_0 .
- At 5 percent significance, there is no sufficient evidence to conclude that the true proportion of YouTube users who become a paying HBO GO subscriber is greater than that of a Facebook user.

CONFIDENCE INTERVAL (ESTIMATING PROPORTIONS DIFFERENCE)

- Assume 95% confidence

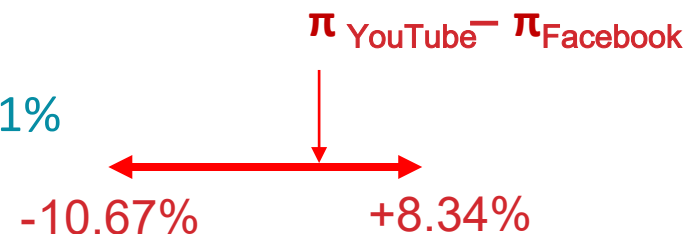
- Point Estimate:

$$p_{1=\text{YouTube}} - p_{2=\text{Facebook}} = 71.50\% - 72.67\% \\ = -1.17\%$$

- Standard Error:

$$SE = 4.85\%$$

- Critical z-value $(CI = 95\%) = 1.96$.
- Margin of Error = Critical Value \times SE = 9.51%
- Confidence Intervals



- At 95% confidence, no conclusive results could be drawn regarding the true difference in proportion of YouTube and Facebook users who become paying HBO GO subscribers.

	YouTube	Facebook
x	143	109
n	200	150

CASE STUDY

PROMOTIONAL MARKETING CAMPAIGN EFFECT

- Suppose **Menulog** wishes to conduct a test to determine whether a “a free-delivery for every 5 **Menulog** transactions” increases the average number of times users may use **Menulog** to order takeaways.
- Such analysis **could be** conducted using paired samples.
- This means that the **same users** will be offered the promotion and then, the number of transactions made by these users will be recorded both **before** and **after** the promotion campaign.



INTERVAL ESTIMATION AND HYPOTHESIS TESTS FOR PAIRED SAMPLES

- Paired samples are **dependent** samples.
- Samples that are selected in such a way that **values in one sample are matched with the values in the second sample** for the purpose of controlling for extraneous factors.

Examples

- Testing difference in gas mileage comparing regular and premium gas.
- Testing difference in employment rate before and after implementing a new wage policy.

PAIRED SAMPLES ESTIMATION AND H-TESTING

- Paired Differences

$$d = x_1 - x_2$$

x_1 and x_2 – Values from samples 1 and 2, respectively

- Point Estimate for the population mean paired difference μ_d :

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

d_i - i^{th} paired difference
 \bar{d} - Mean paired difference
 n - Number of pairs

PAIRED SAMPLES ESTIMATION AND H-TESTING

- t -Test Statistic for Paired-Sample Test

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Where:

\bar{d} = Mean paired difference

μ_d = Hypothesised population mean paired difference

s_d = Sample standard deviation for paired differences

WHAT ARE THE HYPOTHESES IN THE CASE STUDY?

- The mean number of orders users put via Menulog has increased after running the promotion campaign.

H_0 : The mean difference (i.e. before – after) in the number of Menulog orders is **greater than or equal** to zero.

$$H_0: \mu_{d=\text{before} - \text{after}} \geq 0$$

That is, $\mu_{\text{before}} \geq \mu_{\text{after}}$ (**decrease or no change in the number of transactions**)

H_1 : The mean difference (i.e. before – after) in the number of Menulog orders is **less than** zero.

$$H_1: \mu_{d=\text{before} - \text{after}} < 0$$

That is, $\mu_{\text{before}} < \mu_{\text{after}}$ (**increase in the number of transactions**)

H-TESTING (PAIRED SAMPLES)

- Lower tail test
- Assume 5% significance.
- $t_{\text{critical}} (df = n - 1) = -2.015$

User	Before	After	d_i
1	6	12	-6
2	12	8	+4
3	2	17	-15
4	9	18	-9
5	4	9	-5
6	5	20	-15

- If $t_{\text{statistic}} < t_{\text{critical}}$, then reject H_0 . Otherwise DO NOT reject H_0 .
- The mean paired difference (sample statistic): $\bar{d} = \frac{-46}{6} = -7.66$
- The standard deviation for the paired differences: $s_d = 7.14$
- Test statistic: $t = \frac{-7.66 - 0}{\frac{7.14}{\sqrt{6}}} = -2.628$

H-TESTING – CONT'D

(PAIRED SAMPLES)

- The sample statistic of $t = -2.628 < \text{critical value of } -2.015$ and therefore lies in the rejection region. We reject H_0 .
- At 5 percent significance, there is sufficient evidence to conclude that the true average number of transactions made through Menulog has increase after running the promotion campaign. In other words, the campaign has been successful.

CONFIDENCE INTERVAL – PAIRED SAMPLES (ESTIMATING MEAN DIFFERENCE)

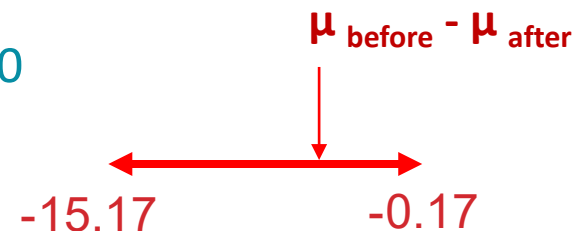
- Assume 95% confidence
- Point Estimate (The mean paired difference):

$$\bar{d} = -7.66$$

- Standard Error:

$$SE = \frac{s_d}{\sqrt{n}} = 2.92$$

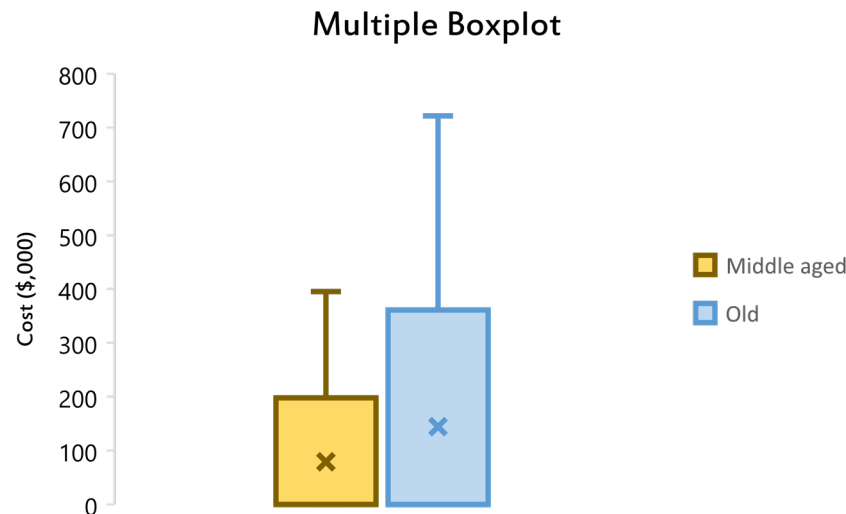
- Critical t-value $(df = n-1, CI = 95\%) = 2.57$.
- Margin of Error = Critical Value \times SE = 7.50
- Confidence Intervals



- At 95% confidence, we can conclude that the true average increase in the number of transactions via Menulog is between 0.17 to 15.17 transactions after running the campaign. Therefore, the promotional campaign has been successful.

REALITY IS NOT AS SIMPLE AS YOU THINK!!!

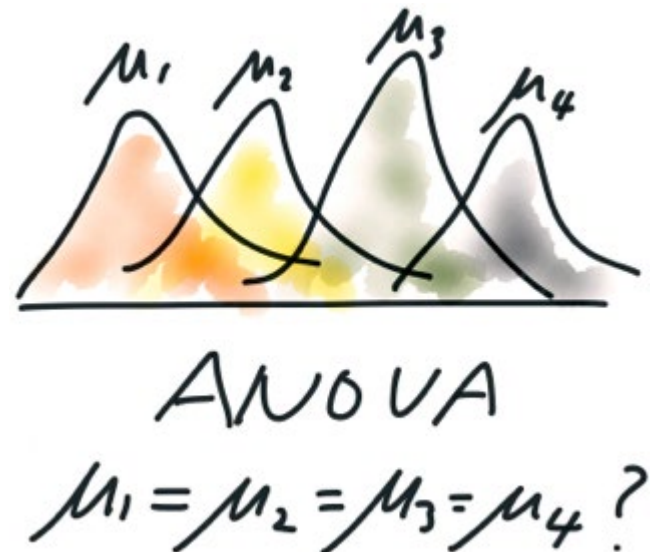
- I have a data set with **tens of thousands of observations** of medical cost data. This data is **highly skewed** to the right.
- It looks like this for **two sets of people** (in this case two **age bands** – **middle aged v. old** – with > 3000 observations each)



- Can I use two independent samples t-test to estimate means difference between these two groups?

FOOD FOR THOUGHT

- An analyst is interested in determining whether there are differences in **leg strength** between **amateur**, **semi-professional** and **professional** AFL players.
- What method can s/he use?
- Multiple t-test?
- One-way ANOVA is the answer!



QUESTIONS?