10 Inference About Means and Proportions with Two Populations

Two populations: we may want to develop an interval estimate of the difference between the mean starting salary for a population of men and the mean starting salary for a population of women.

In this case we are talking about two populations and not one population we were using till now. Hence we need one sample out of each population which projects its own population moment with its own sampling distribution. We see how two sampling distributions interact and make inference challenging in this chapter.

(a) Inferences About the Difference Between Two Population Means: σ_1 and σ_2 Known – Do men earn significantly higher from women?

(b) Inferences About the Difference Between Two Population Means: σ_1 and σ_2 Un-Known

(c) Inferences About the Difference Between Two Population Means: Matched Samples

10.1 Inferences About the Difference Between Two Population Means under Known σ

Notation: μ_1 be mean of population 1, and μ_2 be population mean of for the second.

- ▷ Focus on $(\mu_1 - \mu_2)$ i.e. the difference in population averages in population 1 and 2, we need n_1 random sample from Pop1 and n_2 from Pop2 - samples taken independently called 'Independent Simple Random Samples'

- ▷ Known population standard deviation case for both populations i.e. σ_1 and σ_2

Margin for error and interval estimate for difference in mean performance $\mu_1 - \mu_2$

Example of Big Bazar LLC: operates two stories – one in Grater Kailash and another in Haiderpur Badli

- Observation: products that sell well in one store do not always sell well in the other. Why? Demographics of customers.

- Poorer customers live in Badli and richer in GK

- μ_1: the mean age of all customers who shop at the GK store and μ_2: as the the mean age of all customers who shop at the Badli store

10.1.1 Interval Estimation

Difference of means: $\mu_1 - \mu_2$ can be approximated by difference in mean of the two samples i.e. $\bar{x}_1 - \bar{x}_2$

- Point estimate of $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$.
- What is the standard error that describes the variation in the sampling distribution of the estimator?

Two independent random samples imply the following standard error:

$$\sigma_{x_1 - x_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Normality of population or large sample size allow us to assume that sampling distribution of $\bar{x}_1 - \bar{x}_2$ will have a normal distribution with mean $\mu_1 - \mu_2$

Interval Estimate: is point estimate adjusted for margin for error

$$\bar{x}_1 - \bar{x}_2 \pm \text{ Margin for error}$$

Margin for Error for difference of sample means is as follows:

$$z_{\alpha/2}\sigma_{x_1-x_2} = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $1 - \alpha$ is the confidence coefficient.

Example of Big Bazar LLC: $\sigma_1 = 9$ years and $\sigma_2 = 10$ years
Sample Data indicates the following:

$$n_1 = 36 \qquad n_2 = 49$$
$$\bar{x}_1 = 40 \text{ years} \qquad \bar{x}_2 = 35 \text{ years}$$

Find the point estimate for the above problem. $\bar{x}_1 - \bar{x}_2 = 5$ years.

Inference: from these estimates tells us that the customers from inner city are older in sample than our conjecture.

Find the margin for error for the above problem: under 95% confidence interval $z_{\alpha/2} = 1.96$

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$5 \pm 1.96\sqrt{\frac{9^2}{36} + \frac{10^2}{49}}$$

$$5 \pm 4.06$$

Inference: The average age difference the two populations in inner city and suburbs could be as low as 1 year (0.94) and as high as 9 years (9.06)

- How do I find out if this gap is significant or not? Do the numbers tell me if they are?

10.1.2 Hypothesis Tests

Notation: Let D_0 be the hypothesized difference between first population's mean μ_1 and second population's mean μ_2.

Three forms of hypothesis are

$$H_0 : \mu_1 - \mu_2 \geq D_0 \quad H_0 : \mu_1 - \mu_2 \leq D_0 \quad H_0 : \mu_1 - \mu_2 = D_0$$
$$H_a : \mu_1 - \mu_2 < D_0 \quad H_a : \mu_1 - \mu_2 > D_0 \quad H_a : \mu_1 - \mu_2 \neq D_0$$

Typically we are interested in difference being there or not. This implies $D_0 = 0$ and it also implies

$$H_0 : \mu_1 - \mu_2 = 0$$

while

$$H_a : \mu_1 - \mu_2 \neq 0$$

Large sample then normality approximation implies that the test statistic would be z-statistic:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example of the difference in GRE Coaching Centers – To evaluate the differences in education quality between two training centers – one in GK and another in Badli

How do we evaluate these differences ?:

- μ_1 is the average test score of students in test center in GK and μ_2 from test center in Badli
- Tentative assumption: $H_0 : \mu_1 - \mu_2 = 0$ which implies an alternative of their average scores not being equal.
- Potential Conclusion: If difference exists, then understand the factor which causes these differences and rectify them.

Example of the difference in GRE Coaching Centers – independent random samples were taken. Following characteristics denoted the differences:

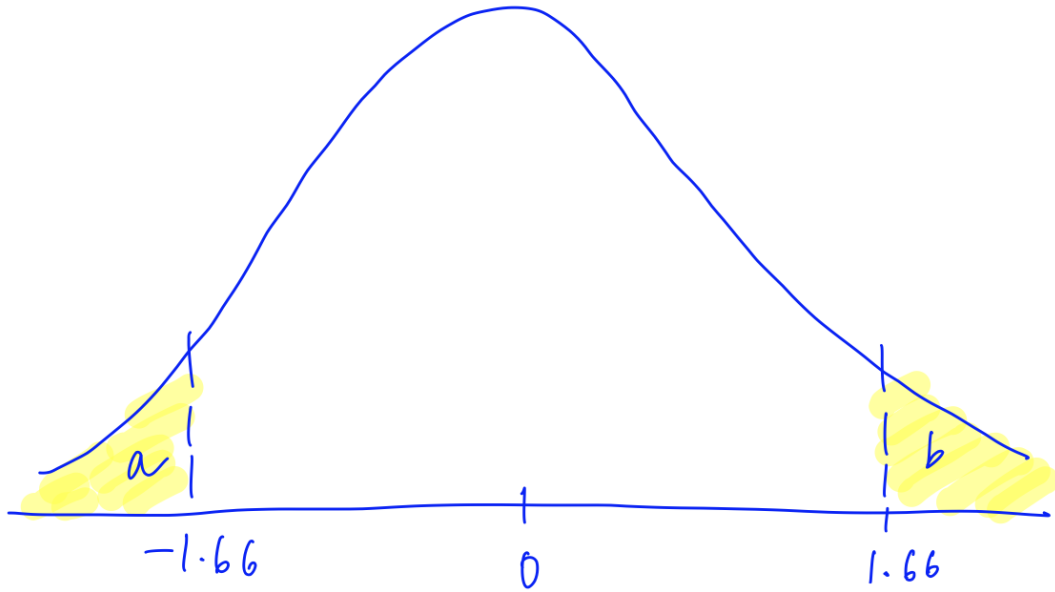
GK GRE Coaching	Badli GRE Coaching
$x_1 = 82$	$x_2 = 78$
$n_1 = 30$	$n_2 = 40$
$\sigma_1 = 10$	$\sigma_2 = 10$

Compute the test statistic with this information:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(82 - 78) - 0}{\sqrt{\frac{10^2}{30} + \frac{10^2}{40}}} = 1.66$$

Find the p-value: $z_{\text{left area}} = 0.9515$

Find the p-value: $z_{\text{left area}} = 0.9515$



area (a) = area (b).

0.0485. = area (a)

Two tailed ⇒ area (a) + area (b) = 2(0.0485) = 0.097 = p*

Rejection rule: says that if the so computed p^* value is less than chosen α, we reject the H_0

- $p^* = 0.097 > 0.05$ implies that we cannot reject the null at the pre-specified 5% level.

Conclusion: sample results do not provide sufficient evidence to conclude the training centers differ in quality.

- Note that n_1 and n_2 both are greater than 30, which is adequate size
- Certainly, it would be better if you collect a larger sample.

10.2 Inferences About Difference between Means based on Unknown σ

Quick summary: like how we approached unknown σ case in chapter 10, we do so here.

- We use the sample standard deviation s_1 and s_2 to proxy the unknown population standard deviation

- We use a t-distribution since we do not know whether population is normally distributed.

10.2.1 Interval Estimates

Example of the State Bank of India: has two branches, one in Connaught Place and another is South Extension -II.

Interested Experiment: SBI wants to estimate the difference between average savings account balance maintained by student depositors at CP branch and at South Ex branch

Sample statistics: indicated a difference of about 100 Rs. We want to know whether this difference is credible or not.

$$\begin{aligned} n_1 &= 28 & n_2 &= 22 \\ \bar{x}_1 &= 1025 & \bar{x}_2 &= 910 \\ s_1 &= 150 & s_2 &= 125 \end{aligned}$$

Interval Estimate: implies adjusting the point estimate with margin for error. In the case of unknown population standard deviation, we use $t_{\alpha/2}$ to weight the standard error

Definition 10.1. Interval Estimate in Unknown σ

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

where $(1 - \alpha)$ is the confidence coefficient.

What would be the degree of freedom in the case of a combined 't'?

The degree of freedom for t-statistic constructed to evaluate population mean differences is obtained as

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

$$n_1 = 28 \qquad n_2 = 22$$

$$\bar{x}_1 = 1025 \qquad \bar{x}_2 = 910$$

$$s_1 = 150 \qquad s_2 = 125$$

State Bank of India's test statistic degree of freedom could be found out as follows:

$$df = \frac{\left(\frac{150^2}{28} + \frac{125^2}{22}\right)}{\frac{1}{28-1}\left(\frac{150^2}{28}\right)^2 + \frac{1}{22-1}\left(\frac{125^2}{22}\right)^2} = 47.8$$

State Bank of India's test statistic so obtained is $t_{0.025}$ with 47 degree of freedom from the t-table for a two sided hypothesis is 2.012 Interval for SBI Branch savings differences:

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 115 \pm 78$$

Result: We obtain a 95% confidence interval between $[37, 193]$ which is the range of average differences in savings account balance between the two SBI branches.

- In other words, if you choose repeated samples and calculate the difference between the two average savings balances, then 95% of these difference would lie between the mentioned intervals.

10.2.2 Hypothesis Testing

Example of Matlab vs. Python coding times: coding times for a macro model vary a lot. Python claims that it tends to be faster since the codes are relatively easy to write and the modeller need not worry about semantics.

Research Hypothesis: want to say that the average coding time for a macro model in python μ_2 is less than corresponding time in matlab μ_1

$$H_0 : \mu_1 \leq \mu_2 \qquad H_a : \mu_1 > \mu_2$$
$$H_0 : \mu_1 - \mu_2 \leq 0 \qquad H_a : \mu_1 - \mu_2 > 0$$

Let us take an $\alpha = 0.05$. Under unknown σ for both the populations, we use the t-statistic. We collect a sample with following characteristics:

$$\begin{aligned} n_1 &= 12 & n_2 &= 12 \\ \bar{x}_1 &= 325 \text{ hours} & \bar{x}_2 &= 286 \text{ hours} \\ s_1 &= 40 & s_2 &= 44 \end{aligned}$$

T-statistic would be the following:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 2.27$$

We can calculate the degree of freedom like before to obtain $df = 21$

For a one tailed test with $\alpha = 0.05$, we get the critical t-value as follows:

$$t_{0.025} = 2.080 < t^* < t_{0.01} = 2.518$$

This implies that the p-value for our test statistic is between 2% and 1%.

Rejection rule would imply that we reject H_0 if the p-value is less than α, which it is in this case.

From this we infer that for a macro model, python takes less time to code than Matlab.

Caution: These tests are robust so that even a smaller sample could do i.e. $n_1 + n_2 \geq 20$. Also, it is advisable to have equal sample sizes, even if they are smaller.

10.3 Inference About The Difference Between Two Population Means: Matched Samples

Uber Example: wants to optimize delivery time from a restaurant to a regular corporate client.

$$\mu_1 : \quad \text{Average time for route 1}$$
$$\mu_2 : \quad \text{Average time for route 2}$$

Nature of Hypothesis: we continue to assume that the time is similar which leads to the following:

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_a : \mu_1 - \mu_2 \neq 0$$

Independent random samples: till now we collect sample n_1 for route 1 and sample n_2 for route 2 with independence being the only requirement.

Limitation of Independent Samples: Even though the two random samples collected separately are independent, there might be variation in the average routes due to other factors which are not captured in the hypothesis.

- For Instance: sample might be biased because when the route time was recorded, the route 1 had lane closures. So those delivery guys with cars took longer than those with bikes.

Idea of Matched Samples: test the times under similar conditions and then compare the differences. This would result in smaller standard error.

- Collect a random sample of delivery guys with bikes
- Delivery guys with bike used for testing both routes - one by one
- Randomization: some delivery guys asked to take route 1 first and then route 2 while others opposite.

Uber Delivery Times: Sample collected shows the following times:

TABLE 10.2 TASK COMPLETION TIMES FOR A MATCHED SAMPLE DESIGN

Worker	Completion Time for Method 1 (minutes)	Completion Time for Method 2 (minutes)	Difference in Completion Times (d_i)
1	6.0	5.4	.6
2	5.0	5.2	−.2
3	7.0	6.5	.5
4	6.2	5.9	.3
5	6.0	6.0	.0
6	6.4	5.8	.6

Hypothesis can then be constructed as

$$H_0 : \mu_d = 0 \text{ rejected if the population mean time differs}$$
$$H_a : \mu_d \neq 0$$

Uber Delivery Times: throws up the following sample statistics: $\bar{d} = \Sigma d_i/n = 0.30$, i.e. 30 minutes.

- The standard deviation is 0.335 which is 33 minutes
- Test of sample size $n = 6$ requires us to assume that population difference follow normal even for using t-tests.

The test statistic is a t-statistic for testing hypothesis in matched samples.

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

- Under $\alpha = 0.05$ and degree of freedom $= 5$, we obtain the test statistic as $t = (0.30 - 0)/(0.335/\sqrt{6}) = 2.20$.

Uber Delivery Times: throws up a t-statistic for which the critical range can be evaluated as:

$$t_{0.10} = 2.015 < t^* < t_{0.05} = 2.571$$

Inference: using the above t-value range in which our calculated t -value, t^* lies tells us that the our p-values would be between 5% to 10%.

Rejection Rule: tells us that $p^* < \alpha$, will allow us to reject the null hypothesis H_0

Uber Delivery Times: we can also compute an interval estimate using our information of t-statistics so computed as follows:

$$\bar{d} \pm t_{0.025} \frac{s_d}{\sqrt{n}} = 0.3 \pm 0.35$$

Inference: There is a difference between average time in the two Uber routes from the local restaurant to the corporate office. It can take as less as 5 minutes to as much as 65 minutes to travel the distance by the same person – due to traffic.

10.4 Inference About Difference Between Two Population Proportions

- Consider two population proportions

$$p_1 : \text{ proportion in populaiton1}$$
$$p_2 : \text{ proportion in populaiton2}$$

- How significant are the difference between these proportions $(p_1 - p_2)$ when chosen from an independent sample of size n_1 and n_2?

Example of Jindal Tax Filing Center: Jindal Tax Center has two branches – Let us call these branches Jindal Branch and OP Branch.

Population Unobservable: p_1: int he total proportion of inaccurate tax filled by Jindal Branch and similarly p_2 is the inaccurate tax filled by the OP Branch.

10.4.1 Interval Estimate on Difference in Population Proportions

Sample Estimates: $\bar{p}_1$ and $\bar{p}_2$ be their respective sample proportions.

- Difference of Interest: is $(\bar{p}_1 - \bar{p}_2)$ which is a point estimator of $(p_1 - p_2)$
- Standard error of $\bar{p},_1 - \bar{p}_2$ is

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Normal approximation of the sampling distribution of $\bar{p}_1 - \bar{p}_2$ is possible if $n_1 p_1 > 5, n_1(1-p_1) > 5$ and similarly for the other sub-sample

Interval Estimates can be written as the point estimate adjusted for margin for error i.e. $(\bar{p}_1 - \bar{p}_2 \pm \text{ margin})$

- When normal, then the margin for error can be written as: $z_{\alpha/2}\sigma_{\bar{p}_1-\bar{p}_2}$
- Without knowledge of (p_1 and p_2) we cannot use $\sigma_{\bar{p}_1-\bar{p}_2}$
- Margin for error could use sample proportions instead, as done before Margin $= z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$

Interval estimate of $\bar{p}_1 - \bar{p}_2$ implies

$$\bar{p}_1 - \bar{p}_2 \pm margin$$

Jindal Tax Center collected two samples from its branches

$$n_1 = 250 \qquad n_2 = 300$$
$$\bar{p}_1 = \frac{35}{250} = 0.14 \qquad \bar{p}_2 = \frac{27}{300} = 0.09$$

Point estimate is the difference between sample proportions i.e. $\bar{p}_1 - \bar{p}_2$

- Margin for error can be calculated based on $\alpha = 0.10$. We obtain $z_{\alpha/2} = 1.645$
- Margin for the point estimate is

$$z_{0.05}\sqrt{\frac{0.14(1-0.14)}{250} + \frac{0.09(1-0.09)}{300}}$$

Jindal Tax Center Interval Constructed: If we obtain difference samples from the two branches, then almost 95% of these point intervals would lie between $[0.005, 0.095]$

10.4.2 Hypothesis Tests on Difference in Population Proportions

Types of Possible Hypothesis and we consider the hypothesized difference to be zero i.e. $D_0 = 0$

$$H_0 : p_1 - p_2 \geq 0 \qquad H_0 : p_1 - p_2 \leq 0 \qquad H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 < 0 \qquad H_a : p_1 - p_2 > 0 \qquad H_a : p_1 - p_2 \neq 0$$

Standard Errors for the test statistic would be similar as in the interval margins

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Under the assumption that null is true, we know this for a fact: $p_1 = p_2 = p$. This simplifies the standard error:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{p(1-p)\left(\frac{1}{n_1}\frac{1}{n_2}\right)}$$

Standard Errors for the test statistic would be similar as in the interval margins

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

How do we obtain even this modified $\sigma_{\bar{p}_1 - \bar{p}_2}$ under the null?

How do we obtain even this modified $\sigma_{\bar{p}_1-\bar{p}_2}$ under the null?

Pooled Point Estimate for proportion is obtained by combining the individual sample point estimates with its population weights to obtain $\bar{p}$ as

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$$

Test statistic can be the z under normality conditions which can be checked as before to obtain

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

From the sample difference obtained, can we conclusively say something about the actual difference between the tax filing errors of the two branches?

$$H_0 : p_1 - p_2 = 0 \qquad H_a : p_1 - p_2 \neq 0$$

If the null is rejected, then we can conclude that the two offices are different, and the managers can be swapped to improve performance.

Sample estimates are as follows: $\bar{p}_1 = 0.14$, $\bar{p}_2 = 0.09$, $n_1 = 250$ and finally $n_2 = 300$

What is its pooled tax estimate?

Jindal Tax Center with its pooled tax estimate is

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = 0.1127$$

Test statistic becomes

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = 1.85$$

Rejection Rule: Consider the $\alpha = 0.10$ under a two tailed test, we obtain the p-value as $p^* = 2(0.0322)$.

- This implies that $p^* < \alpha$ so we can reject the null hypothesis to conclude that the branches differ in the error rates of tax filing.