Inf2 – Foundations of Data Science 2024

Task: Preparation for Semester 2 Week 6 Workshop

22nd January 2025

1. **Logistic Regression**
   Suppose that you are an editor at a scientific journal, who receives hundreds of submissions a day, and has to decide whether to reject the papers or send them to other academics to review. The volume of submissions is increasing, and you don't have time to read all the submissions properly. You identify 3 key words or phrases that you think make you more or less likely to accept or reject a paper: "world-beating", "confidence interval" and "bootstrap". You then run a program to count the number times each occurs in your archive of rejected and non-rejected papers. The first few rows of the resulting dataset look like this:

"world-beating"	"confidence interval"	"bootstrap"	Rejected
0	5	3	No
6	1	1	Yes
...	...	...	...

   You identify the variables as:

   - $x^{(1)}$ Number of occurrences of phrase "world-beating"
   - $x^{(2)}$ Number of occurrences of phrase "confidence interval"
   - $x^{(3)}$ Number of occurrences of phrase "bootstrap"
   - y Whether the paper was rejected (Rejected="Yes") or sent out for review (Rejected="No").

   You then run logistic regression on the dataset and find the following coefficients:

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
1.0	0.5	−0.5	−0.1

   (a) In your dataset, what are the log odds of a paper being rejected that contains none of these words or phrases? Is the paper more likely to be rejected or not rejected?

   (b) In your dataset, what are the odds of a paper being rejected that contains none of these words or phrases?

   (c) In your dataset, what is the probability of a paper being rejected that contains none of these words or phrases?

   (d) Suppose a paper contains the phrase "world-beating" 5 times, and 0 occurrences of "confidence interval" or "bootstrap". What is the predicted probability of rejection?

   (e) Suppose a paper contains the phrase "world-beating" 0 times, and 3 occurrences of "confidence interval" and 2 of "bootstrap". What is the predicted probability of rejection?

(f) Suppose you decide to use these coefficients for a classifier, in which you will reject papers that have a probability of rejection above 0.5. Would you reject an article with 1 occurrence of "world-beating", 3 occurrences of "confidence interval" and 1 of "bootstrap"?

(g) You are still getting too many papers, so you decide to reject papers that have a probability of rejection above 0.25. Would the paper in part 1f get rejected now?

(h) You receive a phone call from a graduate of a prestigious ancient university complaining that his paper "World-beating method of predicting pandemics with absolute certainty" was rejected from the journal, and demanding an explanation. He doesn't know anything about logistic regression. If you were feeling nice, how would you explain to him how the algorithm works, in plain language?

2. **A/B testing**
Suppose you set up a social media advertising campaign for a holiday company in which you are trying to persuade people to come on holiday to Portobello, a suburb of Edinburgh with a beach. You have 2 images that look good: a picture of someone stretched out on a sun lounger on an empty beach and a picture of the beach filled with people partying. To make a quick decision about which image is best, you decide to target your ads in the UK, and, over a period of an hour, show 500 ads with one picture and 500 ads with the other.

(a) You got 224 responses from the picture with someone on a sun lounger, and 150 responses from the picture of the beach filled with people partying. Estimate the difference between the proportion of responses for each picture, and calculate a 95% confidence interval for the difference. What can you conclude about which picture was better?

(b) You decide to choose the picture of the sun lounger for worldwide distribution, but discover that over the next week the proportion of responses is much lower than in your trial. Why could this be?

3. **Hypothesis testing**
This question is based on the information in a report by the Minecraft Speedrunning team, though it does not go into all the depth of the report itself. There is also a lot of controversy online, referenced here.
In November 2020, a Minecraft player and streamer called *Dream* appeared to have especially good luck on a number of speed runs. In this context, a speed run is an attempt to obtain 12 *Eye of Enders* as quickly as possible, to open up the *Ender portal*. Many other players were suspicious and the Minecraft Speedrunning Team investigated to see if Dream was cheating.

(a) Part of the process of obtaining an Eye of Ender requires Minecraft players to obtain an *Ender Pearl* by trading with a creature called a *Piglin*. There is a 4.73% chance that a Piglin will give the player an Ender Pearl in return for a *Gold Ingot*. In six consecutive livestreams of speedruns, Dream received Ender Pearls on 42 out of the 262 times they traded a Gold Ingot with a Piglin. If we want to assess if Dream had exceptionally good luck trading with Piglins, what is the null hypothesis that we should test?

(b) Use the normal approximation to the binomial distribution to determine if the null hypothesis should be rejected, and if so at what level? Hint: you may need to use the `scipy.stats.norm.sf()` function in Python.

(c) Use the binomial distribution to determine if the null hypothesis should be rejected, and if so at what level? How does this compare to your answer to Question 3b using the normal approximation? Hint: you may need to use a Python function like `scipy.stats.binom.sf()`.

(d) Another part of the process of obtaining an Ender Pearl requires players to obtain *Blaze Powder*. To obtain Blaze Powder the player needs to kill a *Blaze Mob*, which produces Blaze Powder with a probability of 50%. Dream killed 305 Blaze mobs and obtained 211 Blaze Powders. Should we reject the null hypothesis that they obtained that many Blaze Powders by chance?