Chapter 10

Two-Sample Problems

10.1: Comparing Two Means

Assume we have two independent populations, and suppose we are to make statistical inferences about the difference between the two population means: $\mu_1 - \mu_2$

Example:
Let one population be the population of all male students, and the second be the population of all female students.

We might be interested in making inferences about the difference between the mean IQ of male students and the mean IQ of female students.

Point Estimate

We take a simple random sample of n_1 subjects from the first population (in our example it's a sample of n_1 males) and an independent simple random sample of n_2 subjects from the second population (in our example it's a sample of n_2 females).

A point estimate of $\mu_1 - \mu_2$ is the difference in the sample means: $\bar{x}_1 - \bar{x}_2$

Sampling Distribution

Assumptions:

(1) Suppose we have two independent simple random samples.

(2) (i) Either both populations are normally distributed: $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$

or (ii) The populations are possibly non-normal but both sample sizes are large enough such that the central limit theorem applies

Sampling Distribution

The sampling distribution for the difference in the sample means, $\bar{x}_1 - \bar{x}_2$, is approximately normal with mean μ_1 - μ_2 and standard deviation $\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}$

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

This assumes that σ_1 and σ_2 are both known.

The Z-Score Transformation is

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

Test for $\mu_1 - \mu_2$

We hypothesize that the difference between the population means equals some quantity say μ_0, we use data from our samples to test whether this value is reasonable or whether the mean difference is actually greater than μ_0, less than μ_0, or not equal to μ_0.

1. Hypotheses

Null Hypothesis: H_0: $\mu_1 - \mu_2 = \mu_0$

Alternative Hypothesis: $\begin{cases} H_a: \mu_1 - \mu_2 > \mu_0 \\ H_a: \mu_1 - \mu_2 < \mu_0 \\ H_a: \mu_1 - \mu_2 \neq \mu_0 \end{cases}$

Test for $\mu_1 - \mu_2$

In many problems $\mu_0 = 0$, and hence the hypotheses are often written as follows:

H_0: $\mu_1 - \mu_2 = 0$	H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 - \mu_2 > 0$	H_a: $\mu_1 > \mu_2$
H_a: $\mu_1 - \mu_2 < 0$	H_a: $\mu_1 < \mu_2$
H_a: $\mu_1 - \mu_2 \neq 0$	H_a: $\mu_1 \neq \mu_2$

Test for $\mu_1 - \mu_2$

Assumptions:

(1) Suppose we have two independent simple random samples.

(2) (i) Either both populations are normally distributed: $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$

or (ii) The populations are possibly non-normal but both sample sizes are large enough such that the central limit theorem applies

(3) The population standard deviations σ_1 and σ_2 are known.

Test for $\mu_1 - \mu_2$

2. Test Statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

3. p-value

The p-value is calculated same as before

4. Conclusion

Reject H_0 if p-value $< \alpha$

Example

The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in *Vital and Health Statistics*. Independent samples of 39 male patients and 35 female patients gave sample means of 7.9 and 7.11 days respectively. At the 5% significance level, do the data provide sufficient evidence to conclude that , on the average, the length of stay in short-term hospitals by males and females differ? Assume that σ_1=5.4 and σ_2=4.6 days.

Exercise

The registrar at Purdue is comparing the GPA of married and unmarried students. They find that 100 married students have a mean GPA of 2.85, while a random sample of 100 unmarried students have a mean GPA of 2.73. At 0.01 level of significance, do married students have a higher GPA? Assume that $\sigma_{married}$=0.4 and $\sigma_{unmarried}$=0.3.

Comparing Two Means
Small Samples and σ's unknown

Suppose we have two independent populations with population means μ_1 and μ_2, respectively. We are interested in making statistical inferences (confidence interval and significance tests) about the difference in the population means: $\mu_1 - \mu_2$.

Earlier we assumed that the population standard deviations σ_1 and σ_2 were known. Now we will discuss what to do when the population standard deviations are unknown.

If the population standard deviations are unknown, we calculate the sample standard deviations and use the t distribution instead of Z.

Confidence Interval for $\mu_1 - \mu_2$

When σ_1 and σ_2 are unknown, the Z statistic that results from the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is not appropriate.

Instead we must use a t-statistic

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

which follows a t distribution with degrees of freedom equal to the smaller of n_1 - 1 and n_2 - 1.

Significance Test on $\mu_1 - \mu_2$

Assumptions:

(1) Suppose we have two independent simple random samples.

(2) (i) Either both populations are normally distributed: $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$

or (ii) The populations are possibly non-normal but both sample sizes are large enough such that the central limit theorem applies

(3) The population standard deviations σ_1 and σ_2 are unknown.

Example

In a sampling study conducted by the Clearview National Bank, two independent samples of checking account balances for customers at two Clearview branch banks yielded the following results:

Bank Branch	Number of Checking Accounts	Sample Mean	Sample Standard deviation
Cherry Grove	12	$1000	$15
Beechmont	10	$920	$12

Develop a 90% confidence interval for the difference between the mean checking account balances at the two branch banks.

Significance Test on $\mu_1 - \mu_2$

We hypothesize that the difference between the population means equals some specified value μ_0 (=0 in most cases), we want to test whether this value is reasonable or whether the mean difference is actually greater than μ_0, less than μ_0, or not equal to μ_0.

1. Hypotheses

Null Hypothesis: $H_0: \mu_1 - \mu_2 = \mu_0$ $(H_0: \mu_1 = \mu_2)$

Alternative Hypothesis: $\begin{cases} H_a: \mu_1 - \mu_2 > \mu_0 & (H_a: \mu_1 > \mu_2) \\ H_a: \mu_1 - \mu_2 < \mu_0 & (H_a: \mu_1 < \mu_2) \\ H_a: \mu_1 - \mu_2 \neq \mu_0 & (H_a: \mu_1 \neq \mu_2) \end{cases}$

Suppose the population standard deviations σ_1 and σ_2 are unknown.

Significance Test on $\mu_1 - \mu_2$

2. Test Statistic
If the assumptions are satisfied then the test statistic is:

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

which follows a t distribution with df=min($n_1 - 1$, $n_2 - 1$).

3. P-value
The p-value of the test depends on the alternative hypothesis and is based on the t distribution.

4. Conclusion
The decision and conclusion are determined the same as for all other tests.

Example: Faculty Salaries

Independent random samples of 25 faculty members in public institutions and 30 faculty members in private institutions yielded the statistics in the following table

	Sample size	Sample Mean	Standard deviation
Public	25	57.5	20
Private	30	66.4	18

At the 5% significance level, do the data provide sufficient evidence to conclude that mean salaries for faculty in public and private institutions differ?

Example

A researcher was interested in comparing the amount of time spent watching television by women and by men. Independent samples of 14 women and 17 men were selected and each person was asked how many hours he or she had watched television during the previous week. The summary statistics are as follows

	Sample size	Sample Mean	Standard deviation
Men	17	16.9	4.7
Women	14	11.3	4.4

Do the data provide sufficient evidence to conclude that mean time for women is less than mean time for men? Perform a t-test at the 5% significance level.

Exercises

1. A random sample of 17 third graders who read poorly has a mean IQ of 98 with a standard deviation of 10; a random sample of 10 third graders who read well has mean IQ of 101 with a standard deviation of 9. At 0.05 level of significance is the mean IQ of good readers higher that the mean IQ of poor readers? Assume that the IQ scores for both groups is normally distributed.
2. A high school teachers' group is investigating summer work patterns. It finds that the mean monthly income of 20 randomly selected teachers who teach in the summer is $600 with a standard deviation of $100, while a random sample of 10 teachers who will sell real estate during the summer is $700 with a standard deviation of $50. The teachers' group believes the pay for both kinds is normally distributed with different variances. At 0.05 level of significance, is there any difference in the earning of the two groups?

Sampling Distribution
Equal Variances

When σ_1 and σ_2 are unknown and assumed equal then the statistic

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

follows a t distribution with degrees of freedom (df) equals to $n_1 + n_2 - 2$ where the pooled variance s_p^2 is given by

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

Example: Specific Motors

Specific Motors of Detroit has developed a new automobile known as the M car. 12 M cars and 8 J cars (from Japan) were road tested to compare miles-per-gallon (mpg) performance. The sample statistics are:

	Sample #1 M Cars	Sample #2 J Cars
Sample Size	n_1 = 12 cars	n_2 = 8 cars
Mean	$\bar{x}_1$ = 29.8 mpg	$\bar{x}_2$= 27.3 mpg
Standard Dev.	s_1 = 2.56 mpg	s_2 = 1.81 mpg

Test to see if the miles-per- gallon (mpg) performance is different between M cars and J cars, assuming that the distributions of the populations are normal with equal variances.

Significance Test on $\mu_1 - \mu_2$

Pooled Procedure

If the population standard deviations σ_1 and σ_2 are unknown and assumed equal (σ_1=σ_2), then the test statistic will be

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

which follows a t distribution with df=$n_1 + n_2 - 2$ where

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

Exercises

3. Recently, a local newspaper reported that part time students are older than full time students. In order to test the validity of its statement, two independent samples of students were selected. The following shows the ages of the students in the two samples. Using the following data, test to determine whether or not the average age of part time students is **significantly more than** full time students. Use an Alpha of 0.05. Assume the populations are normally distributed and have equal variances.

Full-time	**19**	**18**	**17**	**22**	**18**	**19**	**20**
Part-time	**21**	**17**	**25**	**19**	**20**	**18**	