INTRODUCTION TO STATISTICAL MODELLING

STAT2507D

Chapter 8 - 3

Large Sample Estimation

FALL 2014: STAT 2507D
Carleton University

DIFFERENCE BETWEEN TWO POPULATION MEANS

- Sometimes we are interested in comparing the means of two populations:
  - The average growth of plants fed using two different nutrients
  - The average scores on MCAT for students whose major was biochemistry and those whose major was biology
  - There are two populations for each of these samples

Difference between two Population Means

	Population 1	Population 2	Sample 1	Sample 2
Mean	μ_1	μ_2	$\bar{x}_1$	$\bar{x}_2$
Variance	σ_1^2	σ_2^2	s_1^2	s_2^2
Sample size	-	-	n_1	n_2

DIFFERENCE BETWEEN TWO POPULATION MEANS

- To make this comparison,
  - A random sample of size n_1 drawn from population 1 with mean μ_1 and variance σ_1^2
  - A random sample of size n_2 drawn from population 2 with mean μ_1 and variance σ_2^2

DIFFERENCE BETWEEN TWO POPULATION MEANS

- We compare the two averages by making inferences about ($\mu_1 - \mu_2$), the difference in the two population averages
  - If the two population averages are the same, then $\mu_1 - \mu_2 = 0$
  - The best estimate of $\mu_1 - \mu_2$ is the difference in two sample means $\bar{x}_1 - \bar{x}_2$

Sampling Distributions of

- The mean of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$, and the difference in population means can be estimated by $\bar{x}_1 - \bar{x}_2$
- The standard error of $\bar{x}_1 - \bar{x}_2$, $SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- When the sample sizes are large, the SE can be estimated by $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

SAMPLE DISTRIBUTIONS CONT'D

- If the sample population are normally distributed, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is exactly normally distributed regardless of the sample size
- If the sampled populations are not normally distributed, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normally distributed when n_1 and n_2 are both 30 or more, due to Central Limit Theorem

SAMPLE DISTRIBUTIONS CONT'D

- When n_1 and n_2 are large, $\bar{x}_1 - \bar{x}_2$ is unbiased estimator of $(\mu_1 – \mu_2)$ with approximately normal distribution.
- Sampling distribution of difference of sample mean

$$(\bar{x}_1 - \bar{x}_2) \sim N\left((\mu_1 - \mu_2), \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

SAMPLING DISTRIBUTION CONT'D

- Standardizing

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

$$P\left[(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$
$$= 1 - \alpha$$

ESTIMATING μ$_1$ – μ$_2$

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal (z) distribution.
- Point estimate for μ$_1$ – μ$_2$: $\bar{x}_1 - \bar{x}_2$
- Margin of Error : $\pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

ESTIMATING $\mu_1 - \mu_2$

- Confidence interval for $\mu_1 - \mu_2$:

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

EXERCISE

8.11) Independent random samples are selected from populations 1 and 2. The sample sizes, means, and variances are as follows:

	Population	
	1	2
Sample size	35	49
Sample mean	12.7	7.4
Sample Variance	1.38	4.14

EXERCISE

a) Find the 95% confidence interval for estimating the difference in the population means (μ_1-μ_2).

b) Based on the confidence interval in part (a), can you conclude that there is a difference in the means for the two populations?

EXERCISE

8.12) In an attempt to compare the starting salaries of university graduates majoring in education and social sciences, random samples of 50 recent university graduates in each major were selected and the following information was obtained.

Major	Mean	SD
Education ($)	40,554	2225
Social Sciences ($)	38,348	2375

EXERCISE

a) Find a point estimate for the difference in the average starting salaries of university students majoring in education and social sciences. What is the margin of error for your estimate?

b) Based on the results of part a, do you think that there is a significant difference in the means for the two groups in the general population? Explain

EXERCISES

8.13) The means and standard deviations of 65 males and 65 females are shown below:

	Men	Women
Sample mean	36.7°C	36.98°C
Standard deviation	0.70°C	0.75°C

Find 95% confidence interval for the difference in the average difference in the average body temperatures for males versus females. Based on this interval, can you conclude that there is a difference in the average temperatures for males versus females? Explain.

DIFFERENCE BETWEEN THE BINOMIAL PROPORTIONS

- A simple extension of the estimation of a binomial proportion p is the estimation of the difference between two binomial proportions.
- Sometimes we are interested in comparing the proportion of “successes” in two binomial populations:
  - Proportion of defective items manufactured in two production lines
  - The germination rates of untreated seeds and seeds treated with a fungicide.

DIFFERENCE BETWEEN THE BINOMIAL PROPORTIONS

- To make this comparison:
- Independent random samples are drawn from populations 1 and 2.
- A random sample of size n_1 drawn from binomial population 1 with parameter p_1.
- A random sample of size n_2 drawn from binomial population 2 with parameter p_2

Difference between the Binomial Proportions

- The sample estimates $\hat{p}_1$ and $\hat{p}_2$ are calculated.
- The unbiased estimator of the difference (p_1 – p_2) is the sample difference $\hat{p}_1 - \hat{p}_2$

DIFFERENCE BETWEEN THE BINOMIAL PROPORTIONS

- We compare the two proportions by making inferences about $p_1 – p_2$, the difference in the two population proportions
  - If the two population proportions are the same, then p_1- p_2 = 0
  - The best estimates of p_1- p_2 is the difference in the two sample proportions $\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$

PROPERTIES

- The sampling distribution of the difference between sample proportions has these properties: $\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$

1. $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
2. SE = $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$; $\widehat{SE} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$
3. The sampling distribution of $\hat{p}_1 - \hat{p}_2$ can be approximated by a normal distribution when n_1 and n_2 are large, due to CLT

REMARKS

- The range of single proportion (p_1 or p_2) is from 0 to 1, the difference between two proportions ranges from -1 to 1.
- To use the normal distribution to approximate the distribution of $\hat{p}_1 - \hat{p}_2$ both $\hat{p}_1$ and $\hat{p}_2$ should be approximately normal; that is $n_1\hat{p}_1 > 5, n_1\hat{q}_1 > 5$ and $n_2\hat{p}_2 > 5, n_2\hat{q}_2 > 5$

DIFFERENCE OF PROPORTIONS

- Large sample point estimation of (p_1- p_2)
  - Point Estimator: $\hat{p}_1 - \hat{p}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$
  - (1-α)100% Margin of Error: $Z_{\frac{\alpha}{2}} * (SE) =$

$$Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

DIFFERENCE OF PROPORTIONS

- A (1-α)100% large sample confidence interval for $(p_1 - p_2)$: $(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$
- Assumption: $n_1\hat{p}_1 > 5, n_1\hat{q}_1 > 5$ and $n_2\hat{p}_2 > 5, n_2\hat{q}_2 > 5$

EXERCISES

8.14) Independent random samples of n_1= 800 and n_2=640 observations were selected from binomial populations 1 and 2, and x_1=337 and x_2 = 374 successes were observed.

a) What is the best point estimator for the difference (p_1-p_2) in the two binomial proportions?

b) Calculate the approximate standard error for the statistic used in part a.

EXERCISE

c) What is the margin of error for this point estimate?

d) Find a 90% confidence interval for the difference ($p_1 – p_2$) in the two population proportions. Interpret the interval.

e) What assumptions must you make for the confidence interval to be valid? Are these assumptions met?

EXERCISE

8.15) In a study to compare the effects of two pain relievers, it was found that of n_1= 200 randomly selected individuals instructed to use the first pain reliever, 93% indicated that it relieved their pain. Of n_2 = 450 randomly selected individuals instructed to use the second pain reliever, 96% indicated that it relieved their pain.

EXERCISE

a) Find a 99% confidence interval for the difference in the proportions experiencing relief from pain for these two pain relievers

b) Based on the confidence interval in part a, is there sufficient evidence to indicate a difference in the proportions experiencing relief for the two pain relievers? Explain.