MIS771 Descriptive Analytics and Visualisation

Topic 4 Tutorial – (Brief Solution)

Introduction

This tutorial consists of two parts.

Part A: In the previous tutorial, we used one way ANOVA to test the differences between the means of several groups. In this tutorial, we extended the analysis of variance to the two-factor factorial design, in which two factors are simultaneously evaluated. In this week's tutorial, we learn how to apply the two-way analysis of variance (ANOVA) and interpret the results of the analysis.

Part B: In previous tutorials, we used the hypothesis-testing procedure to analyse both numerical and categorical data in two or more samples/groups. This tutorial extends the hypothesis testing to analyse the difference between population proportions in two or more populations using chi square test.

Similar to the previous week, we will complete most of the work using Excel and we need to organise the data for analysis in a particular way.

Scenario

Conrobar is a manufacturing company with over 3,000 employees. Management is concerned about the wide variation in productivity between employees. The company has collected data on these and related issues.

PART - A

Open the data file and install the Data Analysis Tool Pak

a) Download the file ***ConrobarT4.xls*** from Cloud Deakin. Please **save it** to the hard drive.

b) Open the file in Excel.

c) Install the Data Analysis Toolpak. [See the previous tutorial for instructions].

Q1. TWO-WAY ANOVA

(a) What are the advantages of Two-way ANOVA?

In the one-way ANOVA, we classify populations according to one categorical variable/factor. In Two-way ANOVA, there are two factors, each with several levels. When we are interested in the effects of two factors, a two-way design offer the following advantages:

- It is more efficient to study two factors simultaneously rather than separately.
- We can reduce the residual variation in a model by including a second factor to influence the response.
- We can investigate the interaction between factors.

(b) For each of the following statements, explain what is wrong and why.

i. The two-way ANOVA is used when there are two dependent (outcome) variables.

<<A two-way ANOVA is used when there are two factors (explanatory variables), not outcomes.>>

ii. In a 2 X 3 ANOVA, each level of factor A appears with only two levels of Factor B.

<<Each level of Factor A appears with each level of Factor B.>>

iii. You can perform a two-way ANOVA only when the sample sizes are the same in each cell.

<<The sample size in each cell can be different>>

(c) What assumptions do we make about data when using ANOVA?

- We have an independent random sample
- The independent variable should be approximately normally distributed for each of the combination of the groups of the two independent variables.
- Homogeneity of variances for each combination of the groups of the two independent variables.

Q2. Job satisfaction, Gender and Departments

The management team would like to broaden the scope of the analysis to compare the employee job satisfaction from different Departments and Gender of the employee.

(a) What is the dependent "response" variable in this scenario?

Response variable = "Job Satisfaction"

(b) Identify the "factor" variables and describe the different factors/groups.

Factor variables and levels = Gender: Male and Female; Department: Admin, Prod'n and Distr'n

(c) Write the null and alternative hypothesis for this scenario using both notation and words for the three distinct statistical tests.

1. To test the hypothesis of no difference due to factor A:

$H_0: \mu_1 = \mu_2 = \cdots \mu_r$ $\quad$ $where\ 1 = Male, 2 = Female$

$H_1: Not\ all\ \mu_i\ are\ equal$ $\quad$ $where\ i = 1,2$

H_0: Mean Job satisfaction between male and female employees is no different

H_1: Mean Job satisfaction between male and female employees is different.

2. To test the hypothesis of no difference due to factor B:

$H_0: \mu_1 = \mu_2 = \mu_3$ $\quad$ $where\ 1 = Admin, 2 = Prod'n, 3 = Dist'n$

$H_1: Not\ all\ \mu_j\ are\ equal$ $\quad$ $where\ j = 1,2,3$

H_0: Mean Jobs satisfaction of employees between the three departments is no different

H_1: At least one of the departments differ in terms of mean job satisfaction.

3. To test the hypothesis of no interaction between A and B:

H_0: The interaction of A and B is equal to Zero

H_1: The interaction of A and B is not equal to Zero

H_0: There is no interaction between gender and department

H_1: There is an interaction between gender and department

(d) Perform a Two-way ANOVA test using Excel (Use α= 5%) – See Appendix-1 for steps.

(e) Based on the computer output created in (d), briefly advise Conrobar management about the findings.

Anova: Two-Factor With Replication						
SUMMARY	Admin	Dist'n	Prod'n	Total		
Male						
Count	7	7	7	21		
Sum	106	100	105	311		
Average	15.14286	14.28571	15	14.80952		
Variance	5.809524	7.904762	4	5.461905		
Female						
Count	7	7	7	21		
Sum	76	97	77	250		
Average	10.85714	13.85714	11	11.90476		
Variance	2.142857	5.142857	2	4.790476		
Total						
Count	14	14	14			
Sum	182	197	182			
Average	13	14.07143	13			
Variance	8.615385	6.071429	7.076923			
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	88.59524	1	88.59524	19.68783	0.00	4.113165
Columns	10.71429	2	5.357143	1.190476	0.32	3.259446
Interaction	32.33333	2	16.16667	3.592593	0.04	3.259446
Within	162	36	4.5			
Total	293.6429	41				

To understand the main effect – constructing marginal means (we can describe the main effects by the differences between marginal means.

	Admin	Dist'n	Prod'n	Total
Male	15.14	14.28	15	14.8
Female	10.85	13.85	11	11.9
Total	13	14.07	13	

Main effect of Gender:
Do the marginal means of 14.8 (Male) and 11.9 (female) differ?
The decision rule: if the p-value < 0.05 Reject the H_0.
Because the p-value is 0.00 which is less than the α =0.05. We reject the H_0
There is sufficient evidence to conclude that mean job satisfaction is different between male and female employees.

Main effect of Department:
Do the marginal means of 13 (Admin), 14.07 (Dist'n) and 13 (Prod'n) differ?
The decision rule: if the p-value < 0.05 Reject the H_0.
Because the p-value is 0.32 which is greater than α =0.05. We do not reject the H_0
There is insufficient evidence to conclude that the mean job satisfaction of employees between departments differs.

Interaction effect:

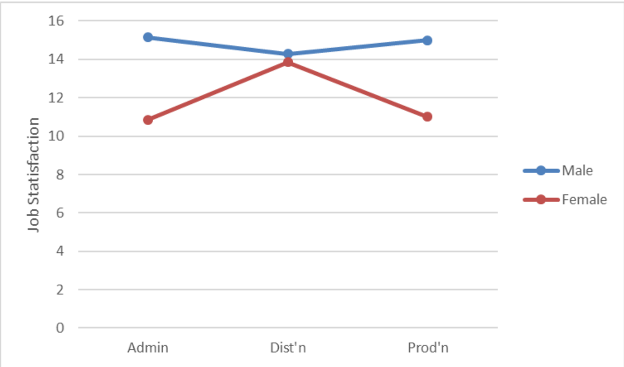
The decision rule: if the p-value < 0.05 Reject the H_0.

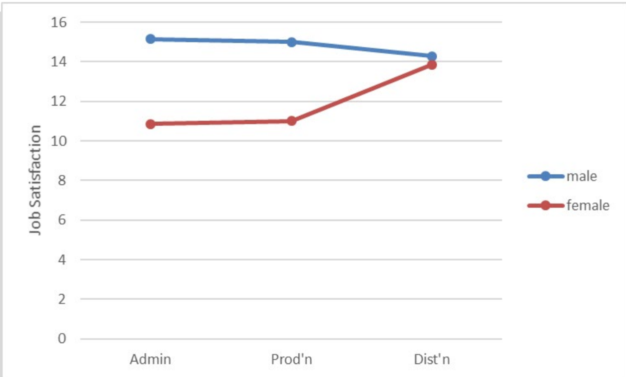
Because the p-value is 0.04 which is less than the α =0.05. We reject the H_0

We have sufficient evidence to conclude that there is an interaction between Gender and Department.

(f) To understand the interaction effect better, draw a cell means plot See Appendix-2 for steps.

For better visualisations changed the order☺





(g) Based on the computer output created in (f), now refine your response to Conrobar management.

For Admin and Production departments, Male employees rated higher job satisfaction but for the distribution department, the job satisfaction for males and females were not different.

Q3. Unpaid Overtime, Gender and Job Security

The effect of the employees' gender and their perception of Job security on unpaid overtime were being studied.

<<To interpret the results of our experiment, start by testing whether there is an interaction effect between Gender and Department. If the interaction effect is significant, further analysis will only refer to this interaction.

If the interaction effect is not significant, we can focus on the main effects – potential differences in Gender and potential differences in Job security>>

(a) Is there a significant interaction between Gender and Job Security?

Anova: Two-Factor With Replication						
SUMMARY	Insecure	Secure	Very Secur	Total		
Female						
Count	7	7	7	21		
Sum	45.2	50.1	47.6	142.9		
Average	6.457143	7.157143	6.8	6.804762		
Variance	10.77952	5.342857	10.66333	8.121476		
Male						
Count	7	7	7	21		
Sum	63.5	48.5	39.5	151.5		
Average	9.071429	6.928571	5.642857	7.214286		
Variance	2.395714	7.322381	5.559524	6.683286		
Total						
Count	14	14	14			
Sum	108.7	98.6	87.1			
Average	7.764286	7.042857	6.221429			
Variance	7.920934	5.85956	7.847967			
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	1.760952	1	1.760952	0.251186	0.62	4.113165
Columns	16.68619	2	8.343095	1.190076	0.32	3.259446
Interaction	27.02905	2	13.51452	1.927739	0.16	3.259446
Within	252.38	36	7.010556			
Total	297.8562	41				

Interaction effect:
The decision rule: if the p-value < 0.05 Reject the $H_{0.}$
Because the p-value is 0.16 which is greater than the α =0.05. We do not reject the H_0
We have insufficient evidence to conclude that there is an interaction between Gender and Job Security.

(b) Is there an effect due to Job Security?

To understand main effect – constructing marginal means (we can describe the main effects by the differences between marginal means.

	Insecure	Secure	Very Secure	Total
Female	6.45	7.15	6.8	6.8
Male	9.07	6.9	5.6	7.2
Total	7.76	7.04	6.22	

Main effect of Job Security:
Do the marginal means of 7.76 (insecure), 7.04 (secure) and 6.22 (very secure) differ?
The decision rule: if the p-value < 0.05 Reject the $H_{0.}$
Because the p-value is 0.32 which is greater than the α =0.05. We do not reject the H_0
There is insufficient evidence to conclude that mean unpaid overtimes are different between the three categories of job security.

(c) Is there an effect due to Gender?

Main effect of Gender:
Do the marginal means of 6.8 (female) and 7.2 (Male) differ?
The decision rule: if the p-value < 0.05 Reject the $H_{0.}$
Because the p-value is 0.62 which is greater than the α =0.05. We do not reject the H_0
There is insufficient evidence to conclude that mean unpaid overtime hours are different between male and female employees.

(d) What can you conclude about the effect of Gender and the perception of Job Security on employee's Unpaid Overtime?

HINT: follow the steps in Q2.

PART – B

Open the data file and install the Data <u>Analysis Tool Pak</u>

a) Download the file ***BLITZT5.xls*** from Cloud Deakin. Please **save it** to the hard drive.

b) Open the file in Excel.

c) Install the Data <u>Analysis Toolpak</u>. [See the previous tutorial for instructions].

Q1. Chi-Square Test for the difference between two or more proportions (independent sample)

(a) Discuss the advantages and disadvantages of using z-test and chi-square test for the difference between two proportions.

Chi-Square test can be applied to both small and large samples.
If we are specifically interested in determining whether there is evidence of a directional difference such as $\pi_1 > \pi_2$, then we should use Z test.

(b) Discuss each of the following statements.

i. For the chi-square test, we must have equal sample size for each group/population.

No this is not true. Can have different sample size for each cell.

ii. The expected frequency is at least five for each cell in the table.

For the Chi-Square test to give accurate results for a 2 X 2 table, we must have the assume that each expected frequency is at least 5. If this assumption is not satisfied then we have to use an alternative test (Fishers exact test).

Q2. Awareness of the Loyalty program across the three cities

The BLITZ management team would like to broaden the scope of the analysis to investigate whether there is a significant difference in the proportion of loyalty program awareness between the three cities.

(a) Write the null and alternative hypothesis for this scenario using both notation and words.

$$H_0 = \pi_1 = \pi_2 \ldots \pi_c$$

$$H_1 = Not\ All\ \pi_j\ are\ equal\ (where\ j = 1,2, \ldots .C)$$

H_0 = Proportions of loyalty program awareness is same for all three cities
H_1 = Not all proportions of loyalty program awareness are equal.

(b) Conduct a hypothesis test (Chi-Square Test Use α= 5%) that will determine if the population proportion of loyalty program awareness is the same for all three cities – See Appendix-3 for answer.

(c) Based on the computer output created in (b), briefly advise BLITZ management about the findings.

The decision rule: if the p-value < 0.05 Reject the H_0.
Because the p-value is 0.02 which is less than the α =0.05. We reject the H_0
There is sufficient evidence to conclude the three cities have different loyalty program awareness.

(d) Use the multiple comparison procedure to determine which population proportions differ significantly (if). See Appendix-4 for steps.

Marascuilo Procedure for loyalty program awareness			
Level of Significance	0.05		
Square Root of Critical Value	2.4477		
Group Sample Proportions			
1: Mel	0.5700		
2: Perth	0.7300		
3: Syd	0.5600		
MARASCUILO TABLE			
Proportions	Absolute Differences	Critical Range	
| Group 1 - Group 2 |	0.1600	0.1628	Not significant
| Group 1 - Group 3 |	0.0100	0.1716	Not significant
| Group 2 - Group 3 |	0.1700	0.1630	Significant

(e) Based on the computer output created in (d), now refine your response to BLITZ management.

It appears that the Sydney and Perth has a significantly different awareness.

Q3. Online Shopping across the three cities

Is there a significant difference in the proportion of online shoppers between the three cities?

(a) Write the null and alternative hypothesis for this scenario using both notation and words.

$$H_0 = \pi_1 = \pi_2 \ldots \pi_c$$

$$H_1 = Not\ All\ \pi_j\ are\ equal\ (where\ j = 1,2, \ldots .C)$$

H_0 = Proportions of online shoppers between cities are no different.
H_1 = Not all proportions of online shoppers are equal.

(b) Conduct a hypothesis test (Chi-Square Test Use α= 5%) that will determine if the population proportion of BLITZ online shoppers is the same for all three cities .

Chi-Square Test							
	Observed Frequencies						
	Hotel				Calculations		
Online Shopper	Mel	Perth	Syd	Total	fo - fe		
Yes	20	22	24	66	-2.0000	0.0000	2
No	80	78	76	234	2.0000	0.0000	-2
Total	100	100	100	300			
	Expected Frequencies						
	Hotel						
Online Shopper	Mel	Perth	Syd	Total	(fo - fe)^2/fe		
Yes	22.0000	22.0000	22	66	0.1818	0.0000	0.1818
No	78.0000	78.0000	78	234	0.0513	0.0000	0.0513
Total	100	100	100	300			

Data	
Level of Significance	0.05
Number of Rows	2
Number of Columns	3
Degrees of Freedom	2

Results	
Critical Value	5.9915
Chi-Square Test Statistic	0.4662
p -Value	0.7921
Do not reject the null hypothesis	

Expected frequency assumption is met.

(c) Based on the computer output created in (b), briefly advise BLITZ management about the findings.

The decision rule: if the p-value < 0.05 Reject the $H_{0.}$
Because the p-value is 0.79 which is greater than the α =0.05. We Do Not reject the H_0
There is insufficient evidence to conclude the three cities have different proportions of online shoppers.

(d) Use the multiple comparison procedure to determine which population proportions differ significantly (if). See Appendix-4 for steps.

Not required as the proportions are not significantly different.

(e) Based on the computer output created in (d), now refine your response to BLITZ management.

Not required as the proportions are not significantly different.