An Intro to Bayesian Statistics

1 Derivation

Let’s start at the very beginning (a very good place to start)

Conjugate Probability

INDEPENDANT

P(A AND B) = P(B)P(A)

IF B DOES NOT depend on A (LIKE ROLLING A DICE – what's the probability of rolling a 5 than a 6?)

DEPENDANT

P(A AND B) = P(B|A)P(A)

IF B DEPENDS on A (Like the probability it rains today and tomorrow where A is the probability it rains today)

P(A) = Probability A is true

P(A|B) = Probability of event A given B is true

BAYES THEOREM

GIVEN:

P(B AND A) = P(A AND B)

THEN:

P(A|B)P(B) = P(B|A)P(A)

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

P(A AND B) = P(B AND A) *(is commutative)*

Example 1: THE COOKIE JAR

Jar 1 has 30 vanilla and 10 chocolate
Jar 2 has 20 vanila and 20 chocolate

Given 1 Vanilla cookie, what's the probability it came from jar 1?

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

P(1|V) = P(V|1)*P(1)/P(V)
P(1|V) = (30/40)*(½)/(50/80)
P(1|V) = 3/5

2

Imposing our Beliefs

The true brilliance of Bayesian Thinking

When the facts change, I change my mind. What do you do sir?

J M Keynes

Diachronic Interpretation

We can use Bayes' Theorem to update our hypothesis (A) in light of new data (X)

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

$P(A|X)$ = posterior probability
$P(A)$ = prior probability
$P(X|A)$ = likelihood
$P(X)$ = normalizing constant

$$P(X) = \sum_i P(X|A_i)P(A_i)$$

Normally we specify a set of i hypotheses that are:

MUTUALLY EXCLUSIVE *(only one can be true)* AND *collectively exhaustive*

Example 1: THE COOKIE JAR (REDUX)

Jar 1 has 30 vanilla and 10 chocolate
Jar 2 has 20 vanila and 20 chocolate

Given 1 Vanilla cookie, what's the probability it came from jar 1 or jar 2?

JAR	Prior P(A)	Likelihood P(X|A)	P(A)P(X|A)	Posterior P(A|X)
1	1/2	30/40	3/8	3/5
2	1/2	20/40	2/8	2/5

P(X) 5/8

Conditional Probability Table

Bayesian Thinking

Frequentists

Ascribe to the classical version of statistics, that probability is the long run frequency of events.

Bayesians

Interpret probability as a measure of belief in an event occuring.

$$\text{BAYESIAN} \xrightarrow[\lim N \to \infty]{} \text{FREQUENTIST}$$

Example 2: Tank Problems

During the war, the Allies discovered the Germans serialized their tanks by production number. Given that you have captured tank #60 – estimate the size (N) of the German tank fleet.

TANKING STATS

ASSUMPTIONS

Assume that ***at most*** the country will have 1000 tanks

Tank probability is uniform in that range –

our PRIOR then is:

$P(N) = (1/N)$

Evaluate:

$$P(N|X) = \frac{P(X|N)P(N)}{\sum_i P(X|N_i)P(N_i)}$$

N	Prior P(N)	Likelihood P(60|N)	P(N)P(60|N)	Posterior P(N|X)
5	1/5	0	1/5	0
60	1/60	1	1/60	.006
120	1/120	1	1/120	.003

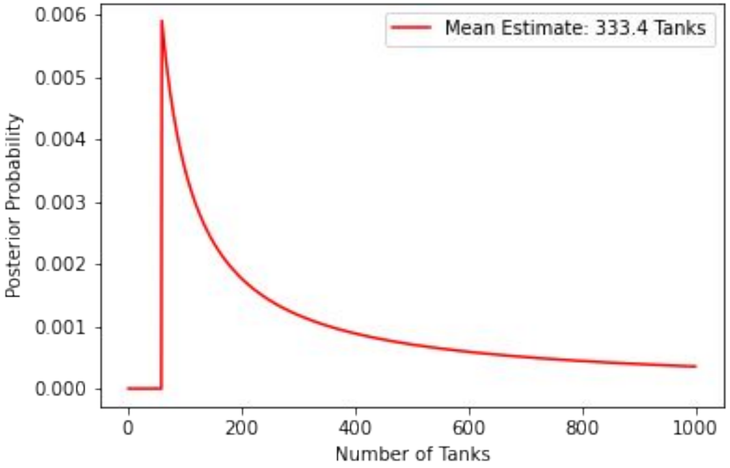
⋮

Here i is from 1 to 1000

$P(X|N) = 0$ *if* $N<X$

1 *if* $X>N$

TANKING STATS



We can estimate ***333*** tanks!

But if we observe just two more say we have now see [60,190,110]

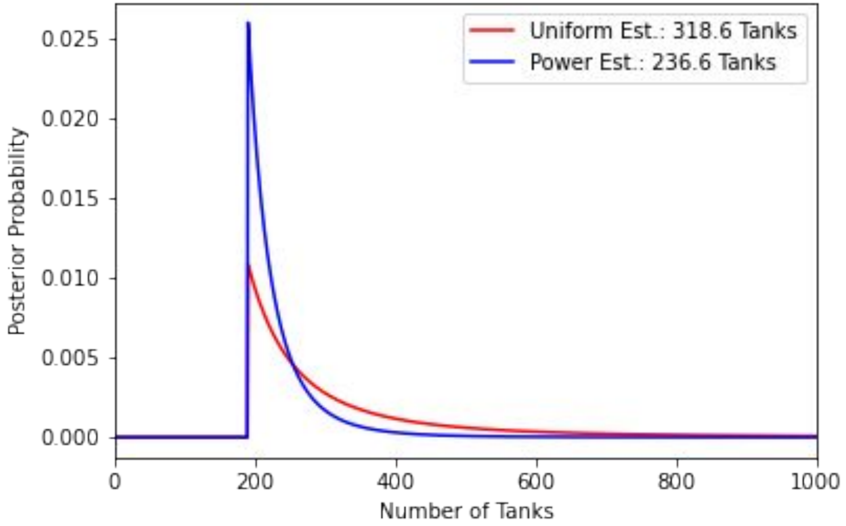
N	Obs(1)	Obs(2)	Obs(3)
500	207	296	275
1000	333	389	319
2000	552	493	346

TANKING STATS

Can we form a more accurate estimate of the prior?

Studies of manufacture rates of heavy goods typically observes a power law ~ 1/N

N	Obs(1)	Obs(2)	Obs(3)
500	143	258	233
1000	179	275	236
2000	215	282	237



Improving the Prediction

Alternative Prior

Model the Data Better

Adding More data

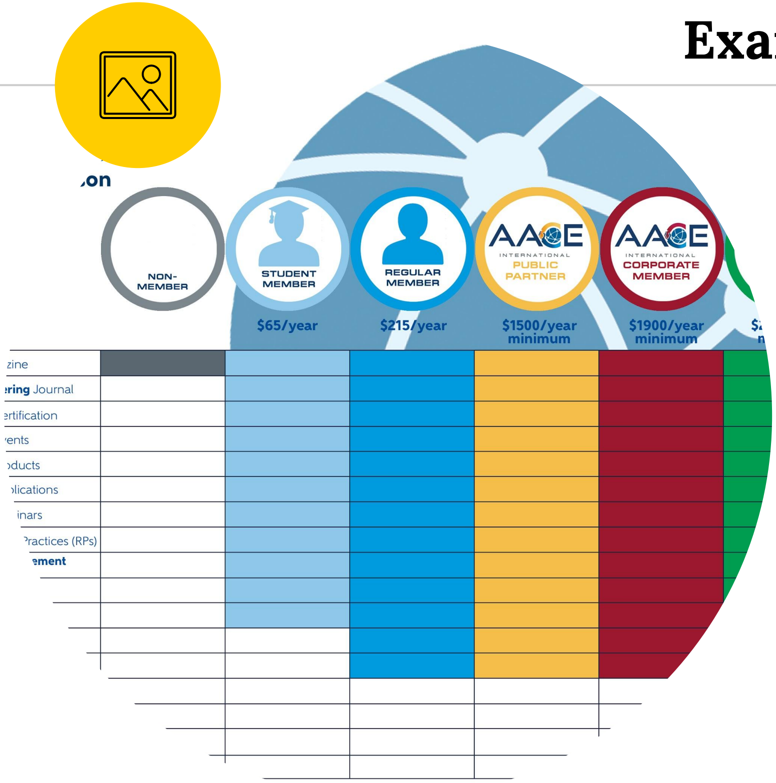
Swamp the Prior

3

A/B Testing

Experimentation The Bayesian Way

Example 3: Web A/B Testing



A company has 3 different plans and wants to optimise their revenue {$79,$49,$25}. They conduct two experiments to determine which website leads to the best performance:

Plan	Subscriptions
79	10
49	46
25	80
0	851

Plan	Subscriptions
79	45
49	73
25	165
0	1451

Prior Priorities

ASSUMPTIONS

We need a prior to model the probabilities of subscription to each plan.

Choose a *Dirlichlet distribution* (a multivariate beta distribution) which has the benefit that its probabilities sum to 1.

Conjugate Priors

Given that our Posterior and Prior distributions are Dirichlet, and the data has a binomial distribution they have a **CONJUGATE RELATIONSHIP** which means that we don't have to use a MCMC (since the posterior is known in closed form)!

This means:

If we choose Dirlichlet_Prior = D(1,1,1,1)   (this samples from a uniform distribution)

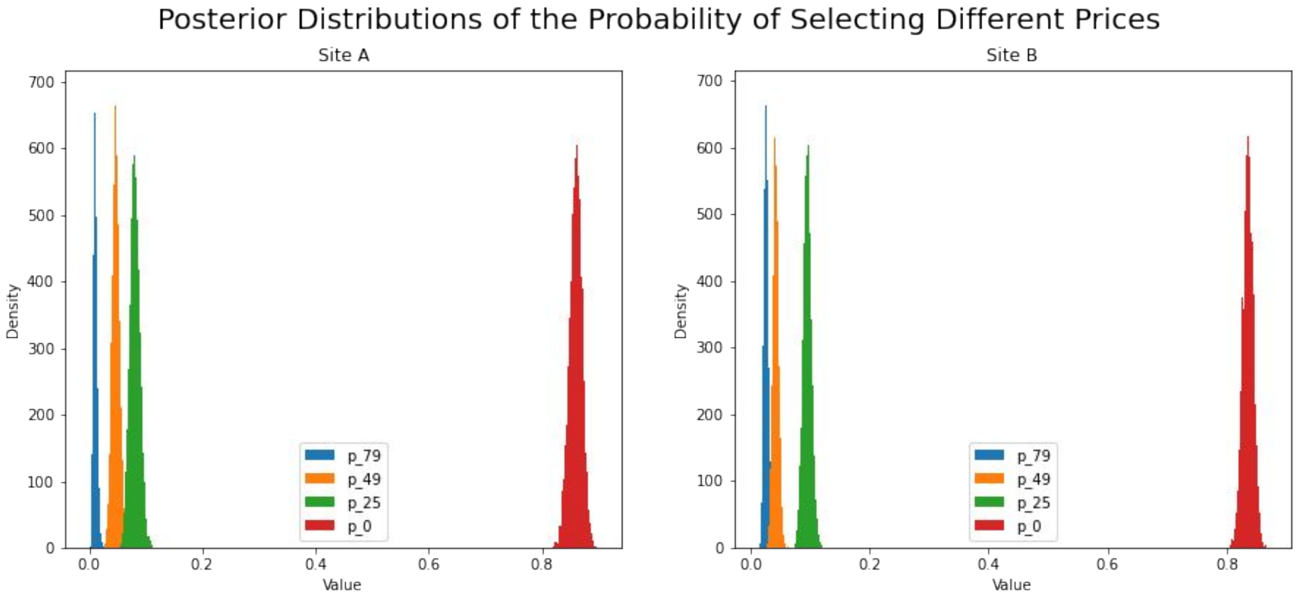
In our case Dirlichlet_Post = D(1+N79,1+N49,1+N25,1+N0) from this conjugate relationship!

For more on conjugate relationships, check out this website: https://www.johndcook.com/blog/conjugate_prior_diagram/

NB. Conjugate priors are useful only in lower dimensional problems and where a subjective prior is required.

Posting the Results

Our Bayesian thinking allows us to compute distributions as opposed to bar graphs for each test case.

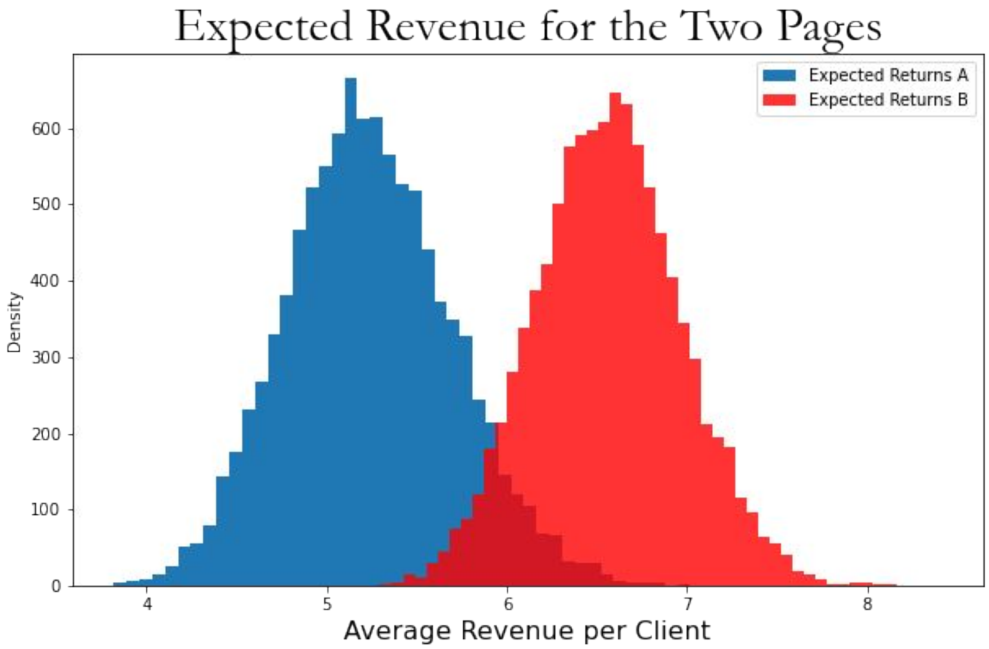


Posting the Results

We can translate the probabilities into expected revenues for each website visitor.

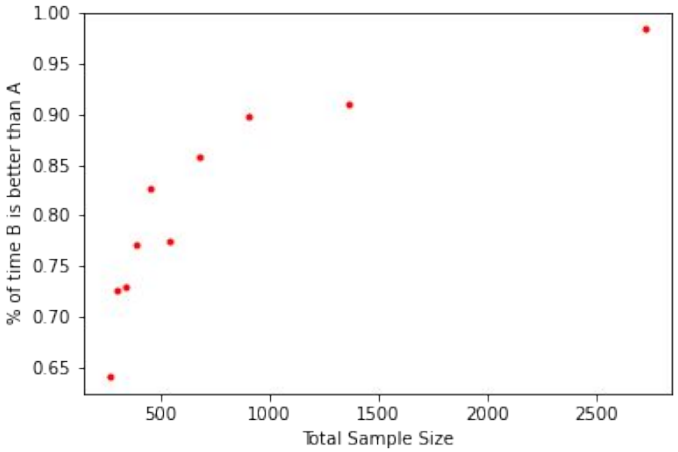
We can then ask the question: How often is website B better than website A?

By randomly drawing from the distributions, we can answer 98.3% of the time.



Sample Size

With Bayesian thinking, sample size consideration are “pre-baked” into our model.

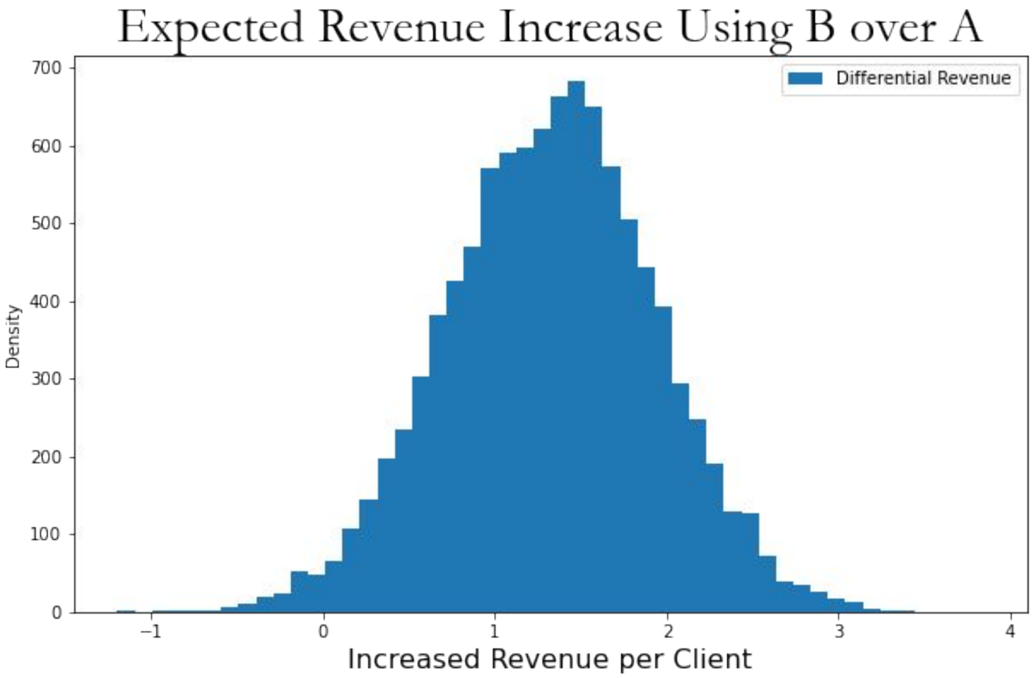


Posting the Results

We can also look at this in terms of the likelihood of the expected increase of using B over A.

We can see that by using B we're likely to make over a $1 more per visitor (maybe even $2).

Also, there a very very slim possibility that we'll lose anything and even if we did, it would be less than a dollar.



Lift

Often business types like to know the relative increase A over B.

One way is to take the mean of both posteriors and compute the lift, but this:

- Loses all the uncertainty about the true values.
- Can lead to crazy values if the results are close to zero.

$$lift = \frac{\hat{r}_b - \hat{r}_a}{\hat{r}_a}$$

Lift

What do we report?

mean

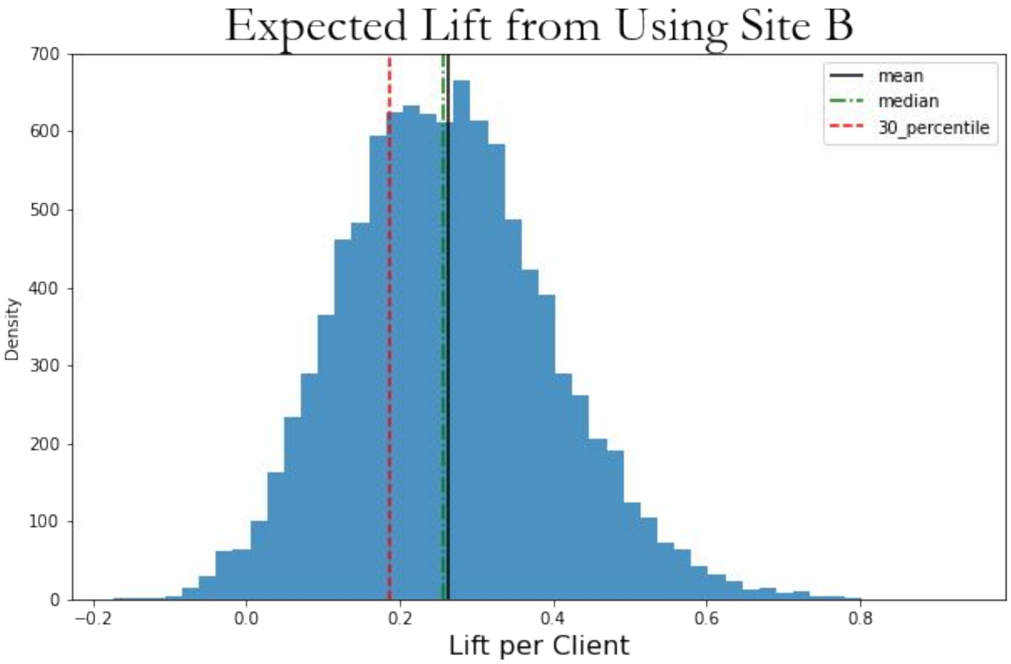
(poor if skewed)

median

30th percentile

-penalises over estimates

-converges to median with more data

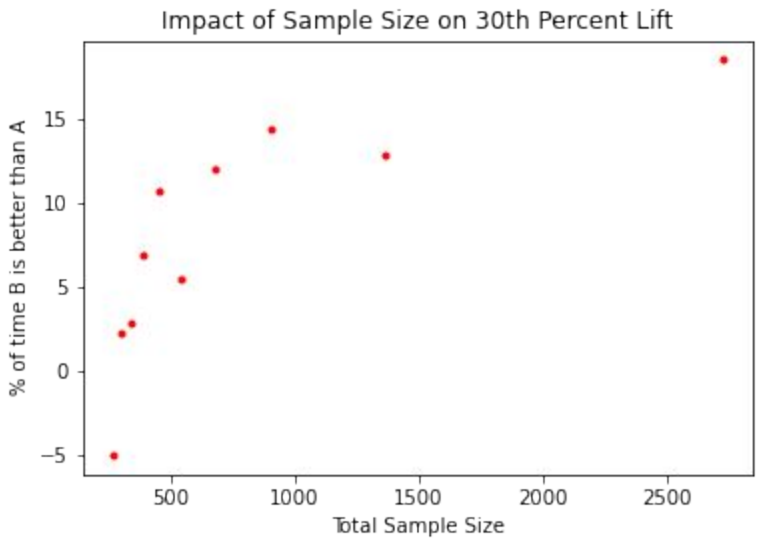


Lift

Looking at the impact of sample size on lift demonstrates:

–why reporting conservatively (30th percentile) makes sense

–Why lift is not always the best statistic (if you ignore uncertainty)



REFERENCES

The code for this lecture:
https://github.com/jmcmummey/IntrotoBayes

http://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/

https://www.johndcook.com/blog/conjugate_prior_diagram/

Think Bayes: Bayesian Statistics in Python by Allen Downey