Workshop 8 - Sample answers

STATISTICS FOR BUSINESS AND FINANCE (BUS5SBF)

Topics: Hypothesis Testing: Comparison of Two Population Means and Variances

Objective:

To make inference about difference between unknown population means both in small and large samples using step-by-step hypothesis testing approach and confidence interval estimation.

Learning Outcomes

By the end of this workshop, you would be able to;

- Perform hypothesis testing to compare two population means
- Distinguish between independent samples and paired observations
- Identify the appropriate test statistic in case of small and large samples
- Compute the p-value of the tests
- Infer about population parameter using confidence intervals
- Apply these concepts to real world problems

Task 1: Is monthly rent on residential properties greatly different in two suburbs?

Rental rates on residential properties have significantly increased in Australia in past few years, whilst it is also observed that rental rates vary greatly across suburbs in a same region. Najam, who is looking for a rental property in either of the two Western suburbs of Altona and Sanctuary Lakes, collects information about monthly rental rates from realestate.com.au. He conjectures that the average monthly rental rate is higher in Sanctuary Lakes than in Altona by $250. To test this belief, he collects monthly rental rates of 25 properties in both suburbs and the data is available in Excel sheet *Monthly Rent* in file Workshop 8.xlsx.

Is there any statistical evidence for Najam's conjecture of differences in average rental rates in those suburbs more than $250? What assumptions are required to use an appropriate test statistic? Use α =0.10.

Solution:

First, we use F-test to determine whether the two populations have same variances. The outcome of the F-test indicates that there is a weak statistical evidence (0.10 > p-value > 0.05) to reject the null hypothesis of equal population variances at the 10% significance level. Hence we use t-test assuming unequal variances.

Table 1: F-Test Two-Sample for Variances

	Santuary Lakes	*Altona*
Mean	1699.84	1353.16
Variance	24763.3067	11432.55667
Observations	25	25
df	24	24
F	**2.1660**	
P(F<=f) one-tail	**0.0321**	
F Critical one-tail	1.702	
Critical values for two-sided test		
Lefthand side critical F	**0.5041**	=F.INV(0.05,24,24)
Righthand side crtitical F	**1.9838**	=F.INV(0.95,24,24)
p-value	0.0641	=2*F.DIST.RT(2.166,24,24)

NB: Excel Data Analysis gives an output for **one-sided test**. That is, **H_0: Var(Sanctuary Lakes) <= Var(Altona) vs H_A: Var(Sanctuary Lakes) > Var(Altona)**.
The decision rule is, **Reject the H_0 if the computed F > the one-tail critical F value**.
Pr(F <= critical F) gives the p-value, i.e. the probability that the H_0 is true.

One-sided test may be reasonable here, as Var(Sanctuary Lakes) is clearly larger than Var(Altona). But, quite often, two variances are not too different from each other. In such a case, we should conduct **two-sided test**, i.e.
H_0: Var(Sanctuary Lakes) = Var(Altona)
H_A: Var(Sanctuary Lakes) not equal to Var(Altona)
In this case, the p-value and critical F value in the Excel output aren't applicable. We need to get the lower and upper critical F values, and compute the p-value.
Reject the H0 if the computed F < Lefthand side critical F
or the computed F > Righthand side critical F.
Remember **Leftthand side critical F = 1/Righthand side critical F**.

t-Test: Two-Sample Assuming Unequal Variances		
	Santuary Lakes	***Altona***
Mean	1699.84	1353.16
Variance	24763.31	11432.55667
Observations	25	25
Hypothesized Mean Difference	250	
df	**42**	
t Stat	**2.540839**	
P(T<=t) one-tail	**0.007421**	
t Critical one-tail	**1.302**	
P(T<=t) two-tail	0.014842	
t Critical two-tail	1.682	

Kurtosis	-0.0058	0.0116
Skewness	0.4397	0.5589

This test requires the assumption of both samples taken from normally distributed populations that are independent. This assumption is supported by low skewness and excess kurtosis close to zero in each case.

1. $H_0 : \mu_{SL} - \mu_A <= 250$ *vs.* $H_1 : \mu_{SL} - \mu_A > 250$ (right-tailed test)

2. $\alpha = 0.1$

3. $t - test\ with\ unequal\ \mathrm{var}iances$

$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad with\ df = 42$$

$$Since\ \ df > 30, t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim_{app} N(0,1)$$

4. Reject $H_0\ if\ \ t_{obs} > t_{(0.1,42\ )} = 1.302$ (as given in Excel output)

$or\ \ if\ \ \ p - value <$ 0.10

$t_{obs} = 2.54 \qquad p - val = 0.0074$

Because observed value of the test statistic falls in the rejection region, there is sufficient statistical evidence to infer that the difference between average monthy rents in Sanctuary Lakes and Altona is higher than $250.

Task 2: Will new-tech design tyres last longer than the existing design tyres?

To determine whether new steel-belted radial tyres last longer than a current model, the manufacturer conducts the following experiment.

A pair of tyres of each type were installed on the rear wheel of 20 randomly-selected cars. Each car was sampled twice, thus creating a pair of observations. Drivers drive in their usual way until the tyres are worn out. The number of kilometers (in thousands) until wear out was recorded as provided in the Excel sheet *Tyres*.

a. **Can the manufacturer infer that the new-tech design tyre will last longer on average than their existing model? Assume that the difference in kilometers of tyre lifetimes are normal.**

   This is the case of t-test with dependent samples.

1. $H_0 : \mu_d <= 0 \quad vs.\ H_1 : \mu_d > 0$
2. $\alpha = 0.5$
3. $t = \dfrac{\bar{d} - \mu_d}{s_d / root\_n} \qquad v = n\text{-}1 = 19$

   $t_{obs} = 2.31 \;\; p - val = 0.016$
4. Reject H_0 $if \;\; p - value < 0.05$

5. Conclusion: Since p-value is less than 5%, we reject the null and hence conclude that there is sufficient statistical evidence that the new-tech design tyre will last longer than the existing design tyre.

t-Test: Paired Two Sample for Means

	New Model	*Current Model*
Mean	74.7	70.35
Variance	236.2211	235.3974
Observations	20	20
Pearson Correlation	0.850176	
Hypothesized Mean Difference	**0**	
df	19	
t Stat	**2.31428**	
P(T<=t) one-tail	**0.016001**	
t Critical one-tail	**1.729**133	
P(T<=t) two-tail	0.032001	
t Critical two-tail	2.093024	

b. **Construct the 95% confidence interval for the mean difference of kilometers travelled by cars.**

$$95\% CI: \left[\bar{d} - \frac{s_d}{\sqrt{n}} t_{(0.025,19)},\ \bar{d} + \frac{s_d}{\sqrt{n}} t_{(0.025,19)}\right] = \left[4.35 - \frac{8.41}{\sqrt{20}} 2.09, 4.35 + \frac{8.41}{\sqrt{20}} 2.09\right] = [0.415, 8.28]$$

We are 95% confident that the true population mean of differences will lie between 1000x[0.415, 8.28] = [415km, 8284km].

c. **Using the above confidence interval, can we reject the null hypothesis that the mean difference in kilometers travelled by paired cars is 8000km?**

   Since the interval includes 8000, we cannot reject the null hypothesis. Hence, we maintain the hypothesis that the mean difference in kilometers travelled by cars with existing and new-tech tyres is 8000km.

Task 3: Are female students outperforming male students in BUS5SBF?

Dr. Ishaq is worried about poor performance of his male students compared to the female students in Statistics. He believes that this is mainly because the average hours spent on doing part-time work each week by male students is higher than the female students. To test his belief, Dr. Ishaq takes a random sample of 14 male and 12 female students and collects information about number of hours worked last week. The data is available in sheet Hours Worked in Excel file Workshop 8.xlsx.

Test Dr. Ishaq's claim at 5 % level of significance. Specify the test statistic and any underlying assumptions required to use that statistic.

[Example] * Do F-test for the equality of variances from two samples first! You'll find you cannot reject the H_0.

1. $H_0: \mu_M <= \mu_F$ vs. $H_1: \mu_M > \mu_F$ (right-tailed test)

2. $\alpha = 0.05$

3. $t-test$ with equal variances

Assuming normally distributed data of hours worked by male and female students with df < 30, the appropriate test statisitc is

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \quad with\ df = 24 \text{(and has Student t-distribution as n < 30)}$$

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

4. Reject H_0 if $t_{obs} > t_{(0.05,24)} = 1.71$ or if $p-value < 0.05$

$t_{obs} = 2.064$ $\quad p-val = 0.016$

5. Conclusion

Since p-values is less than 0.05, there is sufficient statistical evidence to infer that average hours per week by male students are greater than those by female students, but this analysis does not tell us whether doing more hours of part-time work is the cause of poor performance.

t-Test: Two-Sample Assuming Equal Variances

	Males	***Females***
Mean	28.92857143	22.83333333
Variance	52.07142857	39.24242424
Observations	14	12
Pooled Variance		46.19146825
Hypothesized Mean Difference	0	
df	**24**	
t Stat		**2.27970031**
P(T<=t) one-tail		**0.015902732**
t Critical one-tail		**1.71088208**
P(T<=t) two-tail		0.031805465
t Critical two-tail		2.063898562

Task 4: Is there a difference in prices of jeans by Levis and Lee?

The prices of two famous brands of jeans, Levis (L) and Lee at various retail outlets in Australia are normally distributed such that $X \sim N(\mu_L, \sigma_L^2)$ and $Y \sim N(\mu_{Lee}, \sigma_{Lee}^2)$. What will be the sampling distribution of means of X and Y variables?

What will be the sampling distribution of means of X and Y variables?

$\bar{X} \sim N(\mu_L, \sigma_L^2/n_L)$ and $\bar{Y} \sim N(\mu_{Lee}, \sigma_{Lee}^2/n_{Lee})$

a) A random sample of 50 Levis and 40 Lee jeans was taken with the following information. Test the hypothesis that the average price of Levis jeans is greater than the average price of Lee jeans at 5%significance level.

	Levis	Lee
n	50	40
$\bar{X}$	95	85
s	14	19

1. $H_0 : \mu_L <= \mu_{Lee}$ vs. $H_1 : \mu_L > \mu_{Lee}$ $(right-tailed test)$

2. $\alpha = 0.05$

3. $t-test$ with unequal variances since the sample size is large enough. (pooling won't matter here)

$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$Since\ df > 30, t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim_{app} N(0,1)$$

4. Reject H_0 if $t_{obs} > z_\alpha = 1.645$ (in case you use normal-distribution as approximation)

$or\ \ if\ \ p-value < 0.05$

$t_{obs} = 2.77 \qquad p-val = 0.003$

5. We reject the null hypothesis since there is an overwhelming evidence of larger average price of Levis than of Lee.

b. Using information in part (a), can we infer that the mean price of Levis jeans is $10 more than the average price of Lee jeans?

1. $H_0 : \mu_L - \mu_{Lee} <= 10$ vs. $H_1 : \mu_L - \mu_{Lee} > 10$ (right-tailed test)

2. $\alpha = 5\%$

3. $t-test$ with unequal variances since the sample size is large enough. (pooling won't matter here)

$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$Since\ df > 30, t = \frac{\overline{X}_1 - \overline{X}_2 - 10}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim_{app} N(0,1)$$

4. Reject H_0 if $t_{obs} > z_\alpha = 1.645$ (in case you use normal-distribution as approximation)

5. We cannot reject the null hypothesis since $t_{obs} = 0 < 1.645$.

The End