MAST90007: Statistics for Research Workers 2020

1,500 word assignment

Due: **5 pm, Friday 7 August 2020**

Submission

Submit an electronic copy of the assignment via the LMS.

IMPORTANT: All students in this subject are required to complete the online plagiarism declaration form for the subject as a whole, covering all work. You will find a link to the form on Canvas. **If you do not include complete the online plagiarism form your assignment will not be accepted.**

This assignment contains three (3) questions worth a total of 20 marks. There is some general advice on the assignment at the end of this document, on page 6.

The questions are based on case studies described on the website:

http://realstat.science.unimelb.edu.au

The material on this site was developed by the Statistical Consulting Centre and is based on real consulting projects.

Case study 1

This is based on: **Study of graft arteries – Professor Brian Buxton**

Read the Introduction, the Background and the Study design. (You may also wish to read about other aspects of the study, or have a look at other studies on the website.)

Go to the Data page, and download the Excel file.

Question 1 [6 marks]

Cut and paste the data into Minitab. Check to see if you need to rename any of the variable labels.

In these analyses, you will consider the association between smoking and various measures of the health of arteries.

(a) Produce a summary table to describe the risk factors of the study participants.

(b) Consider the three main indices of severity of disease for the radial artery that are obtained from the morphometric analysis. Produce suitable visual display(s) to allow a comparison of the distributions of these measures in describing the results of the morphometric analysis of radial arteries, according to whether or not a patient ever smoked.

You might consider using Minitab 19's layout tool to combine graphs into one layout. You can find this under the Editor menu when you have a graph open. (Mac users, see the Advice at the end of the assignment.)

(c) Carry out appropriate analyses to compare those who never smoked with those who smoked on the three main indices of severity of disease in the radial arteries obtained from the morphometric analysis. Provide one or more suitable tables that includes the summary statistics and inferential statistics.

(d) Discuss and justify the assumptions underlying your choice of analysis.

(e) Write a summary of the analyses you have carried out explaining the results of all the comparisons you have made. Write the summary for a doctor interested in the practical application of the study results.

(f) Carry out a logistic regression analysis predicting intimal abnormality in the internal thoracic artery from smoking status. Write a summary of the results, again suitable for a doctor interested in the findings.

(g) A doctor you're working with wants to know: "Are any of the results statistically significant? Let's publish any significant findings; don't worry about the rest: the journal won't be interested in them." Respond to this comment.

Case study 2

This is based on: **Growing better pinot noir – Vincent Lakey**

Read the Introduction, and then the Background up to the section headed Harvest. (You may also wish to read about other aspects of the study, or have a look at other studies on the website.) You have examined some of the data from this study in the exercises during the subject.

Go to the Data page, and download the Excel file.

Question 2 [7 marks]

For this question, you will use the average yield in 2002 (as measured by the average weight of bunches per vine, in kg) and the average number of bunches per vine in 2002.

Look at the data in the worksheets *Trunk diameter* and *Harvest yield.* Cut and paste the data from the worksheet *Harvest yield* into Minitab. Note that the variable names will need to be corrected in Minitab.

It is possible to count bunches of grapes long before they are harvested. The grower would like to know if the average weight of the bunches harvested, based on 9 vines in each case, can be reasonably predicted from the average number of the bunches. Consider the relationship between the average number of bunches and the average weight of bunches harvested, without considering the different treatments used.

(a) Produce an appropriate graph showing the relationship between the average number of bunches harvested and the average weight harvested, for 2002.

(b) Describe the relationship between the two variables, and give a suitable summary statistic.

(c) Describe an appropriate statistical model for examining the grower's question, and fit the model in Minitab. Provide an appropriate summary table and give a plain language explanation of the estimates of the parameters of the model.

(d) Examine appropriate diagnostic plots, and comment on anything that challenges the assumptions of the model you have used.

(e) Find a 95% prediction interval for the average weight when the average number of bunches is 35. Explain its meaning.

(f) Find the predicted value of the average weight when the average number of bunches is zero.

(g) Suppose the grower looks at the analysis that you have carried out and comments: "I don't understand. When you plug in zero bunches in your estimated model, you don't get a prediction of zero weight. How come? That can't be right!"

(h) The grower would like to improve the prediction of the average weight of the bunches. Explain, in principle, a possible approach. You do not need to implement the approach.

(i) This analysis has been done on averages (across 9 vines) for the number of bunches and the weight of the bunches harvested. Would you expect an analysis based on the individual vines to be similar? Why, or why not? (Explain your answer.)

Question 3 [7 marks]

In this question, we are interested in the growth of the vines.

Cut and paste the data from the worksheet *Trunk diameter* into Minitab.

There were three treatments used: 1: herbicide; 2: compost; 3: straw. The experiment used blocking; each treatment was used in each of six blocks. A research question of interest was: does the treatment influence the growth of the vines?

The grower carried out an analysis but has asked you for advice. He has examined the trunk diameters in 2003, three years after the trial began. To carry out the analysis required in the remainder of the question, you will need to average the trunk diameter in 2003 across the nine vines for each treatment and block combination.

In Minitab, use Stat > Basic Statistics > Store Descriptive Statistics.
Enter *Trunk diameter in 2003 (cm)* in the "Variables:" box.
Enter Treatment, EWblock, Aspect, Block in the "By variables (optional):" box.
Under "Statistics…" tick 'Mean'.
Under "Options…" tick *only* 'Store distinct values of By variables'.

You should obtain columns at the right-hand end of the worksheet that contain the variables; provide suitable names for all of them, and use the data in these created columns for the rest of this question.

(a) Produce a graph of the data that allows a comparison of the trunk diameter in 2003 according to type of treatment.

(b) Comment on any differences between treatments, based on the graph.

(c) Describe a suitable statistical model for examining the research question which uses these data and assumes that differences between parameters are of key interest. Describe the factor or factors that you will include in your model. Explain why you have included these factor(s) in your model.

(d) State one assumption required for analysing the data using the model you have suggested. Describe this assumption in concrete terms in relation to Vincent's study, rather than in abstract form.

(e) Use Minitab to fit the model involving treatment that you have specified above. Provide a summary table of the Analysis of variance, and give a plain language explanation of the meaning of the *P*-value. Again, use concrete terms in relation to Vincent's study, rather than in abstract form.

(f) Find the predicted treatment means; report them in a suitable table and provide a plain language explanation of the pattern you observe.

(g) Consider the assumption you described above in (e). State if the assumption is reasonable and provide relevant evidence.

(h) Find 95% confidence intervals for comparing the mean growths for each pair of treatments. Provide a suitable report of these confidence intervals, including a plain explanation in concrete terms in relation to Vincent's study.

(i) Write a brief summary of your recommendations about the use of different treatments in relation to vine growth for the grower, based on the analysis and confidence intervals you have provided.

(j) The grower asks you if there are other ways he might examine the research question. Comment on the use of the 2003 trunk diameter for measuring growth. Propose a suitable alternative.

Advice

Here is some advice to follow when preparing your assignment.

- The purpose of the assignment is to relate the statistical theory and practice learned in *Statistics for Research Workers* to real world data. The essential feature is that you must demonstrate understanding and application of statistical ideas covered in SRW to real world practice.
- The presentation of results should be consistent with the principles for presenting graphics and tables discussed in the course. For those using Minitab for Mac, we are aware that you cannot edit graphs. You can indicate in a note at the beginning of your assignment that you are using Minitab for Mac and also footnote graphs to indicate changes you would have made, if you could.
- In general, you are not required to provide Minitab output in the assignment, with the exception of graphs.
- The word limit for the assignment is 1,500 words. From our point of view, this is an upper limit for the assignment and you should aim to submit between 1,400 and 1,500 words. The word count does not include graphs and tables. University policy allows for a 10% deduction of marks once a written assignment exceeds 10% of the specified word limit. As the 1,500-word assignment is worth 20% of your final mark, you could lose 2% from your final mark if your assignment was, for example, 1,670 words.
- Your answers should be on no more than twelve (12) A4 pages of standard sized writing. This includes any graphs. Twelve pages is a generous limit for the assignment; this document is on six pages, with a lot of white space, and it contains over 1,500 words.
- You do not need to reproduce the questions in your assignment.