Two Samples

MATH 2441, BCIT

Technical Mathematics for Food Technology

April 11, 2018

Two Proportions

For the first population, let p_1 be the population proportion. The sample proportion is $\hat{p}_1$, which is often calculated by $\hat{p}_1 = x_1/n_1$, where n_1 is the sample size and x_1 is the number of successes. As usual, $\hat{q}_1 = 1 - \hat{p}_1$.

The corresponding notations $p_2, n_2, x_2, \hat{p}_2, \hat{q}_2$ apply to the second population.

The pooled sample population is denoted by $\bar{p}$ and calculated using

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{1}$$

Not surprisingly, $\bar{q} = 1 - \bar{p}$.

Two Proportions Test Statistic

In order to use the following test statistic, two requirements need to be met:

1. There are two SRS (simple random samples) which are independent—they must not be related or paired with each other.
2. For each of the two samples, there are at least five successes and at least five failures. This is equivalent to $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

The test statistic z has a standard normal distribution if the null hypothesis is true and $p_1 = p_2$.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \tag{2}$$

Two Proportions Confidence Interval

The confidence interval estimate of the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E \tag{3}$$

where the margin of error E is given by

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \tag{4}$$

Exercise

Exercise 1: 89 undergraduate business students from two different colleges were randomly assigned to two different groups. In the "dollar bill" group, 46 subjects were given dollar bills. The "quarter" group consisted of 43 subjects given quarters. Then the two groups were herded through a candy store. Is there a significant difference between the two spending patterns? Use a significance level of $\alpha = 0.05$.

	Group 1	Group 2
	dollar bill	quarter
spent the money	12	27
number of subjects	46	43

Exercise

Exercise 2: In the largest clinical trial ever conducted, 401,974 children were randomly assigned to two groups. The treatment group consisted of 201,229 children given the Salk vaccine for polio, and the other 200,745 were given a placebo. Among those in the treatment group, 33 developed polio, and among those in the placebo group, 115 developed polio. Claim the hypothesis that the Salk vaccine protected children from polio. Use a significance level of 0.05. Then construct a confidence interval with a confidence level of 95%.

Exercise

Exercise 3: Test the claim that the rate of left-handedness among males is less than the rate of left-handedness among females at a 0.01 significance level, given the following data from a sample:

	Males	Females
left	23	65
right	217	455

Two Means: Independent Samples

Here is some notation.

population 1	**population 2**	**meaning of notation**
μ_1	μ_2	population mean
σ_1	σ_2	population standard deviation
n_1	n_2	sample size
$\bar{x}_1$	$\bar{x}_2$	sample mean
s_1	s_2	sample standard deviation

Two Means: Independent Samples

Here are some requirements.

1. The values of σ_1 and σ_2 are unknown, and we do not assume that they are equal.
2. The two samples are independent.
3. Both samples are simple random samples.
4. Either or both of these conditions is satisfied:
   (i) The two sample sizes are both large ($n_1 > 30$ and $n_2 > 30$).
   (ii) Both samples come from a normally distributed population.

Two Means: Independent Samples

If the requirements are met, then the following score has a Student-t distribution:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{5}$$

$\mu_1 - \mu_2$ is often assumed to be 0 by the null hypothesis. There are two methods for determining the degree of freedom. We will use the simpler and more conservative method

$$df = \min\{n_1 - 1, n_2 - 1\} \tag{6}$$

Statistics software sometimes uses a less conservative but more accurate degree of freedom following a complicated formula.

Two Means: Independent Samples

The confidence interval estimate of the difference $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \tag{7}$$

with

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{8}$$

Exercise

Exercise 4: Researchers at UBC conducted trials to investigate the effects of colour on creativity. Subjects were given creative tasks when the background of the room was either red or blue. The researchers made the claim "blue enhances performance on a creative task." Test that claim using a 0.01 significance level.

Creativity Scores			
Red Background	$n = 35$	$\bar{x} = 3.39$	$s = 0.97$
Blue Background	$n = 36$	$\bar{x} = 3.97$	$s = 0.63$

Two Means: Independent Samples

If the population variances are known, the procedure changes as follows. The test statistic is now

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{9}$$

and it has a standard normal distribution. The confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E \tag{10}$$

with

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{11}$$

Two Means: Independent Samples

If the population variances (unknown) are assumed to be equal, the procedure changes as follows. Use a degree of freedom $df = n_1 + n_2 - 2$ and define the pooled sample variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{12}$$

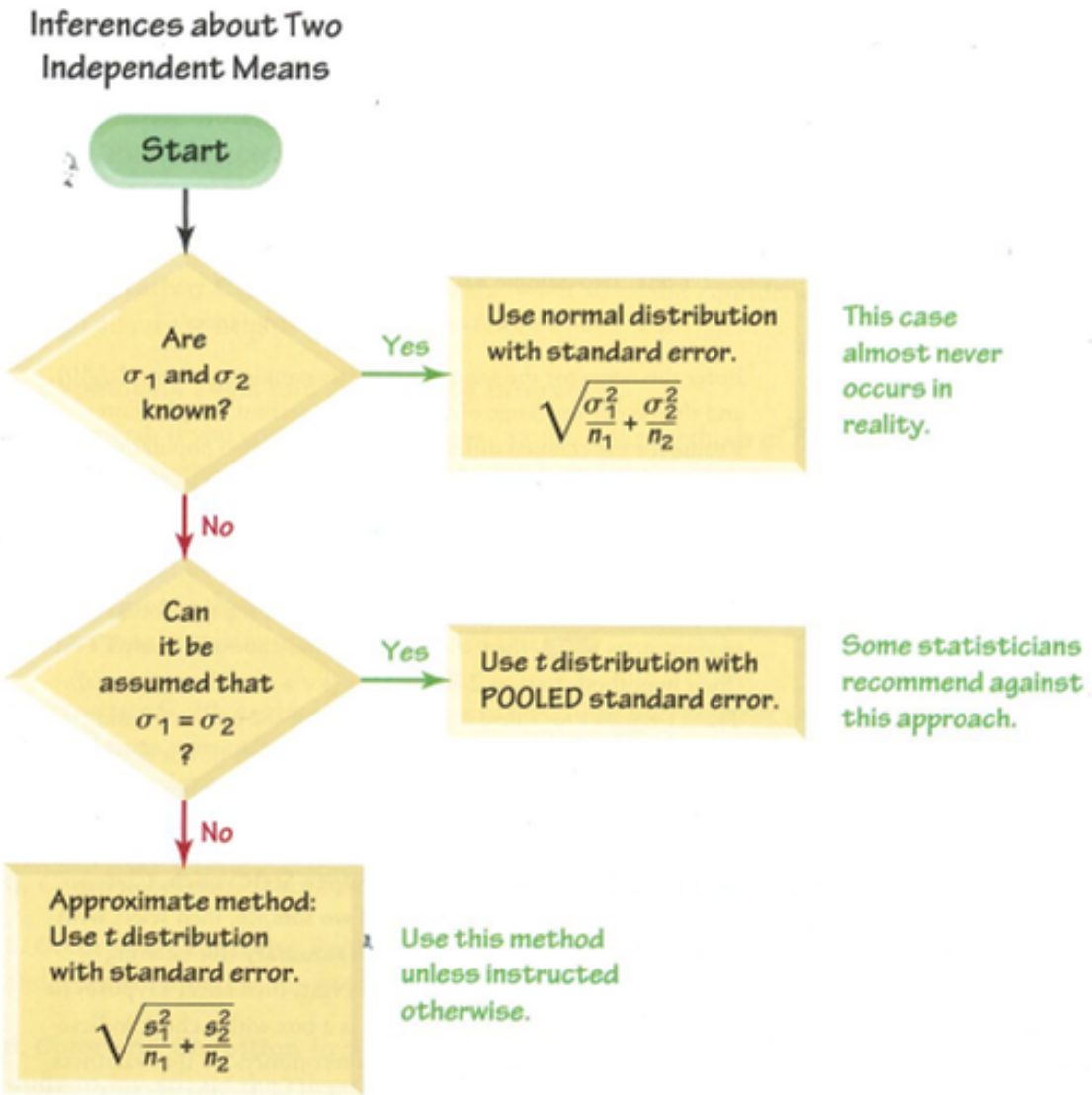
Then the test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{13}$$

has a Student-t distribution. The confidence interval is as in (10) with an error

$$E = t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \tag{14}$$

Two Means: Independent Samples



Two Means: Dependent Samples (Matched Pairs)

Paired data is usually more informative than independent samples. Here is some notation.

notation	**meaning of notation**
d	individual difference for pair
μ_d	population mean for differences
$\bar{d}$	sample mean for differences
s_d	sample standard deviation for differences
n	number of pairs

Two Means: Dependent Samples (Matched Pairs)

Here are some requirements.

1. The sample data are dependent (matched pairs).
2. The samples are simple random samples.
3. Either or both of these conditions is satisfied:
   - (i) The number of pairs of sample data is large ($n > 30$).
   - (ii) The pairs of values have differences that are from a normally distributed population.

Two Means: Dependent Samples (Matched Pairs)

Use degree of freedom $df = n - 1$. Then the following score

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \tag{15}$$

is distributed according to a Student-t distribution. The confidence interval is

$$\bar{d} - E < \mu_d < \bar{d} + E \tag{16}$$

with

$$E = t_{\alpha/2} \frac{s_d}{\sqrt{n}} \tag{17}$$

Exercises

Exercise 5: Listed below are brain volumes (cm^3) of twins. Construct a 99% confidence interval estimate of the mean of the differences between volumes for the first-born and the second-born twins. What does the confidence interval suggest?

First Born	1005	1035	1281	1051	1034	1079	1104	1439	1029	1160
Second Born	963	1027	1272	1079	1070	1173	1067	1347	1100	1204

Exercise 6: In a study of proctored and nonproctored tests in an online Intermediate Algebra course, researchers obtained the data for test results given below. Use a 0.01 significance level to test the claim that students taking nonproctored tests get a higher mean than those taking proctored tests.

Group 1 (proctored)	$n = 30$	$\bar{x} = 74.30$	$s = 12.87$
Group 1 (nonproctored)	$n = 32$	$\bar{x} = 88.62$	$s = 22.09$

Exercises

Exercise 7: A study was conducted to determine the proportion of people who dream in black and white instead of color. Among 306 people over the age of 55, 68 dream in black and white, and among 298 people under the age of 25, 13 dream in black and white (based on data from "Do We Dream in Color?" by Eva Murzyn, Consciousness and Cognition, Vol. 17, No. 4). We want to use a 0.01 significance level to test the claim that the proportion of people over 55 who dream in black and white is greater than the proportion for those under 25. Test the claim using a hypothesis test.

Exercises

Exercise 8: A study investigated survival rates for in-hospital patients who suffered cardiac arrest. Among 58,593 patients who had cardiac arrest during the day, 11,604 survived and were discharged. Among 28,155 patients who suffered cardiac arrest at night, 4139 survived and were discharged. Use a 0.01 significance level to test the claim that the survival rates are the same for day and night.

Exercises

Exercise 9: A data set lists full IQ scores for a random sample of subjects with low lead levels in their blood and another random sample of subjects with high lead levels in their blood. The statistics are summarized below. Use a 0.05 significance level to test the claim that the mean IQ score of people with low lead levels is higher than the mean IQ score of people with high lead levels.

Low Lead Level	$n = 78$	$\bar{x} = 92.88462$	$s = 15.34451$
High Lead Level	$n = 21$	$\bar{x} = 86.90476$	$s = 8.988352$

Exercises

Exercise 10: The herb ginkgo biloba is commonly used as a treatment to prevent dementia. In a study of the effectiveness of this treatment, 1545 elderly subjects were given ginkgo and 1524 elderly subjects were given a placebo. Among those in the ginkgo treatment group, 246 later developed dementia, and among those in the placebo group, 277 later developed dementia (based on data from "Ginkgo Biloba for Prevention of Dementia," by DeKosky et. al., Journal of the American Medical Association, Vol. 300, No. 19). We want to use a 0.01 significance level to test the claim that ginkgo is effective in preventing dementia.

1. Test the claim using a hypothesis test.
2. Test the claim by constructing an appropriate confidence interval.

Based on the results, is ginkgo effective in preventing dementia?

Exercises

Exercise 11: A simple random sample of front-seat occupants involved in car crashes is obtained. Among 2823 occupants not wearing scat belts, 31 were killed. Among 7765 occupants wearing seat belts, 16 were killed (based on data from "Who Wants Airbags?" by Meyer and Finney, Chance, Vol. 18, No. 2). We want to use a 0.05 significance level to test the claim that seat belts are effective in reducing fatalities. Test the claim using a hypothesis test. We want to use a 0.01 significance level to test the claim that the survival rates are the same for day and night.

Exercise 12: We know that the mean weight of men is greater than the mean weight of women, and the mean height of men is greater than the mean height of women. A person's body mass index (BMI) is computed by dividing weight (kg) by the square of height (m). Given below are the BMI statistics for random samples of males and females. Use a 0.05 significance level to test the claim that males and females have the same mean BMI.

Male BMI	$n = 40$	$\bar{x} = 28.44075$	$s = 7.394076$
Female BMI	$n = 40$	$\bar{x} = 26.6005$	$s = 5.359442$

Exercises

Exercise 13: The accompanying table gives results from a study of the words spoken in a day by men and women. Use a 0.01 significance level to test the claim that the mean number of words spoken in a day by men is less than that for women.

Men	**Women**
$n_1 = 186$	$n_2 = 210$
$\bar{x}_1$	$\bar{x}_2$
$s_1 = 8632.5$	$s_2 = 7301.2$

Exercises

Exercise 14: Listed below are body temperatures of four subjects measured at two different times in a day.

Body Temperature (°F) at 8 a.m.	98	97.0	98.6	97.4
Body Temperature (°F) at 12 p.m.	98	97.6	98.8	98.0

Use the sample data to test the claim that there is no difference between body temperatures measured at 8 a.m. and at 12 p.m. Use a 0.05 significance level.

Exercises

Exercise 15: Listed below are the numbers of years that popes and British monarchs (since 1690) lived after their election or coronation. Treat the values as simple random samples from a larger population. Use a 0.01 significance level to test the claim that the mean longevity for popes is less than the mean for British monarchs after coronation.

```
Popes: 2 9 21 3 6 10 18 11 6 25 23
 6 2 15 32 25 11 8 17 19 5 15 0 26
```

```
Kings and Queens: 17 6 13 12 13 33
                  59 10 7 63 9 25 36 15
```

Exercises

Exercise 16: Listed on the next slide are the numbers of words spoken in a day by each member of six different couples. Use a 0.05 significance level to test the claim that among couples, males speak more words in a day than females. The mean sample difference (words of males minus words of matched females) is -1867.107. The standard deviation of the differences for the paired sample data is 8955.155. There are 56 pairs.

Exercises

27531	20737	19153	6017	13560	21261	18821	17646
15684	24625	1411	18338	18876	12964	14069	16255
5638	5198	20242	23020	13825	33789	16072	28838
27997	18712	10117	18602	9274	8709	16414	38154
25433	12002	20206	16518	20547	10508	19017	25510
8077	15702	16874	13770	17190	11909	37649	34869
21319	11661	16135	29940	10578	29730	17427	24480
17572	19624	20734	8419	14821	20981	46978	31553
26429	13397	7771	17791	15477	16937	25835	18667
21966	18776	6792	5596	10483	19049	10302	7059
11680	15863	26194	11467	19377	20224	15686	25168
10818	12549	10671	18372	11767	15872	10072	16143
12650	17014	13462	13657	13793	18717	6885	14730
21683	23511	12474	21420	5908	12685	20848	28117

End of Lesson

Next Lesson: Linear Regression