Bayesian Statistics the Fun Way

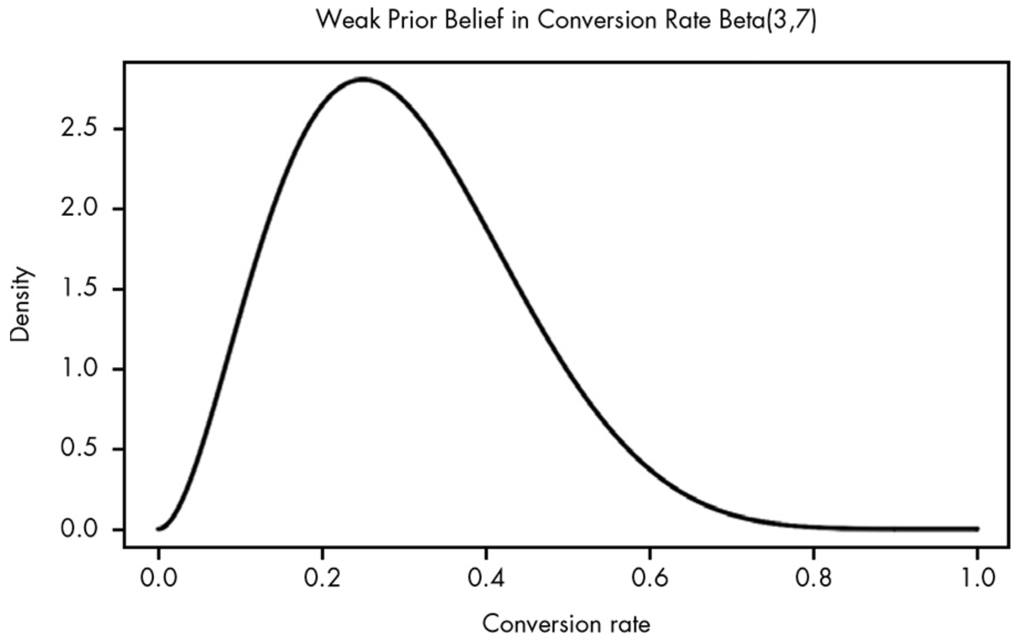
Chapter 15. From Parameter Estimation to Hypothesis Testing: Building a Bayesian A/B Test

Example: A/B test to see if removing an image from an email will increase the *click-through rate (CTR)* against the belief that removing it will hurt the CTR.

Note: "*A/B tests can be performed using classical statistical techniques such as t-tests, but building our test the Bayesian way will help us understand each part of it intuitively and give us more useful results as well*" [source]

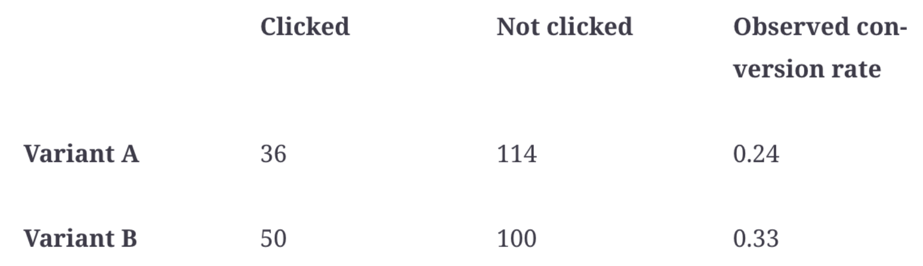
Setup

- How many users? 300 users, 150 per group
- What is our prior probability? From previous email campaigns, the expected CTR is 30%. To make things simple, we use the same prior for both variants. We also choose a pretty weak version of our prior distribution, meaning that it considers a wider range of conversion rates to be probable; We don't really know how well we expect B (the treatment group) to do compared to A (the control group), and this is a new email campaign, so other factors could cause a better or worse conversion. We'll settle on Beta(3,7) for our prior probability distribution (3 clicks, 7 non-clicks 30% CTR). This distribution allows us to represent a beta distribution where 0.3 is the mean, but a wide range of possible alternative rates are considered.



- Now we need our likelihood, which means we need to collect data

Table 15-1: Email Click-through Rates

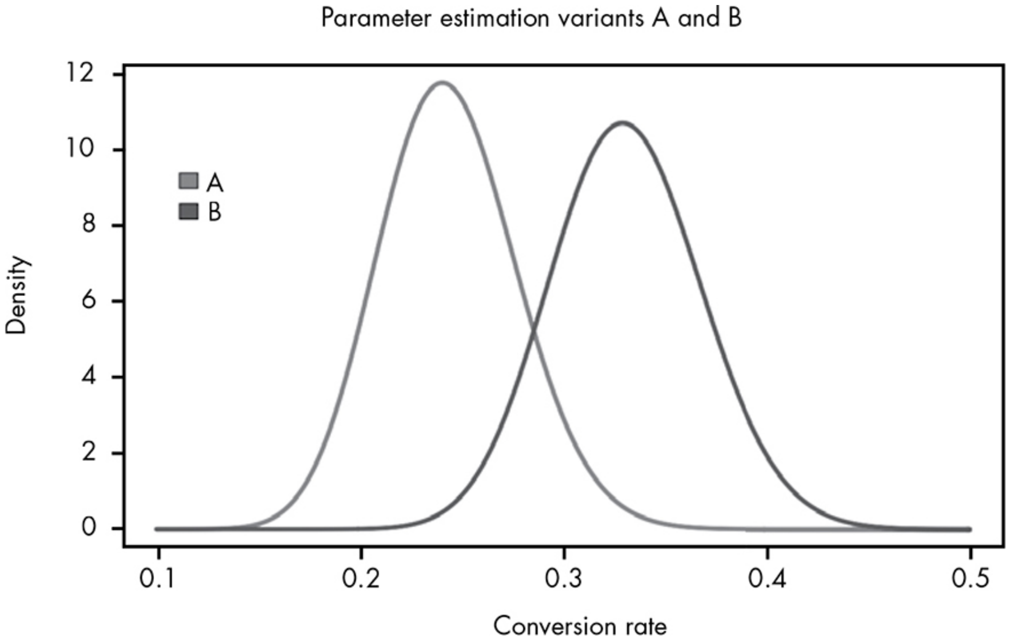


	Clicked	Not clicked	Observed conversion rate
Variant A	36	114	0.24
Variant B	50	100	0.33

We can treat each of these variants as a separate parameter that we're trying to estimate. In order to arrive at a posterior distribution for each, we need to combine both their likelihood distribution and prior distribution. For the likelihood of each, we'll again use the beta distribution, making the number of times the link was clicked through and the number of times it was not.

$$\text{Beta}({}_{\text{posterior}},{}_{\text{posterior}}) = \text{Beta}({}_{\text{prior}} + {}_{\text{likelihood}},{}_{\text{prior}} + {}_{\text{likelihood}})$$

Variant A will be represented by Beta(36+3,114+7) and variant B by Beta(50+3,100+7).



- How can we interpret the results? Our data suggests that variant B is has a higher CTR. However, we can also see here that there’s an overlap between the possible true conversion rates for A and B. What if we were just unlucky in our A responses, and A’s true conversion rate is in fact much higher? What if we were also just lucky with B, and its conversion rate is in fact much lower? How sure can we be that B is the better variant? Monte Carlo simulation

  A *Monte Carlo simulation* is any technique that makes use of random sampling to solve a problem. In this case, we’re going to randomly sample from the two distributions, where each sample is chosen based on its probability in the distribution so that samples in a high-probability region will appear more frequently.

  We can imagine that the posterior distribution represents all the worlds that could exist based on our current state of beliefs regarding each conversion rate. Every time we sample from each distribution, we’re seeing what one possible world could look like. The more frequently we sample, the more precisely we can tell how often B is the better variant. From our samples, we can measure the ratio of samples where B is the best and get an exact probability that B is in fact greater than A.

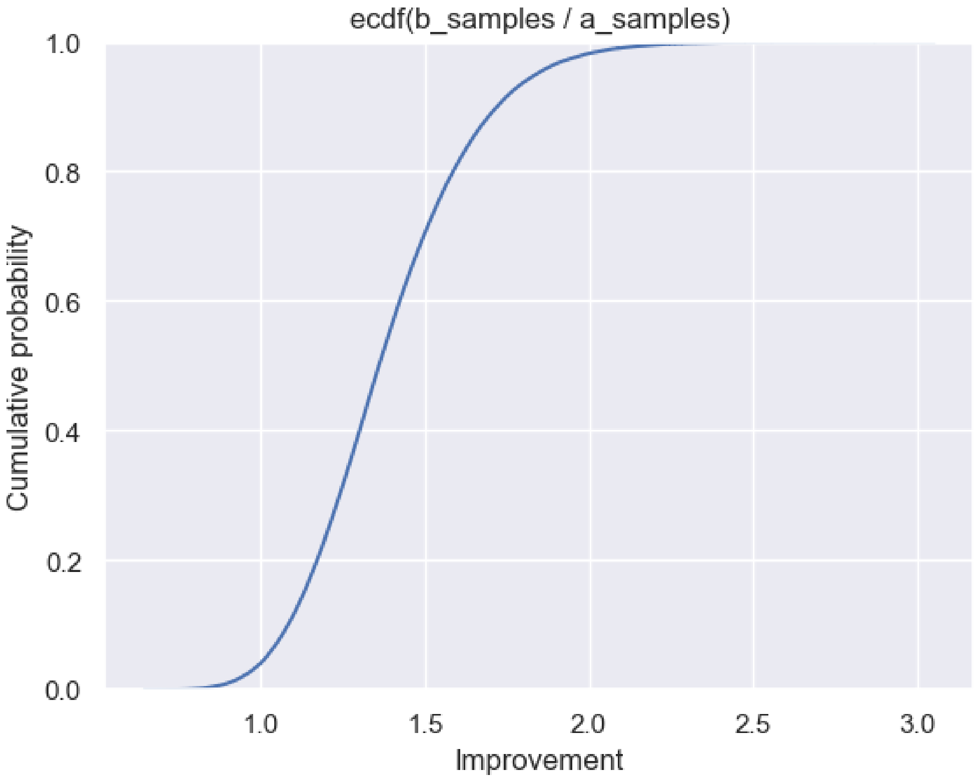
- Monte Carlo simulation setup:
  - We can consider each comparison of two samples a single trial. The more trials we run, the more precise our result will be. We start with 100,000 trials
  - Next we’ll put in our prior alpha and beta and observed data
  - Then we need to collect samples from each variant's beta distribution e.g. using numpy
  - Finally, we compare how many times the group B samples are greater than the group A samples
  - What we see here is that in 96% of the 100,000 trials, variant B was better

Monte Carlo simulation

```
import numpy as np

rng = np.random.default_rng(123)

n_trials = 100_000
prior_alpha = 3
prior_beta = 7

# NOTE: A and B are the groups in the A/B test,
# not to be confused with the alpha and beta from the Beta distribution below
a_clicks, a_no_clicks = (36, 114)
b_clicks, b_no_clicks = (50, 100)

a_samples = rng.beta(
    a=a_clicks + prior_alpha, b=a_no_clicks + prior_beta, size=n_trials
)
b_samples = rng.beta(
    a=b_clicks + prior_alpha, b=b_no_clicks + prior_beta, size=n_trials
)
p_b_greater = np.mean(b_samples > a_samples)
```

    - Comparing this to the *t*-test, this is roughly equivalent - if we used a Beta(1,1) prior - to getting a *p*-value of 0.04 from a single-tailed *t*-test (often considered “statistically significant”)
- How much better is B than A? We can take the exact results from our last simulation and look at compute `b_samples / a_samples`. This will give us a distribution of the relative improvements from variant A to variant B

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.set()
sns.ecdfplot(improvement)
plt.xlabel("Improvement")
plt.ylabel("Cumulative probability")
plt.title("ecdf(b_samples / a_samples)")
```



The results show:

- There is a small chance that A is better, but even if it is better, it’s not going to be by much. The chance that B is 20 percent worse is roughly the same that it’s 100 percent better.
- There’s about a 25% chance that variant B will give a 50% or more improvement over A (cumulative probability == 0.75)

Exercises:

- **Q1. Suppose a director of marketing with many years of experience tells you he believes very strongly that the variant without images (B) won’t perform any differently than the original variant. How could you account for this in our model? Implement this change and see how your final conclusions change as well.**

  A1. You can account for this by increasing the strength of the prior. For example:

  `prior_alpha = 300`

  `prior_beta = 700`

  This will require much more evidence to change our beliefs. Our new p_b_superior is 0.74, which is much lower than our original 0.96.

- **Q2. The lead designer sees your results and insists that there's no way that variant B should perform better with no images. She feels that you should assume the conversion rate for variant B is closer to 20% than 30%. Implement a solution for this and again review the results of our analysis.**

  A2. Rather than using one prior to change our beliefs, we want to use two - one that reflects the original prior we had for A and one that reflects the lead designer's belief in B. Rather than use the weak prior, we'll use a slightly stronger one:

  ```
  a_prior_alpha = 30

  a_prior_beta = 70

  b_prior_alpha = 20

  b_prior_beta = 80
  ```

  And when we run this simulation, we need to use two separate priors:

  ```
  a_samples = rng.beta(a_clicks + a_prior_alpha, a_no_clicks + a_prior_beta, n_trials)

  b_samples = rng.beta(b_clicks + b_prior_alpha, b_no_clicks + b_prior_beta)

  p_b_superior = np.mean(b_samples > a_samples)
  ```

  The `p_b_superior` this time is 0.66, which is lower than before, but still slightly suggests that B might be the superior variant.

References:

- Bayesian Statistics Book (O'Reilly)

Useful links:

- Python code (not used but seems useful)
- PyMC (not used but might be an alternative)