Analysis of Scientific Data –Summer Semester, 2022

Week 7 Tutorial Solutions

PART A - Two-Way ANOVA

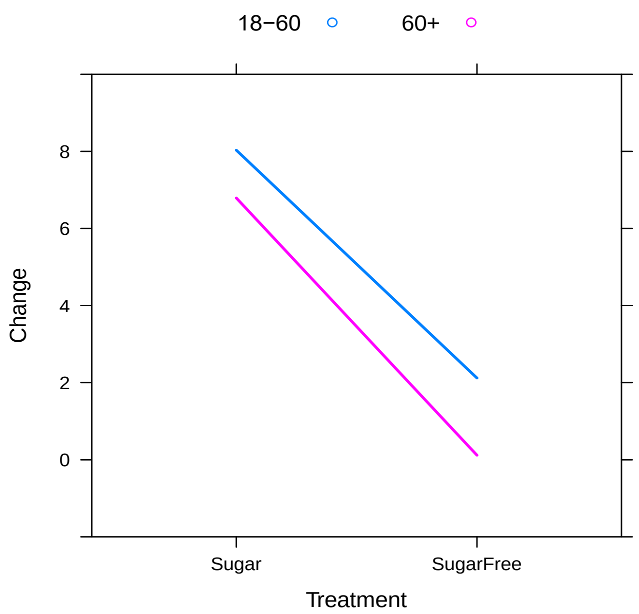
A project within the Islands investigated whether consuming sugar influences memory performance, and if any effects were the same for elderly subjects.

A sample of 40 subjects was randomly chosen from houses in Helluland, with 20 subjects aged 18-60 and 20 aged over 60. Each subject played a pairs memory game with 30 cards as quickly as they could, with the results recorded in seconds. Within each age group, half the subjects were randomly assigned to consume 50 g of lollies, while the other half consumed 50 g of sugar-free lollies. After 15 minutes, they repeated the memory game task.

a) The table below shows the summary statistics for the decrease in memory game time between the first and second plays:

Age	Treatment	n	Mean (s)	SD (s)
18-60	Sugar	10	8.03	5.054
	Sugar-free	10	2.12	8.140
60+	Sugar	10	6.79	6.289
	Sugar-free	10	0.12	5.547

Draw an interaction effects plot for this data. Use the vertical axis for the mean change in game time and the horizontal axis for the treatment.



b) Briefly discuss the potential interaction effect from your plot.
There seems to be a difference between the treatments and some difference between the age groups. However, since the lines are roughly parallel, there does not seem to be an interaction between the two factors in the effect on change.

c) To compare the interaction effects of treatment and age group on the decrease in memory game time, a two-way analysis of variance is conducted, giving the following results in R:

```
summary(aov(Change ~ Treatment*Age, data=sugar))
              Df Sum Sq Mean Sq F value  Pr(>F)
Treatment      ?  395.6   395.6   9.761 0.00351 **
Age            ?   26.2    26.2   0.647 0.42630
Treatment:Age  ?    1.4     1.4   0.036 0.85135
Residuals      ? 1459.2    40.5
```

What are the degrees of freedom for each of the variance components?

With four factor combinations, the Residuals degrees of freedom is $40 - 4 = 36$.

With two levels each, Treatment and Age will each have 1 degree of freedom, with the interaction then 1 as well.

Adding these together gives a total of 39 degrees of freedom, which matches $n - 1$.

d) Calculate and interpret the R^2 value of this two-way ANOVA model.

$$R^2 = \frac{\text{SS Model}}{\text{SS Total}} = \frac{395.6 + 26.2 + 1.4}{395.6 + 26.2 + 1.4 + 1459.2} = 0.2248$$

So, the combination of treatment and age group explains around 22% of the variability in change in memory game time.

e) Interpret the three p-values from the ANOVA table. What do you conclude about the effect of consuming sugar on memory performance, and how it relates to age?

There is strong evidence of a difference in mean change in memory game time between the sugar and sugar-free treatments ($F_{1,36} = 9.761, p = 0.00351$).

There is no evidence of a difference in mean change in memory game time between the age groups ($F_{1,36} = 0.647, p = 0.426$), and no evidence of an interaction between sugar treatment and age group on the mean change in memory game time ($F_{1,36} = 0.036, p = 0.851$).

Given the lack of evidence of an interaction, it may be useful to try an additive model, without the interaction term.

Part B: Two-Way ANOVA using RStudio (Based on a past exam paper)

The Maryland Biological Stream Survey conducted a study to investigate the effects of physical and chemical water characteristics on the abundance of Longnose Dace (Rhinichthys Cataractae, a freshwater minnow native to North America). The researchers sampled 54 segments of streams, each 50m long. In each 50m segment, they recorded the number of fish, sulfate concentration, and dissolved oxygen level of the stream water. The datafile "DaceFish.csv" contains the following variables.

Sulfate – Sulfate concentrate in three levels (S1 – Less than 20mg/litre, S2 – Less than 40mg/litre, S3 – more than 40 mg/litre).

Oxygen – Dissolved oxygen in three levels (O1 – Less than 8mg/litre, O2 – Less than 10mg/litre, O3 – More than 12g/litre).

Fish – Number of Longnose Dace.

a) What are the variable types in the data?

   Sulfate and Oxygen are categorical variables.
   Fish is a quantitative variable.

b) How many segments had Sulfate concentration less than 20mg (S1) and dissolved oxygen level less than 10g/litre (O2)

   6

c) The study uses a Two-Way ANOVA model to investigate the interaction effect of sulfate and oxygen concentrations on the mean number of longnose daces. What are the numerator and denominator degrees of freedom of the F-test statistic to test the statistical significance of the interaction effect?

   Numerator df = 4
   Denominator df = 45

d) What do you conclude about the interaction effect?

   p-value = 0.83321. There is no evidence to conclude that there is an interaction between Sulfate and oxygen on mean number of fish.

Part A – Chi-square test

Question 1

In semester 2 2020, students were asked in a Zoom poll to "toss a coin twice and indicate your outcome". The combined counts from the two streams are shown in the following table:

HH	HT	TH	TT
63	80	64	42

a) Supposing these were independent tosses of fair coins, how many counts would you expect in each of these four groups?

The total is $63 + 80 + 64 + 42 = 249$. The probability of each outcome is 0.25, so we would expect $0.25 \times 249 = 62.25$ in each group.

a) Calculate the statistic to compare the observed counts to the expected count. Is there any evidence that this process did not work as intended?

$$\chi^2 = \frac{(63 - 62.25)^2}{62.25} + \frac{(80 - 62.25)^2}{62.25} + \frac{(64 - 62.25)^2}{62.25} + \frac{(42 - 62.25)^2}{62.25} = 11.71$$

With $4 - 1 = 3$ degrees of freedom, p-value is $P(\chi_3^2 \geq 11.71) = 0.0084$, strong evidence to suggest the observed counts do not come from a uniform distribution, suggesting that the process did not work as intended.
p-value using R: 1-pchisq(11.71, df=3)

Question 2

The following table shows counts of students in the 2018 and 2020 surveys who lived at home or in a shared flat/house:

Year	Home	Shared Flat/House	Total
2018	391 (426.1)	185 (149.9)	576
2021	550 (514.9)	146 (181.1)	696
Total	941	331	1272

Is there any evidence of a change in living situations between 2018 and 2020?

Start by calculating totals, as shown. Expected value for 'Home' in '2018' is $\frac{576 \times 941}{1272} = 426.1$, with similar calculations for remaining expected values (or by subtraction). This gives

$$\chi_1^2 = \frac{(391 - 426.1)^2}{426.1} + \frac{(185 - 149.9)^2}{149.9} + \frac{(550 - 514.9)^2}{514.9} + \frac{(146 - 181.1)^2}{181.1} = 20.31$$

P-value is $P(\chi_1^2 \geq 20.31) < 0.0001$, very strong evidence against the null hypothesis of no association, suggesting that the distribution of residence has changed between 2018 and 2021.

p-value using R: 1-pchisq(20.31, df=1)