MA50259: Statistical Design of Investigations

Coursework 2 (2024)

09618

Part 1: Barley Experiment

The data in table below show the yields (measured in bushels per acre) of five varieties of barley in an experiment carried out in a rural area in the US.

Place	Year	Excel	Compana	Drummond	Conlon	Kindred
1	1971	71.0	95.4	109.7	119.5	88.3
1	1972	90.7	102.3	79.4	86.2	80.2
2	1971	132.6	121.0	140.7	171.5	135.7
2	1972	99.4	105.5	120.2	157.7	102.1
3	1971	87.3	73.3	79.4	128.6	85.3
3	1972	102.5	111.4	100.5	131.9	122.6
4	1971	129.8	111.3	121.5	148.8	114.8
4	1972	96.9	60.3	83.2	118.5	72.4
5	1971	92.3	81.4	72.1	84.8	99.1
5	1972	65.2	48.9	91.5	72.8	77.4
6	1971	84.3	72.1	76.9	109.8	90.0
6	1972	63.3	68.4	62.7	97.8	92.2

(a) What would be the purpose of running these experiments in different locations and years? Also, comment on the design that would be appropriate for analysing the data above.

Environmental Variability: Different locations have varying climatic conditions, soil types, and ecosystems that can significantly impact crop yields. By testing in multiple locations, researchers can assess how each barley variety performs under diverse environmental conditions.

Temporal Variability: Testing across multiple years allows researchers to see how the yields of each variety are affected by annual fluctuations in weather patterns, such as rainfall, temperature, and potential pest or disease outbreaks. This helps in evaluating the consistency and reliability of each variety over time.

Generalization of Results: By collecting data from a range of locations and years, the experiment aims to provide more generalized and robust conclusions about which barley varieties perform best, under what conditions, and their resilience to different environmental stresses.

Interaction Effects: It is also essential to study how location and year interact with barley varieties. Some varieties might perform exceptionally well in one location but poorly in another, or some might have good years and bad years depending on external conditions. Understanding these interactions can help in breeding programs and agricultural planning.

Appropriate Design for Analysing the Data

Given the structure of the data and the nature of the experiment, the following experimental design considerations are critical:

```
Split-Plot Design: The main plot treatments could be the locations, with the subplot treatments being th

Repeated Measures: Since the data involve measurements from the same experimental setup across differen

Randomized Block Design: If each location is considered a block, this design can be useful, especially i

Analysis of Variance (ANOVA): This statistical method would be appropriate to analyze the data, allowing

Mixed-Effects Models: Considering that some factors like location and year might have random effects ac
```

(b) Organise the data in an appropriate dataframe in R with barley yields as the response variable, places as the blocks and the five varieties of barley as the treatment effects. Use an appropriate model which includes block effects and treatment effects to perform the ANOVA and determine if there is a significant difference across barley varieties. Clearly state the assumptions you may need in the model and write the null and the alternative hypotheses considered in the ANOVA.

(c) Omit the block effects in part (b) and use an appropriate model to perform the ANOVA and determine if there is a significant difference across barley varieties. Clearly state the assumptions you may need in the model and write the null and the alternative hypotheses considered in the ANOVA.

(d) Which of the two designs represented by the two models in part (b) and part (c) would be more efficient. Justify your answer.

Part 2A: Chronic Respiratory Disease Study

An increase in deaths due to chronic respiratory disease (CRD) was observed in certain parts of the UK after the millennium. A case-control study was carried out to investigate the possible association between prescription of one particular medication, namely X12 and CRD deaths.

Both cases and controls were chosen among persons who were admitted to hospital for CRD. The cases comprised 257 persons who died of CRD; the controls were 570 persons who did not die of CRD. The data are presented in the following table.

X12 prescribed	Cases	Controls
Yes	130	200
No	127	370
Total	257	570

(a) Obtain the odds ratio between X12 prescription and CRD deaths and the corresponding 95% confidence interval from the above table.

(b) Test the null hypothesis of no association between X12 prescribed and CRD deaths. Interpret the results in the context of the study.

(c) There was a concern that cases and control may have differed according to the underlying severity of their CRD. Indeed, disease severity may be associated with a lifestyle habit such as smoking, and hence is a potential confounder. Accordingly, the data were stratified by variables associated with severity. One such indicator of severity is smoking habit in the previous year. The data, stratified by this variable, is shown in the following table

Non-smokers			Smokers		
X12 prescribed	Cases	Controls	X12 prescribed	Cases	Controls
Yes	74	170	Yes	56	30
No	100	267	No	27	103
Total	174	437	Total	83	133

Estimate the odds ratio and calculate the corresponding 95% confidence interval for each stratum.

(d) Calculate the overall regression summary odds ratio (together with its corresponding 95% confidence interval) when adjusting for smoking. Interpret the results in the context of the study.

(e) Compare the odds ratios you obtained in parts (a) and (d). How did confounding by smoking affect the apparent direction of association between X12 and CRD deaths?

(f) Which other variables that you can think of could have been used as confounders to have a better picture of the association between X12 and CRD deaths? You should justify your answer.

Part 2B: Low Birth Weight Study

Consider a cohort study on birth weight (BW) of singleton babies and lifestyle habits (smoking, drinking, exercise etc.) of their mothers. Style 1 relates to unhealthy habits such as smoking/drinking and little exercise, and Style 2 relates to healthy habits including no smoking, no drinking and sufficient exercise. The number of babies born with low birth weight (LBW) within a 1-year period to Style 1 and Style 2 are 34 and 10 respectively. The total number of mothers who predominantly follow Style 1 and Style 2 are 335 and 320 respectively.

(a) Calculate the risk difference and the relative risk of LBW in the general population for the two lifestyle habits.

(b) Find a 95% confidence interval for the risk difference and the relative risk of LBW in the general population for the two lifestyle habits. Interpret the results in the context of the study.

(c) Perform a chi-square test to determine whether there is a significant association between lifestyle habits and the birth weight of singleton babies.