Stats assignment – mid Hilary Term 2024

This assignment is designed as revision of the topics covered this term. You should be able to find the answers to all conceptual questions in the lecture slides and/or the course notes on the website.

Please complete your answers in a Jupyter notebook.

Where questions require a verbal answer please use markdown cells not code cells!

Please answer each question part in a separate markdown cell and label each answer with the question number/letter.

Power

1. Which effect size requires more participants to detect, d=0.4 or d=0.5?
2. For an effect size of d=0.5, which alpha value requires more participants to achieve 80% power, 0.05 or 0.01?
3. Testing for an effect size of r=0.37, with a sample size of 50 pairs, what is the power (at alpha=0.05) for
   a. a one-tailed test and
   b. a two tailed test?

   Briefly explain the difference in power.
4. The heights of 10 English adult men and 10 Dutch adult men are measured. The means and standard deviations for the groups are as follows:
   - English: mean = 177cm, sd = 7.7cm
   - Dutch: mean = 184cm, sd = 6.5cm

   a. What is the effect size for the difference of means?
   b. What is the t-value (you can work it out using the equations!)
   c. A two-tailed t-test for difference of means is carried out at the alpha=5% level. The critical t-value is t=2.1 (that is, the test is significant if it is greater than 2.1)
      i. Assuming the null hypothesis is true, what is the probability of a false positive?
      ii. Assuming the population effect size is the same as the effect size in the sample, what is the probability of a false negative?
5. It is sometimes said that significant effects in studies with small sample sizes (underpowered studies) are more likely to be false positives than effects in larger studies. Can you explain why? Key terms to include in your answer are 'false positive', 'power' and 'file drawer effect' – if unsure about the last one (which was mentioned only briefly in the lecture), you can try Googling it!

Binomial and Normal distribution

1. A factory produced bags of mini eggs which contain 60 chocolate eggs. We can reasonably assume that the distribution of weights of the bags of eggs is Normal. Explain why, with reference to the central limit theorem.
2. The heights of British women are normally distributed with mean 166cm and standard deviation 6cm. What is the probability that, in a group of 50 women, at least three are over 175cm tall? (You need to use both the binomial and the normal distributions here)

3. The normal distribution is a good approximation to the binomial when np and nq are both greater than 5. Explain why the approximation breaks down when these conditions are not met (ie, when n is low, or when p is close to 0 or 1).
4. How do we define the 'best fitting normal distribution' when fitting a normal approximation to the binomial?

Probability Theory and Bayes

For a certain rare disease, 0.1% of the population are infected at any given time. A screening test for the disease measures the level of a certain protein in the blood of the patient. In healthy individuals, the protein level is normally distributed with mean 14300 and standard deviation 1540. In infected individuals, the protein level is normally distributed with mean 27000 and standard deviation 3320.

It is decided that if the protein level is above a certain cut-off, patients will be called back for a second test, which is conclusive but involves a painful biopsy.

1. Sketch the distribution of protein levels in both healthy and infected individuals on separate graphs.

2. Say I want to set the cut-off point such that 99% of infected individuals will be called back for further testing. What is the cut-off point?

3. Given the cut-off point defined in part b, what proportion of healthy individuals will be called back for further testing?

4. A given patient is called back for further testing. What is the probability he has the disease?

5. If one million patients are screened, how many would we expect to get an incorrect result from the screening test?

Consider the screening as a hypothesis testing problem in which the null hypothesis is that the patient is not infected:

6. Define a type I and type II error in this context

7. What would happen to the number of type I and type II errors if a lower protein level was used as the cut-off? Is this desirable?

8. Without calculation, state which would affect the greater number of patients, the increase in type I or type II errors? Give a reason.

9. What else would you need to know to decide whether to lower the cutoff point?

10. Say the prevalence of the disease in the population was 10%. Without calculation: would the proportion of individuals being mis-diagnosed change? Give a reason.