Data Science in Medicine: Tutorial 3

Hypothesis testing

Semester 1, 2020-2021

- Please attempt all questions on this worksheet in advance of the tutorial, and bring with you all work, including printouts of code and other results. Tutorials cannot function properly unless you do the work in advance.
- You are welcome to bring along any questions you may have from the lectures, textbook, etc.
- Assessment is formative, meaning that tutorials do not contribute to your final grade.
- Attendance is compulsory. If you have good reasons to miss a session, you should contact your year coordinator in advance to arrange to attend a different session.

Introduction

In this tutorial you are given two simple (entirely fictitious) datasets and you are asked to argue about the relationship between different variables in the datasets. You are expected to summarise/visualise the data and carry out appropriate hypothesis testing. You do not have to use R for this tutorial, but you are welcome to give it a try once you've had your Lab 2 session.

Part 1: Correlation between two numerical variables

The local community health centre has been running a study to investigate the relationship between Body Mass Index (BMI) and weekly hours of exercise. They have recruited 12 participants and they have collected the data shown in Table 1. (Note that you can also obtain the corresponding dataset `tut3_BMIexercise.csv` from the Tutorials page on the course website.)

BMI	Weekly hours of exercise
26	1.6
21	3
28	1.6
30	0.7
23	2.7
22	2.5
31	0.6
17	3.5
18	3.3
29	0.8
24	2.1
19	3.2

Table 1

You have been asked to assist the team with the data analysis and help them decide whether the two variables are correlated.

(1) Draw a scatterplot for the two variables, including a line of best fit. Based on this graph, does there appear to be any correlation between BMI and weekly hours of exercise? If so, is it positive or negative?

(2) Based on this sample, we estimate the correlation coefficient between BMI and weekly hours of exercise for the wider population. Its value is -0.9829309. Is this an indication of a strong correlation? Is it positive or negative?

(3) Use the statistic provided above to carry out hypothesis testing. What are the null and alternative hypotheses? What are the results of the test? What conclusions can you draw based on this test and your analysis for the previous two questions?

(4) Upon presenting the results of your analysis to the local community health centre, someone from their team says: “So an increase in hours of exercise causes a decrease in BMI!” Would you agree or disagree with this statement, and why?

Part 2: Association between two categorical variables

We are conducting a genome-wide association study, and we would like to investigate the association between being homozygous for the minor allele (aa) and having a certain disease, say D. You are asked to analyse the dataset `tut3_GWAstudy.csv` (which you can find on the Tutorials page of the course website and open with a text editor of your choice), to help with this investigation[1].

(1) Below you can find the observed contingency table for this dataset. How can you interpret the results?

	case		
allele	**0**	**1**	
0	17	17	34
1	15	23	38
	32	40	72

(2) Calculate the expected frequencies and the chi-square statistic.

(3) We carry out chi-square hypothesis testing to argue about the association between the two variables. What are the null and alternative hypotheses? As part of the hypothesis test, we find that the chi-square statistic calculated corresponds to p-value 0.3695. What conclusions can you draw based on this test and your analysis for the previous two questions?

Part 3: Discussion – Translating hypothesis testing into health service improvement

Hypothesis testing can be used for improving the health and wellbeing of our patients (Part 1 can be seen under this lens), for better understanding disease (an example was provided in Part 2), as well as for improving healthcare service delivery. Suppose that you have been given a dataset about hospital admissions and patient satisfaction across Scotland in 2018 (a sample of national data), an extract of which can be seen in Table 2.

date	**age**	**sex**	**marital_status**	**hospital**	**specialty**	**length_of_stay**	**overall_satisfaction**
09/11/2018	55	M	married	Pink	Cardiology	5	73
12/11/2018	34	F	divorced	Rose	General Medicine	12	67
04/05/2018	87	M	widowed	Pink	Geriatric Medicine	36	40
22/09/2018	86	M	married	Green	Geriatric Medicine	65	78
…	…	…	…	…	…	…	…

Table 2

[1] This is a variation of a dataset produced by Rafael A Irizarry and Michael I Love for the course “Data Analysis for the Life Sciences” offered on edX: https://www.edx.org/xseries/data-analysis-life-sciences

How could you perform hypothesis testing on this data in order to improve the delivery of healthcare services? In answering this question, please cover the following points:

a) Who would benefit from the service improvement (e.g. the whole of Scotland, a particular hospital, a subgroup of Scottish patients, etc.)?
b) Which variable(s) would you focus on, and what would the research hypothesis be?
c) What hypothesis test would you use, and what would be the null and alternative hypotheses?
d) Supposing the results of your analysis were found to be statistically significant, what would be the next steps for turning this knowledge into service improvement?