Analysis of Scientific Data – Summer Semester, 2022

Week 8 Tutorial Solutions

Part A – Chi-square test and Logistic regression using RStudio

A study was conducted to determine whether endovascular therapy in patients with stroke improves the likelihood of a favourable outcome relative to standard medical therapy. Patients were randomly assigned within 6 hours of the estimated onset of a stroke to receive either endovascular therapy or standard medical care. Information was record on the patients Prestroke health, the time between the estimated onset of stroke and the commencement of treatment, and whether the patient experienced a favourable outcome at 90 days after the stroke.

Download the Stoke.CSV data from the Week 8 folder under Learning Resources.

The data contains the following variables:

Therapy	"Endovascular" or "Standard" medical therapy
Outcome	Favourable outcome ("Yes"/"No")
Time	Time between stroke and commencing therapy (in hours)
Prestroke	Prestroke health score (L0 – good health, L1 – minor health issues, L2 – major health issues)

a) What proportion of patients in the Endovascular therapy group experienced a favourable outcome 90 days after the stroke?

65/178 = 0.3652

b) If there was no association between Prestroke health and therapy group, what is the expected count for the number of patients in the Endovascular therapy group with a Pre-stroke health score L1?

This can be answered in two ways.
Method 1: Generate a contingency table using addmargins(table()) function in RStudio for the two variables Prestroke and Therapy.

	Endovascular	Standard	Sum
L0	138	100	238
L1	23	20	43
L2	17	28	45
Sum	178	148	326

Then find the expected count as follows.

Expected count = (178*43)/326 = 23.4785

Method 2: Use the function chisq.test() in RStudio and then obtain the expected count.

c) The researchers want to test that the therapy groups are similar in terms of their pre-stroke health. What is the $\chi 2$ statistic used to test for an association between Pre-stroke health score and therapy group?

6.2577

d) Ignoring any possible effect due to the therapy group, use logistic regression to determine the effect of time to therapy on the patient's outcome. Based on this, what is the estimated odds for a favourable outcome for the patient if the time between stroke and commencing therapy is 4 hours?

The Maximum Likelihood estimates from the Logistic Regression are as follows.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0006696  0.3679505   0.002   0.9985
Time        -0.2013850  0.1073613  -1.876   0.0607 .
```

Then the estimated model is

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.0006696 - 0.201385Time$$

Substitute Time= 4

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.0006696 - 0.201385 \times 4 = -0.8062096$$

$$\frac{\hat{p}}{1-\hat{p}} = \exp(-0.8062096) = 0.4471459$$

Thus, the estimated odds for a favourable outcome for the patient if the time between stroke and commencing therapy is 4 hours is 0.4471.

Otherwise, you can use the predict() in RStudio and first estimate ln(odds) and then estimate odds. See the Week 8 -Part A.R file for the codes.

e) Based on the estimated Logistic Regression Model, what is the estimated probability of a favourable outcome for the patient if the time between stroke and commencing therapy is 4 hours?

From part d), $\frac{\hat{p}}{1-\hat{p}} = 0.4471459$

Rearranging,

$$\hat{p} = \frac{0.4471495}{1 + 0.4471495} = 0.3089847$$

The estimated probability of a favourable outcome for the patient if the time between stroke and commencing therapy is 4 hours is 0.3090

PART B – Sign Test and Signed Rank Test

The following table gives the stroke volume index (mL/beat/m^2) for eight patients, suffering from chronic severe anemia, before and after a treatment designed to lower the stroke index:

Patient	1	2	3	4	5	6	7	8
Before treatment	109	57	53	57	68	72	51	65
After treatment	56	44	55	40	52	46	49	41
Reduction	53	13	-2	17	16	26	2	24
Rank	8	3	1.5	5	4	7	1.5	6

a) Carry out a sign test to determine if there is any evidence that patients tend to have lower stroke volume index after the treatment.

First complete the Reduction row in the table (shown in blue)

Let p be the probability of a reduction after treatment for a patient with this condition. We want to test $H_0: p = 0.5$ vs $H_1: p > 0.5$. If H_0 is true then the count of positive reductions, X, will have a Binomial$(8, .5)$ distribution, so p-value is

$$P(X \geq 7) = P(X = 7) + P(X = 8) = 0.0352,$$

Using RStudio, p-value is:

```
 sum(dbinom(7:8, prob=0.5, size=8))
[1] 0.03515625
```

There is moderate evidence to conclude that the stroke volume index tends to be lower after the treatment.

b) Carry out a signed-rank test to determine if there is any evidence that patients tend to have lower stroke volume index after the treatment. (If you encounter any ties then average the ranks between the tied values.) Do you get the same results as in (a)?

First complete the Rank row in the table (shown in blue) based on the absolute values of the reductions.
'-2' and '2' are tied in magnitude. They would have had ranks 1 and 2, so we average and give them both 1.5.

The test statistic is $S = 8 + 3 + 5 + 4 + 7 + 1.5 + 6 = 34.5$. If the null hypothesis is true, the expected value and standard deviation are

$$E(S) = \frac{8 \times 9}{4} = 18 \quad \text{and} \quad \text{sd}(S) = \sqrt{\frac{8 \times 9 \times 17}{24}} = 7.141428,$$

so p-value is $P(S \geq 34.5) \approx P\left(Z \geq \frac{34.5-18}{7.141428}\right) = P(Z \geq 2.310462) = 0.0104313$

Using RStudio p-value is:

```
1-pnorm(2.310462)
[1] 0.0104313
```

slightly stronger evidence than the sign test that stroke volume index tends to be lower after the treatment.

PART C – Revision

Substantial research has established that the effect of alcohol on the human body depends on the blood alcohol concentration (BAC). Blood alcohol concentration depends on many factors including the number of standard drinks, sex, and body mass (kg). A study assigned a target alcohol dosage (Low, Medium and High) to participants. The actual standard drinks consumed was recorded, along with the blood alcohol concentration (g/dL) from a urine sample after a fixed waiting period.

a) The mean and standard deviations of the blood alcohol concentration for males and females are as follows.

Group	n	Mean (g/dL)	SD (g/dL)
Males	12	0.0632	0.0208
Females	9	0.0754	0.0224

Assuming population standard deviations are not equal, calculate the standard error of the difference in the sample BAC means.

$$\text{se}(\bar{x}_M - \bar{x}_F) = \sqrt{\frac{0.0208^2}{12} + \frac{0.0224^2}{9}} = 0.00958 \text{ g/dL}$$

b) What is the t-statistic to test whether the population mean BAC is lower for males than for females?

$$t = \frac{(0.0632 - 0.0754) - 0}{0.00958} = -1.273$$

c) What are the degrees of freedom to be used to calculate the p-value (not using the Welch approximation)?

We can use $\text{df} = \min(n_M - 1, n_F - 1) = \min(11,8) = 8$

d) What do you conclude for the test in part (b)?

The p-value is $P(T_8 \leq -1.273) = 0.119$, giving no evidence to suggest that mean BAC is lower for males than for females.

Using RStudio, p-value is:

pt(-1.273, df=8)

[1] 0.1193815