Business Experiments

Week 5: Bayesian A/B Tests

Elea McDonnell Feit

Agenda

Simpson's Paradox

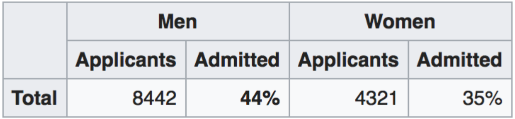
Bayesian Analysis of A/B Tests

- Binary Data: Beta-Binomial Model
- Continuous Data: Normal Normal Model

Profit-Maximizing Sample Size

Simpson's Paradox

UC Berkley admissions data (1973)

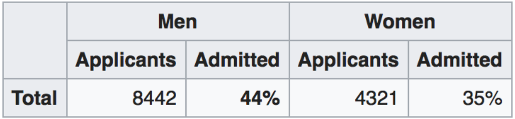


	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

Is this discrimination?

Simpson's Paradox

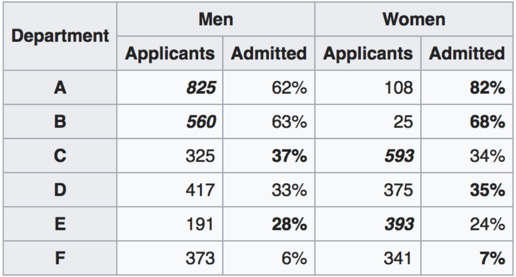
UC Berkley admissions data (1973)



	Men		Women	
	Applicants	**Admitted**	**Applicants**	**Admitted**
Total	8442	**44%**	4321	35%

Is this discrimination?

No! Women tend to apply to highly-competitive programs.



Department	**Men**		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

How do we avoid Simpson’s Pardox in our own analysis?

How do we avoid Simpson's Pardox in our own analysis?

Randomize

Bayesian Analysis of A/B Tests

When analyzing an experiment, our goal is to determine if we have enough data to make a decision.

Most experimenters use confidence intervals and hypothesis tests to analyze the data from A/B tests. This approach is called frequentist or classical.

An less popular (but useful!) alternative is Bayesian analysis, which is an entirely different approach to determining if we have enough data to make a desision.

Motivating question

Jon and Mark are both candidates for a sales position.

Jon

Talks to 3 customers and gets 2 of them to buy.

Mark

Talks to 3 customers and gets 1 of them to buy.

Does Jon have a better success rate than Mark?

Motivating question

Jon and Mark are both candidates for a sales position.

Jon

Talks to 3 customers and gets 2 of them to buy.

Mark

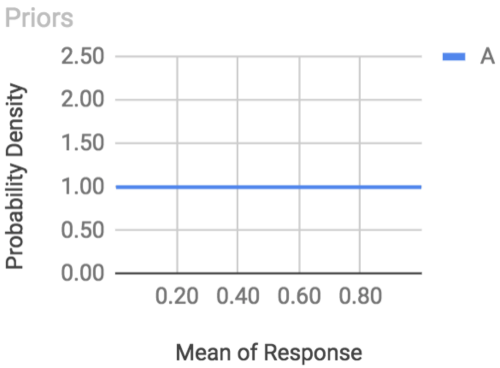
Talks to 3 customers and gets 1 of them to buy.

Does Jon have a better success rate than Mark?

Prior distribution for Mark

We want to know Mark's long-run average success rate p. We begin with a prior distribution that tells us likely different success rates are.

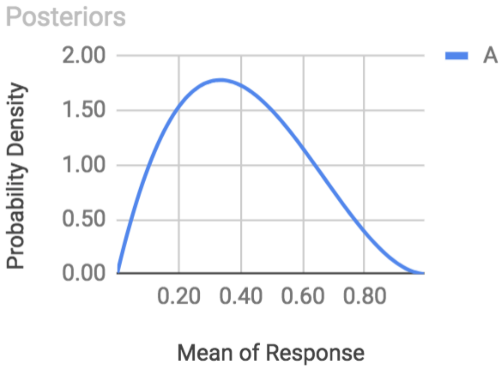
This prior says that (before we see the data) all values of the success rate for Mark are equally likely.



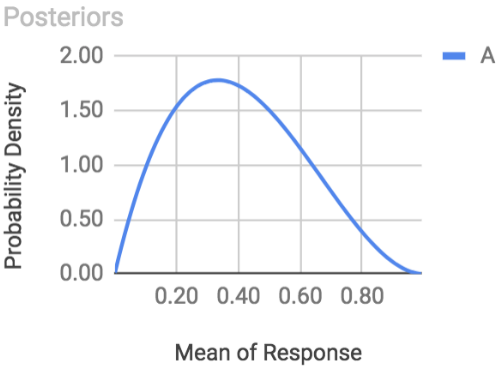
Posterior distribution for Mark

After we observe that Mark had one success in two tries, we can update our distribution describing how likely different success rates are.

The updating uses Bayes' rule, but we will not go into the details of how this works.



What can we do with the posterior?



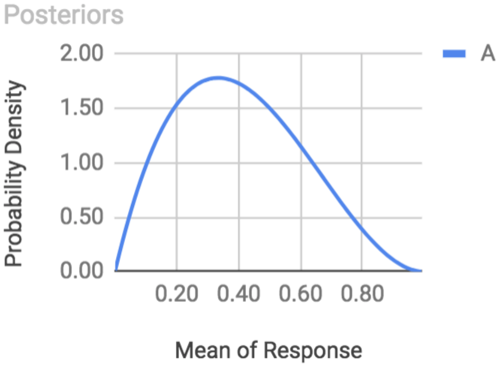
We can compute the credible interval, which is the range of values that Mark's true success rate is likely to take. The 95% credible interval for Mark's success rate is [0.072, 0.796]. In business language, there is a 95% chance that Mark's success rate is between 7.2% and 79.6%.

Demo: Bayesian Analysis in Google Sheets

Posteriors for Mark and for Jon

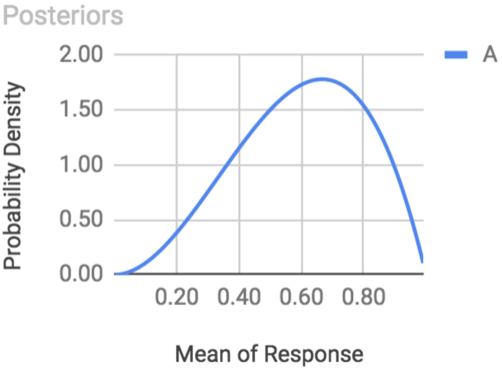
We can also compute a posterior for Jon and compare to Mark.

Posterior of Mark's success rate



95% CI = [0.072, 0.796]

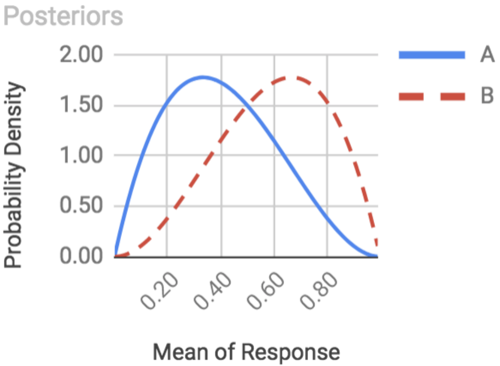
Posterior of Jon's success rate



95% CI = [0.207, 0.924]

Comparing Mark and Jon

Our business question is "Is Jon better than Mark?". To answer this question, we can use the posteriors to compute the probability that Jon is better than Mark. The probability that Jon is better than Mark is 75%.



More updating

Knowing that there is a 75% chance that Mark has a better success rate than Jon, the business can decide if they are comfortable making a decision about who to hire at this point.

If the business wants higher certainty before making a decision, you can collect more data and update your distributions again (and again and again).

It is a bad idea to conduct a classical hypothesis test over and over again. Doing so invalidates the p-value.

Where do priors come from?

One thing some people don't like about Bayesian analysis is that you have to set the priors and these are subjectively determined by the analyst. We can set priors in several ways:

- Use "Diffuse" or "flat" priors that put equal weight on all reasonable values of the success rate.
- Base priors on previous experience.
    - For instance, click through rates on ads are rarely more than 10% and often very small. When analyzing click-through rates, I can set my prior accordingly.

Either way, when you have a lot of data, the prior becomes unimportant.

Demo: Bayesian Analysis in Google Sheets

Practice Exercise

The A version of an email was sent to 500 people can had an open rate of 5%, while the B version was sent to 100 had an open rate of 6%. What is the probability that version B is better?

Beta-Binomial Model Details [optional]

Model:

$$y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } (1-p) \end{cases}$$

Prior:

$$p \sim \text{Beta}(\alpha_0, \beta_0)$$

Posterior:

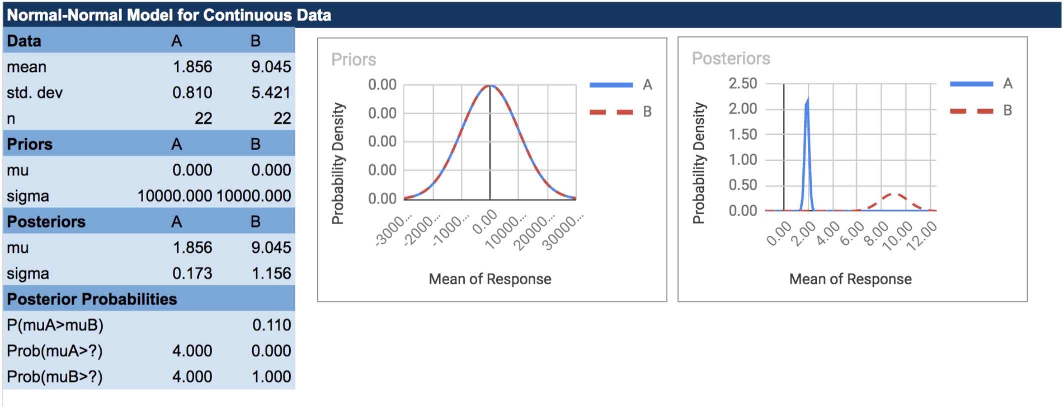
If we observe s success in n observations, the updated posterior is:

$$p \sim \text{Beta}(\alpha_0 + s, \beta_0 + (n - s))$$

You can use `BETA.DIST` in Google Sheets or `dbeta()` in R to make plots of the posterior or compute credible intervals. To compute $P(p_A > p_B)$, you can take random draws from each posterior and count how often your draw for A is greater than your draw for B.

Continuous Data

We can do similar prior-posterior updating for the mean of a continuous response variable.



Normal-Normal Model for Continuous Data

Data	**A**	**B**
mean	1.856	9.045
std. dev	0.810	5.421
n	22	22
Priors	**A**	**B**
mu	0.000	0.000
sigma	10000.000	10000.000
Posteriors	**A**	**B**
mu	1.856	9.045
sigma	0.173	1.156
Posterior Probabilities		
P(muA>muB)		0.110
Prob(muA>?)	4.000	0.000
Prob(muB>?)	4.000	1.000

Practice Exercise

Two different ads were shown at the begining of a YouTube video. Some users saw ad A, while others saw ad B (at random). The number of minutes each viewer watched the video after seeing the ad was recorded. What is the probability that ad A makes people watch the video longer?

Ad	Viewers	Average Minutes Viewed	St.Dev
A	100	28.2	5.3
B	100	20.7	8.8

Normal-Normal Model Details [optional]

Model:

$$y_i \sim N(m, s) \qquad s \text{ known}$$

Prior:

$$m \sim N(\mu_0, \sigma_0)$$

Posterior:

If we observe s success in n observations, the updated posterior is:

$$m \sim N(\mu_1, \sigma_1)$$

$$\sigma_1 = \sqrt{\left(\frac{1}{\sigma_0^2} + \frac{n}{s^2}\right)^{-1}} \qquad \mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{s^2}\right)$$

You can use `NORM.DIST` in Google Sheets or `dnorm()` in R to make plots of the posterior or compute credible intervals.

Pros and Cons of Bayesian Analysis

Pros:

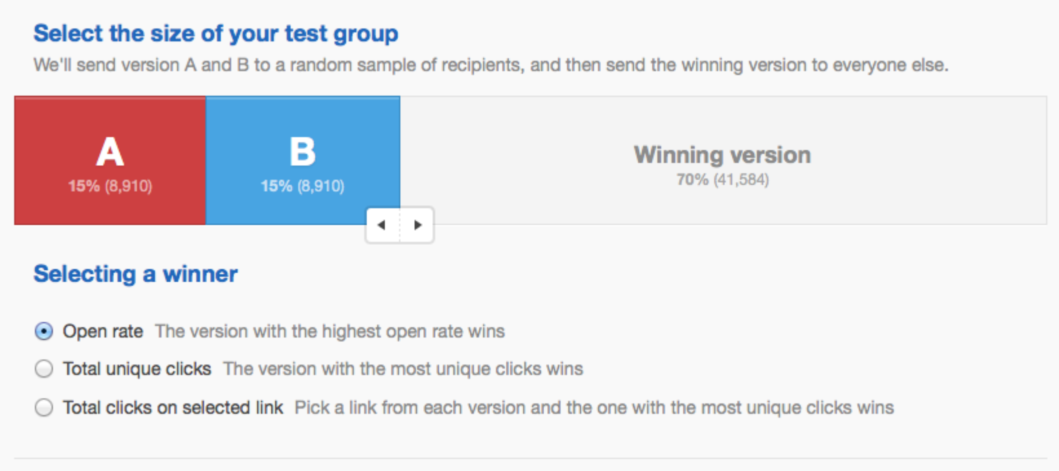
- Gives a direct answer to the question “What is the chance that treatment A is better than treatment B?”
- Allows you to analyze the data as it comes in and stop when you’ve got satisfactory confidence in your decision.

Cons:

- You have to set priors based on your own judgement.
- Many analysts are decision makers are more familar with classical methods.

Demo: Bayesian Analysis in R

Test Planning



Select the size of your test group

We'll send version A and B to a random sample of recipients, and then send the winning version to everyone else.

Selecting a winner

● Open rate The version with the highest open rate wins

○ Total unique clicks The version with the most unique clicks wins

○ Total clicks on selected link Pick a link from each version and the one with the most unique clicks wins

Source: Zapier.com

Sample Size Recommendation for Hypothesis Test (Classical)

$$n_1 = n_2 \approx 2(z_{(1-\alpha)/2} + z_\beta)^2 \left(\frac{s_1^2 + s_2^2}{\delta^2} \right)$$

s_1, s_2: s.d. of response
δ: difference to detect
α: chance of false positive (significance)
β: chance of a false negative if the true difference is δ

Example Sample Size Calculation

95% confidence (α)
80% power (β)
likely conversion rate = 5%
minimum detectable effect = 0.5% (10% lift)

$$n_1 = n_2 \approx 2(1.96 + 0.84)^2 \left(\frac{2(0.05(1 - 0.05))}{0.005} \right) \approx 31,000$$

Profit-Maximizing Sample Size

The goal of a tactical A/B test is to maximize total profit earned, which is the sum of the profit earned in the test and in the deploy stage:

$$\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}$$

Profit-Maximizing Sample Size

The goal of a tactical A/B test is to maximize total profit earned, which is the sum of the profit earned in the test and in the deploy stage:

$$\text{Profit}_{\text{Test}} + \text{Profit}_{\text{Deploy}}$$

In the test stage we know we are giving some people a worse treatment, so we want to keep the test small.

But, if the test is bigger, we are more less likely to deploy the wrong treatment.

Profit-Maximizing Sample Size: The Idea

We can set the sample size to make the optimal trade-off between the opportunity cost of the test and the risk of making a deployment error.

$$\max_{n_1, n_2} E[P_{\text{Test}} + P_{\text{Deploy}}]$$

You can only compute the expected profit if you are Bayesian and have priors.

Profit-Maximizing Sample Size: Formula (continuous data)

The optimal sample size is:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2$$

Profit-Maximizing Sample Size: Formula (continuous data)

The optimal sample size is:

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2$$

Where:

- N is the size of the total customer population
- s is the standard deviation of the response $(s = s_1 = s_2)$
- σ is the standard deviation of the prior on the mean response for both treatments in the normal-normal model

Bayesian Sample Size Formula: (continuous data)

$$n_1^* = n_2^* = \sqrt{\frac{N}{4}\left(\frac{s}{\sigma}\right)^2 + \left(\frac{3}{4}\left(\frac{s}{\sigma}\right)^2\right)^2} - \frac{3}{4}\left(\frac{s}{\sigma}\right)^2$$

Profit maximizing sample size:

- is higher when the data is noisy (s)
- is higher when the population size (N) is larger
- is smaller when there is likely to be a greater difference in peformance between treatments (σ)

Example

In previous website tests the, mean response rate was distributed normally with mean 0.68 and standard deviation $\sigma = 0.03$. We can approximate the response variance as $s = 0.68 * (1 - 0.68)$.

Example

In previous website tests the, mean response rate was distributed normally with mean 0.68 and standard deviation $\sigma = 0.03$. We can approximate the response variance as $s = 0.68 * (1 - 0.68)$.

Plugging those values into our sample size formula we get

$$n_1^* = n_2^* = 2,284$$

Demo in spreadsheet

Comparison to hypothesis test sample size

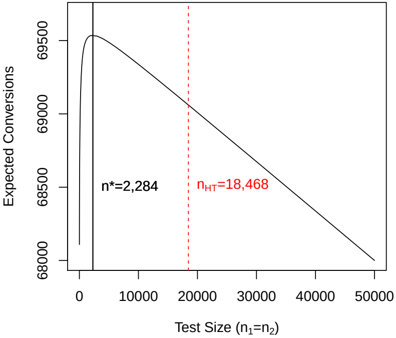
Bayesian sample size:

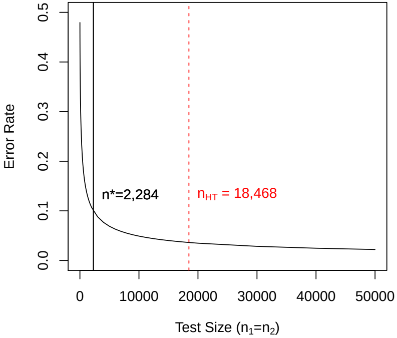
$$n_1^* = n_2^* = 2,284$$

Hypothesis Test (Classical): $\alpha = 0.05$, $\beta = 0.8$ and $\delta = 0.68 * 0.02 = 0.0136$ (2% lift) requires a sample size of:

$$n_{HT} = n_{HT} = 18,468 \tag{1}$$

Profit and Error Rate for Example





Even small tests improve profit and decrease error rate substantially.

As tests get bigger, the error rate goes down, but the total profit also goes down.

Other Bayesian sample sizes

For binary data (beta-binomial model) there is no formula for sample size. It can only be estimated using a computer program.

Profit-Maximizing Sample Size: Pros and Cons

Pros:

- Based on the total population available
- Smaller than that that needed for a classicial hypothesis test

Cons:

- Most analysts are unfamilar with this approach