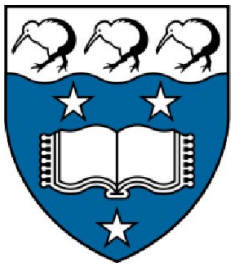
Binomial and Survival Models for Bayesian AB Testing



Hannah Jamieson

Department of Statistics

The University of Auckland

Supervisor: Brendon J. Brewer

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistics, The University of Auckland, 2020.

Abstract

Bayesian AB testing is becoming a popular alternative to Frequentist methods and p-values. Current methods of Bayesian AB testing in industry underutilise the ability to fully model prior distributions and likelihoods to the experiment problem. In this report, we propose using a dependent prior for conversion parameters between treatment groups and develop likelihoods to model time to event outcomes. Additionally, we demonstrate how using the prior expectation of the posterior expected loss can be used to estimate the "cost" of picking a poor prior distribution in small sample sizes.

Acknowledgements

I would like to express my deepest gratitude to Brendon Brewer for supervising my Masters project this year. Thank you for helping me turn my many ideas into Bayesian inference. I have been extremely lucky to have a supervisor who was very invested and involved in my work. Sorry about all the bugs.

To my dearest fiancé who once said to me *"Stop making things boring with statistics"*, this is for you. I hope it doesn't bore you too much. I couldn't have done it without you.

Last but not least, mum and dad. Thank you for always supporting my love of science.

Contents

Chapter 1

Introduction

In the software industry, to *test, learn and iterate* has become a strategic pillar for software development. This is an ideal that can be applied across a range of uses in a company. From the data practitioner's standpoint, the heart of the *test, learn, iterate* method is the AB test; the ability to conduct an experiment to derive causal results from the change made.

An AB test is a type of Randomized Control Trial (RCT), with this particular name generally denoted for used in software applications. In industry the AB test is typically used to gain understanding of the effect a change has a user. This could be the users' experience in the software or their status as a customer. This could include, but is not restricted to, changes to the user interface, the features of a software product, improvements in machine learning predictions or the methods of outbound communication.

Bayesian AB testing has been a become popular alternative to Frequentist hypothesis testing and p-values. However it is often limited to:

- Conjugate examples with very little explanation of priors.
- Single point estimates such as proportions or averages.

This project endeavours to bring additional value to Bayesian AB testing by developing models that look more deeply into prior specification and utilise time-to-event outcomes. We also give more context on AB testing by showing the variety of use cases and examples these models might apply to, plus the ability to customise models to match the design of the AB test.

1.1 Randomized Controlled Trials

A randomized controlled trial (RCT) is an experiment in which subjects are randomly assigned to one of two groups: one (the experimental group) receiving the intervention that is being tested, and the other (the comparison group or control) receiving an alternative treatment. The two groups are then followed up to see if there are any differences between them in outcome. The results and subsequent analysis of the trial are used to assess the effectiveness of the intervention (Kendall [2003]). Figure 1.1 shows a diagram of the RCT design:

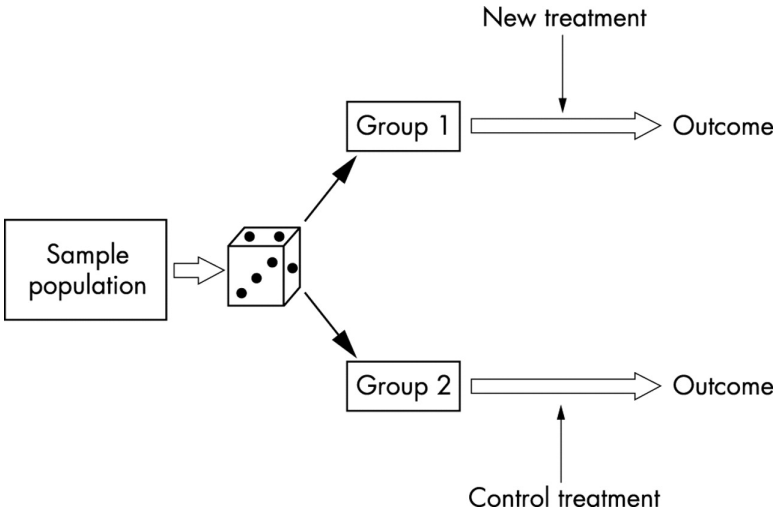


Figure 1.1: *The randomised control trial (Figure from Kendall [2003])*

An RCT is the most rigorous way to determine whether a cause-effect relation exists between the intervention and the outcome. As such RCTs are used in clinical trials as the best way to evaluate the safety and efficacy of new treatments. This application is so well known that terms RCT and clinical trials are almost used interchangeably. Other study designs, including non-randomised controlled trials, can detect associations between an intervention and an outcome, but they cannot rule out the possibility that the association was caused by a third factor linked to both intervention and outcome (Sibbald and Roland [1998]).

The RCT achieves the aforementioned rigour by controlling for as much bias as possible. The two methods that used are:

1. **Randomisation** - A valid comparison between treatment groups depends on the groups being as similar as possible, except for the treatment under investigation.

The best way to achieve this is randomization in which a chance mechanism determines the treatment assignment (Stanley [2007]).

2. **Blinding** - If the treatment assignment is known to the investigator or patient then bias can be introduced for a variety of reasons; investigators treat groups differently, or patients may observe or experience effects differently due to the placebo effect, subjects don't comply with treatment. Double blinding is the practice to control this bias by neither the investigator or patient knowing what treatment they have received (Kendall [2003]).

Together these make up the double-blinded RCT which generally accepted as the gold standard for clinical trials.

Frequentist Analysis of RCTs

Frequentist analysis of RCTs involve single parameter tests such as t-test, chi-square test or statistical models such as regression or ANOVA if there are additional covariates in the study. Each of these proposes a null and alternative hypothesis and calculates the Frequentist p-value: *"The probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is correct"* to determine if the difference between groups is due to the treatment of interest and not just by chance.

At the very beginning of an RCT, the study organisers will decide on a minimum detectable effect. The minimum detectable effect is the effect size that represents the relative minimum improvement we hope to see over the control. The idea of "clinical relevance" determines this what effect size for RCT in clinical trials should be. A clinically relevant intervention is the one whose effects are large enough to make the associated costs, inconveniences, and harms worthwhile (Armijo-Olivo [2018]).

The p-value depends on effect size, sample size and the power of the test. Statistical power is the probability of a hypothesis test of finding an effect if there is an effect to be found. As such, a test must have sufficient power to detect a given effect size. Power analysis is done at the start of the clinical trial to estimate the minimum sample size required for the effect size at the desired significance level (usually at least 0.95) and statistical power (usually at least 0.8). If the required sample size is met any effect size smaller than the minimum detectable effect will not be statistically significant, and any effect size greater will be statistically significant (barring the loss of sample due to non-compliance and loss to follow up).

False Positive Rate & Interim Analysis

When a statistical hypothesis test produces significant results, there is always a chance that it is a false positive. The significance level, α, controls the false positive rate. An α of 0.05 means that if we ran this experiment 100 times, five or 5% of the trials we obtain a statistically significant p-value by chance alone. We would unknowingly reject a null hypothesis that is true and falsely conclude that an effect exists in the population when it does not. We only get to run our trial once, so we are usually satisfied by a false positive rate of 5%.

AB tests and clinical trials meet sample sizes through enrolling subjects into the trial over time. It can often take a significant amount of time to meet the required sample size, which makes it tempting to analyse results before the end of the trial. The problem is that every comparison we make has a chance of showing a false positive. As such if we make any additional comparisons during the experiment, we increase the probability of getting a false positive.

However, there are methods for interim analysis to control for the false positive rate and maintain an overall significance level. The methods described by Pocock (Pocock [1977]) and O'Brien & Fleming (O'Brien and Fleming [1979]) are popular implementations of group sequential testing for clinical trials. Study organisers will plan for the interim analysis, so we are not penalised by looking at results early by seeing differences that exist just due to chance. Figure 1.2 demonstrates this by plotting the p-values 100 simulated experiments from days 50 to 100.

1.2 AB Testing

An AB test is a type of RCT used for web or software applications (app). A standard definition is an experiment where two or more variants of a webpage or app are shown to users at random to determine which one performs better (Optimizely [2019]). In this report, we will extend the definition of AB testing to any RCT a software company would conduct to see how a change would affect its users.

AB tests help to make better business and product decisions by testing interventions before they are published site-wide. Only the interventions that are proven to be effective are worth implementing. The most common use for AB testing is to test small incremental changes to optimize an outcome. As such, AB tests are run for short durations at a high frequency.

The primary outcomes for AB tests are commonly called conversion metrics. A conversion can refer to any desired action that the business would want the user to take.

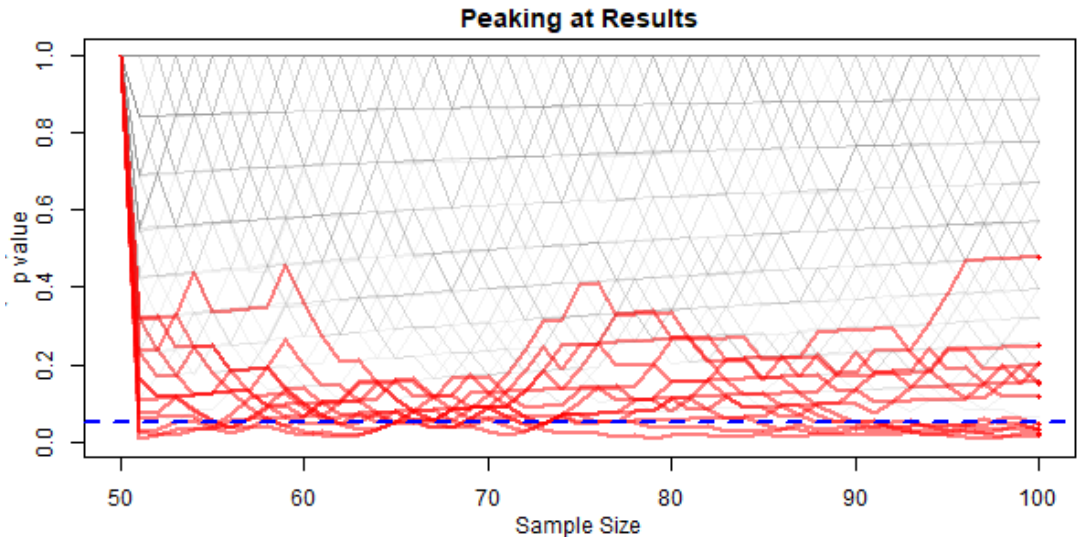


Figure 1.2: *One hundred simulations of two treatment groups, each with a success probability of* 0.5*. This plot shows how the p-value changes as more subjects enter our experiment. If we calculate a p-value at halfway through the experiment at* $N = 50$ *in each group, the red lines are effect sizes we would have accepted at the* 5% *significance level. By the end of the experiment with* $N = 100$*, we calculated the p-value again, and conclude only 5 (as expected from* $\alpha = 0.05$*) to be statistically significant. In this simulation, if we stopped early at* $N = 50$ *we were twice as likely to get a false-positive result.*

These actions can include anything from a click on a button to making a purchase and becoming a customer. A standard measure for this outcome is a Conversion Rate; the total number of conversions divided by the total sample size.

Applications of AB testing

Here we introduce our motivating example to guide how one might use our Bayesian AB testing models in a company.

Motivating Example

Company XYZ is a software website offering a cloud based services. They charge a monthly subscription to provide unlimited access to content and services. Company XYZ also offers a free 1 month trial before signing up. Company XYZ, like most businesses, wants to increase revenue and total customers. The steps that users take to move from being a prospect to a customer is called a marketing funnel. The stages for a marketing funnel typically include:

1. Awareness - *landing on the website*
2. Consideration - *Signing up to a free trial*
3. Conversion - *Becoming a paying customer*
4. Retention - *Renewing subscription each month*

At each step of the funnel the number of prospective customers is expected to drop. AB testing provides a method to test what interventions or changes improve the conversion rate at each funnel step.

Resources on how to conduct an AB test are almost exclusively for web-based experiments only. However, any steps that users take to move from being a prospect to a customer can be experimented on, and it is standard to call any such experiment an AB test. The web-based AB test design does not fit for an RCT done using outbound contact (such as email marketing or sales calls) because it is (a) randomised and allocated into groups differently; and (b) subject to non-response.

We will propose two classifications of AB tests based on their application:

1. **Web Based AB Testing** - We define this as users entering an experiment by landing on a web page with randomisation to a variant done upon landing. The experiment continues until there has been enough traffic (web users who visit a website, usually measured in visits or sessions) to meet the sample size required or a winning variant can be determined.
2. **Outbound AB Testing** - We define as preselecting a sample from a cohort of users and randomising into treatment groups. Then the intervention is sent out to all users in the treatment group (or none if the control group is to get no treatment).

Table 1.1 highlights the general difference between the two applications. Of course,

there can be exceptions and hybrid designs.

Table 1.1: *Differences between Web and Outbound AB Testing*

	Outbound	**Web**
Treatment	Contact such as phone calls or emails	Changes to user interface or feature
Audience	Target & Personalization	What works best for everyone
Primary Outcomes	Conversions/Actions	Clicks/product use
Choice of Control	Existing treatment or No treatment	Existing Experience
Treatment Allocation	Ahead of time	When users land on page
Non Response	Yes	No
Duration	When sample has been contacted	When traffic meets sample size
Outcome Event	Can happen any point in lifetime	Cannot happen before experiment. Should happen soon after landing on page.

1.3 Bayesian Inference

When we observed a data set sampled from a larger population, there is variability in the data and uncertainty about the truth. Statistical inference is about applying consistent reasoning to infer the truth when not all the data is available. Bayesian statistics is a particular approach to applying probability to make statistical inferences. We can express our uncertainty as probability and update our subjective beliefs in light of new data or evidence (O'Hagan et al. [2004]). In particular Bayesian inference interprets probability as a measure of believability an individual has about the occurrence of a particular event or the plausibility of a hypothesis. This is in contrast to Frequentist statistics, which assumes that probabilities are the long-run frequency of events from repeated trials.

In order to carry out Bayesian inference, we utilise Bayes Rule. To derive Bayes' rule, we start with the definition of conditional probability, which gives us a rule for determining the probability of an event A, given the occurrence of another event B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.1}$$

This states that the probability of A occurring given that B has occurred is equal to the probability that they have both occurred, relative to the probability that B has occurred. This is a rearrangement of the product or chain rule $P(A \cap B) = P(B|A)P(A)$. For the joint probability of A and B both occurring we can write:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ P(B \cap A) &= P(B|A)P(A) \\ \therefore P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \end{aligned} \tag{1.2}$$

Bayes' Rule for Bayesian Inference

To use Bayes Rule for Bayesian inference we use a modified version of equation (1.2) above that represents the process of stating prior beliefs and updating them in the face of new data.

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{P(D)} \tag{1.3}$$

- We start with a parameter of interest θ which we specify prior distribution for. The prior distribution represents our initial belief about the parameters and can be written as $p(\theta)$[1].
- Then we must consider the data D and the probability of seeing the data as generated by a model with parameter θ. This is expressed as the likelihood[2] and is written as $p(D|\theta)$.
- Lastly the probability of the data D itself is called the marginal likelihood or evidence $P(D)$. This is determined by summing (or integrating) across all possible values of θ, weighted by our prior beliefs about for each value of θ[3].

[1] Here lower case p represents a probability distribution, while uppercase P represents a single probability, i.e. the chance of an event occurring. We will use this notation throughout the report.

[2] The likelihood can be thought of as the probability getting the data we have assuming our hypothesis is true. A likelihood function associates the probability of an observed value $X = x$ for a value of θ.

[3] This is the same as the expectation $E[p(D|\theta)]$.

From this we get the posterior distribution $p(\theta|D)$. This is the updated strength of our beliefs in the possible values of θ once the evidence D has been taken into account.

Calculating the Posterior

The evidence or marginal likelihood $P(D)$ is formally written as:

$$p(D) = \int p(\theta)p(D|\theta)\ d\theta \tag{1.4}$$

When we have one or two parameters we can calculate the posterior analytically either by integrating the marginal likelihood or using conjugate priors[4]. In higher parameter spaces it can be difficult to obtain $P(D)$ so the posterior is often simplified to the un-normalised posterior.

$$P(\theta|D) \propto P(\theta)P(D|\theta) \tag{1.5}$$

However, without the normalising constant we cannot make the posterior distribution an actual probability distribution (that integrates to one). In Bayesian statistics, the posterior distribution has to be a probability distribution, from which one can derive moments like the posterior mean. When analytical methods are not appropriate, we can instead use simulation methods to generate samples from our posterior distribution for us to make inference from.

1.3.1 MCMC

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution. When the posterior distribution cannot be solved analytically, we can use MCMC to generate random samples of parameter values drawn from the posterior distribution, $p(\theta|x)$ which we will call the "target" distribution. The MCMC algorithms work by simulating a Markov Chain, whose stationary distribution is $p(\theta|x)$. The term stationary just means that in the long run, the samples from the Markov chain look like the samples from $p(\theta|x)$.

[4]If the posterior distribution is in the same family of the prior distribution, then the prior and posterior are called conjugate distributions, and the prior is called the conjugate prior to the likelihood function. This gives us a a closed-form expression for the posterior avoiding numerical integration.

Metropolis Hastings

Metropolis-Hastings is one type of MCMC algorithm. If we know a function $f(\theta|x)$ is proportional to the desired posterior distribution (i.e. the un-normalised posterior) then Metropolis-Hastings can draw samples from the probability distribution $p(\theta|x)$. Metropolis Hastings iteratively produces sample values with the distribution of the next sample being dependent only on the current sample value (thus making the sequence of samples into a Markov chain). As the algorithm produces more sample values, the closer the distribution of values approximate the desired distribution. The iterative method is:

1. The iteration starts by the algorithm picking a candidate $\boldsymbol{\theta^*}$ for where to go next given that you are at $\boldsymbol{\theta_n}$
2. Accept the move to $\boldsymbol{\theta^*}$ at $n+1$ with acceptance probability $\alpha(\boldsymbol{\theta^*}|\boldsymbol{\theta_n})$ or reject $\boldsymbol{\theta^*}$and stay where you are (in which the current value is reused in the next iteration).

The acceptance probability is determined by comparing the values of the function $f(\boldsymbol{\theta}|x)$ of the current and candidate sample values with respect to the desired distribution $p(\boldsymbol{\theta}|X)$. The acceptance probability for a symmetric proposal distribution is:

$$\alpha(\boldsymbol{\theta^*}|\boldsymbol{\theta_n}) = \min\left\{1,\ \frac{f(\boldsymbol{\theta^*}|x)}{f(\boldsymbol{\theta_n}|x)}\right\} \tag{1.6}$$

For our examples in chapter 2 we will use MCMC using the `JAGS` program in `R` to do Bayesian inference.

Posterior Summaries

The result of running MCMC is instead of having a probability distribution of values of θ and corresponding densities, there is a large dataset of sampled parameter values. The more probable regions will contain more data points so we can look at the posterior distribution by plotting a histogram of samples. We can summarise θ by calculating descriptive statistics such as the mean, median or standard deviation of the samples. Probabilities like $P(\theta \geq 0.5)$ are calculated by counting all the samples with $\theta \geq 0.5$ and dividing by the total number of samples.

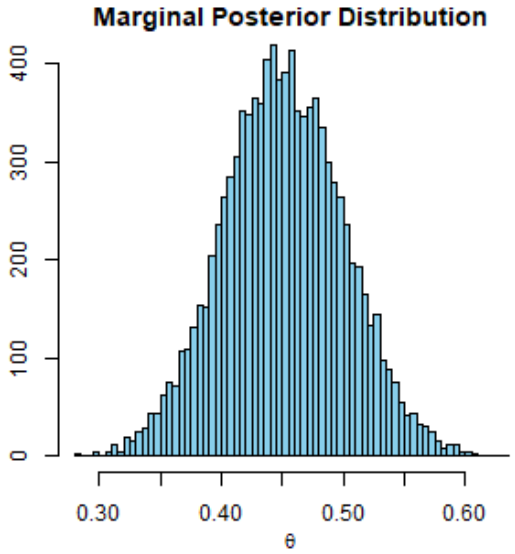


Figure 1.3: *Histogram of posterior samples from MCMC*

This method works well on more complex problems when we have many parameters. Usually if we wanted to infer two parameters θ_1 and θ_2 we would have the joint posterior distribution:

$$p(\theta_1, \theta_2|D) \propto p(\theta_1, \theta_2)p(D|\theta_1, \theta_2)$$

and inferring the value of θ_1 on its own would require the marginal posterior distribution for θ_1 (that is, the posterior distribution for a on its own, not the joint distribution with θ_2), which we can get by summing over all values of θ_2:

$$p(\theta_1|D) = \int p(\theta_1, \theta_2|D)\ d\theta_2 \tag{1.7}$$

Having posterior samples makes the process of marginalisation much easier as MCMC already returns the marginal distribution of each parameter for if all other parameter were ignored. We can easily plot the distribution or make inference about a single parameter using the marginal samples provided.

1.3.2 Nested Sampling

Nested sampling by physicist John Skilling (Skilling [2004]) is another algorithm which we can use to get posterior samples. The actual goal of nested sampling is the marginal

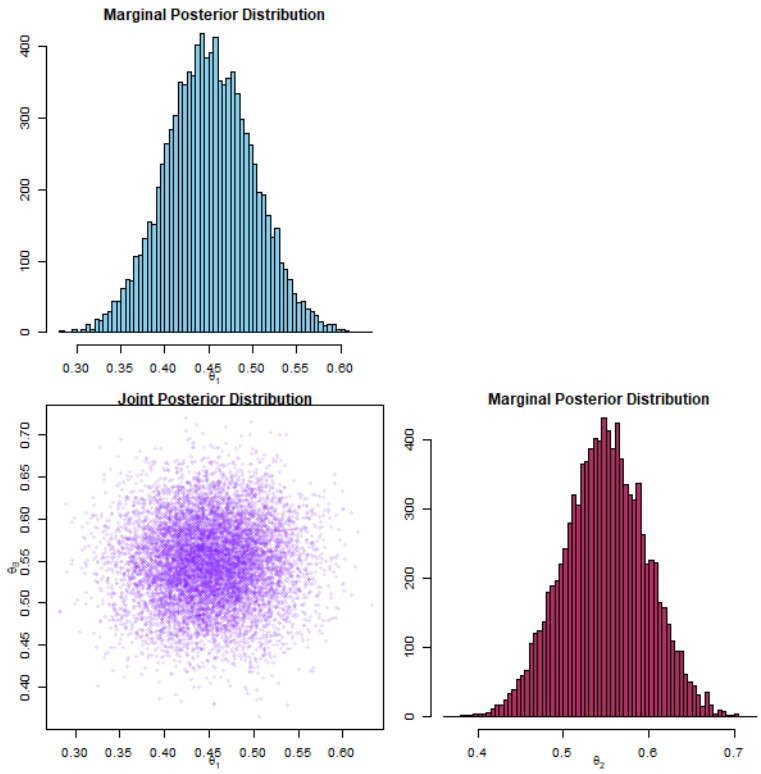


Figure 1.4: *Histogram of marginal distribution and plot of joint distribution from MCMC samples*

likelihood or evidence. Samples from the posterior distribution are a useful by-product of the algorithm, meaning it can be used for model comparison as well as for inference.

When discussing nested sampling, the posterior distribution is written as

$$Z = \int \pi(\theta)\ L(\theta)\ d\theta \tag{1.8}$$

Where Z is the marginal likelihood, $L(\theta)$ is the likelihood function and $\pi(\theta)$ is the prior distribution.

Nested sampling uses the idea that a high dimensional problem can be mapped on to onto the 1-D problem, using the CDF of likelihood values implied by $\pi(\theta)$. Nested sampling then uses this to estimate how the likelihood function relates to prior mass.

Then from approximating a weighted average of likelihood values, the area under the

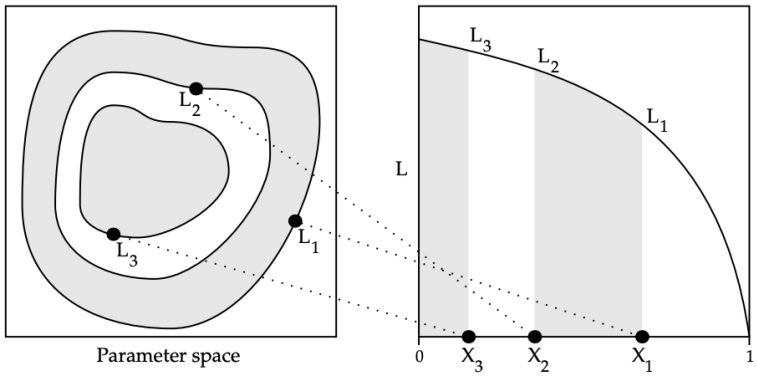


Figure 1.5: *Figure fromSkilling [2004]: Nested likelihood contours are sorted to enclosed prior mass X.*

curve can be summed, and the marginal likelihood is obtained. The marginal posterior distributions for θ can also be obtained by taking weighted samples of θ from the nested sampling run (Skilling [2004]).

Model Selection

The marginal likelihood is useful for model selection —when we wish to compare the strength of evidence of one model over another based on observed data D. We can use Bayes' theorem to express the posterior probability $P(m|D)$ of a model m given data D

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)}$$

We can extend Bayes rule in equation 1.3 to more formally consider the assumptions in the model we are using as part of the posterior calculation. Consider two models m_1 and m_2 with parameters θ_1 and θ_2 respectively. The marginal likelihoods are:

$$\begin{aligned} P(D|m_1) &= \int p(\theta_1|m_1)p(D|\theta_1, m_1)\ d\theta_1 \\ P(D|m_2) &= \int p(\theta_1|m_2)p(D|\theta_2, m_2)\ d\theta_2 \end{aligned} \tag{1.9}$$

The models can be compared using Bayes Factor—the ratio of the likelihood of one particular hypothesis to the likelihood of another.

$$\begin{aligned}\frac{P(D|m_1)}{P(D|m_2)} &= \frac{\int p(\theta_1|m_1)p(D|\theta_1,m_1)\ d\theta_1}{\int p(\theta_1|m_2)p(D|\theta_2,m_2)\ d\theta_2} \\ &= \frac{\frac{P(D|m_1)P(m_1)}{P(D)}}{\frac{P(D|m_1)P(m_1)}{P(D)}} \\ &= \frac{P(m_1|D)}{P(m_2|D)}\frac{P(m_2)}{P(m_1)}\end{aligned} \tag{1.10}$$

The Bayes Factor can be interpreted as a measure of the strength of evidence in favour of one model among two competing models. It can also be derived through the posterior odds, which is the prior odds times the Bayes factor.

$$\begin{gathered}\frac{P(m_1|D)}{P(m_2|D)} = \frac{P(m_1)}{P(m_2)} \times \frac{P(D|m_1)}{P(D|m_2)} \\ \text{(posterior odds)} = \text{(prior odds)} \times \text{(bayes factor)}\end{gathered} \tag{1.11}$$

We can see that if we use nested sampling, we will have marginal likelihoods readily available to do model comparison. Presenting the marginal likelihood, along with the results, means that any future models can be compared with the current one, without having to redo any analysis. The marginal likelihood is also beneficial to provide in published research as it allows the reader to perform model comparison between studies.

In chapter 3 we will use nested sampling to get samples from our posterior distributions to do Bayesian inference. We also provide the marginal likelihoods for any future model comparison.

Chapter 2

The Binomial AB Testing Model

2.1 Bayesian AB Testing

Now we are ready for Bayesian AB tests. Suppose we have an experiment with two treatments, control group (group A) and treatment group (group B), we can compute a posterior distribution for the parameters describing the conversion rates for each treatment, $P(\theta_A|D_A)$ and $P(\theta_B|D_B)$. However, as we are analysing these two parameters together we can actually specify a joint prior and joint posterior:

$$\begin{aligned} p(\theta_A, \theta_B) &\propto \dots \\ p(\theta_A, \theta_B|D) &\propto p(\theta_A, \theta_B)p(D|\theta_B, \theta_A) \end{aligned} \tag{2.1}$$

Most AB testing methods will use independent priors and treat the two marginal posterior distributions, $P(\theta_A|D_A)$ and $P(\theta_B|D_B)$ as independent distributions (this just means the joint distribution is the product of the two marginal distributions). In the prior we propose in section 2.2, we show how we can use a joint prior to specify dependence between θ_A and θ_B. The joint distribution lets us understand the relationship between two variables. We can use both the joint and marginal distribution to calculate various quantities of interest.

Binomial Likelihood

A binomial distribution models the probability getting x number of successful outcomes in an experiment that has N number of independent trials:

$$p(x|\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{2.2}$$

- N stands for the number of trials in the experiment.
- x stands for how many successes.
- θ represents the probability success for one specific outcome $1-\theta$ represents the probability of failure (i.e. that outcome does not happen).

For example: you could toss a fair coin 100 times and record the two possible outcomes (heads or tails) to calculate the probability of getting extactly 50 heads. Or a classroom taking a test can have two possible outcomes (pass or fail) and the teacher can calculate the proportion of students that pass the test. Both are examples of binomial likelihoods.

It is natural then that conversion metrics that measure the proportion of users who do a conversion event are also binomial –as the user will either do the event or not (e.g. sign up, page click). So in any Bayesian AB test that has a binary outcome we will use a binomial likelihood.

Sample Size

In the Bayesian framework we can perform analysis with with any number of data points and still make correct inferences. However, this is not a magical solution to sample sizes required in Frequentist AB testing. The fewer data points we have the wider our posterior distributions from each group will be. This can affect how certain we are that one treatment is better than the other. We may still need a minimum sample size to obtain the required certainty to declare one treatment a winner.

AB testing is different from other analysis because we collect data over the duration of the experiment, rather than having a fixed data set. Our estimate of the conversion rate may vary over this duration. However, as more data is collected we can expect the conversion rate to converge closer to the true value with a narrower posterior as we have more information on θ.

Some criticism against Bayesian statistics is the influence of a subjective prior on the posterior distribution. The great thing is, the more data we have the less the prior information becomes relevant and the more it relies on the likelihood (and observed data). We do not need to be as concerned about priors with larger data sets as the results will be less sensitive to the choice of prior distribution.

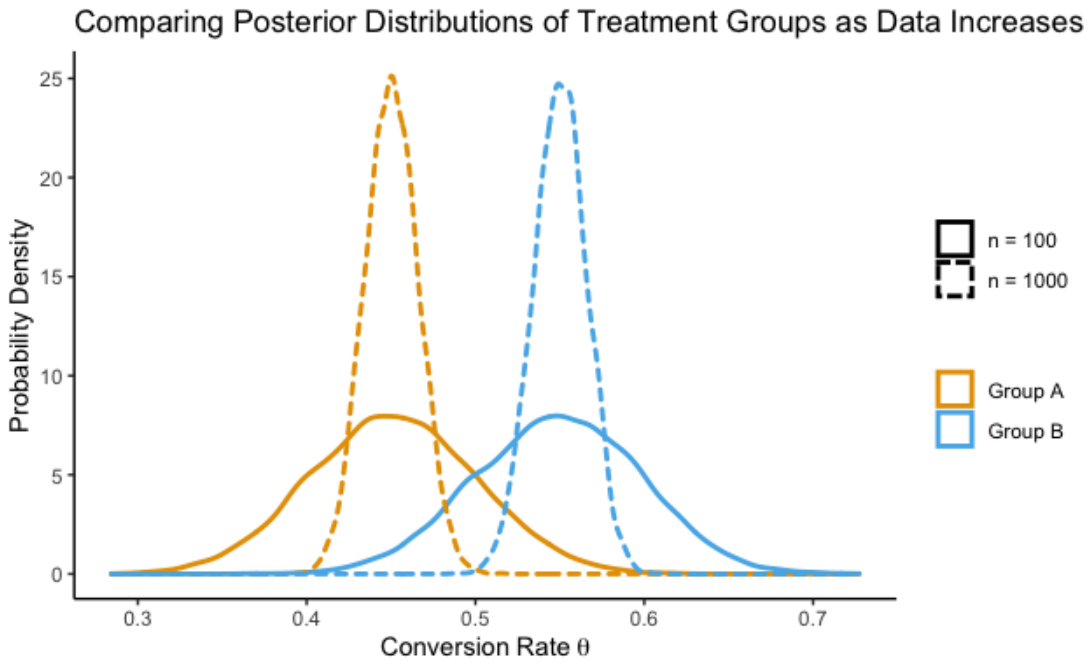


Figure 2.1: *Posterior distributions of θ_A and θ_B. With less data points our posterior distributions are wider and overlap each other making it harder to declare a clear winner. With a larger sample size we can see that θ_B is always greater than θ_A and we should pick treatment B as the winner.*

2.2 Prior

To begin our Bayesian inference we must decide our prior distributions for θ_A and θ_B; $p(\theta_A)$, $p(\theta_B)$. In this report we propose a dependent prior between treatment groups. Because one treatment is usually a variation of the other, we use this prior to represent our beliefs that:

1. That we expect the success probability of a new treatment to be an improvement or worsening of the old treatment and hence dependent on each other.
2. We expect that it is much more likely that θ_A and θ_B will be close and that differences between them to be small, then seeing large differences.

We set a normal prior distribution for the logit of of θ_A. Then we use a T distribution for its heavy tails to model the difference between treatment groups such that θ_B becomes

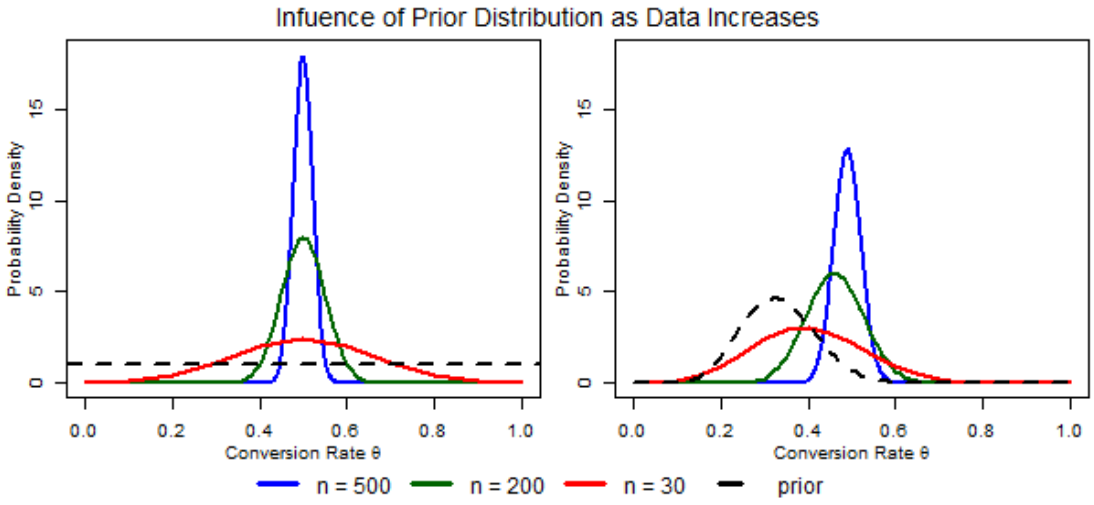


Figure 2.2: *Simulations of estimating θ at different type points. The posterior becomes sharper as data streams in.*

θ_A plus the difference. The logit transformation to accounts for any boundary issues outside of $[0, 1]$. The prior distribution is shown in equation 2.3.

$$\begin{aligned} \text{logit}\Big(\theta_A\Big) &\sim N(\mu = 0, \sigma = 1) \\ \text{logit}\Big(\theta_B\Big) &\sim \text{logit}(\theta_A) + t(\nu = 1, \lambda = 0.1) \end{aligned} \tag{2.3}$$

$$\begin{aligned} \theta_A &= \text{expit}\Big(\text{logit}\Big(\theta_A\Big)\Big) \\ \theta_B &= \text{expit}\Big(\text{logit}\Big(\theta_B\Big)\Big) \end{aligned}$$

We used a $N(0, 1)$ for the logit prior and a shape of $\nu = 1$ and scale of $\lambda = 0.1$ for the T distribution. We use this as standard set up for this prior because it reflects our beliefs for what we expect in our AB testing examples. For their own analysis the practitioner can adjust the distribution of the θ's and the differences to reflect their own prior beliefs.

Usually in Bayesian AB testing with a binomial likelihood uniform or beta priors are used. A uniform prior puts the same probability density of θ being anywhere from 0 to 1. We can adjust the shape and scale of a beta prior to put the peak of the density

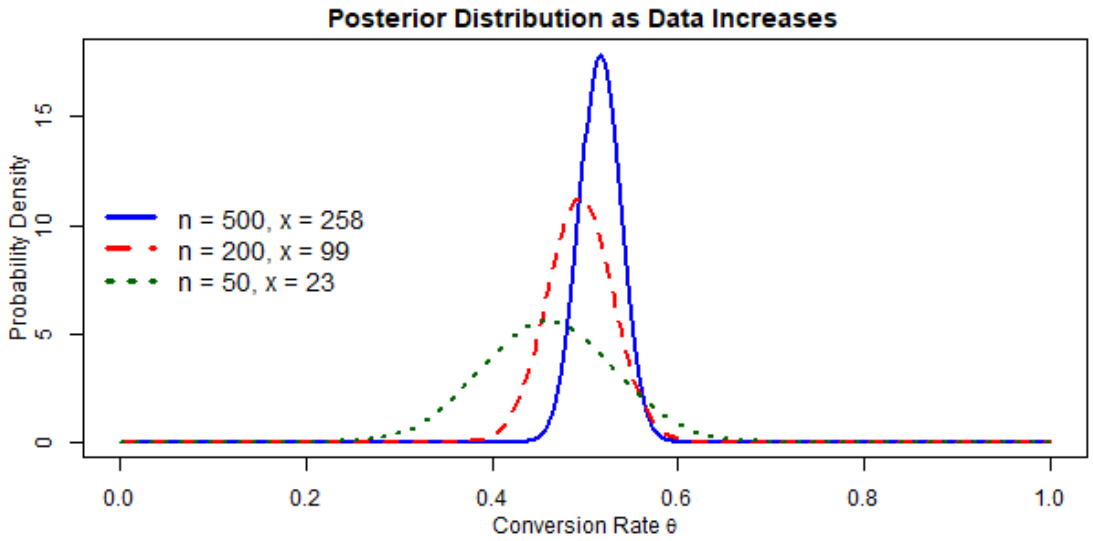


Figure 2.3: *The posterior distribution for two different priors for* θ *with a binomial likelihood each with the same success probability of* 0.5*. One is a uniform distribution putting equal probability of* θ *being between* 0 *and* 1*, the second is a Beta distribution putting the most prior probability around* $\theta = 0.3$*. With a large sample size both posterior distributions end up about the same.*

where we expect θ to be and the width to how certain we are about it. However, using independent priors still puts a greater amount of probability on larger differences than we deem realistic. When perform an AB test we generally have a decent idea of the effect sizes we can expect and that they will be small - otherwise we have been previously doing something very wrong! Figure 2.4 shows the shape of our prior distribution compared to uniform priors in figure 2.5 and beta priors in 2.6. Our dependent priors put a lot more prior probability on small differences (i.e. on the diagonal) compared to pair of beta priors that are effectively the same. The uniform prior simply treats all differences, large and small, as the same.

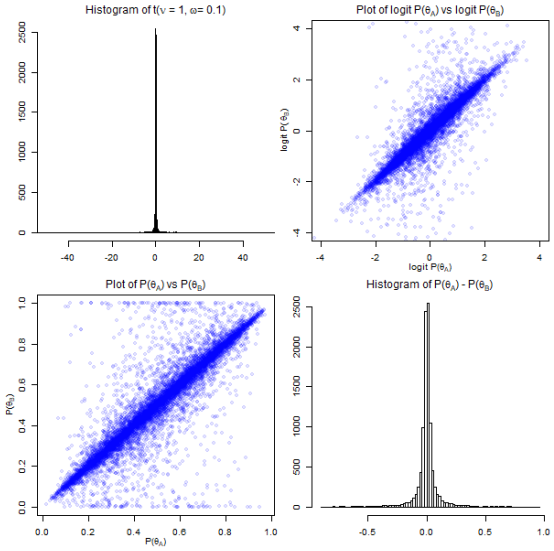


Figure 2.4: *Plots showing the prior from equation (2.3) for our beliefs on the joint distribution between θ_A and θ_B and the difference between them.*

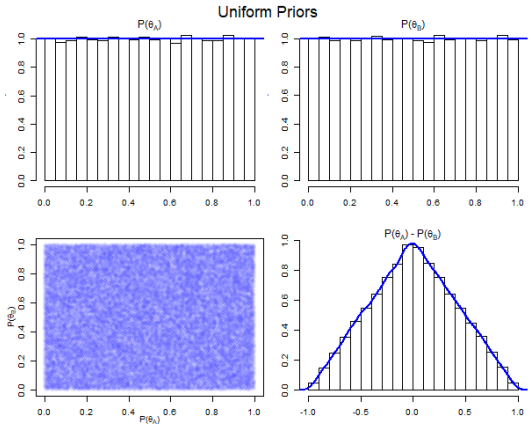


Figure 2.5: *Uniform(0,1) priors*

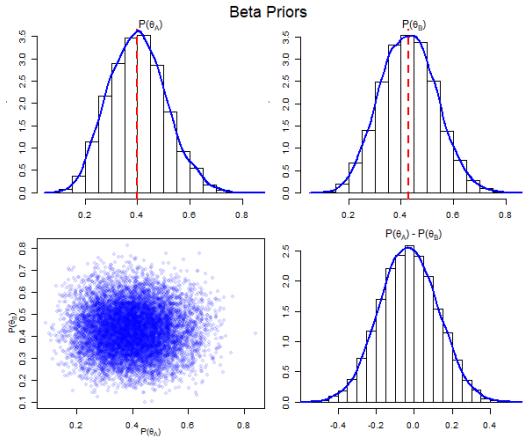


Figure 2.6: *Beta(8,12) and Beta(9,12) priors*

2.3 Example 1

Example 1 - Landing Page Experiment

Experiment 1 takes place on the home page for Company XYZ. They wanted to see if placing the video content on a different part of the homepage would get more views. They also wanted to see if more users view the video, that moving the video placement also resulted a greater chance of a conversion (sign up) in a session.

In this AB test the variants are

- Variant A - Old home page
- Variant B - New home page

In this experiment the experimental unit was a user session. The primary outcomes were proportion viewed content and proportion signed up. Users were randomized to a version when they landed on the page (user id was tracked so those who came back were allocated to the same page). The experiment ran for four weeks until there was $\sim 11,000$ in each group.

The results were

	Variant A	**Variant B**
N	11,196	11,019
Viewed content	2,001	3,052
Sign up	1,198	1,210

To analyse this experiment for each outcome, we used our prior from equation (2.3) and binomial likelihoods. We used MCMC simulations using `JAGS` software in `R` to get the posterior distributions.

2.3.1 Posterior Inference

In Frequentist AB testing we would conduct a hypothesis test for a difference between the two groups. In Bayesian AB testing we have the joint and marginal posterior distributions for the parameters of each of the treatment groups. The posterior distribution describes

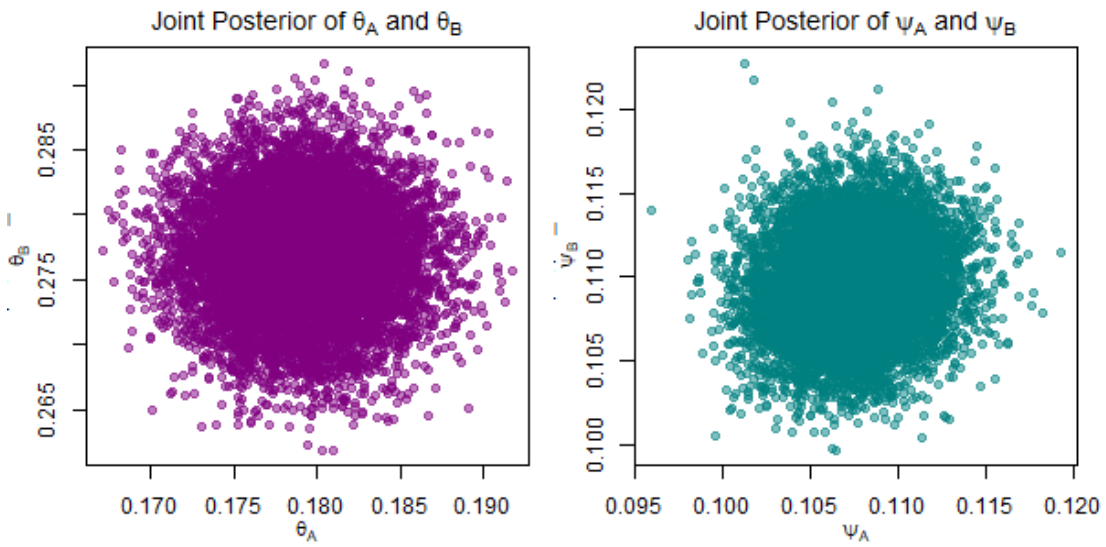


Figure 2.7: *Joint distributions from example 1*

knowledge and our uncertainty about the parameters and there are potentially many different ways of making inference from our data. We denoted the parameter to estimate the conversion rate for viewing content as θ and the conversion rate for signing up as ψ. We can start by inspecting posterior visually by plotting the marginal and joint distributions from each group:

- Both the joint distributions from figure 2.7 look uncorrelated with each other
- In figure 2.8 we have two narrow posterior distributions that are far apart and do not overlap for θ_A and θ_B. They are centred at 0.18 and 0.28. This tells us that the effect size is about 10% and we are pretty certain that the success rate for viewing content is better in treatment B.
- We have two wider posterior distributions that overlap for ψ_A and ψ_B. While ψ_B is centred higher it is not always better than ψ_A so we have a bit more uncertainty about which treatment is the better one for sign-ups.

In addition we could describe the posterior distribution by point estimates such as the mean, median or mode. When our posterior looks somewhat normal it is common to summarize as:

$$\theta = \text{posterior mean} \pm \text{posterior standard deviation} \tag{2.4}$$

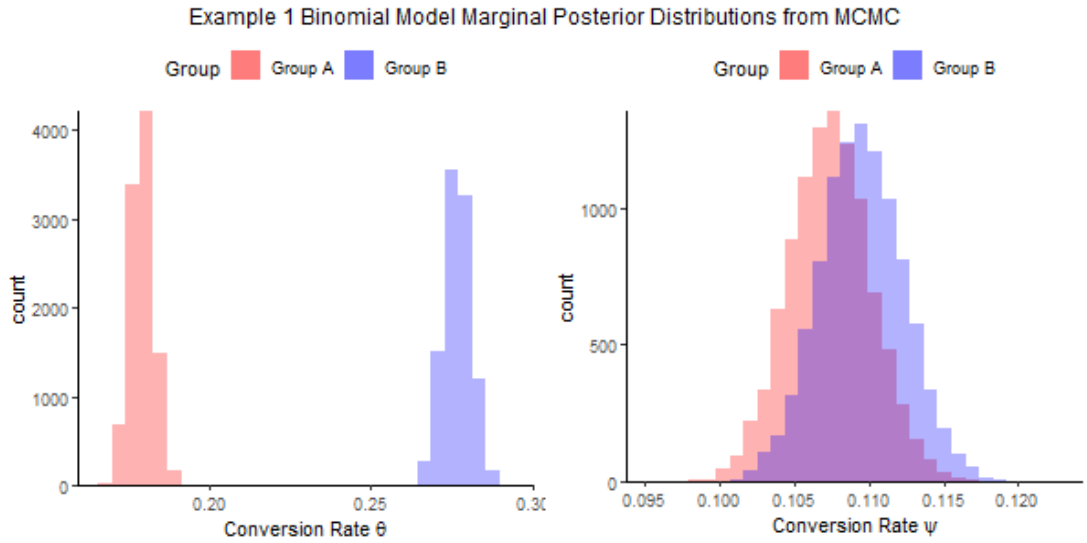


Figure 2.8: *Marginal distributions from example 1*

In our case we could summarize the individual parameters or the difference between them:

- $\theta_A = 0.179 \pm 0.004$, $\theta_B = 0.276 \pm 0.004$, $\theta_B - \theta_A = 0.097 \pm 0.006$
- $\psi_A = 0.107 \pm 0.003$, $\psi_B = 0.110 \pm 0.003$, $\psi_B - \psi_A = 0.003 \pm 0.004$

We can also calculate probabilities from the posterior. For a single parameter θ_A we could calculate:

$$P(\theta_A \geq 0.15) = \int_{0.15}^{1} p(\theta_A|D) \tag{2.5}$$

From our posterior samples it would be:

```
mean(thetaA >= 0.15)
```

We could also calculate the probability that $\theta_B \geq \theta_A$ and $\psi_B \geq \psi_A$:

```
mean(thetaB > thetaA) = 1
mean(psiA > psiB) = 0.76
```

This is known as the error function, probability to be the best (PTBB) or probability to beat control (PTBC). The opposite $\theta_A \geq \theta_B$ and $\psi_A \geq \psi_B$ would just be the compliment.

```
mean(thetaA > thetaB) = 0
mean(psiA > psiB) = 0.24
```

Declaring a Winner

While we can summarise the posterior distribution in many ways, the main questions AB testing is our experiment conclusive? If so, who is the winner? In some cases like comparing θ_A and θ_B in our example we can be pretty sure just from looking at the posterior that variant B is the better version for content clicks. For which variant is better for sign-ups, the results are much closer and it can be harder to make a call. In this case, before we start our experiment we can set a decision rule to help us determine the outcome of the experiment.

In this section we will share two methods to evaluate the difference between variants; Expected loss and highest density interval (HDI). With each of these we can specify ahead of time the difference in variants that would be practically relevant to our AB test. We then use this to make the decision on if there is a conclusive winner. This usually boils down to the minimum difference that makes the cost of the change worthwhile. In the real world we know a true difference may exist, but not all differences are practically relevant, e.g. running on a treadmill –one machine may measure the distance as 5km where the true difference was 5.05km - but it is not going to change how tired we feel after our run!

2.3.1.1 Expected Loss

Suppose we choose treatment B. The error function can be used to express the probability that we made a mistake and we can use the error function as a decision rule by picking the variant with the least chance of making a mistake (i.e. the one with the highest probability of being the better treatment). While it is easy to understand and calculate, its flaw is that it treats all errors equally as bad.

An alternative is using a loss function. The loss function corrects the error function by treating small errors as less bad than big ones. The method proposed by C. Stucchio (Stucchio [2015]) is the loss function is the amount of uplift that one can expect to be lost by choosing a given variant, given particular values of θ_A and θ_B:

$$\begin{aligned} Ł(\theta_A, \theta_B, A) &= \max(\theta_B - \theta_A, 0) \\ Ł(\theta_A, \theta_B, B) &= \max(\theta_A - \theta_B, 0) \end{aligned} \tag{2.6}$$

To use the loss function on our posterior distribution we calculate the expected loss of our joint posterior.

$$\begin{aligned} E[Ł(\theta_A, \theta_B, A)] &= \int_0^1 Ł(\theta_A, \theta_B, A) P(\theta_A, \theta_B | D) d\theta_A \theta_B \\ E[Ł(\theta_A, \theta_B, B)] &= \int_0^1 Ł(\theta_A, \theta_B, B) P(\theta_A, \theta_B | D) d\theta_A \theta_B \end{aligned} \tag{2.7}$$

This looks tricky, but we can actually get this very easily by setting up a deterministic node in our MCMC model. For `BUGS/JAGS` we can use:

```
loss.a = step(theta.b - theta.a)*(theta.b-theta.a)
loss.b = step(theta.a - theta.b)*(theta.a-theta.b)
```

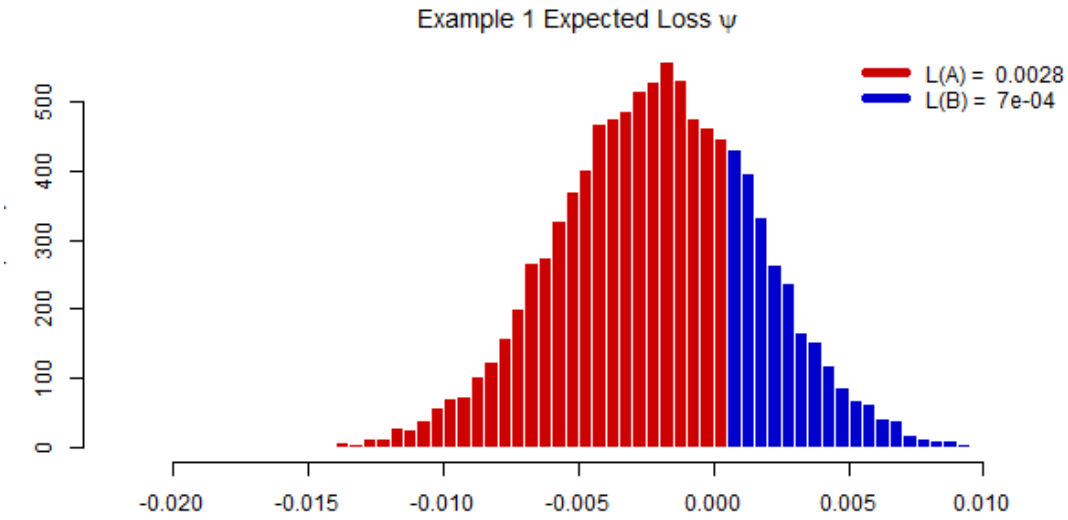


Figure 2.9: *Example 1 expected loss for the sign up conversion* ψ

The expected loss calculates cost associated in mistakenly declaring variant A (or variant B) as the winner. We set the decision rule for this method through a "threshold of caring" –for when the loss is low enough we do not care. We set a threshold of

caring for analysis of the sign-ups at any loss greater than 0.01. The results were $Ł(\psi_A) = 0.0035$ and $Ł(\psi_B) = 0.0006$ meaning we could pick either variant and still be under the threshold. By this rule, the result for sign-ups is inconclusive. We could still pick variant B because it has a higher click rate and at least we know that it is no worse than variant A for sign ups.[1]

2.3.1.2 Highest Density Interval

A credible interval can sometimes be thought of as the Bayesian version of a confidence interval. It is an extension of our previous posterior summaries, but this time we want to find an entire interval $[a, b]$ that contains α amount of the posterior probability.

$$P(a \leq \theta \leq b|D) = \alpha \tag{2.8}$$

It can be useful to summarise the posterior with statements like "There is 95% probability the parameter is between 0.53 and 0.57" with $\alpha = 0.95$ being a popular choice, analogous to the 95% confidence interval (though they have very different interpretations). There are many ways to get a legitimate, credible interval - any interval containing 95% of the posterior probability is a 95% credible interval. Two commonly used ones are:

- Central credible interval - The interval with equal probability on each side.
- Highest density interval (HDI) - The interval that spans the distribution such that every point inside the interval has higher credibility than any point outside the interval[2].

The Region Of Practical Equivalence (ROPE) as defined by (Kruschke [2013]) is the interval around the null value (i.e. difference of 0) that includes all values of the parameter that are for practical purposes negligibly different from the null value. We can use ROPE as an interval of differences between our two treatment parameters, where we consider them practically the same. We calculate the HDI and compare if the ROPE sits inside or outside of this interval as our decision rule.

[1] We did not need tp calculate the expected loss for θ because variant B was the clear winner. The expected losses were $L(\theta_B) = 0$ and $L(\theta_A) = 1$

[2] The HDI is not invariant under transformation, unlike the central credible interval. However, in AB testing, we use it exclusively to test the difference between two parameters, so no transformations need to take place.

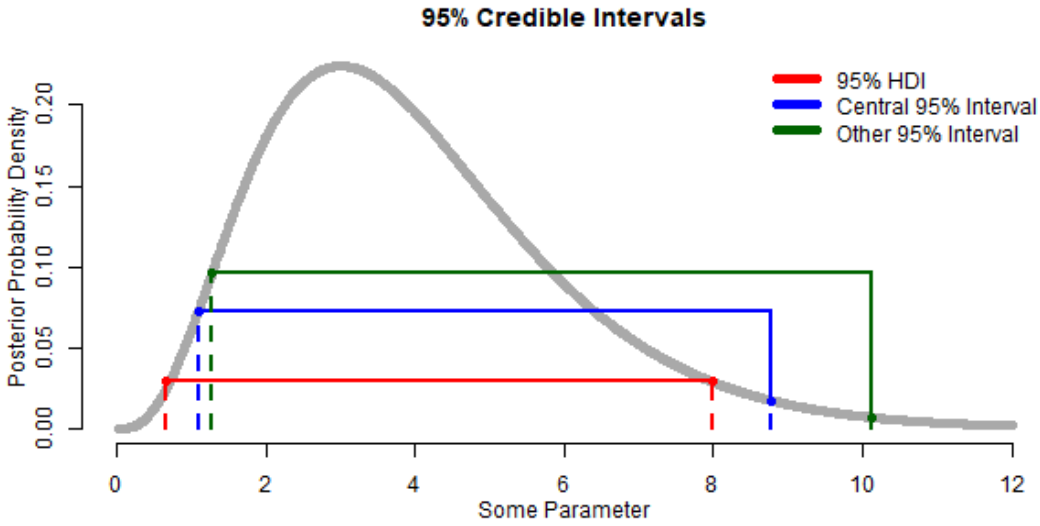


Figure 2.10: *Examples of different Bayeisan 95% credible intervals*

For our analysis, we set ROPE at ± 0.01. Our credible interval says that 95% of the probability places variant B being 1.1% higher and 0.5% lower than variant A. In figure 2.11 the results are again inconclusive because the credible interval contains 0 and the span of the credible interval encloses the ROPE.

2.3.2 Cost of Picking a Prior

We previously looked at a loss function for picking the wrong treatment. This loss function was the expectation of the posterior loss. Our prior makes up one part of our posterior distribution, the other being the likelihood. Assume we have a well thought out prior, $p(\theta)$ that we believe to be correct or good. We may want to compare this to some quick approximate prior $q(\theta)$ to understand the cost of using a 'bad' prior if our good prior reflected the true value of the parameter. This cost or 'badness' is the prior expectation of the posterior expected loss. We go into further detail in chapter 4. In this section, we only present the results of our chosen prior for this experiment.

In this example we have a large sample size of $\sim 11,000$ in each treatment. For this sample size the expected loss of picking a bad prior is the same as our good prior. This matches what we would expect from figure 2.3, that with a large data set our posterior is much less sensitive to the choice of prior distribution. In this case we have a large enough dataset not to worry about the cost of the prior that we picked.

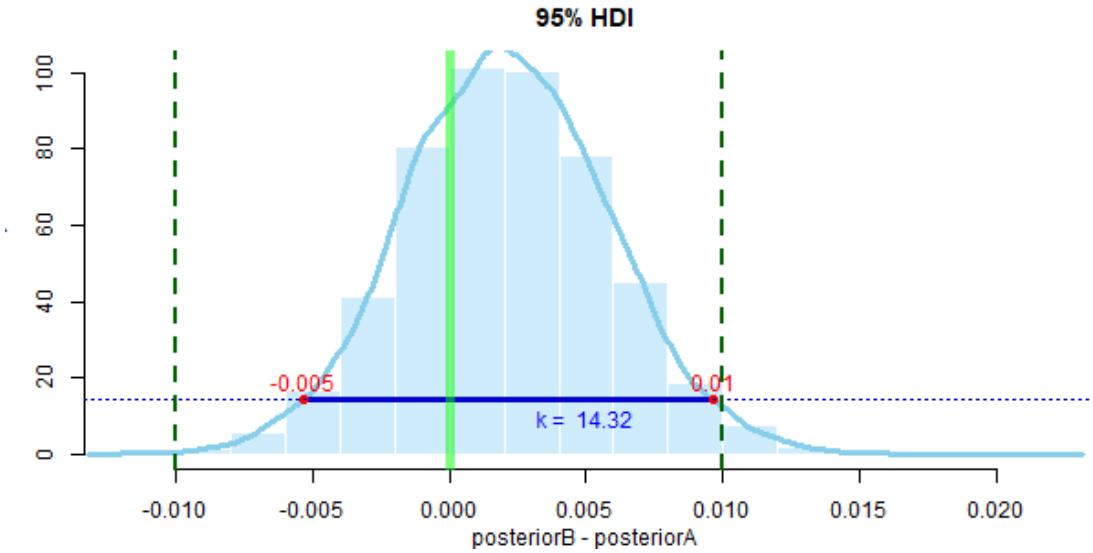


Figure 2.11: *Example 1 95% HDI for $\psi_B - \psi_A$ with ROPE of ± 0.01*

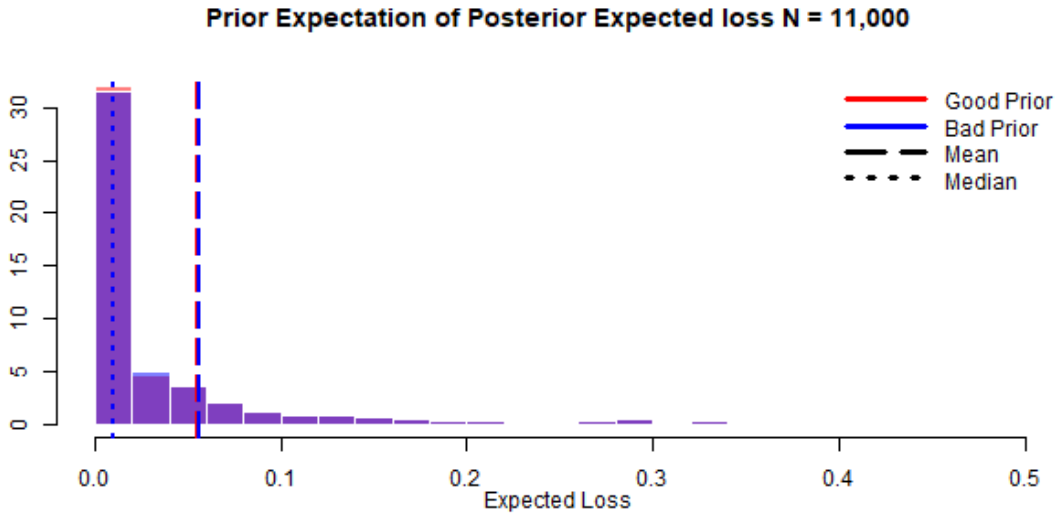


Figure 2.12: *Example 1 comparison of prior expectation of the posterior expected loss between a uniform 'bad' prior and our prior in section 2.2*

2.4 Example 2

Example 2 - First Action

Experiment 2 takes place on the first time a customers signs in to the website after signing up. As there is a 30 day free trial, of about ~ 600 new sign ups a week only half actually log in to the website. Even fewer complete an action in the product before the end of the trial. It was proposed that showing a pop up with instructions to get started could help users complete the first main action in the product and increase the conversion rate of users staying on after the trial period. As traffic was low it would take a long time to see if the change increased the conversion rate, so an initial experiment was done over 30 days to see if the proportion of new users doing a first action was increased.
In this AB test the variants are:

- Variant A - No Instructions (existing experience)
- Variant B - Instructions

In this experiment the experimental unit was the user. The primary outcomes were proportion who did an action during the experiment (θ). Users were randomized to a version when they landed on the page (cookies were tracked so those who came back were allocated to the same page).

The results were

	Variant A	**Variant B**
N	500	500
Did Action	352	340

Again our results from the posterior distribution and HDI look inconclusive:

- There is possibly some slight correlation between θ_A and θ_B seen in the joint posterior in figure 2.13.
- The posterior distributions for θ_A and θ_B overlap with θ_A being slightly higher. The posterior summaries for the mean are very close together:
  - $\theta_A = 0.696 \pm 0.017$
  - $\theta_B = 0.686 \pm 0.018$

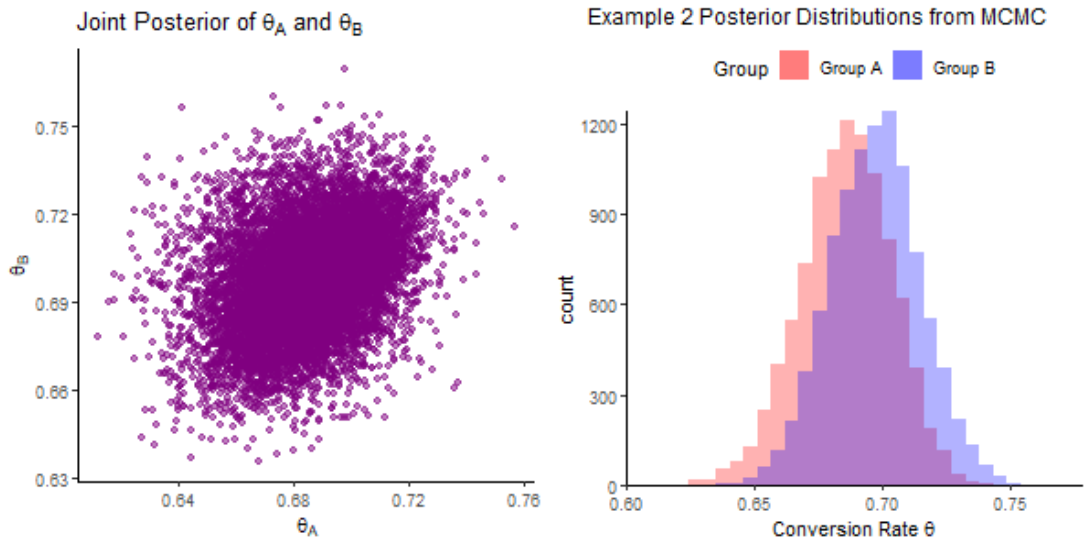


Figure 2.13: *Joint and marginal posterior distributions for example 2*

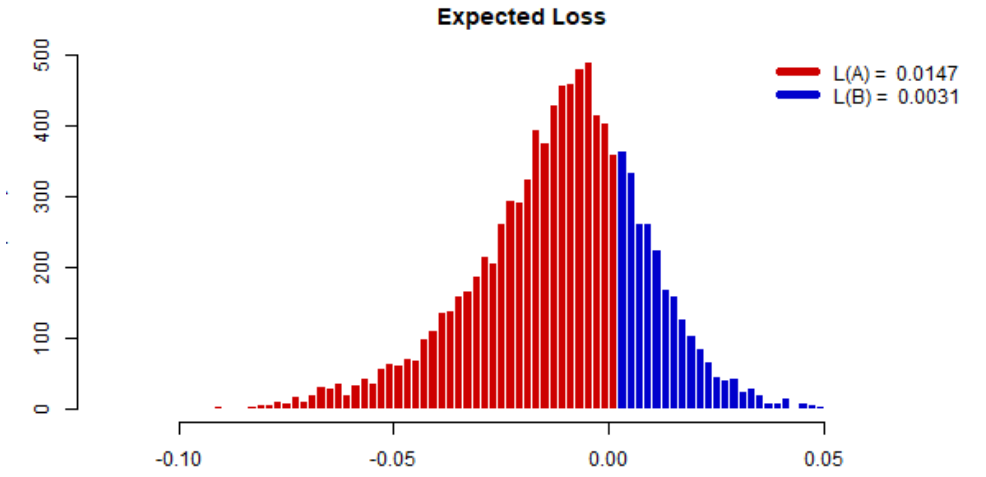


Figure 2.14: *Example 2 expected loss for θ*

- In figure 2.15 the 95% HDI is -0.0053 to 0.0029 with our ROPE ±0.01 enclosed inside
- In figure 2.14 $L(A) = 0.0147$ and $L(B) = 0.0032$ for our threshold of caring of 0.01 would suggest not to pick variant A as it sits just above this threshold.

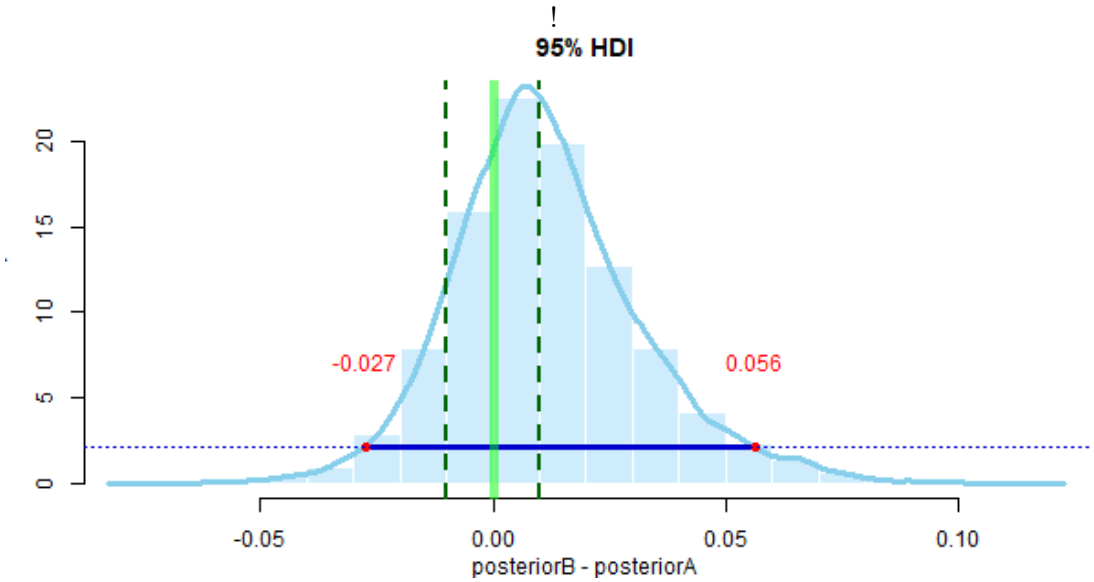


Figure 2.15: *Example 2 95% HDI for θ_B-θ_A with ROPE = ±0.01*

Unfortunately, it looks like the difference in effect size between variant A and variant B is pretty inconclusive. It is hard to tell if showing the pop up in variant B is having any effect on the rate of completing the first action compareds to not having it there at all. If we do decide to pick variant B, we do know that it is not any worse and has a lower expected loss on uplift compared to variant A.

This time we have a much smaller sample size when we estimate the cost of our prior. Figure 2.21 shows our expected loss is decreased with the 'good' prior that we used, compared to the 'bad' prior. Even though we weren't able to make a conclusive decision, it seems the prior we used would be better for modelling the small effect sizes we expected to see compared to the quick uniform approximation.

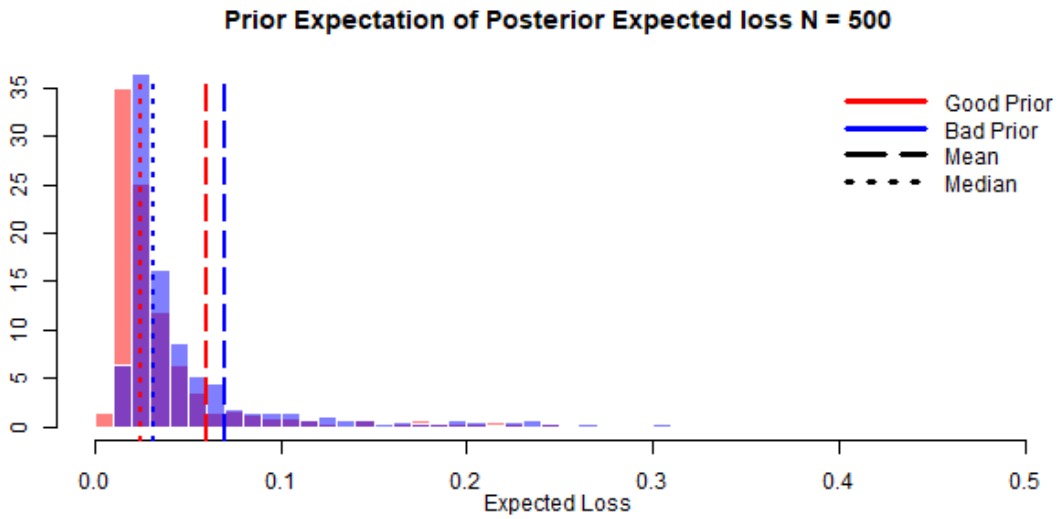


Figure 2.16: *Example 2 comparison of prior expectation of the posterior expected loss between a uniform 'bad' prior and our prior in section 2.2*

2.5 Example 3

Example 3 - EDM propotion

Company XYZ has recently launched a new feature on the website and wants to tell its customers about it. They select a sample of customers to test two different versions of an email before deciding on what final version to send to the rest of the customer base[a].
In this AB test the variants are

- Variant A - Email A (Content with text information about feature)
- Variant B - Email B (Content with video demo of new feature)

This set up is a bit different than in our previous two examples because the treatment groups are preselected before the start of the experiment and every user got sent the email on the same day. They then waited 30 days to see what proportion of users in each group ended up using the new feature.

[a]This type of email is called Electronic Direct Mail marketing (EDM). It is a form of marketing that companies can use to communicate with customers to build customer loyalty.

The results were

	Email A	**Email B**
N	1,000	1,000
Opened	256	260
Did Action	287	321
Did Action & Opened	178	208

Notes on Non-Response

Two aspects make this experiment different from our previous two examples:

1. While we sent emails to 2,000 users, we only had 511 ($\sim$ 25%) open them to "receive" the treatment. This means we have a large amount of non-response (75%) that could cause bias.
2. Some subjects completed the action without even opening the email/getting the treatment.

These are attributes of the outbound AB test we introduced in chapter 1.2. Our previous two examples were web-based AB tests, so it was not possible to have non-response or complete the action without getting the treatment by design.

So how do we deal with these issues?

In clinical trials, subjects are normally analysed within the group to which they were allocated, irrespective of whether they experienced the intended intervention (intention to treat analysis, ITT) (Sibbald and Roland [1998]). While ITT ensures randomisation is kept, it will be the most conservative estimate of the effect size.

Sometimes we wish to break randomisation to get an idea of what the best possible effect size could be. Per protocol (PP) is a subset of the ITT comparison of treatment groups that includes only those subjects who completed the treatment as initially allocated. Analysing only those who completed the treatment as planned provides an estimate of the true efficacy of an intervention (Tripepi et al. [2020]). We have to be more cautious with PP as it breaks randomisation. We do not know if the subjects who did not respond to email A are the same that would not respond to email B. However, as the email subject lines were the same and the open rates similar, it is

probably safe to make this comparison in this example. In other AB tests, the reason for non-response between two treatments may be different and the PP comparison less fair (e.g. comparing one treatment group that received phone calls and another that received emails –the non-response will be caused by different mechanisms; not answering the phone due to being busy/unknown caller vs. not checking or opening emails).

The other issue with non-response is that it decreases our sample size, which may make us less sure about the result (wider posterior). However, it is accepted that only a fraction of emails that get sent out will be read. It is standard practice to send emails to a larger sample size to make up for the non-response. The consequence is that an unopened EDM means noise or clutter for the customer's inbox and that can have detrimental effects such as unsubscribing. In Company XYZ it is known that for every 100 emails they send, approximately 25 will be opened by a customer. Furthermore, of those 25, only a fraction will end up doing the intended action. Performing ITT analysis is the best for understanding the rate of return for each email they send.

Lastly, if we want to understand the performance of a new EDM, it is beneficial to have a withholding group (treatment of no email) as part of the experiment. It is important in email marketing to compare the difference between the intervention and sending no email at all. We can use this information to determine what minimum uplift is required to make sending any email worthwhile (i.e. even if we can detect an uplift of 0.1% over no email, it may be not worth the cost of potentially irritating customers). Whether we find a difference or not between two email variants, this threshold should also be part of the decision rule with a third option to not have either as a winner.

Posterior Inference

For the posterior distribution, θ was the conversion rate for our ITT analysis, and ψ was the conversion rate for our per-protocol analysis (for those who opened the email).

- The joint distributions in 2.17 look uncorrelated.
- In figure 2.18 the posterior distributions for θ_A and θ_B overlap with θ_A being slightly higher. The posterior summaries for the means are:
  - $\theta_A = 0.295 \pm 0.013$
  - $\theta_B = 0.314 \pm 0.014$
- The posterior distributions for ψ_A and ψ_B only slightly overlap with θ_A being higher. The posterior summaries for the means are:
  - $\psi_A = 0.692 \pm 0.029$

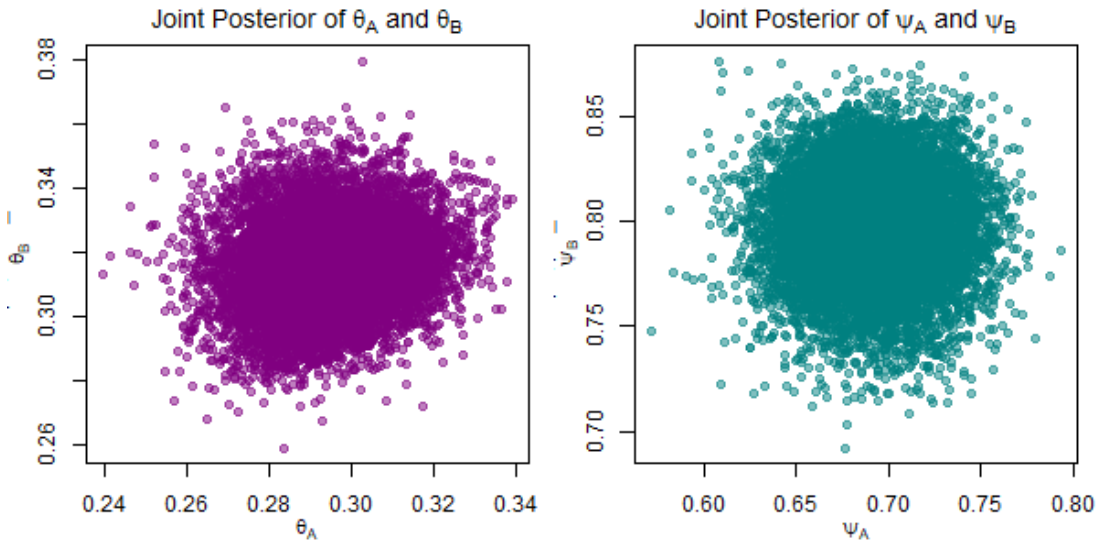


Figure 2.17: Joint distribution for example 3

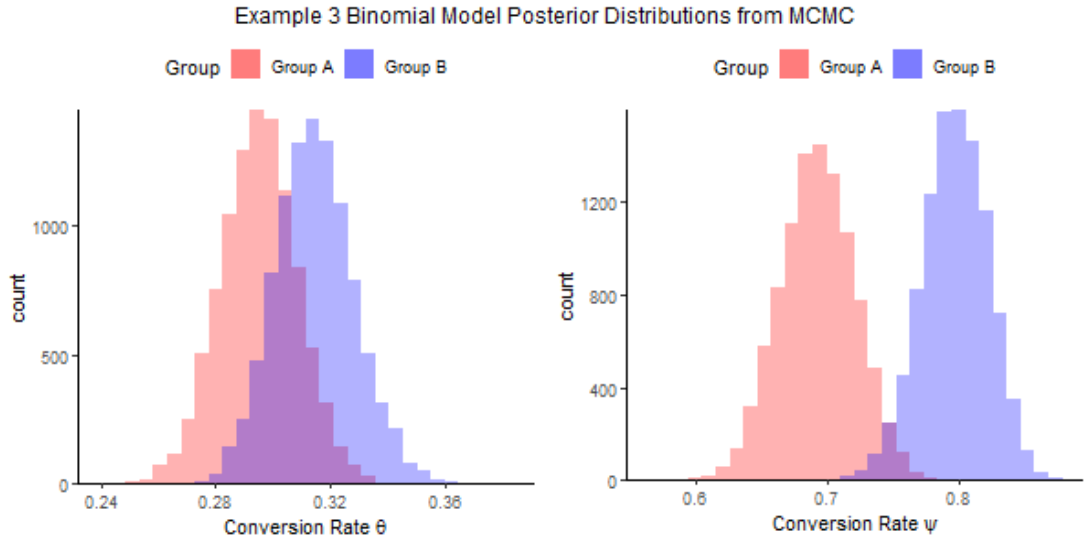


Figure 2.18: Marginal distributions for example 3

  - $\psi_B = 0.796 \pm 0.025$

- In 2.19 $L(\theta_A) = 0.0203$ and $L(\theta_B) = 0.001$. For our threshold of caring of 0.01, would suggest not to pick variant A and should pick Variant B. The loss for $L(\psi_B)$

was practically 0 so was not plotted. Compared to $L(\psi_A) = 0.105$ we should pick variant B over variant A.

- In figure 2.20 the 95% HDI for $\theta_B - \theta_A$ is -0.013 to 0.055 with our ROPE ± 0.01 enclosed inside indicating no practical difference. The 95% HDI for $\psi_B - \psi_A$ is 0.03 to 0.181 with our ROPE ± 0.01 sitting outside this interval. This indicates that 95% of the posterior probability of $\psi_B - \psi_A$ sits above our ROPE.

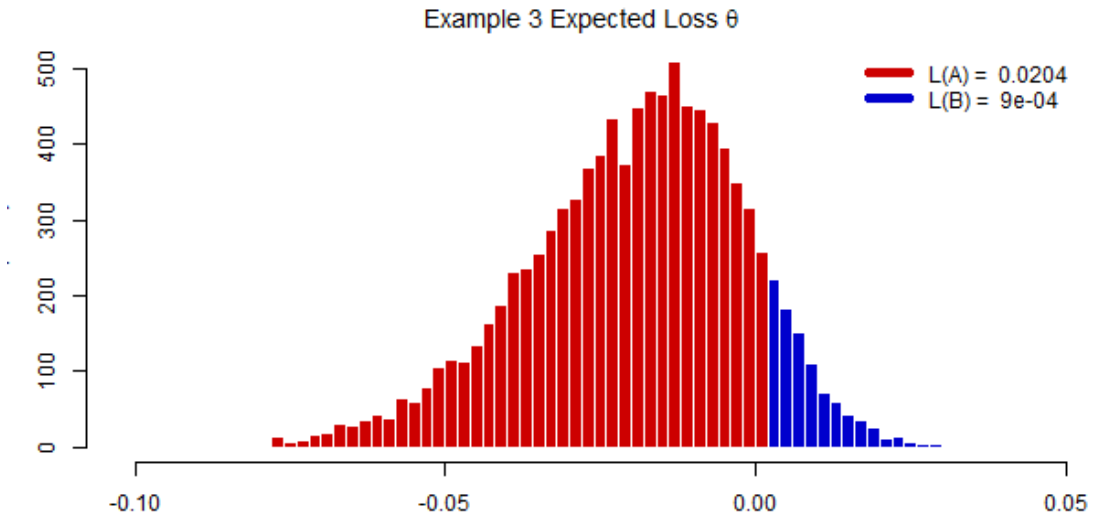


Figure 2.19: Expected loss for example 3

Ahead of time we decided that we needed at least 50% conversion in those that opened the email to make sending an email to the rest of our customer base worthwhile. We are relieved that both variants have met this criteria, so now we must just pick the winner. Given these decision rules in our per-protocol analysis, we would declare variant B to be the winner. In our ITT analysis, we may need a slightly larger sample size to make a conclusive decision for the effect size between θ_A and θ_B. If we needed to pick a single variant, then based on the expected loss we would be better to pick variant B.

In figure 2.21 it looks like we have a large enough sample size that the expected loss between our chosen prior and a uniform prior is a slightly less, but likely makes no practical difference on our inference.

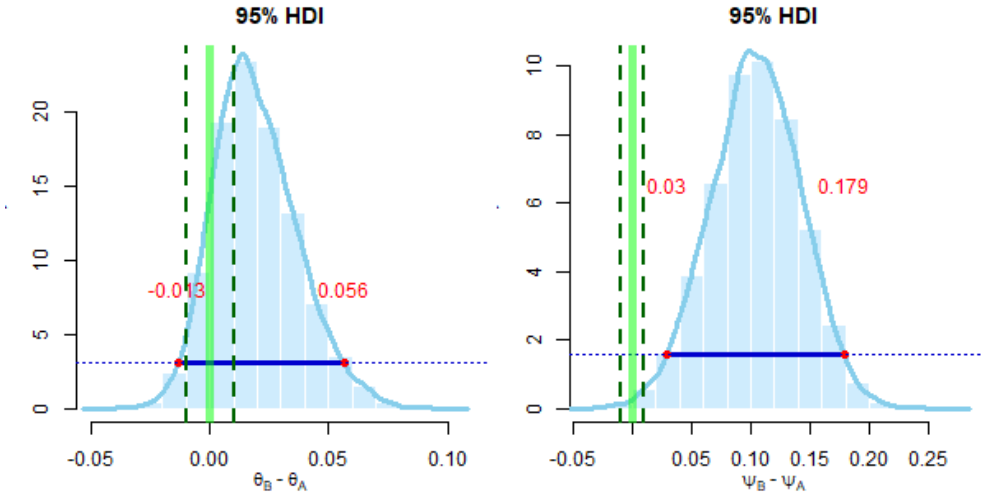


Figure 2.20: 95% HDI for example 3

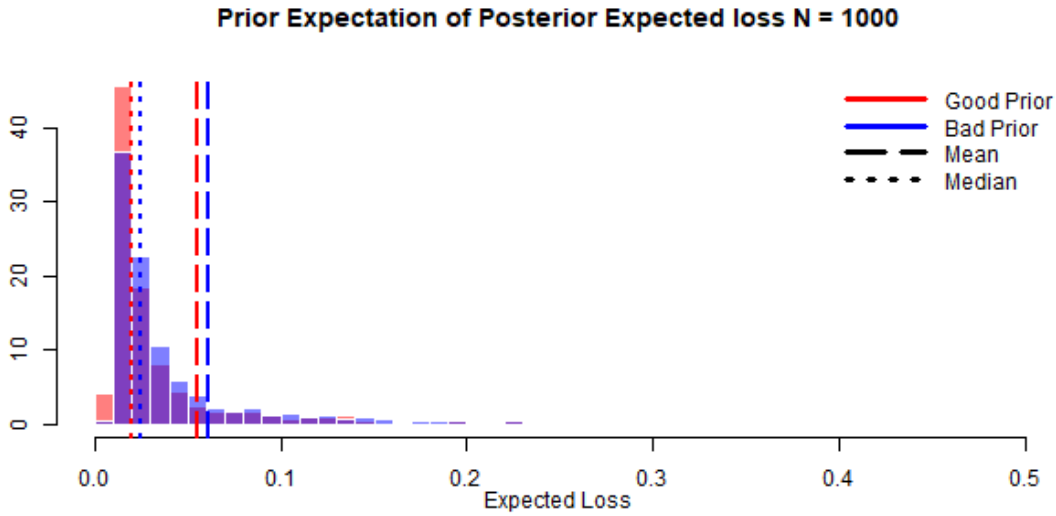


Figure 2.21: *Example 3 comparison of prior expectation of the posterior expected loss between a uniform 'bad' prior and our prior in section 2.2*

Summary

In this chapter, we demonstrated applications of the binomial AB testing model using our dependent prior in three different examples. Even with a binary outcome measure,

there is much complexity we can model when we include multiple conversion parameters and issues of non-response in the analysis. As such, there are further ways we can customise the binomial model to account for multiple parameters than we have had time to cover here. In examples 2 and 3, the binary outcome of the proportion of conversions in each group was a simplification of a time to event outcome. In the next chapter, we explore what happens when we instead analyse the same data using the time the event happens rather than a binary success or failure at the end of the experiment. This time our successes have an observed time of event completion, and non-successes become censored observations.

Chapter 3

The Survival AB Testing Model

3.1 Introduction to Survival analysis

Survival analysis studies time-to-event data that consists of a specific start time and end time. Time to event outcomes are common in biology and medical studies where death is considered an "event", hence the term "survival". If one looks closely, time to event outcomes also govern many user actions in a web-based product or service (e.g. time to first purchase or time to unsubscribe from a product/service –also known as churn). The event can be considered either a good or bad thing, depending on the situation.

In this section, we will outline our survival-based model for AB testing time to event outcomes. There are two key reasons why survival analysis is so vital for the data practitioner to have in their repertoire:

1. Binomial models with outcomes of proportions can be a simplification of a more complex sampling distribution with a time to an event outcome.
2. Incorrectly summarising censored data.

For example: If we were to experiment with an email campaign targeted to get leads to sign up to a web service, the binary outcome would tell us the overall success rate over a given period, say 30 days. However, what if the proportions of the two groups were the same? One might conclude that the treatment was no better than the control, or one might ask the question *"If leads are not converting at a higher rate, is the treatment at least getting leads to convert earlier?"*.

If a lead is converting earlier, then there is an opportunity for further intervention in the customer funnel. If there is a monetary value in a customer signing up, then

getting money earlier will increase the total value of the customer over their lifetime. Powerful decision making information is masked when we stick to binary outcomes.

In an attempt to answer the question above, one might naively try to summarize information as average days to conversion, dropping out those who do not have an event. The consequence of ignoring the censored data (see next section) is that this estimate will be skewed. Censored subjects still provide information, so they must be appropriately included in the analysis.

Censoring

A critical aspect of survival analysis is censoring, which is when the measurement of an observation is only partially known. Right censoring occurs when the subject has not experienced the event of interest by the end of data collection (i.e. end of the experiment).

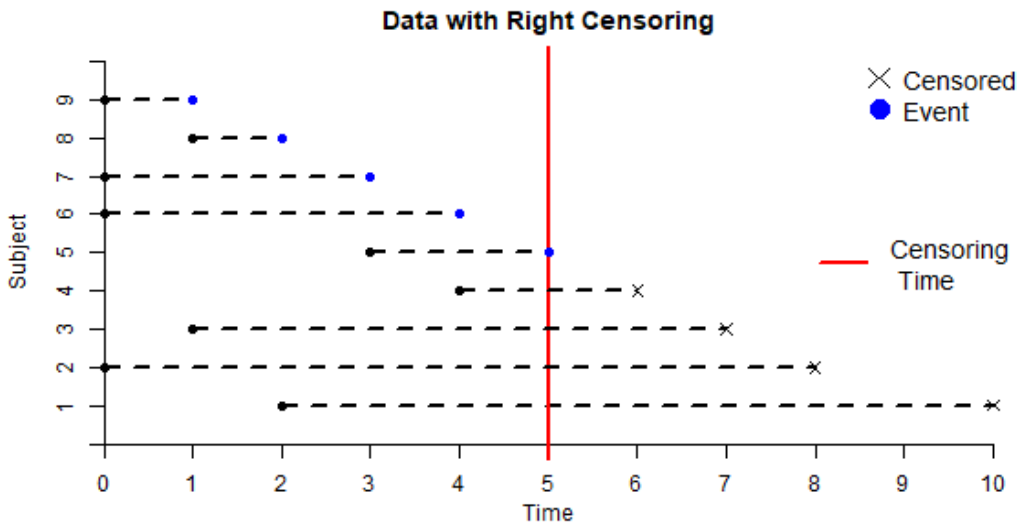


Figure 3.1: *Black dots mark when the subject entered the experiment, blue dots mark the observed event. By the censoring time at $t = 5$ we still have 4 subjects who we have not observed the event for. For some events such as death, it may happen for at some unknown time after the experiment ends. For other events it may be never.*

Left censoring and interval censoring are also possible, that is when the event happens before the trial starts, or the event happens between two intervals such that exact the date of the event is unknown. While methods exist to handle this, it is uncommon for

AB testing, so models will be limited to right censoring only.

We will also assume that only a single event occurs for each subject, which traditionally assumes the subject is dead. In our case, it will be because: (1) the event can only happen once, e.g. sign up to a website or (2) we are only interested in the first occurrence of the event in the duration of the experiment.

Survival function

The survival function is a function that gives the probability that a user will survive beyond any specified time. For a given probability density function $f(t)$, with cumulative probability distribution $F(t) = P(t \leq T)$, the survival function is:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(t)\ dt. \tag{3.1}$$

Survival probability at a particular time, $S(t)$, is the conditional probability of surviving beyond that time, given that an individual has survived just before that time. We can estimate this as the number of subjects who are "alive" at that time, divided by the number of patients who were alive just before that time.

We can determine median survival from the survival function. For example, in the top left in figure 3.2 50% of the subjects survive 15 days. In some cases, median survival cannot be determined from the graph if we have observed the event in less than 50% of subjects. The bottom right example in figure 3.2 more than 50% of the subjects survive longer than the observation period of 30.

Kaplan-Meier Curve

The Kaplan-Meier estimate of survival probability is the product of these conditional probabilities up until that time. At time t_0, the survival probability is 1, i.e. $S(t_0) = 1$. The estimator of the survival function $S(t)$ (the probability that life is longer than t) is given by:

$$\hat{S} = \prod_{i=1;t_t \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.2}$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals are known to have survived (have not yet had an event or been censored) up to time t_i. We can use a plot of the Kaplan-Meier estimate against time to look at survival over time for one group or compare survival

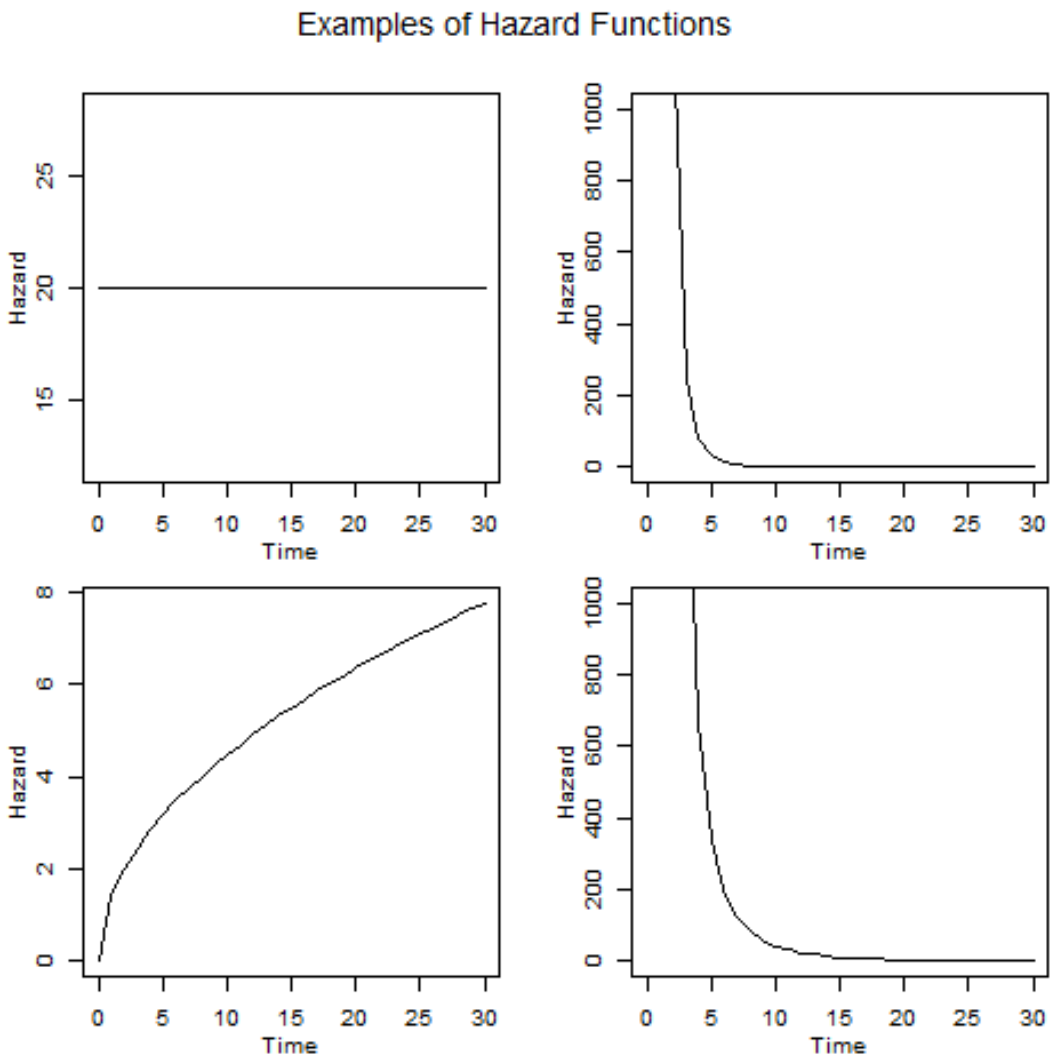


Figure 3.2: *Examples of survival functions generated from Weibull distributions with different shape and scale parameters. The red line shows the median survival.*

between two groups. On the plot, small vertical tick-marks state individual patients whose survival times have been right-censored.

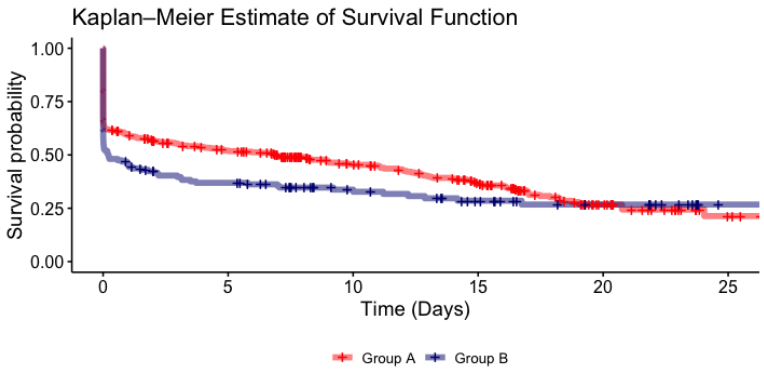


Figure 3.3: *The Kaplan-Meier curve comparing survival between two groups.*

Hazard

The hazard function, λ, is defined as the event rate at time t conditional on surviving until time t or later. It gives the instantaneous risk that the event of interest happens, within a very narrow time frame.

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T \leq t + dt)}{dt \; S(t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)}$$

The hazard function models which time periods have the highest or lowest chances of an event. It can be useful to think about what the shape of the hazard function for real phenomena might be. If we were studying infant mortality, we might expect $\lambda(t)$ to be decreasing in the first year of life, as chances of survival increase with age. For modelling hazard of engines, we might expect $\lambda(t)$ to increase, as the older the engine gets, the greater the chance of catastrophic failure. Some things we can model with constant hazard throughout their lifetime, such as the time to win the lotto, being in a plane crash or getting eaten by a shark. The chances of such events do not increase or decrease with time.

3.2 Global Survival Model

First, let us define the likelihood of a single treatment arm in an experiment. Let us imagine we have N subjects and we are interested in the time between a start time t_0 and the time at which the conversion event E happens. We denote this as t_i for observation i.

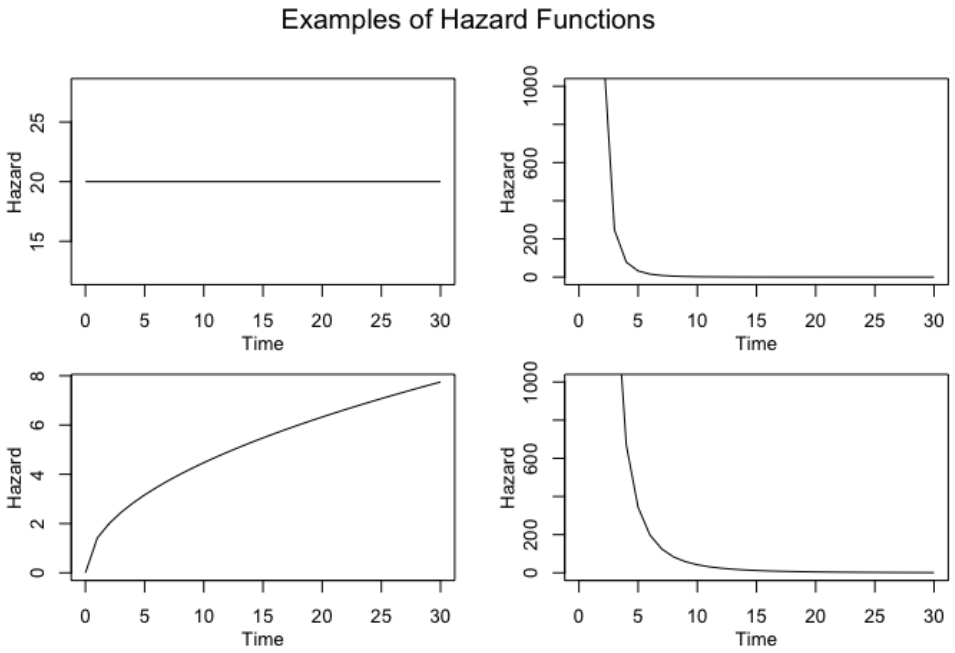


Figure 3.4: *Hazard functions of survival curves from 3.2. Top right shows a constant hazard function. Left plots show a hazard function decreasing over time. Bottom left shows hazard increasing over time.*

We let our experiment run for D days, which at this point, we have x subjects who have completed the event and $N - x$ subjects who have not. If subject i completed the event at time t_i before the end of the experiment, then its contribution to the likelihood function is its density at that duration[1]:

$$L_i = f(t_i|\gamma) \ . \tag{3.3}$$

If subject i did not complete the event at the end of the experiment, then their observation is censored. All we know is that this time exceeds the last observed point of t_i for this subject. The probability of this event is:

$$L_i = 1 - F(t_i|\gamma) = S(t_i|\gamma) \ . \tag{3.4}$$

[1]Where γ represents the parameters for the probability density

Therefore the product of each subjects contribution to the likelihood is:

$$L = \prod_{i=1}^{N-x} S(t_i|\gamma) \prod_{i=1}^{x} f(t_i|\gamma) \ . \tag{3.5}$$

This model is assuming that t_i is different for each surviving user, i.e. entered experiment at different times and are in the experiment for different durations. If surviving users all start and end the experiment at the same time, they will all have the same duration and $t_i = t^*$ hence the likelihood simplifies to:

$$L = S(t^*|\gamma)^{N-x} \prod_{i=1}^{x} f(t_i|\gamma) \ . \tag{3.6}$$

The model likelihood can be considered complete on its own. However, unlike the event of death which is certain, we know from experience that we have subjects that will never have a conversion event. In survival analysis this would just be represented as $\lim_{t\to\infty} t_i = \infty$, but this is not practically relevant for the analysis of AB test.

To fully develop the likelihood for an AB test, let us adjust the standard parametric Bayesian model we have in equation 3.5 and assume there are two types of subjects: (1) those that will eventually convert if we observe them for long enough and, (2) those that will never convert.

Let us use ϕ to denote the probability a customer will convert eventually and $1 - \phi$ for the probability a customer will never convert. Then the probability that a customer does not convert before time t can be written using the product rule:

$$\begin{aligned} P(\text{no conversion before } t) =& P(\text{no conversion before } t|\text{does not convert ever})P(\text{does not convert ever}) \\ &+ P(\text{no conversion before } t|\text{convert ever})P(\text{convert ever}) \\ =& 1 * (1 - \phi) + S(t_i|\gamma) * \phi \\ =& 1 - \phi * F(t_i|\gamma) \ . \end{aligned} \tag{3.7}$$

This captures the probability that we did not see a customer convert either because they did not convert before the end of the experiment. Same can be done for the probability a customer converts before time t:

$$\begin{aligned} P(\text{conversion before } t_i) =& P(\text{conversion before } t_i|\text{does not convert ever})P(\text{does not convert ever}) \\ &+ P(\text{conversion before } t_i|\text{convert ever})P(\text{convert ever}) \\ =& 0 * (1 - \theta) + f(t_i|\phi) * \theta \\ =& \theta * f(t_i|\phi) \ . \end{aligned} \tag{3.8}$$

Together the likelihood function becomes:

$$L = \prod_{i=1}^{N-x} \left[1 - \phi F(t_i|\gamma)\right] \prod_{i=1}^{x} \left[\phi f(t_i|\gamma)\right] \tag{3.9}$$

The log likelihood is:

$$l = \sum_{i=1}^{N-x} \Big[\log(1 - \phi F(t_i|\phi))\Big] + x \log(\phi) + \sum_{i=1}^{x} \log f(t_i|\phi) \tag{3.10}$$

Again, if all survivors have the same duration we can simplify this to

$$l = (N - x) \log(1 - \phi F(t^*|\gamma)) + x \log(\phi) + \sum_{i=1}^{x} \log f(t_i|\gamma) \tag{3.11}$$

Then to analyse the posterior distributions of the AB test the joint likelihood is just the product of the likelihoods in each group:

$$\begin{aligned} L &= L_A * L_B \\ &= \prod_{i=1}^{N_A - x_A} \left[1 - \phi_A F(t_i|\gamma_A)\right] \prod_{i=1}^{x_A} \left[\phi_A f(t_i|\gamma_A)\right] \prod_{i=1}^{N_B - x_B} \left[1 - \phi_B F(t_i|\gamma_B)\right] \prod_{i=1}^{x_B} \left[\phi_B f(t_i|\gamma_B)\right] \end{aligned} \tag{3.12}$$

or more simply for any number of treatment arms $k = 1, \ldots, m$

$$L_m = \prod_{k=1}^{m} \left(\prod_{i=1}^{N_k - x_k} \left[1 - \phi_k F(t_i|\gamma_k)\right] \prod_{i=1}^{x_k} \left[\phi_k f(t_i|\gamma_k)\right] \right). \tag{3.13}$$

3.3 Local Survival Model

In medical trials, follow-up may be from years to a lifetime, and so censored observations can have very large values. AB tests are usually for decisions that need to be made quickly, and the duration of the experiment exists on a much shorter time period (usually in days or months at the most). In this case, the censored observations of someone converting in the far off future become irrelevant. When all subjects are in the experiment for the same duration we propose a different way to handle censored observations; here we do not worry about all time and just let θ be the parameter describing the conversion probability before the end of the experiment, t^*.

We start by going back to our model above: We can express the probability a subject converts before t^* from equation (3.8) as the joint probability of that they convert before t^* and have a specific conversion time t_i. This can be decomposed using the product rule of probabilities:

$$\begin{aligned} p(\text{conversion before } t^*, t_i) &= P(\text{converts ever}) * p(\text{conversion before } t^* | \text{converts ever}) * \\ &\quad P(t_i | \text{conversion before } t^*, \text{converts ever}) \\ &= \phi \ * \Big(\int_0^{t^*} f(t|\gamma) \ dt \Big) \ \frac{f(t_i|\gamma)}{\int_0^{t^*} f(t|\gamma) \ dt} \end{aligned} \tag{3.14}$$

We can see that the denominator will cancel out. If we do not worry about all time and just let θ be the parameter describing the conversion probability at time t^*, following a similar derivation to above we get:

$$\begin{aligned} p(\text{converts before } t^*, t_i) &= P(\text{converts before } t^*) \ p(t_i | \text{converts before } t^*) \\ &= \theta \frac{f(t|\gamma)}{\int_0^{t^*} f(t|\gamma) \ dt} \end{aligned} \tag{3.15}$$

For the non converters the probability is just

$$P(\text{does not convert before } t^*) = 1 - \theta \tag{3.16}$$

In this model, we have dropped users with censored observations t_i and instead only care about the proportion that did not convert by the end of the experiment.

Therefore the likelihood is

$$L = (1-\theta)^{N-x} \ \theta^x \ \prod_{i=1}^{x} \frac{f(t_i|\gamma)}{\int_0^{t^*} f(t|\gamma) dt} \tag{3.17}$$

and log likelihood

$$l = (N-x) \log(1-\theta) \ + x\theta \ + \sum_{i=1}^{x} f(t_i|\gamma) - x \int_0^{t^*} f(t|\gamma) dt. \tag{3.18}$$

Equation (3.15) is equivalent to equation (3.8) and equation (3.16) is equivalent to equation (3.7) if $\theta = \phi \int_0^{t^*} f(t|\gamma) \ dt$. Therefore, the model in equation 3.17 is ultimately equivalent to 3.9 except for the parametrisation.

3.4 Example 2 Revisited

Example 2 - First Action Revisited

The previous experiment in section 2.4 was somewhat unorthodox, though a commonly used approach. Subjects entered at different times after the experiment started, so each subject had different follow-up time at the end of the experiment. Our conversion metric θ, was defined by the conversion within the duration of the experiment –not who might convert ever.

This time we analyse the experiment with a time to event outcome. We take the time, t_0, from when the subject signs into the website and gets the treatment (pop-up instructions or nothing) to the time the subject either completes the action or the experiment ends, t_i. If the subject did not complete the action by the end of the experiment the observation is censored.

We can use our global survival model to compare the all-time conversion rate (which we will denote as ϕ) and survival rates between variants. It is of particular interest to know if one variant is better at getting users to do events earlier even if the conversion rate for the experiment, θ, is the same.

This time when we look at the results of the experiment, we can plot the times recorded for each user. Each observation will either be the time the event happened or the censored time. In figure 3.5 we plot the distribution of times. We can see that they are skewed and differ between censored users and those with events. This is one reason why we would want to take the censored data into account for our analysis.

The next way we can plot the data is to use the Kaplan-Meier method to estimate the survival curve. This is how we will plot the data and compare it to our posterior estimates of survival. Figure 3.6 already shows something we did not see in the previous section: while the survival probability at the end of the 30 days is about the same, at the start of the experiment there is a difference between the two groups. It looks like group B has lower survival (in this case lower survival is a positive outcome) in the beginning up to the middle of the experiment. This suggests that variant B (having the pop up) motivates users to do the event earlier on even if the overall conversion rates are the same. In marketing, this could be a winner!

To analyse this example in the Bayesian framework, we will use our the likelihood from our global survival model with a Weibull distribution as the distribution for our event times. The probability density function of the Weibull distribution for a time t is

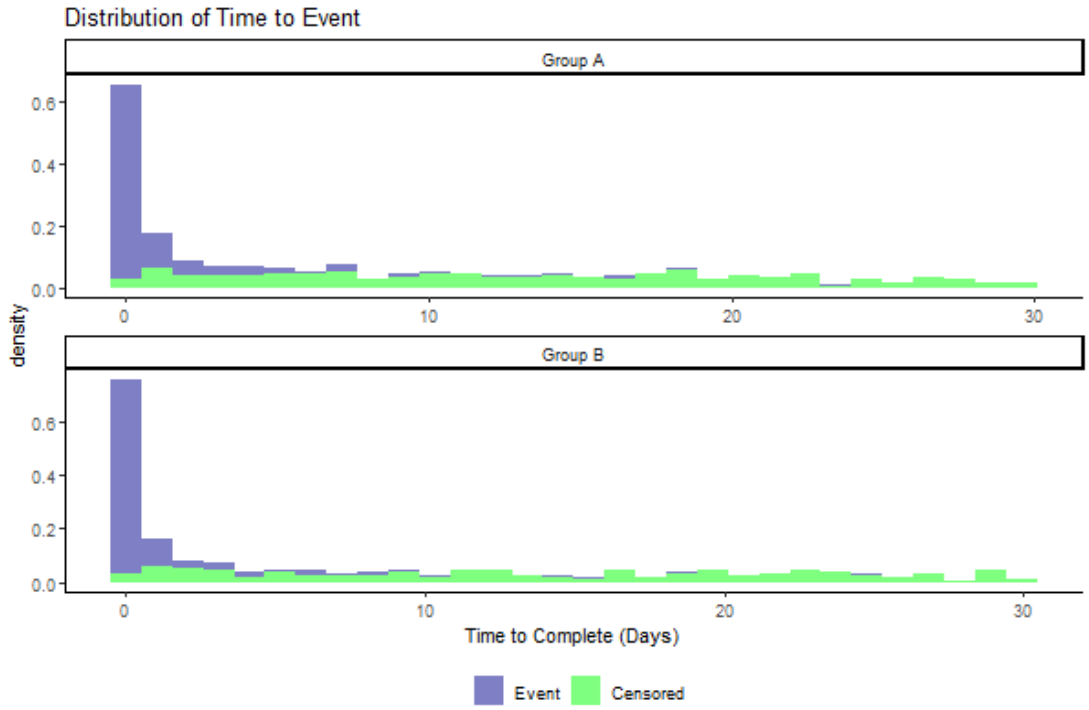


Figure 3.5: *Plot of user times for censored observations and those with events.*

$$f(t;\nu,\lambda) = \begin{cases} \frac{\nu}{\lambda}\left(\frac{t}{\lambda}\right)^{\nu-1} e^{-(t/\lambda)^{\nu}} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{3.19}$$

Where $\nu \geq 0$ is the shape parameter and $\lambda \geq 0$. We chose a Weibull distribution because with $\nu < 1$ we can specify a survival rate that decreases over time. This represents that we expect the hazard function function to be decreasing over time i.e. the greatest hazard is at the start of the experiment then the effect of the treatment wears off.

We will use the parameter ϕ to represent the global conversion from our model and use our dependent prior from section 2.2 to model ϕ_A and ϕ_B. This time we also have to define priors for ν_A, ν_B, λ_A and λ_B. For each $i = A, B$:

$$\begin{aligned} \nu_i &\sim U(0,1) \\ \lambda_i &\sim U(0,10) \end{aligned} \tag{3.20}$$

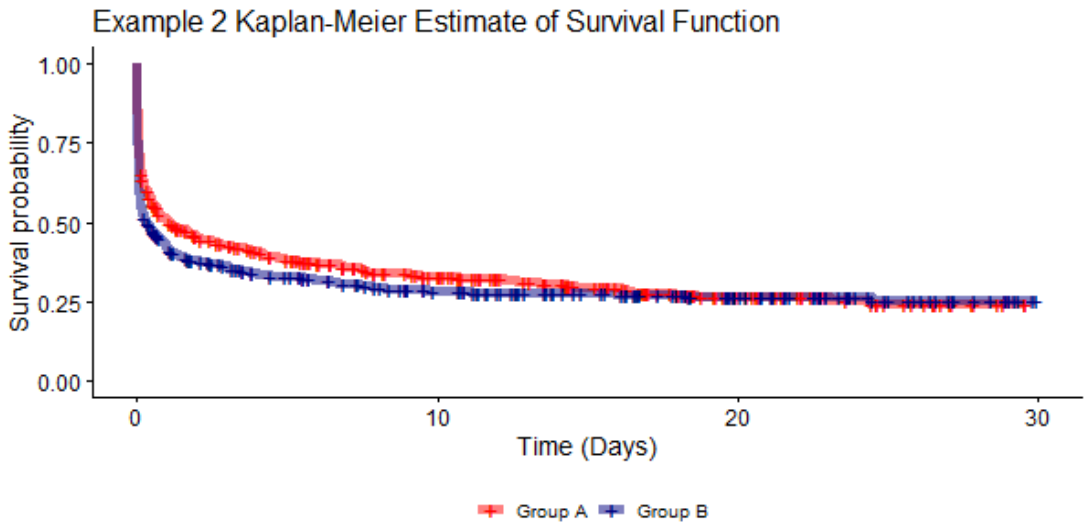


Figure 3.6: *Example 2 Kaplan-Meier estimate of the survival function. The steps in the curve indicate an event. Each "+" is a censored data point.*

For this example we used uniform priors that are not too precise, but represent our belief that ν should be between 0 and 1, and λ will be between 0 and 10. For their own analysis, the practitioner can pick different priors for ν an λ or even a different distribution for the time-to-event outcome (e.g. an exponential distribution if you believe the hazard to be constant over time).

This time we use the nested sampling method to get our posterior samples.

Posterior Inference

Using nested sampling we can calculate the marginal likelihood along side the posterior samples. The marginal likelihood from the nested sampling run was:

```
Marginal likelihood: ln(Z) = 346.1609 +- 0.3557589
```

To do inference on our posterior distribution, we can look at the marginal and joint distributions between the shape, scale and conversion rate between groups. Figure 3.7 shows combinations of joint marginal distributions between parameters. We can see that there is a quite a strong correlation between parameters, especially between conversion parameters ϕ, and scale parameters λ^2.

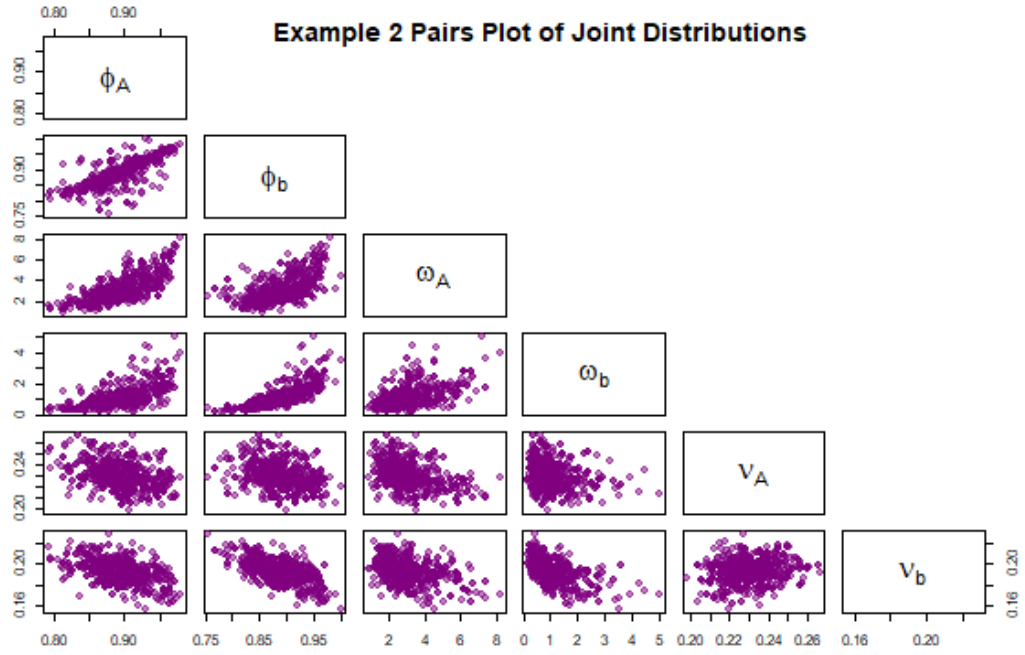


Figure 3.7: *Example 2 pairs of joint posterior distributions for the time to event outcome using the local survival model.*

Figure 3.8 shows the marginal distributions for parameters in each group. The point estimate for the mean and standard deviation for each parameter was:

- $\phi_A = 0.895 \pm 0.037$
- $\phi_B = 0.890 \pm 0.042$
- $\nu_A = 0.230 \pm 0.012$
- $\nu_B = 0.192 \pm 0.011$
- $\lambda_A = 3.080 \pm 1.267$
- $\lambda_B = 1.118 \pm 0.690$

[2]This is one of the reasons we picked nested sampling over JAGS for this example. In MCMC correlations between parameters can lead to extremely slow convergence of sampling chains, and sometimes to non-convergence.

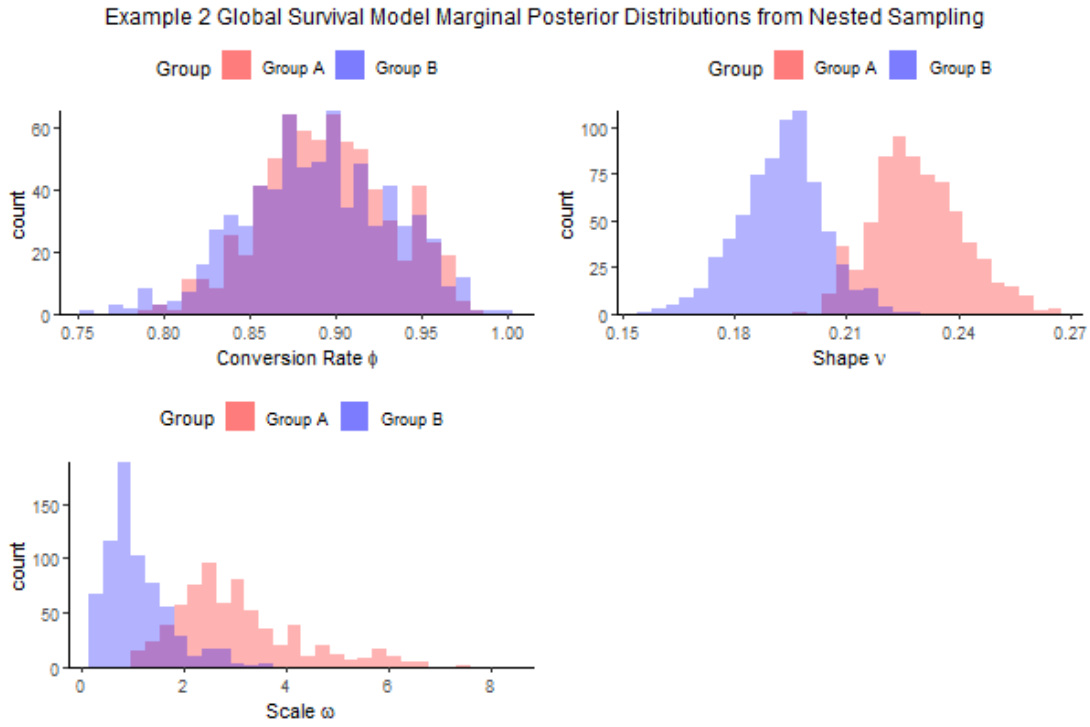


Figure 3.8: *Example 2 marginal posterior distributions for the time to event outcome using the local survival model.*

This time we have calculated the all-time conversion rate ϕ. Compared to our point estimates of $\theta_A = 69.6\%$ and $\theta_B = 68.6\%$ that we calculated in section 2.4, ϕ_A is about 89.5% and ϕ_B is 89.0%. These conversion rates are what we can expect over all-time if we let either variant run permanently. It is quite a big difference, and indicates that only 89% in either variant will ever convert. However this is only useful if it happens in a relevant time frame. If it takes years for everyone to convert ϕ is not as important as θ. From what we can see the there is no difference in global conversion rate in each variant.

Interpreting the shape and scale can be a little bit less intuitive than a success probability, but we can see that variant B has a lower shape and scale that variant A. This means the Weibull distribution of the time to the event gets pushed in towards the left, and its height increases. These values look reasonably separated with little overlap in the marginal distributions. However, both these values are relatively small, which makes visualising the density of the Weibull distribution a little tricky. What we are primarily interested in is if we can see a difference in survival between our two groups.

The survival function for the global survival model is:

$$S(\gamma|t) = 1 - \phi F(\gamma|t) \tag{3.21}$$

We can use our posterior samples to plot the posterior predictive survival distribution for each group and compare the results without original data. We have plotted these results in figure 3.9 and overlayed the Kaplan-Meier estimate of the data. We can see the model has been able to estimate the difference in survival curves between each variant reasonably and the posterior suggests there is a difference in survival between variants.

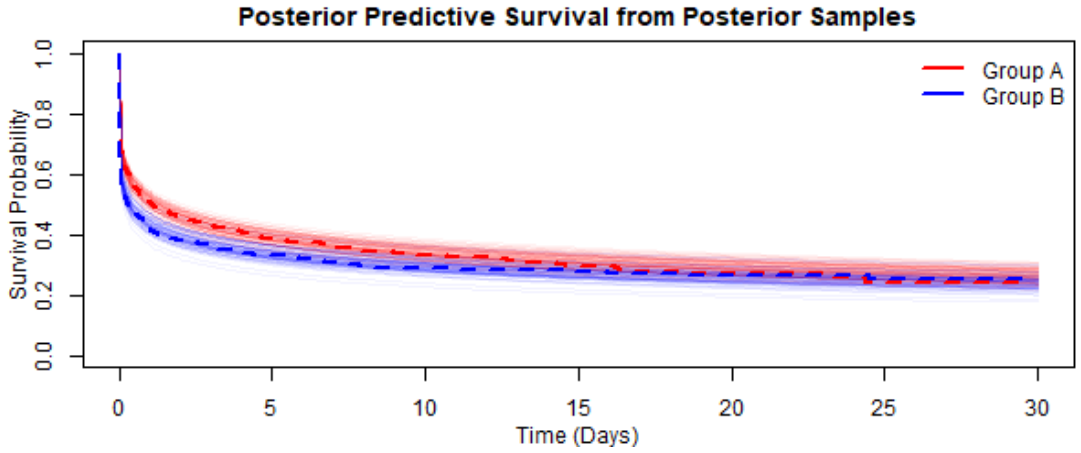


Figure 3.9: *Example 2 posterior predictive survival curves. Solid lines show the posterior samples, dotted lines show the actual data using the Kaplan-Meyer estimate of the survival curve.*

How do we answer our question if users are converting earlier in one variant? The survival curve shows us where the median survival is at the point 50% of users have done the event in one group. We can get a posterior estimate using a quantile function to tell us where this time point is for each group, and also get a 95% central credible interval for it.

The median survival for each variant was:

- Variant A: 0.57 Days, 95% CI (0.28, 1.19)
- Variant B: 0.14 Days, 95% CI (0.05, 0.38)

So in variant B by 1/14th of a day (1.6 hours) 50% of users have completed the event. In variant A by just over half a day (13.68 hours) 50% of users have completed the event.

What is also very helpful about the survival model over the binomial model is that we can estimate how many more conversions for each variant we can expect if we kept the experiment running for longer or dedided to pick one of the variants to go site-wide. Using the cumulative distribution $\phi * F(\gamma|t)$: after 60 days we would expect 77.3% CI (73.2%, 81.4%) to convert in variant A and 79.2% CI (74.8%, 82.8%) in variant B. So we get a $\sim$ 17% increase for continuing to run the experiment for double the length of time. In this case as we already know ϕ_A and ϕ_B, and θ_A and θ_B, are the same it doesn't really change our decision, but it is useful to know what might happen in the future

So while in section 2.4 we didn't see a difference conversion rate between variants, applying the survival model to the results has shown us there is a difference in when the event is happening!

3.5 Example 3 Revisited

Example 2 - First Action Revisited

In this experiment all users got the treatment at the same time and were in the experiment for the same duration. We are not interested in the survival rate after the 30 days of the experiment, as we know the effect of the email will be irrelevant after this time. However, during the duration of the experiment we are interested in if one of the variants is better at getting users to do the event earlier.

We can use our local survival model from section 3.3 to compare the survival rates between converters and the conversion rate for the duration of the experiment.

This time our model only looks at the times of those who do the event, and all censored observations are treated equally as non-converters. In figure 3.10, we can plot the results of the experiment using the Kaplan-Meier curve for those with observed events.

To analyse this data in the Bayesian framework, we will use our the likelihood from our local survival model with a Weibull distribution as the distribution for our event times. We will use the parameter θ to represent the conversion metric for the duration of the experiment (same as in section 2.) We will use the same dependent prior for θ_A

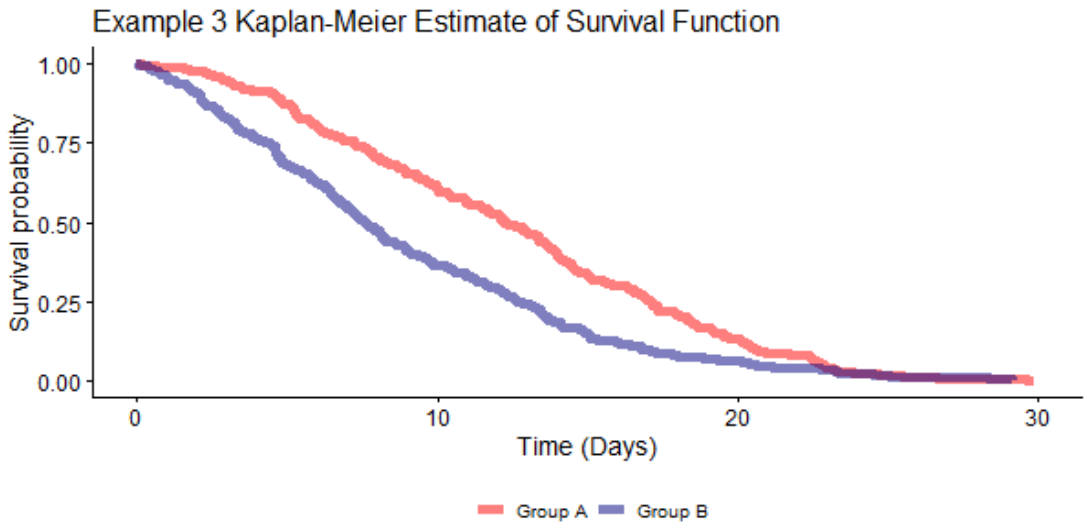


Figure 3.10: *Example 3 Kaplan-Meier estimate of the survival curve for users with observed events.*

and θ_B from section 2.2. Our prior for λ and ν were:

$$\begin{aligned} \nu_i &\sim U(0,3) \\ \lambda_i &\sim U(0,20) \end{aligned} \tag{3.22}$$

This time we expected the the scale to be larger, so we set a wider prior for λ. The uniform(0,3) prior for ν puts prior probability on $\nu > 1$ and well as $\nu < 1$, allowing the hazard to be increasing or decreasing. This describes our prior belief in the possibility that subjects may be interested in the feature when they get the email, but come back to do the event later in the week/month. In this case the hazard may be increasing, so we want prior probability to allow this.

We again used time we use nested sampling method to get our posterior samples.

Posterior Inference

The marginal likelihood from the nested sampling run was:

```
Marginal likelihood: ln(Z) = -3157.71 +- 0.350197
```

The point estimates for the mean and standard deviation for each parameter were:

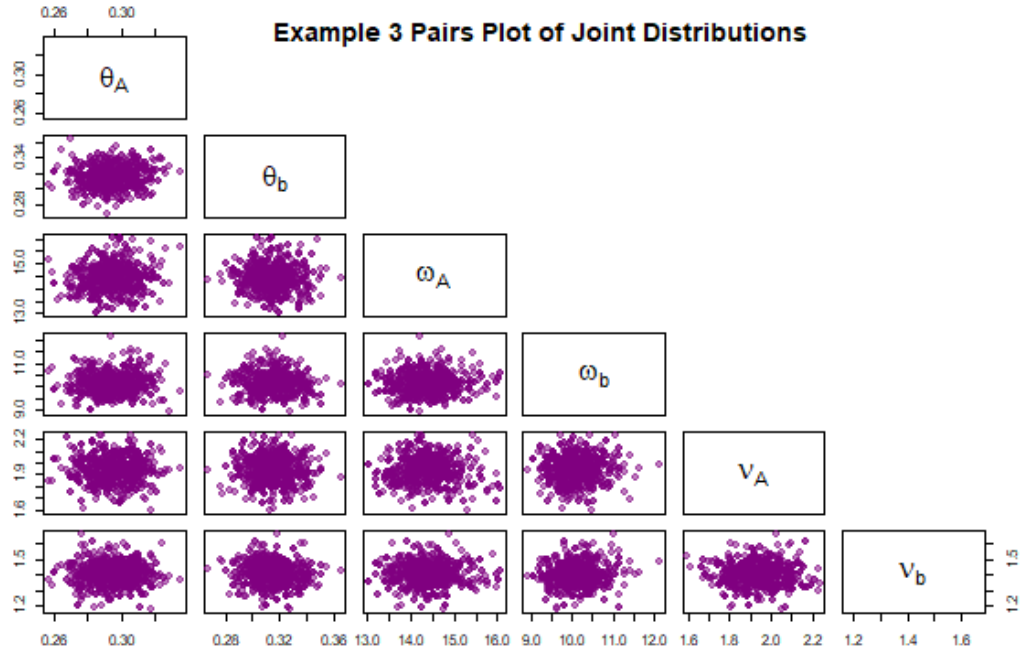


Figure 3.11: *Example 3 pairs of joint posterior distributions for the time to event outcome using the local survival model.*

- $\theta_A = 0.295 \pm 0.013$
- $\theta_B = 0.314 \pm 0.013$
- $\nu_A = 1.888 \pm 0.078$
- $\nu_B = 1.398 \pm 0.071$
- $\lambda_A = 14.366 \pm 0.549$
- $\lambda_B = 10.083 \pm 0.467$

Figure 3.11 shows combinations of joint marginal distributions between parameters. The joint distributions look reasonably uncorrelated. Figure 3.12 shows the marginal distributions for parameters in each group. θ_A and θ_B are similar to our estimates from section 2.5 in figure 2.13 using MCMC. The marginal posterior distributions ν and λ

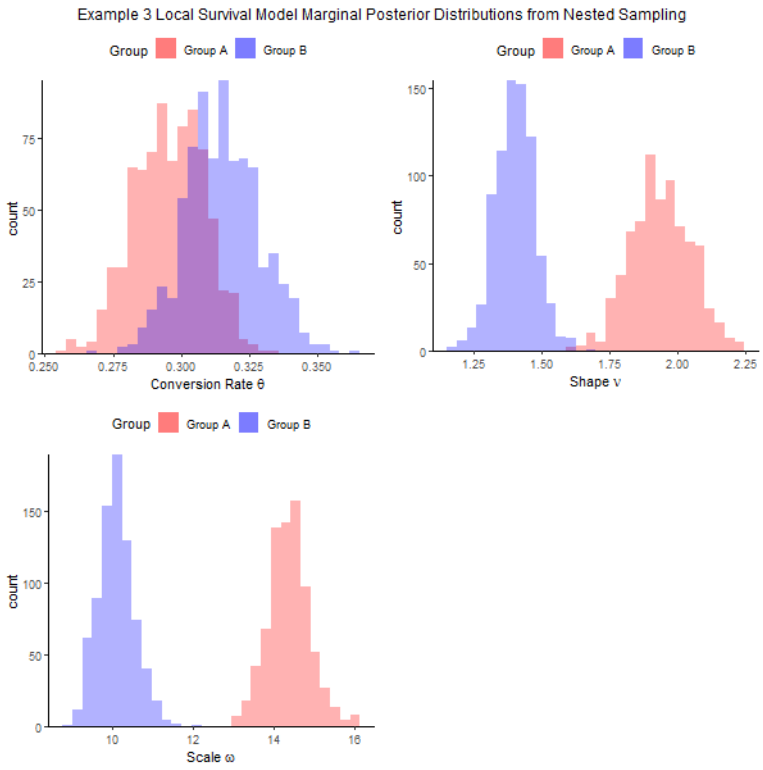


Figure 3.12: *Example 3 marginal posterior distributions for the time to event outcome using the local survival model.*

in each group show narrow distributions with no overlap. We can clearly see that the variant B has a lower shape and scale that variant A.

This means the Weibull distribution of the time to the event gets pushed in towards the left, and its height increases. We can see this in figure 3.13. This means that converters variant B are generally doing the event earlier than in variant A. The consequence of this is that in the 30 day period, there is a higher density of users doing the action, which also gives a higher experiment conversion rate for variant B.

The fact that our posterior distributions for ν_A and ν_B are both greater than one, shows that we were correct in our prior belief about the hazard function. We can plot the posterior predictive hazard function using samples from our posterior. Figure 3.13 shows our hazard function is increasing over 0 to 30 days.

We can also plot posterior predictive survival curves from our posterior distribution. We can see in figure 3.14 our survival estimate for variant B sits lower than variant A.

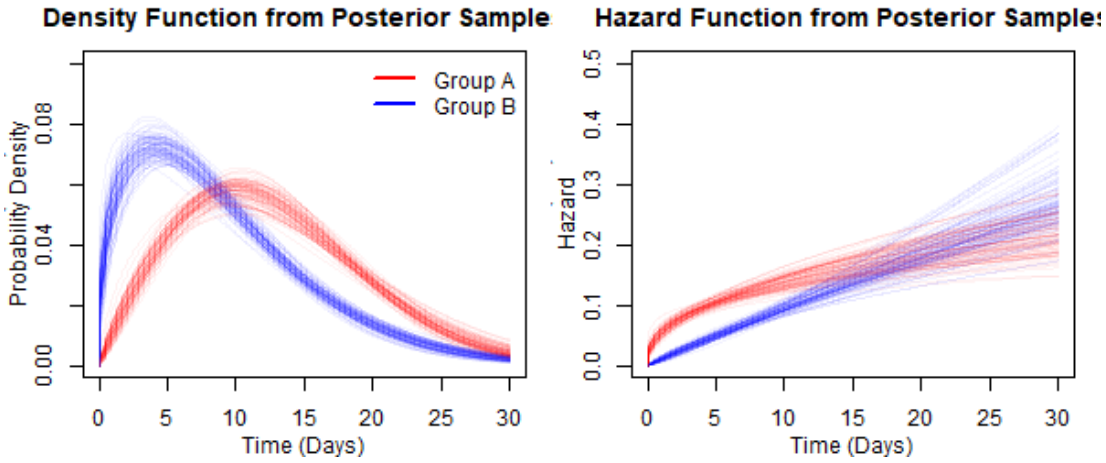


Figure 3.13: *Example 3 posterior predictive density and hazard*

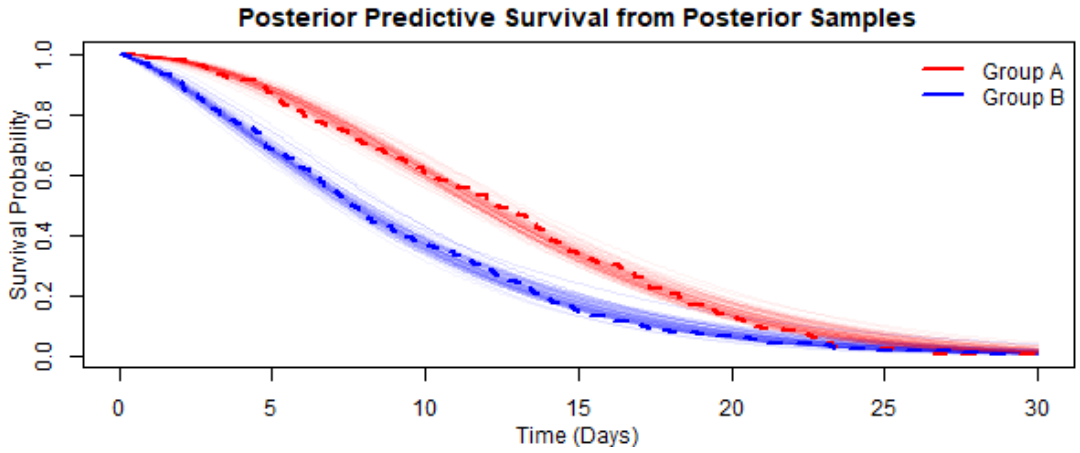


Figure 3.14: *Example 3 posterior predictive survival curves of converters. Solid lines show the posterior samples, dotted lines show the actual data using the Kaplan-Meyer estimate of the survival curve.*

The median survival for each variant was:

- Variant A: 11.08 Days, 95% CI(11.05, 12.79)[3]
- Variant B: 7.75 Days, 95% CI(7.01, 8.50)

[3]In this credible interval and in figure 3.14 we use the central 95% credible interval.

So while our estimate in section 2.5 showed inconclusive results for the difference between θ_A and θ_B, this time we can see a conclusive difference in when the event is happening. We would pick variant B because of its lower median survival rate.

Chapter 4

Expected Loss and Cost of Priors

In chapter 2 we touched on the idea of expected loss to define a loss function we could use to make a decision on what variant to pick at the end of the experiment. In this chapter we will introduce decision theory a bit more and describe a method to evaluate the cost of our chosen prior.

4.1 Expected Loss

One part of Bayesian inference involves using the posterior distribution to describe what we know about a parameter of interest, θ. The other part is using the posterior to make decisions about θ, in AB[1] testing it is usually a decision based on the estimated difference in parameters between our two treatments (i.e. $\theta_{diff} = \theta_B - \theta_A$). In other problems, it is just a decision based on a point estimate of the parameter θ itself. Either way, a poor estimate of θ would inevitably lead to making a worse decision than a good estimate. So to make this decision we need to know what estimate of θ is the best, but how do we know that?

The concept of *utility* gives a measure of how good or bad a particular decision is. The utility function $u(d, \theta)$ represents how good it would have been to make a decision d if the parameter value was θ. Conversely the *loss function* $L(d, \theta)$ (just negative utility) represents how bad it would have been to make a decision d if the parameter value was θ (O'Hagan et al. [2004]). We still have uncertainty about what is true for θ, so we cannot just choose the decision that gives us the greatest utility or smallest loss.

[1] It is essential to note that the AB test always ends in a decision - to pick one variant over the other. Even if we decide that the control or 'no experience' it is better than the new treatment or variant being tested, that is still making a decision.

However, we can find the optimal decision, which will be the one that maximises the posterior expected utility (i.e. minimised expected loss in equation 4.1).

$$E[L(d,\theta)] = \int L(d,\theta)P(\theta|D)d\theta \tag{4.1}$$

Let $\hat{\theta}$ be the decision we make for what estimate of θ to use. One form for the loss function is the absolute loss.

$$L(\hat{\theta},\theta) = |\hat{\theta} - \theta| \tag{4.2}$$

This treats all losses linearly proportional to each other, i.e. an error in the estimate of 10 is ten times worse than an error of 1[2]. The expected loss in equation 3.16 is minimised by the posterior median.

4.1.1 Expected Loss of Experimental Design

So what does this mean for priors? Consider a Bayesian inference problem with parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{x}$. Before we even know the data or the parameters, Bayes' rules tells us the relationship will be:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) \tag{4.3}$$

The joint distribution $p(\boldsymbol{\theta},\boldsymbol{x}) = p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})$ describes the connection between the data and parameters, and the fact that they are not independent. For this situation, we are not analysing a data set, but instead, we would like to know how well inference will perform. The joint distribution of prior and likelihood describes a prior distribution for what the posterior distribution will be.

We describe the expected loss in equation 3.16. However because we don't know the data we need to take another integral over the data space to get a *prior expectation of the posterior expected loss*:

$$E[L(d,\theta)] = \int\int L(d,\theta)P(\theta|D)d\ D \tag{4.4}$$

This expected loss describes how good the experimental design is with respect to the loss function.

[2]Another way might be a quadratic loss function $L(\hat{\theta},\theta) = (\hat{\theta} - \theta)^2$ that would treat being off by 10 100 times worse than if we were off by 1, but we will not be using that loss function here.

4.2 Expected Loss of a Bad Prior

Suppose we have a well thought out prior, $p(\boldsymbol{\theta})$ that we believe is a good representation of our beliefs about the true value. For a fixed experimental design we want to compare this 'good' $p(\boldsymbol{\theta})$ to another prior $q(\boldsymbol{\theta})$, we know be convenient for posterior inference, but a much poorer approximation about our prior beliefs.

We can adjust the expected loss in equation 4.4 to compare the two priors. Our prior $p(\boldsymbol{\theta})$ describes prior knowledge better than $q(\boldsymbol{\theta})$, so first we should take the expectation with respect to p, but execute the posterior expectation from the inner integral using q. We can then compare this expected loss to the expected loss in equation 4.4 using only $p(\boldsymbol{\theta})$. We will denote this difference in the expected loss as the 'cost' having a poor choice in prior.

Method

We used MCMC to simulate the expected loss from 4.4 to use this method for the difference between two treatment group parameters θ_A and θ_B for a given sample size N_1, N_2 in each treatment group. We use our prior from section 2.2 as our $p(\boldsymbol{\theta})$ and a $U(0,1)$ prior for each of θ_A and θ_B as our $q(\boldsymbol{\theta})$. We find that most decisions from AB tests tend to be linearly correlated with revenue, so we choose the absolute loss from equation 4.2 as our loss function.

The general method for simulating the expected loss is as follows (full method using `R` and `JAGS` can be found in the appendix). This can be run for a number of simulations to get a distribution for each expected loss:

```
for i in nsims:

 # Generate a pair of parameters from a sample size of N1 and N2 from the prior p
 pair = ... generate pair ...

 # Calculate the true difference between the pair
 true_diff = abs(pair1  - pair2)

 # Generate a dataset of parameter pairs using MCMC from the prior p
 new_data = ... do MCMC ...

 # Calculate the vector difference from generated pairs
```

```
diff = new_data[pair1] - new_data[pair2]

# Calculate Expected loss
expected_loss = mean(abs(diff - true_diff))

# Now do the same for the bad prior
# Generate a dataset of parameter pairs using MCMC from the prior q
new_data = ... do MCMC ...

# Calculate the vector difference from generated pairs
diff = new_data[pair1] - new_data[pair2]

# Calculate Expected loss
expected_loss_bad = mean(abs(diff - true_diff))
```

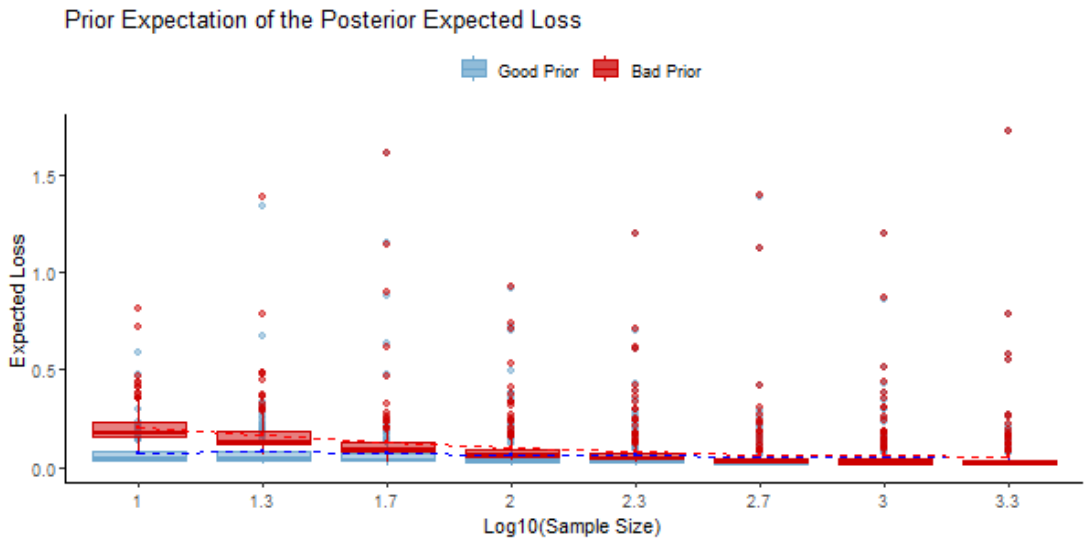


Figure 4.1: *Plot of distribution of prior expected loss between* $p(\boldsymbol{\theta})$ *and* $q(\boldsymbol{\theta})$ *across increasing sample size. The mean is marked by the dotted line.*

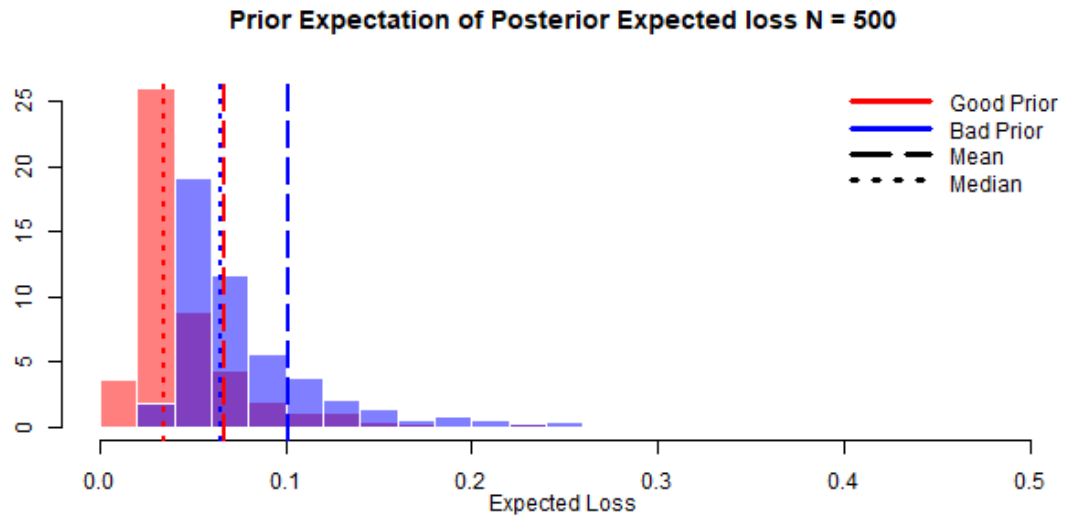


Figure 4.2: *Plot of distribution of prior expected loss between* $p(\boldsymbol{\theta})$ *and* $q(\boldsymbol{\theta})$ *for* $N = 100$

Summary

As we would expect in figure 4.1, the difference between the two priors converges as the sample size becomes large. However, figure 4.2 shows for a small sample size of $N = 100$ in each treatment group, the average expected loss is twice that of our 'bad' prior. This highlights the problem with the uniform prior; that there is not enough weight near the diagonal of the joint distribution of θ_A and θ_B (which corresponds to small effect size). When we are dealing with small effect sizes and limited sample size, there is a substantial advantage to our prior.

We mentioned in the introduction that an advantage of Bayesian inference is that we can use the prior to describe our initial beliefs about our parameters of interest before we even see the data. In AB testing we test small incremental changes and expect small effect sizes between variants. Using a prior that puts the majority of the probability on the difference between parameters being small fully utilises this advantage. As such, in AB testing when we want to specify a prior other than for convenience (i.e. uniform or beta), a dependent prior that models these beliefs is the best choice.

Chapter 5

Discussion & Further Work

The advantage of Bayesian inference is that our priors and likelihoods are fully customisable to the problem at hand. The downside is that this can be somewhat overwhelming to the practitioner unfamiliar with Bayesian inference. This dissertation gives a gentle introduction to what is possible for Bayesian AB testing.

We have proposed three new analysis methods for AB testing:

1. Dependent Priors for conversion metrics
2. Prior Expectation of Posterior Loss to evaluate the choice of prior
3. Survival models for time to event and conversion outcomes

We have brought attention to some of the overlooked design features of RCTs and how they apply to the web based and outbound based AB testing design. We applied our analysis methods to both designs to demonstrate how the practitioner can use them in real industry situations.

However, there are several aspects of further work that would be valuable to investigate further: First, there is value in investigating how to model multiple conversion parameters and the correlation between them in AB tests. Second, additional loss functions to evaluate expected loss for survival models. Lastly, evaluating the prior expectation of the posterior expected loss for the full prior of our survival model.

Bibliography

Susan Armijo-Olivo. The importance of determining the clinical significance of research results in physical therapy clinical research. *Brazilian journal of physical therapy*, 22 (3):175, 2018.

J M Kendall. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20(2):164–168, 2003. ISSN 1472-0205. doi: 10.1136/emj.20.2.164. URL `https://emj.bmj.com/content/20/2/164`.

John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

Anthony O'Hagan, Jonathan Forster, and Maurice George Kendall. Bayesian inference. 2004.

Optimizely. Optimization Glossary a/b testing. `https://www.optimizely.com/anz/optimization-glossary/ab-testing/`, 2019.

Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

Bonnie Sibbald and Martin Roland. Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal*, 316(7126):201, 1998.

John Skilling. Nested sampling. In *AIP Conference Proceedings*, volume 735, pages 395–405. American Institute of Physics, 2004.

Kenneth Stanley. Design of randomized controlled trials. *Circulation*, 115(9):1164–1169, 2007.

Chris Stucchio. Bayesian a/b testing at vwo. *Whitepaper, Visual Website Optimizer*, 2015.

Giovanni Tripepi, Nicholas C Chesnaye, Friedo W Dekker, Carmine Zoccali, and Kitty J Jager. Intention to treat and per protocol analysis in clinical trials. *Nephrology*, 25 (7):513–517, 2020.

Appendix A

Code

Full repository of code and data can be found at
`https://github.com/h-jam/Topics-in-Bayesian-AB-Testing`

Below we include the code to get the posterior distributions for each of the examples included.

Example 1

```
library(rjags)
library(yaml)

# A data set
data <- yaml.load_file("example-1-landing-page/example-1-data.yaml")
df <- data.frame(data)

model = "model {
        # Priors
        logits[1] ~ dnorm(0, 1)
        logits[2] ~ dt(logits[1], (1/0.1)^2, 1)

        logits_[1] ~ dnorm(0, 1)
        logits_[2] ~ dt(logits_[1], (1/0.1)^2, 1)

        for(i in 1:2){
```

```
                # Transform from logit to success prob.
                theta[i] <- 1.0/(1.0 + exp(-logits[i]))
                psi[i] <- 1.0/(1.0 + exp(-logits_[i]))

                x1[i] ~ dbin(theta[i], N[i])
                x2[i] ~ dbin(psi[i], N[i])
        }

        loss[1] <- step(psi[2] - psi[1])*(psi[2] - psi[1])
        loss[2] <- step(psi[1] - psi[2])*(psi[1] - psi[2])

} "

variable_names = c("theta", "psi", "loss")

jm = jags.model(textConnection(model), data)
out = coda.samples(jm, variable.names = var_names, n.iter = 100000, thin=10)
```

Example 2

Binomial Model in Jags

```
# A data set
data <- yaml.load_file("example-2-first-action/example-2-data.yaml")
df <- data.frame(data$data)

data <- list(
        N = table(df$group),
        x = table(df$group, df$event)[,2])

model = "model {
        # Priors
        logits[1] ~ dnorm(0, 1)
        logits[2] ~ dt(logits[1], (1/0.1)^2, 1)

        for(i in 1:2){
```

```
                # Transform from logit to success prob.
                theta[i] <- 1.0/(1.0 + exp(-logits[i]))
                x[i] ~ dbin(theta[i], N[i])
        }

        loss[1] <- step(theta[2] - theta[1])*(theta[2] - theta[1])
        loss[2] <- step(theta[1] - theta[2])*(theta[1] - theta[2])

}"

variable_names = c("theta", "loss")

jm = jags.model(textConnection(model), data)
out = coda.samples(jm, variable.names = var_names, n.iter = 100000, thin=10)
```

Global Survival Model Nested Sampling

Prior and log likelihood functions for nested sampling:

```
names = c("phi_a", "phi_b", "w_a", "w_b", "nu_a", "nu_b")
num_params = 6

us_to_params = function(us){

        # Vector to be returned as the result of the function
        params = rep(NA, num_params)

        # Apply the names
        names(params) = names

        logit1 = qnorm(us[1])
        logit2 =  logit1 + qt(us[2], df = 1)*0.1
        params["phi_a"] = 1.0/(1.0 + exp(-logit1))
        params["phi_b"] = 1.0/(1.0 + exp(-logit2))
        params["w_a"] = qunif(us[3], min = 0, max = 10)
        params["w_b"] = qunif(us[4], min = 0, max = 10)
        params["nu_a"] = qunif(us[5])
        params["nu_b"] = qunif(us[6])
```

```
        return(params)
}

log_likelihood <- function(params){

        ll <- function(x1, x2, phi, scale, shape){

                l1 <- sum(log(phi) + log(shape/scale) +
                        (shape - 1.0)*log(x1/scale) -
                        (x1/scale)^shape)

                l2 <- sum(log(1 - phi*(1-exp(-(x2/scale)^shape))))

                logL <- l1 + l2

                return(logL)
        }

        a <- ll(x1 = converted1, x2 = non_converted1,
                phi = params['phi_a'], scale = params['w_a'],
                shape =  params['nu_a'])

        b <- ll(x1 = converted2, x2 = non_converted2,
                phi = params['phi_b'], scale = params['w_b'],
                shape = params['nu_b'])

        logL <- a + b

        if(is.na(logL)){
                logL = -Inf
                }

        return(logL)

        }
```

Example 3

Binomial Model in Jags

```
# A data set
data <- yaml.load_file("example-3-edm-promotion/example-3-data.yaml")
df <- data.frame(data$data)

data <- list(
        N1 = table(df$group), # total sample in each group
        N2 = table(df$open, df$group)[2,], # opened in each group
        x1 = table(df$event, df$group)[2,], # total did event
        x2 = table(df$event[df$open == 1], df$group[df$open == 1])[2,]
        # total did event that opened
        )

model = "model {
        # Priors
        logits[1] ~ dnorm(0, 1)
        logits[2] ~ dt(logits[1], (1/0.1)^2, 1)

        logits_[1] ~ dnorm(0, 1)
        logits_[2] ~ dt(logits[1], (1/0.1)^2, 1)

        for(i in 1:2){

                # Transform from logit to success prob.
                theta[i] <- 1.0/(1.0 + exp(-logits[i]))
                psi[i] <- 1.0/(1.0 + exp(-logits_[i]))

                x1[i] ~ dbin(theta[i], N1[i])
                x2[i] ~ dbin(psi[i], N2[i])
        }

        loss_theta[1] <- step(theta[2] - theta[1])*(theta[2] - theta[1])
        loss_theta[2] <- step(theta[1] - theta[2])*(theta[1] - theta[2])

        loss_psi[1] <- step(psi[2] - psi[1])*(theta[2] - psi[1])
        loss_psi[2] <- step(psi[1] - psi[2])*(psi[1] - psi[2])
```

```
} "

variable_names = c("theta", "psi", "loss_theta", "loss_psi")

jm = jags.model(textConnection(model), data)
out = coda.samples(jm, variable.names = var_names, n.iter = 100000, thin=10)
```

Local Survival Model Nested Sampling

Prior and log likelihood functions for nested sampling:

```
names = c("theta_a", "theta_b", "w_a", "w_b", "nu_a", "nu_b")
num_params = 6

us_to_params = function(us){

        # Vector to be returned as the result of the function
        params = rep(NA, num_params)

        # Apply the names
        names(params) = names

        logit_thetaa =  qnorm(us[1], mean=0, sd=1)
        logit_thetab =  logit_thetaa + qt(us[2], df = 1)*0.1

        # transform back
        params["theta_a"] = exp(logit_thetaa)/(1 + exp(logit_thetaa))
        params["theta_b"] = exp(logit_thetab)/(1 + exp(logit_thetab))

        params["w_a"] = qunif(us[3], min = 0, max = 20)
        params["w_b"] = qunif(us[4], min = 0, max = 20)

        params["nu_a"] = qunif(us[5], min = 0, max = 3)
        params["nu_b"] = qunif(us[6], min = 0, max = 3)

        return(params)
}
```

```
log_likelihood <- function(params){

        ll <- function(x, censored, theta, scale, shape){

                log_integral <- log(1 - exp(-(30/scale)^shape))

                times <- log(shape/scale) +
                        (shape - 1.0)*log(x/scale) -
                        (x/scale)^shape

                l1 <- sum(log(theta) + times - log_integral)

                l2 <- log(1 - theta)*censored

                logL <- l1 + l2

                return(logL)
                }

        a <- ll(x = times1, censored = censored1,
                theta = params['theta_a'], scale = params['w_a'],
                shape = params['nu_a'])

        b <- ll(x = times2, censored = censored2,
                theta = params['theta_b'], scale = params['w_b'],
                shape = params['nu_b'])

        logL <- a + b

        if(is.na(logL)){
                logL = -Inf
        }

        return(logL)

}
```