Inf2 – Foundations of Data Science 2024
Workshop solution: Semester 2 Week 6 Workshop

25th February 2025

1. **Logistic Regression**
   (a) $\hat{\beta}_0 = 1.0$. More likely to be rejected, since the log odds are above 0. Remember that
   $$\text{Log odds} = \ln \frac{P(\text{Success})}{P(\text{Failure})} = \ln \frac{\text{Reject=“Yes”}}{\text{Reject=“No”}}$$
   Somewhat counterintuitively, here “Success” equates to “Rejection” – but whenever we say “The odds of X”, it is X that counts as success.
   (b) $\exp(\hat{\beta}_0) = 2.718$
   (c) $p(\text{Reject}) = \frac{1}{1+\exp(-\hat{\beta}_0)} = 0.731$ probability of rejection
   (d) Predicted probability of rejection is:
   $$\begin{aligned} p(\text{Reject}) &= \frac{1}{1+\exp(-(\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \hat{\beta}_3 x^{(3)}))} \\ &= \frac{1}{1+\exp(-(1.0 + 0.5\times 5 - 0.5 \times 0 - 0.1 \times 0))} \\ &= \frac{1}{1+\exp(-3.5)} = 0.971 \end{aligned}$$
   (e)
   $$\begin{aligned} p(\text{Reject}) &= \frac{1}{1+\exp(-(\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \hat{\beta}_3 x^{(3)}))} \\ &= \frac{1}{1+\exp(-(1 + 0.5\times 0 - 0.5 \times 3 - 0.1 \times 2))} \\ &= \frac{1}{1+\exp(0.7)} = 0.332 \end{aligned}$$
   (f)
   $$\begin{aligned} \text{Log odds} &= \hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \hat{\beta}_3 x^{(3)} \\ &= 1 + 0.5 \times 1 - 0.5 \times 3 - 0.1 \times 1 \\ &= -0.1 < 0 \end{aligned}$$
   Since the log odds are less than 0 (which corresponds to probability of rejection equal to 0.5), we do not reject the paper.

(g) The probability of rejection equal to 0.25 is equal to log odds of $\ln 0.25/(1-0.25) = -1.097$, so this is the threshold. The log odds computed in the previous part now exceed this threshold, so the paper will be rejected.

(h) You could explain that when you used to review papers yourself a paper that does not contain any of the phrases "world-beating", "confidence interval" or "bootstrap" would have had a probability of 0.731 of being rejected, or, in other words, it was 2.718 times more likely to be rejected than accepted. You could then say that every extra occurrence of the word "world-beating" increased the odds of rejection by 1.65 times (i.e. $e^{\hat{\beta}_1}$), but that the word "confidence interval" reduced the odds by a factor of 1.65 ($e^{\hat{\beta}_2}$), and the word "bootstrap" reduced the odds by a factor of 1.11 ($e^{\hat{\beta}_3}$).

Alternatively, you could say that you've now implemented a scoring system that is implemented by weighting the number of occurrences of each word, and give the weights of each word and the threshold, which is $\ln 1/3 - \hat{\beta}_0 = -2.1$, assuming a probability threshold of 0.25, i.e. an odds threshold of 1/3.

2. **A/B testing**

(a) Let p_1 denote the proportion who responded to the sun lounger picture, and let p_2 who responded to the beach filled with people.

The sample estimates of the true proportions are $\hat{p_1} = 224/500 = 0.448$ and $\hat{p_2} = 150/500 = 0.3$.

The estimator of the difference between the sample proportions is $\hat{d} = \hat{p_1} - \hat{p_2} = 0.148$.

The estimated standard error of the difference is

$$\begin{aligned}\hat{\sigma}_{\hat{d}} &= \sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}} \\ &= \sqrt{\frac{0.448 \times (1-0.448)}{500} + \frac{0.3 \times (1-0.3)}{500}} \\ &= 0.0302\end{aligned}$$

Assuming a 95% CI and using a normal approximation to the binomial we compute the two-sided confidence interval with $\alpha = 0.05$ as:

$$\hat{d} \pm z_{\alpha/2}\hat{\sigma}_{\hat{d}} = 0.148 \pm 1.96 \times 0.0302 = 0.148 \pm 0.0593 = (0.0887, 0.2073)$$

The 95% confidence interval does not contain 0, hence we can conclude that the sample proportions are sufficiently different and that the campaign with the sun lounger picture is more successful.

(b) It might be that the time of day that you ran the initial trial had a different demographic online than for the week as a whole. Or perhaps people in the UK do not represent the worldwide population.

Note: it might still be the case that the sun lounger picture is more popular than the partying picture worldwide, but that the overall response from outwith the UK is lower than that from the UK. Given the average daily high temperature in

Portobello in July is 19.3°C, with an average of 5.5 hours of sunshine a day, it would not be surprising that Portobello is not a popular international destination. (Climate data from the *Gazetteer for Scotland*.)
In any case, more work (e.g. A/B testing on a non-UK and UK audiences is needed. In addition, the testing should be over a representative time period (e.g. a week) to control for potential biases due to time-of-day effects.

3. **Hypothesis testing**

   (a) H_0: The 42 out of 262 trades in which Dream received an Ender Perl arose from each trade having a probability of 4.73% returning an Ender Perl.
   H_a: The trades occurred via cheating which made it more likely that Dream received Ender Pearls.

   (b) The distribution implied by the null hypothesis is a binomial distribution with $p = 0.0473$ and $n = 262$ trials. As n is large we can approximate it by a normal distribution with $\mu = np = 12.3926$ and $\sigma^2 = np(1-p) = 11.8064$, so the standard deviation is 3.4360. We should do an upper tailed test, since the alternative hypothesis suggests that the process returns more Ender Pearls. The value of z is
   $$z = \frac{42 - \mu}{\sigma} = 8.6170$$
   To find the area in the upper tail, we need to compute $1 - \Phi(z) = 1 - \Phi(8.6170)$. A value of z this large isn't to be found in statistical tables. However, the scipy function[1] `scipy.stats.norm.sf()` is equal to $1 - \Phi(z) = 1 - \Phi(8.6170)$, so we compute:

   ```
   scipy.stats.norm.sf(8.6170)
   ```

   in python. The result is 3.446×10^{-18}, which is the chance that if the null hypothesis were true, 42 or more Ender Pearls from 262 trades with Piglins would result.

   (c) With the binomial distribution $b(x; n, p)$, we are looking for the number of successful trades X to be greater than or equal to 42, i.e. $P(X \geq 42; n, p) = \sum_{x=42}^{n} b(x; n, p)$. This is equivalent to $1 - \sum_{x=0}^{41} b(x; n, p)$, which is one minus the cumulative distribution function for the binomial distribution, $B(X \leq 41; n, p)$. The scipy function `scipy.stats.binom.sf()` is exactly $1 - B(X \leq 41; n, p)$, i.e. 1 minus the cumulative distribution function. We compute the value of $1 - B(X \leq 41; n, p)$ in Python like this:

   ```
   scipy.stats.binom.sf(41, n, p)
   ```

   This returns 5.65×10^{-12}, the probability of 42 or more out of 262 trades being successful under the null hypothesis. This probability is about 1.6×10^{6} times higher than the normal approximation, but still very low.
   Note 1: this result shows that the normal distribution is underestimates the weight in the extreme tails of the binomial distribution for this value of $n = 262$. We might expect that there should be good agreement, because of the Central Limit Theorem, but $n = 262$ is not large enough for the convergence guaranteed by the

[1] The "sf" stands for "survival function".

Central Limit Theorem for large n to hold. However, if we were to look at much less extreme tails, for example 2 standard deviations away from the mean, the area in the tails would agree much better between the binomial distribution and its normal approximation.

Note 2: Because n is fairly large, we can also use the Poisson distribution to model approximately the number of Ender Pearls obtained in Gold Ingot-Piglin trades. A Poisson process is meant to model events that can happen in continuous time in a time interval. However, the Poisson distribution can be derived by splitting time into n bins, with a probability of success of $p = \lambda/n$ in each bin; thus the expected number of successes across all trials is $np = \lambda$. Taking the limit of $P(x) = b(x; n, \lambda/n)$ as $n \to \infty$ gives $P(x) \to \mathrm{e}^{-\lambda}\lambda^k/k!$, i.e. a Poisson distribution. Conversely, the Poisson is a reasonable approximation to the binomial with reasonably large n. We can compute the value in the tail above 41 like this:

```
scipy.stats.poisson.sf(41, n*p)
```

giving a value of 3.38×10^{-11}, which is a factor of 6 *higher* than the binomial estimate, but still very low.

(d) With the binomial distribution, we follow the pattern above and calculate $1 - B(211 - 1; 305, 0.5) = 8.8 \times 10^{-12}$.