Ethics Part II - Week 4 Lab

Solution Sheet

Zak Varty

This lab will not focus on the practical analysis of randomised control trial data or of A/B test outcomes, though the latter of these will be covered as introductory material in the learning agents course. Instead, we focus of the ethical aspects of these study designs and what data scientists can learn from the ethical design of clinical trails.

Task 1: Outline the similarities and differences between Randomised Control Trials and A/B test. In doing so you might want to think about:

- The statistical problem being addressed
- How and where these techniques are used
- The ethical implications of each approach.

Model Solution:

Both RCTs and A/B tests aim to compare response values across two or more "treatment" groups to assess the effectiveness of an intervention. While in principle this testing could look for any change in the distribution of the responses, it is usually focused on specific summary statistic of interest: for example the expected response or a given quantile.

Randomised control trials are widely used across scientific disciplines and are particularly popular in medical research, where they are used to assess the validity of new medical treatments. In these cases the recruitment methods used and analyses performed are often strictly controlled and specified in advance. Conversely, A/B testing is more frequently used in business or industrial settings, for example where individual customers are shown one of two (or more) versions of a product or webpage to assess the impact this has on sales or click-through rate. While often also carefully designed, these AB tests often have much lower direct costs associated with them and do not always pre-register an analysis plan.

In both cases, there are incentives (ethical or financial) to allocate individuals to the "best" arm of the study given the information available so far - i.e. to give the most effective treatment or to show the layout that generated the most revenue. This lead to the parallel development of adaptive trail design and learning agent systems. A second ethical issue relating to A/B testing is ensuring users are aware and consenting to take part in the test, which might be less obvious to a participant than in a clinical trail.

Task 2: In a clinical trail an RCT might compare a new treatment to either no intervention, a placebo (a dummy treatment that is known to have no effect) or to the current standard treatment. Briefly describe the ethical implications of each approach and how these same concepts might translate to A/B testing.

Model Solution:

Comparing a new treatment to no treatment allows estimation of the full effect of that treatment, while comparison to a placebo allows the direct effect of the treatment to be measured (controlling for the the placebo effect of improved outcomes when any treatment is delivered, even if that treatment is not itself beneficial). A comparison between these three groups allows the direct and placebo components of the treatment effect to be identified. However, withholding an effective treatment from a patient might be considered unethical depending on the context (e.g. whether the treatment is for cancer or a mild skin condition).

Although the ethical considerations do not directly translate to AB testing in a business setting, we must still be wary of placebo effects. An increased click-through rate after moving a button might be because the new placement is better; however, it could equally be because the novelty of the new location draws attention and the effect will soon fade. In these cases delivering a “Placebo” treatment is not so obvious and we need to be careful to assess the permanence of any changes in click-through rate.

Task 3: In both RCTs and A/B testing there are different types of tests that we might want to conduct. Investigate and explain the differences between the following, giving an example of when each might be used:

- An equivalence trail
- A non-inferiority trial
- A superiority trial.

Solution:

- An **equivalence trial** is designed to demonstrate that two treatments or interventions have the same effect within a predefined margin of difference (an equivalence margin). The goal is to establish that the new treatment performs just as well (not better or worse) than the standard treatment. For this reason a common example is the development of generic versions of existing “branded” drugs. Another example might be comparing two different delivery methods for the same drug, to show that they achieve the same concentration profile in the blood: having either a lower concentration might lower the effectiveness of the treatment while a higher concentration might lead to more severe side-effects.

- A **non-inferiority trial** is conducted to demonstrate that a new treatment is not meaningfully worse than another treatment, often an active control. Similar to a an equivalence trial, this is formalised by defining "non-inferiority" margin. The aim is to establish that the new treatment is at least as effective as the comparator, even if it does not provide a superior outcome. An example of a non-inferiority trail might be the development of a new antihistamine, where we want to show that it is not meaningfully worse than the current offerings.
- A superiority trial is designed to determine whether a new treatment or intervention is meaningfully better than either a standard treatment or placebo. Superiority trials are typically conducted when there is a reasonable expectation that the new treatment will outperform the current best option. A common example might be demonstrating that a new treatment for a currently untreatable condition is better than placebo.

Task 4: When conducting an A/B test or RCT it is important that we consider enough individual cases that we have the statistical ability to detect a difference between groups, if such a difference really exists. (In the live session we will work through such a sample size calculation)

- What aspects of the test will influence the number of observations needed?
- What are the statistical consequences of failing to meet this sample size?
- What are the ethical consequences of failing to meet this sample size?
- What might make obtaining this sample size difficult?

We will discuss these points further during the live session.

Task 5: (Extension - suggest revisiting after live session.)

An online retailer is proposing a change to the location of the 'Buy Now' button on their website. The aim of this change is to increase the proportion of page visits that are converted to a sale by at least some amount $\Delta > 0$. Let p_1 and p_2 respectively denote the proportion of page visits that lead to sales under the current and proposed layouts.

To assess the benefits of the change, the retailer is planning an AB test. This will be applied to n customers, where half of these customers are shown each layout. You have been asked to help formalise this AB test and to advise on the required sample size to meet the power requirements for this trial.

(a) State both in words and mathematically the null and alternative hypotheses for this superiority trial.

Solution:

When considering a superiority trial the null hypothesis is that the two layouts are equivalent, while the alternative hypothesis is that the novel layout increases the sale proportion by at least the superiority margin, $\Delta > 0$.

Letting p_1 and p_2 respectively denote the proportion of page visits converted to sales under the current and proposed layouts:

$$H_0 : p_1 = p_2 = p \quad \text{vs.} \quad H_1 : p_2 \geq p_1 + \Delta.$$

(b) Let N_1 and N_2 denote the number of sales made under each layout within the AB test, while $n_1 = n/2$ and $n_2 = n/2$ denote the number of customers directed to each layout. State the sampling distributions of N_1 and N_2. Hence find expressions for the mean and variance of D, the change in the sample proportion of customers who make a purchase when the proposed layout is used rather than the current layout.

Solution:

To calculate the required sample size, we must consider the value of p_2 that requires the greatest possible sample size. That happens when p_1 and p_2 are most similar, i.e. when $p_2 = p_1 + \Delta$.

Let N_1 and N_2 denote the number of purchases made based on each layout, while n_1 and n_2 denote the number of customers directed to each layout. Then the conversion proportion for layout $i \in \{1, 2\}$ is given by:

$$P_i = \frac{N_i}{n_i} \qquad \text{where } N_i \sim \text{Binomial}(n_i, p_i).$$

Therefore letting $D = P_2 - P_1$, we have that:

$$\mathbb{E}[D] = \mathbb{E}\left[\frac{N_2}{n_2} - \frac{N_1}{n_1}\right] = \frac{n_2 p_2}{n_2} - \frac{n_1 p_1}{n_1} = p_2 - p_1.$$

Similarly,

$$\text{Var}[D] = \text{Var}\left[\frac{N_2}{n_2} - \frac{N_1}{n_1}\right] = \frac{n_2 p_2 (1 - p_2)}{n_2^2} + \frac{n_1 p_1 (1 - p_1)}{n_1^2} = \frac{p_2 (1 - p_2)}{n_2} + \frac{p_1 (1 - p_1)}{n_1}.$$

Since we are allocating customers evenly between layouts, this simplifies to

$$\text{Var}[D] = \frac{2 p_2 (1 - p_2) + 2 p_1 (1 - p_1)}{n}.$$

(c) By applying a Gaussian approximation to the sample sales counts, state the approximate sampling distribution of D under the null hypothesis and the most pessimistic version of the alternative hypothesis.

Solution:

If n, p_1, and p_2 are such that the Gaussian approximation to the binomial distribution is valid for each sample sales count, then the difference in sample sales proportions is also well approximated by a Gaussian distribution, since it is the weighted sum of approximately Gaussian random variables.

It follows that

$$D \mathrel{\dot\sim} \mathrm{N}\left(p_2 - p_1, 2\frac{p_2(1-p_2) + p_1(1-p_1)}{n}\right).$$

Under H_0, $p_1 = p_2 = p$ and this simplifies to:

$$D \mathrel{\dot\sim} \mathrm{N}\left(0, \frac{4p(1-p)}{n}\right).$$

Under the most pessimistic alternative hypothesis $p_2 = p_1 + \delta$, in which case

$$D \mathrel{\dot\sim} \mathrm{N}\left(\Delta, 2\frac{(p_1 + \Delta)(1 - p_1 - \Delta) + p_1(1-p_1)}{n}\right).$$

(d) Calculate an expression of the power of this AB test for given values of the type 1 error rate α, the type 2 error rate β, the sample size n, and the smallest meaningful difference δ. Hence find an expression for the minimum required sample size to achieve this power.

Solution:

To calculate the required sample size, we first construct an expression for the power of our superiority trail and then rearrange this expression to find the smallest required value for n.

$$\begin{aligned}
1 - \beta &= \Pr(\text{reject } H_0 \mid H_1 \text{true}) \\
&\geq \Pr(\text{reject } H_0 \mid p_2 = p_1 + \Delta) \\
&= \Pr\left(\left.\frac{D\sqrt{n}}{\sqrt{4p(1-p)}} > z_{(1-\alpha)}\right| p_2 = p_1 + \Delta\right) \\
&= \Pr\left(\left. D > z_{(1-\alpha)}\sqrt{\frac{4p(1-p)}{n}}\right| p_2 = p_1 + \Delta\right).
\end{aligned}$$

Using the approximate distribution for D under the most pessimistic alternative hypothesis, we have that:

$$
\begin{aligned}
1-\beta &= \Pr\left(Z > \left[z_{(1-\alpha)}\sqrt{\frac{4p(1-p)}{n}}-\Delta\right]\left[\sqrt{2\frac{(p_1+\Delta)(1-p_1-\Delta)+p_1(1-p_1)}{n}}\right]^{-1}\right)\\
&= 1-\Phi\left(\left[\Phi^{-1}(1-\alpha)\sqrt{\frac{4p(1-p)}{n}}-\Delta\right]\left[\sqrt{2\frac{(p_1+\Delta)(1-p_1-\Delta)+p_1(1-p_1)}{n}}\right]^{-1}\right)\\
&= 1-\Phi\left(\frac{\Phi^{-1}(1-\alpha)\sqrt{4p(1-p)}-\Delta\sqrt{n}}{\sqrt{2[(p_1+\Delta)(1-p_1-\Delta)+p_1(1-p_1)]}}\right).
\end{aligned}
$$

Rearranging this expression we find that

$$
n \geq \left(\frac{\Phi^{-1}(1-\alpha)\sqrt{4p(1-p)}-\Phi^{-1}(\beta)\sqrt{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}}{\Delta}\right)^2.
$$

or equivalently:

$$
n \geq \left(\frac{\Phi^{-1}(1-\alpha)\sqrt{4p(1-p)}+\Phi^{-1}(1-\beta)\sqrt{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}}{\Delta}\right)^2.
$$

(e) The retailer currently expects around 7% of page visits to convert to sales and would consider an increase of 0.5% or greater to be a meaningful improvement in sales conversion. Calculate the minimum required sample size for this AB test, given that the retailer will accept a 10% chance of falsely identifying a meaningful improvement and a 5% chance of failing to identify a meaningful improvement when one truly exists.

Solution:

Substituting $\alpha = 0.1$, $\beta = 0.05$, $\Delta = 0.005$ and $p_1 = 0.07$, we find that $n \geq 90841.43$. Therefore this superiority test requires a sample size of at least 90,842 customers to guarantee a power of at least 0.95.