DE MONTFORT UNIVERSITY LEICESTER

# IMAT5169 Statistics

## Solutions to Phase Test 2 2019/20

Date: 8th January 2020 Time

Allowed 1.5 hours

| | |
|---|---|
| **Student Name** | |
| **Student P Number** | |

**Instructions to candidates:**

This is an OPEN BOOK test and you may refer to any of your study materials, books and other appropriate learning resources etc. You may use SAS and other appropriate software as well as a calculator. The use of social media, email or mobile phones etc. is NOT permitted.

This test is worth 60 marks and the marks for each part are clearly shown.

There are two questions and all questions require the use of SAS.

You should attempt **BOTH** questions.

This test contributes 50% to your overall module mark.

A separate Statistical Formulae and Statistical Tables booklet is provided.

Save this Word document and **type your answers in the spaces provided**.

Make sure you save your final version of this test with answers included and upload and submit this as directed on the Blackboard instructions **BEFORE the time limit has been exceeded**.

The version that is uploaded is the one that will be assessed so check carefully that you have uploaded the correct final test document.

Since the test is time constrained if you have any problems during the test please contact the module tutor Dr Elmina Homapour via email at elmina.homapour@dmu.ac.uk or by telephone on +44 (0) 116 2013988 as soon as possible.

**Question 1** [20 marks]

A company has undertaken a survey of its employees, which included asking them about their perceptions of how well they think health and safety policies in the work-place are adhered to by the company. The response to this question was either “Always", “Mostly" or “Rarely". In addition, respondents to the survey were asked to indicate their job role. Data on a total of 90 employees was collected whose job role was recorded as either being an “Engineer" or a “Manager”. The responses are summarised in the contingency table below.

| | Always | Mostly | Rarely | Total |
|---|---|---|---|---|
| Engineer | 39 | 11 | 1 | 51 |
| Manager | 20 | 14 | 5 | 39 |
| Total | 59 | 25 | 6 | 90 |

You are required to use SAS to undertake a Chi-squared test to determine whether the data provide any evidence that the perceptions of adherence to health and safety in the workplace are associated with job role. To do this you should complete and answer each of the following:

a) State the hypotheses to be tested. [2 marks]

My answer:
Ho: Perceptions of adherence to health and safety are NOT associated with job role.
HA: Perceptions of adherence to health and safety ARE associated with job role

b) Enter the data into SAS in a suitable way and use SAS to undertake the Chi-squared test. Copy and paste your SAS code(but not the output!) in this space below: [4 marks]

```
Data Staff;
 input Job $ Opinion $ Count @@;
datalines;
Manager Always 20
Manager Mostly 14
Manager Rarely 5
Engineer Always 39
Engineer Mostly 11
Engineer Rarely 1 ;
run;
proc freq data = Staff;
 tables Job*Opinion / chisq nopercent;
 weight Count;
 run;
```

c) From the SAS output generated, complete the following (ignore any warning messages given by SAS at this stage): [5 marks]

Chi-squared statistic = 7.6819
p-value = 0.0215
My conclusion is (state this in as much detail as possible): There is evidence that perceptions of adherence to health and safety ARE associated with job role. Engineers appear to more often perceive that health and safety as being always adhered to compared to managers

d) By typing in all your working, show how to calculate BY HAND, the expected value for the employees whose job role is "**Engineer**" and their opinion response is "**Always**": [2 marks]

My answer:
For Engineer/Always the calculation is: 51x59/90 = 33.433

e) Use SAS to obtain ALL of the expected values and type them in the following table. Copy and paste your SAS code (but not the output!) in this space below: [2 marks]

```
proc freq
data = Staff;
tables Job*Opinion / chisq expected;
weight Count;
run;
```

| | Always | Mostly | Rarely |
|---|---|---|---|
| Engineer | 33.433 | 14.167 | 3.400 |
| Manager | 25.567 | 10.833 | 2.600 |

f) Comment on what problem(s) you can identify with regard to the analysis. If appropriate suggest how the problem(s) could be overcome, commenting on any additional results from SAS that should be used. [5 marks]

My answer:
The warning message that 33% of cells have expected counts below 5 is a problem. This renders the use of the Chi-square distribution doubtful. Instead we should use the results of Fishers Exact test.

To get this we need to add the statement "Exact" to the tables command as follows: tables Job*Opinion / chisq exact;
This shows the p-value is 0.0213 so the conclusion does not change from that stated earlier.

**Question 2** [40 marks]

A lock manufacturer is considering three possible designs for the packaging of its locks with a choice of two primary colours (red and yellow). The different combinations of design and colour were tested in 24 similar DIY shops as follows:

| Colour | YELLOW | | | | RED | | | |
|---|---|---|---|---|---|---|---|---|
| Design | SALES | | | | | | | |
| 1 | 32 | 40 | 38 | 44 | 14 | 17 | 21 | 20 |
| 2 | 49 | 58 | 53 | 50 | 38 | 27 | 31 | 34 |
| 3 | 50 | 49 | 40 | 48 | 26 | 20 | 24 | 30 |

You are required to use SAS to fit a two-way ANOVA model WITH AN INTERACTION TERM which looks at the effect of colour and design on sales. To do this you should complete and answer each of the following.

a) State why this is a two-way design. . [2 marks]

My answer:
Because we are assessing the effect of two factors "Colour" and "Design" on mean sales

b) State how many replications there are. . [1 marks]

My answer:
There 4 replications of each colour and design combination.

c) Use SAS to fit a two-way ANOVA model WITH AN INTERACTION TERM which looks at the effect of colour and design on sales. Use relevant variable names within your SAS codes. When completed, copy and paste your SAS code (but not the output!): [4 marks]

```
data lock;
input colour$ design sales;
datalines;
R 1 14
R 1 17
R 1 21
R 1 20
R 2 38
R 2 27
R 2 31
R 2 34
R 3 26
R 3 20
R 3 24
R 3 30
Y 1 32
Y 1 40
Y 1 38
Y 1 44
Y 2 49
Y 2 58
Y 2 53
Y 2 50
Y 3 50
Y 3 49
Y 3 40
Y 3 48
```

```
run;
proc glm data= lock;
class colour design;
model sales = colour design colour*design;
run;
```

d) From your SAS output for the above analysis complete the following ANOVA tables by typing in the appropriate values (you only need to use 2 decimal places): [4 marks]

| Source | DF | Sum of Squares | Mean Square | F statistic | p-value |
|---|---|---|---|---|---|
| **Model** | 5 | 3400.21 | 680.04 | 36.69 | <0.001 |
| **Error** | 18 | 333.75 | 18.54 | | |
| **Corrected Total** | 23 | 3733.96 | | | |

| Source | DF | Sum of Squares | Mean Square | F statistic | p-value |
|---|---|---|---|---|---|
| **Colour** | 1 | 2583.38 | 2583.38 | 139.33 | <0.001 |
| **Design** | 2 | 813.58 | 406.79 | 21.94 | <0.001 |
| **Interaction** | 3 | 3.25 | 1.63 | 0.090 | 0.92 |

e) Comment on the results shown in the tables above in part d). [4 marks]

My answer:
The first table indicates that there is evidence ($p<0.001$) that colour and design (the model) do explain at least some differences in mean sales.
There is no evidence ($p=0.092$) of a statistically significant interaction between colour and design.
The results above suggest that mean sales do differ by both colour and design but we should refit the model without the interaction term

f) Explain how the F-statistic for Colour is calculated from the mean square values given in the two tables in part d). [2 marks]

My answer: 139.33 = 2583.38/18.54

g) Explain how the model sum of squares in the first table in part d) relates to the sums of squares given in the second table in part d). [2 marks]

My answer:
The total of the three sums of squares in the second table = 2583.38+813.58+3.25 =3400.21 which is the model sum of squares in the first table.

h) Use SAS to fit a two-way ANOVA model which looks at the effect of colour and design on sales but this time WITHOUT THE INTERACTION TERM. Copy and paste your SAS code but not the output!) in the space below: [2 marks]

```
proc glm data= lock;
class colour design;
model sales = colour design;
run;
```

i) From your SAS output for the above analysis complete the following ANOVA tables by typing in the appropriate values (you only need to use 2 decimal places): [4 marks]

| Source | DF | Sum of Squares | Mean Square | F statistic | p-value |
|---|---|---|---|---|---|
| **Model** | 3 | 3396.96 | 1132.32 | 67.20 | <0.001 |
| **Error** | 20 | 337.00 | 16.85 | | |
| **Corrected Total** | 23 | 3733.96 | | | |

| Source | DF | Sum of Squares | Mean Square | F statistic | p-value |
|---|---|---|---|---|---|
| **Colour** | 1 | 2583.38 | 2583.38 | 153.32 | <0.001 |
| **Design** | 2 | 813.58 | 406.79 | 24.14 | <0.001 |

j) Comment on the results shown in the tables above in part i). [3 marks]

My answer:
The first table indicates that there is evidence ($p<0.001$) that colour or design (the model) do explain at least some differences in mean sales.
There is evidence ($p<0.001$) that mean sales differs due to both colour and design.

k) Discuss any additional results which should be considered such as multiple comparisons and boxplots. You should copy and paste your SAS codes used to do this as well as copy and paste the resulting SAS output and plots and add your comments on these outputs. [6 marks]

Discussion could relate to multiple comparisons, boxplots etc. For example using the following code we can do multiple comparisons using the Ryan-Einot-Gabriel-Welsch Multiple Range Test:

```
proc glm data= lock;
class colour design;
model sales = colour design;
means colour design / REGWQ;
run;
```

The output is as shown here on the right

This indicates that all three designs (1, 2 and 3) are actually statistically significantly different from each other.

The boxplot below and the table on the right indicate how mean sales are highest with Design 2 (42.5) and lowest with Design 1 (28.25).

**The SAS System**

**The GLM Procedure**

**Ryan-Einot-Gabriel-Welsch Multiple Range Test for sales**

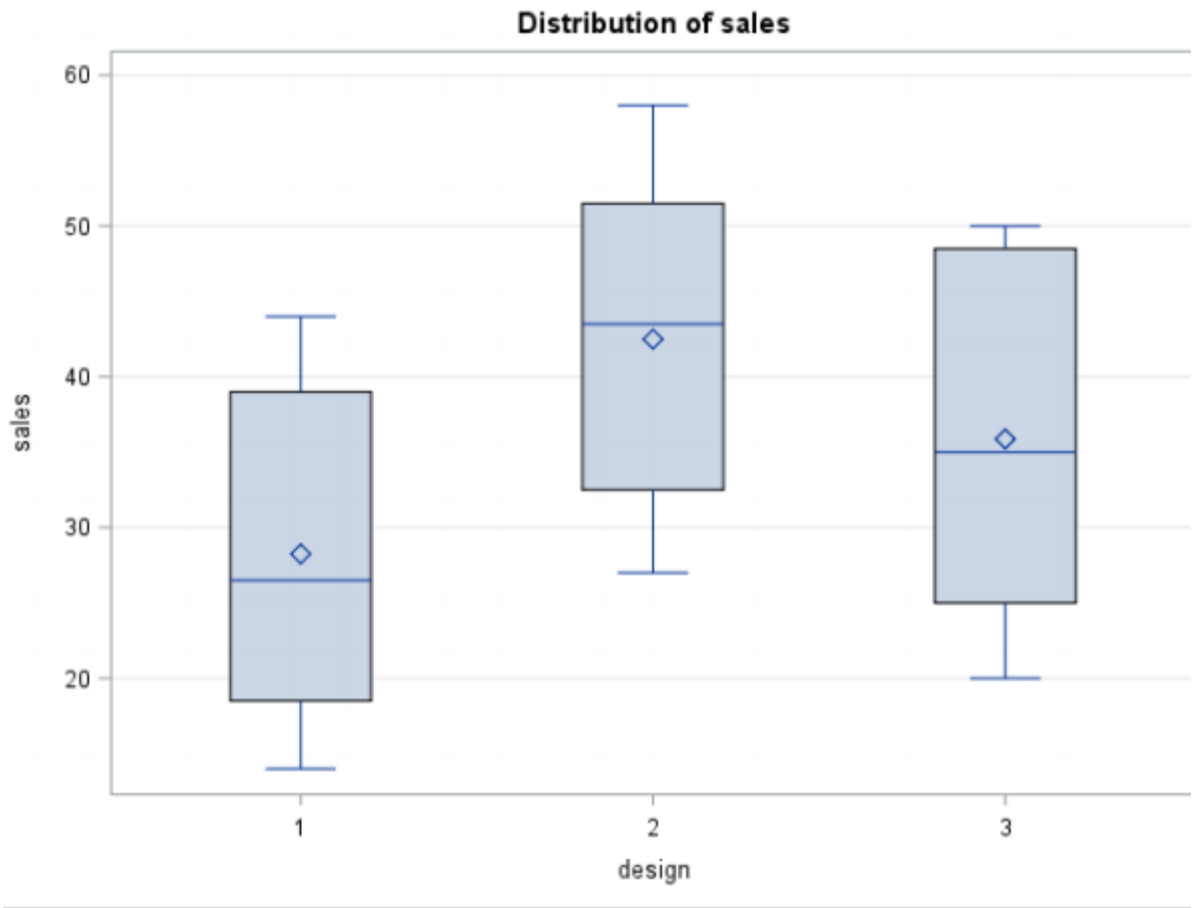
**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 16.85 |

| Number of Means | 2 | 3 |
|---|---|---|
| Critical Range | 4.2812326 | 5.1926302 |

Means with the same letter are not significantly different.

| REGWQ Grouping | Mean | N | design |
|---|---|---|---|
| A | 42.500 | 8 | 2 |
| | | | |
| B | 35.875 | 8 | 3 |
| | | | |
| C | 28.250 | 8 | 1 |



Similar output (see right and box plot below) for Colour indicates that the differences in mean sales are more evident due to colour.

Yellow has the higher mean sales at 45.917 compared to red at 25.167.

**The SAS System**

**The GLM Procedure**

**Ryan-Einot-Gabriel-Welsch Multiple Range Test for sales**

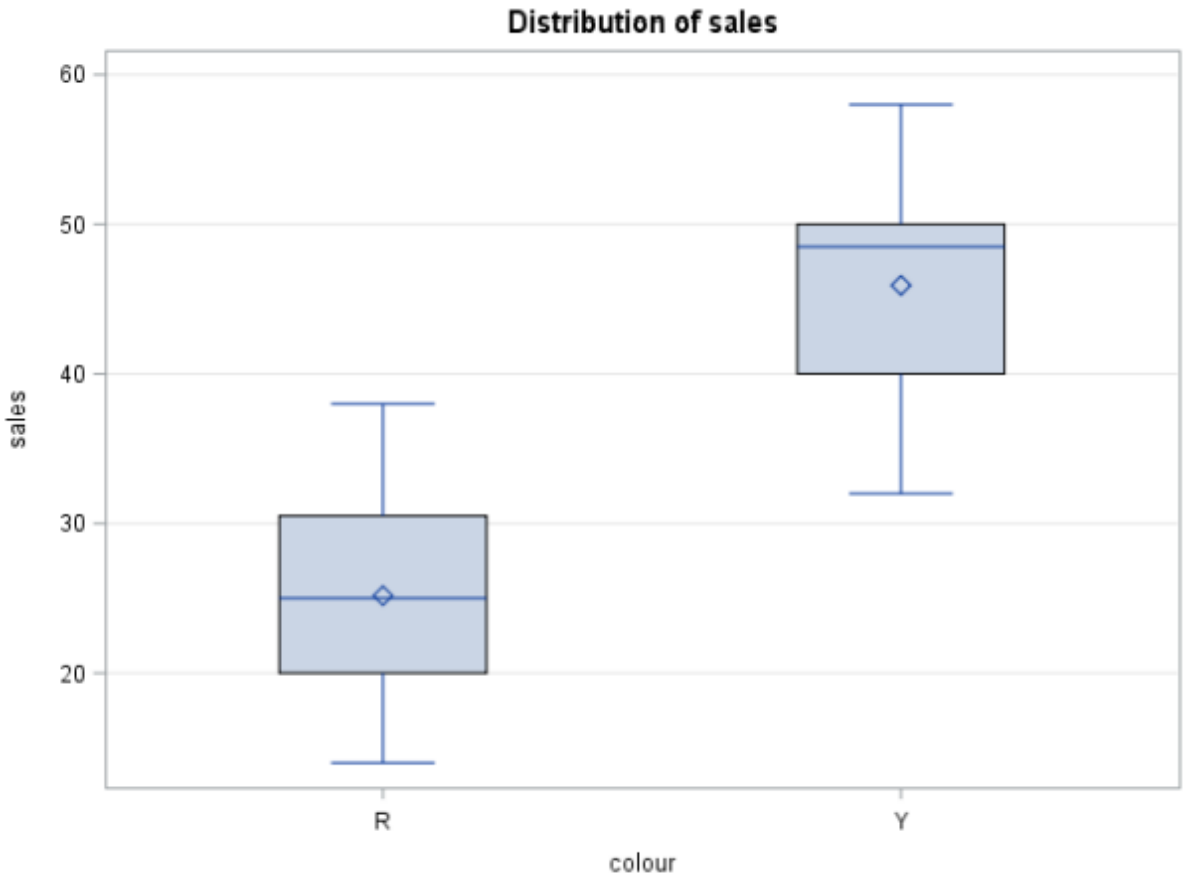
**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 20 |
| Error Mean Square | 16.85 |

| | |
|---|---|
| Number of Means | 2 |
| Critical Range | 3.4956117 |

Means with the same letter are not significantly different.

| REGWQ Grouping | Mean | N | colour |
|---|---|---|---|
| A | 45.917 | 12 | Y |
| | | | |
| B | 25.167 | 12 | R |



l) Use SAS to extract and save the residuals and then explore them to assess the assumption of normality of the residuals. You should copy and paste your SAS codes used to do this as well as copy and paste the resulting SAS output and plots and add your comments on these outputs. [6 marks]

```
proc glm data= lock;
class colour design;
model sales = colour design;
output out=lockfit r=resid;
run;
proc univariate data=lockfit plot normal;
var resid;
run;
```

Comment should include plots such as that shown overleaf and an argument that the assumption of normality seems reasonable.