Bayesian A/B Testing at VWO

Chris Stucchio

Visual Website Optimizer

September 2, 2015

Abstract

Before joining Visual Website Optimizer, I ran A/B testing at AOL Patch and acted as an statistical consultant. Whenever an A/B test concluded, people would approach me and ask questions about the results. One of the most common questions I was asked was "what is the probability that version B is better than version A?" Unfortunately, at the time I was using frequentist testing methods, and I was completely unable to provide an answer to that question. In fact, I was unable to answer most of the questions I was asked, and instead had to give unintuitive alternative statistics that didn't really address the question.

At Visual Website Optimizer, we have solved this problem. We have a new *Bayesian* statistical method which provides concrete, intuitive answers to all these questions. The new method mitigates some methodology errors that many people doing A/B testing make, and we've also modified the tool steer the user in a scientifically sound direction.

1 The Problem

The number one question a user typically asks me at the end of a test is the following: what is the probability that variation B is better than variation A? In mathematical terms, they want to know:

$$P(\lambda_B > \lambda_A) \tag{1.1}$$

where $\lambda_{A,B}$ is the conversion rate of variation A or B.

Unfortunately, the standard frequentist statistical techniques *cannot answer this question.* Instead, a frequentist technique will choose a null hypothesis, e.g. H_0 which represents the claim that $\lambda_B \equiv \lambda_A$. Then a hypothesis test will be run, and a test statistic t_e is computed from the experiment.

Finally, a p-value will be computed. The p-value is defined as:

$$p \equiv P(t \geq t_e | H_0) \tag{1.2}$$

If a p-value is reported, then this means that *if we ran an A/A test* with the same sample size, the probability of seeing a result at least as "extreme" as what we just saw is smaller than p.

To the best of my knowledge, no marketer or web designer has ever asked a statistician for this number.

Moreover, this number is often wrongly misinterpreted as being $P(\lambda_B > \lambda_A)$.

1.1 Representing uncertainty

A staple of frequentist statistics is the *maximal likelihood estimate.* This provides a single number which is often interpreted as being the "most likely" value of a statistic. However, presenting such a number is often misleading. The reality of statistics is that uncertainty is always present. All we, as statisticians, can do is quantify it.

At VWO we typically communicate uncertainty by providing *credible intervals* - a credible interval is a region which has a specified probability of containing the true value. This is described in detail in Definition 3.1 on 3.

1.2 Methodology

Many blog posts, articles and tutorials have been written about how to avoid errors when running A/B tests. In fact, it would be far easier for me to write a document on how to *create* false positives than it would be to write a document on how to avoid them. Most A/B testing tools, including the old version of VWO, made it very easy to make such mistakes.

For example, one can go "fishing" for a result by choosing multiple goals, and then choosing the goal for which a result is demonstrated by changing the primary goal at the end of the test. The new version of VWO will warn the user anytime a goal change is attempted, and all reported results will contain an audit log of changes of this nature.

Because of this, agencies and their clients can have more confidence that the results are being reported based on proper methodology, rather than fishing for something interesting to show the client. This is a real problem as described here: `https://www.chrisstucchio.com/blog/2015/ab_testing_segments_and_goals.html`.

2 Introduction

Since this document concerns A/B testing, let us define a few terms. Suppose we have displayed some variation of the site (for concreteness suppose it is variation A) to n_A users. We suppose that somewhere out in the real world, there is a *true conversion rate* λ_A - this means that for each user we show variation A to, there is a λ_A probability that user will convert.

3 Probability distributions

A probability distribution is a set of possible values for a parameter (e.g. λ_A) together with a function saying how likely any individual parameter is. For our purposes, the only relevant parameter is the click through rate of some variation λ_A, so we'll focus on this. To begin, we'll simply assume there is one variation $\lambda_A = \lambda$.

The probability distribution $P(\lambda)$ represents our *opinion* as to which values the parameter are more likely. Any probability distribution must satisfy the following properties:

$$\forall \lambda, P(\lambda) \geq 0 \tag{3.1a}$$

$$\int_0^1 P(\lambda) d\lambda = 1 \tag{3.1b}$$

The symbol $\forall$ means "for all" - i.e., (3.1a) means that for any specific value of λ, the value of $P(\lambda)$ must be positive.

Definition 3.1 *An* $x-\%$-credible interval *is a region* $[a, b]$ *with the property that:*

$$\int_a^b P(\lambda) d\lambda \geq x \tag{3.2}$$

In words, if the 95% credible interval for a probability distribution is $[0.35, 0.40]$*, it means we are 95% confident that the true value of the parameter is contained in the interval* $[0.35, 0.40]$.

4 Bayesian Statistics

Bayesian statistics is the mathematical study of changing your opinion based on evidence. It provides a rule, namely Bayes Rule, which gives an optimal way of changing your beliefs.

I'll now walk through the process of actually changing your opinion, but before I do that I need a definition.

Definition 4.1 *The* conditional probability $P(A|B)$ *represents the probability of A being true* assuming *that B is true.*

As an example, $P(\text{conversion}|\lambda) = \lambda$ - this means that if we know the value of λ, then the probability of a conversion occuring is known to be λ.

4.1 The Prior - an uneducated opinion

To begin doing Bayesian Statistics, you must come up with a prior. A prior is your uneducated opinion - it's what you believe *before you have evidence*. For example, if I believe all possible values of λ are equally likely, I might choose the function $P(\lambda) = 1.0$. If I believe values closer to zero than to one are more likely, I could choose $P(\lambda) = 2(1 - x)$.

4.1.1 Pulling a prior from your posterior

It is important to emphasize that *there is usually no scientific basis for choosing a prior*. The prior is *completely subjective*, and different people may choose different ones.

This is totally fine. An important fact about Bayesian statistics is that although people may disagree at the beginning, once evidence is gathered they will eventually agree.

This is where the whimsical phrase "pull a prior from your posterior" comes from. It means that it's completely acceptable to make up a silly prior with no data - once you find data, you'll be able to change your opinions.

4.2 Evidence

The next step in doing Baysian Statistics is to gather evidence. This means running an experiment which has an outcome that changes depending on the true value of λ.

In our case, evidence involves displaying a variation to a set of users. So for concreteness, suppose we display variation A to n_A users. We then count the number of successes and observe that c_A users convert.

If λ_A is large, then we expect that the c_A will be nearly as large as n_A - if $\lambda_A = 1.0$ we will have that $n_A = c_A$. Conversely, if $\lambda_A = 0.0$, then $n_A = 0$.

This means that our experiment generates evidence about the true value of λ_A.

4.3 The Posterior - how to change your opinion

This part is where Bayesian statistics gets it's name. To change your opinion after observing evidence, Bayes rule is used. In it's full glory:

$$P(\lambda|\text{evidence}) = \frac{P(\text{evidence}|\lambda)P(\lambda)}{P(\text{evidence})} \tag{4.1}$$

The left side of (4.1) represents your *posterior* - your opinion after observing the evidence.

Lets go through the right side of the equation. The expression $P(\lambda)$ represents your prior - your opinion before observing any evidence. The expression $P(\text{evidence}|\lambda)$ represents the probability of observing the evidence *assuming* you know the true value of λ.

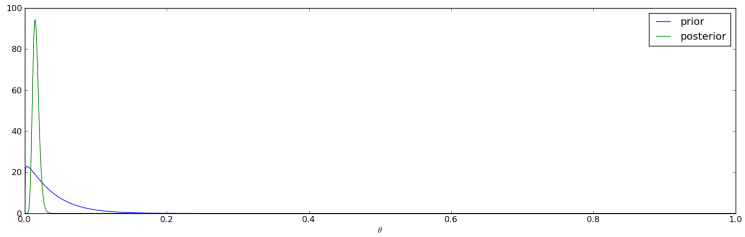

Figure 1: Another example of converting a prior to a posterior, this time in the form of pictures. In this case, the prior was $P(\lambda) = Const \cdot \lambda^{1.1}(1-\lambda)^{30})$. The evidence was displaying the page to 794 users, of whom 12 converted.

Finally, the term $P(\text{evidence})$ is the probability of having observed the evidence you actually did observe. The important fact to note about $P(\text{evidence})$ is that it *does not vary with* λ.

4.3.1 An example

Suppose that to begin with, $P(\lambda) = 1.0$. This represents the case that before any evidence, we believe the conversion rate could be anything. Now suppose we show a single user the test, and this user does NOT convert. Lets compute the posterior.

First, observe that $P(\text{evidence}|\lambda) = 1 - \lambda$. This follows because if the probability of the user converting was λ, then the probability of the user NOT converting must be $1 - \lambda$. We can now compute $P(\text{evidence})$:

$$P(\text{evidence}) = \int_0^1 P(\text{evidence}|\lambda)P(\lambda)d\lambda = \int_0^1 (1-\lambda)\cdot 1 d\lambda = \left[\lambda - \frac{\lambda^2}{2}\right]_0^1 = \frac{1}{2}$$

We can now compute $P(\lambda|\text{evidence})$:

$$P(\lambda|\text{evidence}) = \frac{(1-\lambda)\cdot 1}{1/2} = 2(1-\lambda)$$

5 The Beta Distribution - simplifying the calculation

The Beta Distribution is defined as follows:

$$f(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, 0 \leq x \leq 1 \tag{5.1}$$

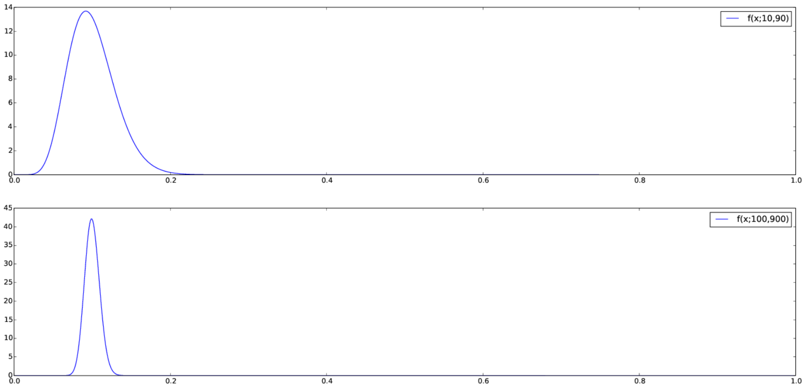

Figure 2: Example of the concentration of the beta distribution for increasing (a, b). The first graph represents $f(x; 10, 90)$, the second $f(x; 100, 900)$.

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{5.2}$$

The beta distribution has a wonderful property which makes it very useful in Bayesian inference.

Theorem 5.1 *Suppose the prior $P(\lambda) = f(\lambda; a, b)$. Suppose that the variant was displayed to n visitors and c converted. Then the posterior is given by:*

$$P(\lambda | n, c) = f(x; a + c, b + (n - c)) \tag{5.3}$$

This theorem is proven in many places, e.g. on wikipedia: `http://en.wikipedia.org/wiki/Beta_distribution#Bayesian_inference`

This theorem allows us to compute a posterior for conversion rates in a simple way, provided the prior is a beta distribution.

Many possible priors can be represented via a beta distribution. Figures 2, 3 and 4 illustrate the diversity of shapes which can be represented this way. A uniform prior is given by the choice $(a, b) = (1, 1)$.

Using the Beta distribution, we need to track only 2 numbers to compute a posterior - n and c. Then, whenever the posterior needs to be manipulated, we can compute the posterior directly:

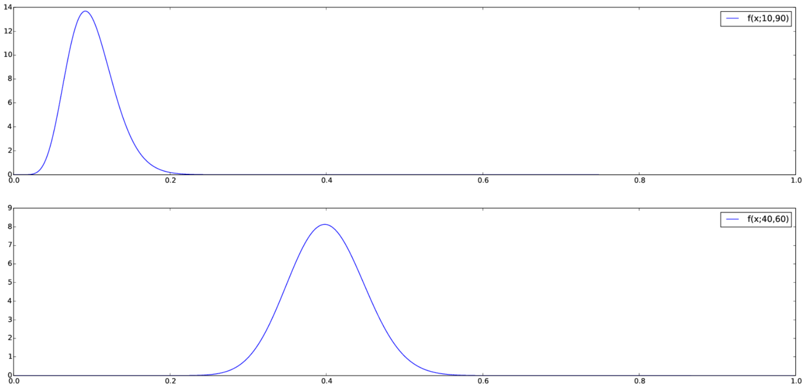

Figure 3: Example of the shift in mean for the beta distribution. The first graph represents $f(x; 10, 90)$, the second $f(x; 40, 60)$.

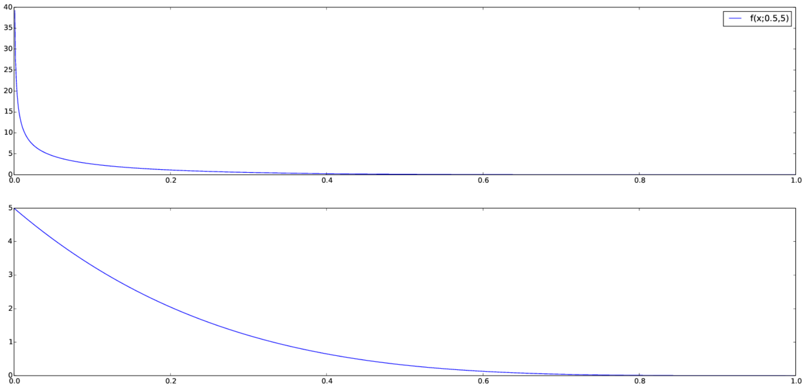

Figure 4: Example of the singular nature of the beta distribution for values of $a \leq 1.0$.The first graph represents $f(x; 0.5, 5)$, the second $f(x; 1.0, 5)$.

5.1 Intuitive interpretation of the parameters

In general, if you want a prior with mean μ and variance σ^2, they can be found via the formulae:

$$\mu = \frac{a}{a+b} \tag{5.4a}$$

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{5.4b}$$

We can solve (5.4) for (a, b) as follows:

$$b = a(1-\mu)/\mu \tag{5.5}$$

Substituting (5.5) into (5.4b) yields:

$$\begin{aligned} \sigma^2 = \frac{a^2(1-\mu)}{\mu(a+b)^2(a+b+1)} &= \frac{a^2(1-\mu)}{\mu(a[1+(1-\mu)/\mu])^2(a+a(1-\mu)/\mu+1)} \\ &= \frac{(1-\mu)}{\mu(1+(1-\mu)/\mu)^2(a/\mu+1)} \end{aligned} \tag{5.6}$$

Rearranging (5.6) yields:

$$a = \frac{1-\mu-\sigma^2\mu^3}{\sigma^2\mu^2} \tag{5.7a}$$

$$b = \frac{(1-\mu)[1-\mu-\sigma^2\mu^3]}{\sigma^2\mu^3} \tag{5.7b}$$

6 The Joint Posterior for 2 variants

At this point we are now ready to discuss actual A/B tests. Suppose we have two page variants - say A and B. Suppose we have run an experiment, displaying variant A and B to n_A and n_B users. At this point we can compute a posterior for each variation - $P_A(\lambda_A)$ and $P_B(\lambda_B)$.

(I'll discuss the case of more page variants later.)

The *joint posterior* of A and B together is:

$$P(\lambda_A, \lambda_B) = P_A(\lambda_A)P_B(\lambda_B) \tag{6.1}$$

The joint posterior can be used to calculate various quantities of interest. The major quantities we are interested in are loss functions - functions which measure what sort of a mistake we will make assuming we choose a variant and stop the test.

In pictures, we can plot the joint posterior as a function of two variables - c.f. Figures 5, 6 and 7.

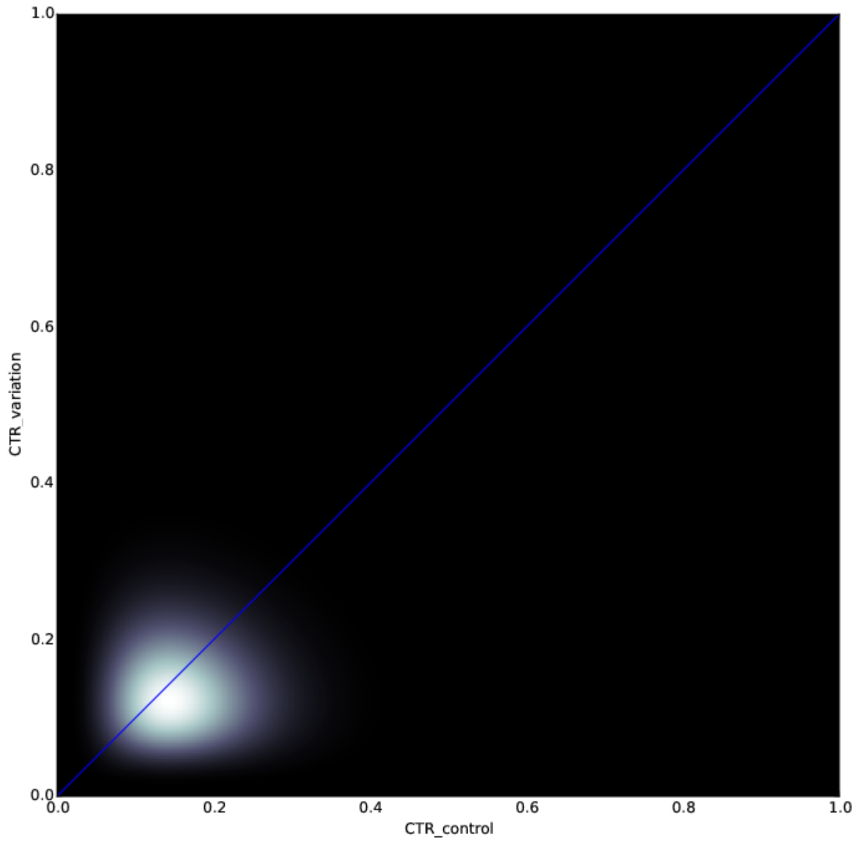

Figure 5: Joint posterior near the start of the test. Points in the grey and white region represent highly likely values of (λ_A, λ_B), while dark regions represent areas of low probability for same. The blue line plots $\lambda_A = \lambda_B$.

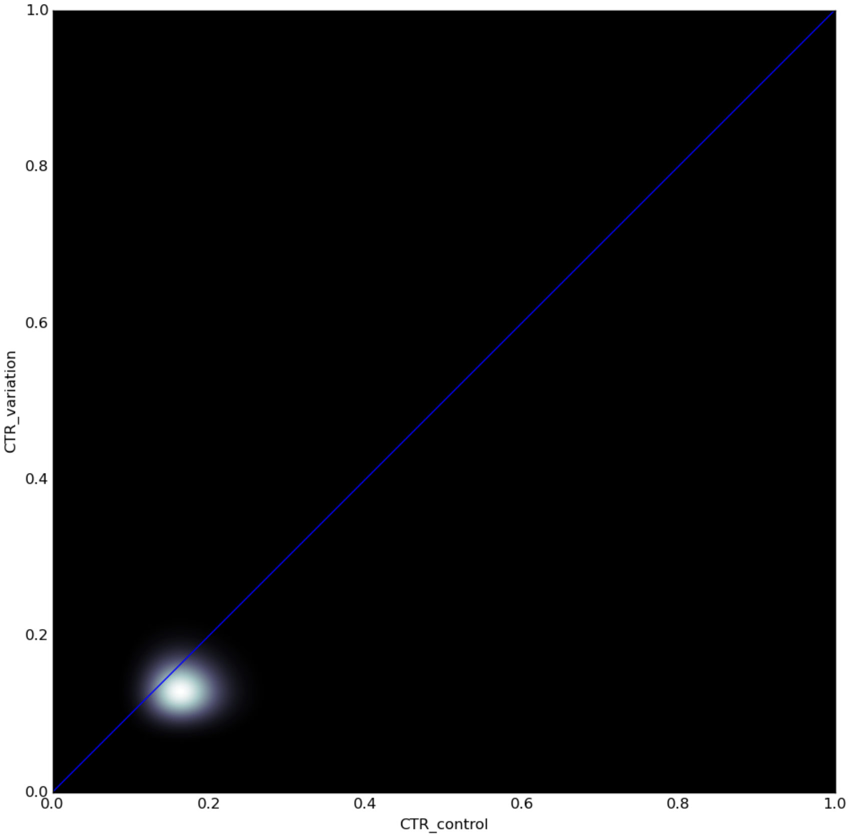

Figure 6: Joint posterior near the middle of the test. The posterior is narrowing.

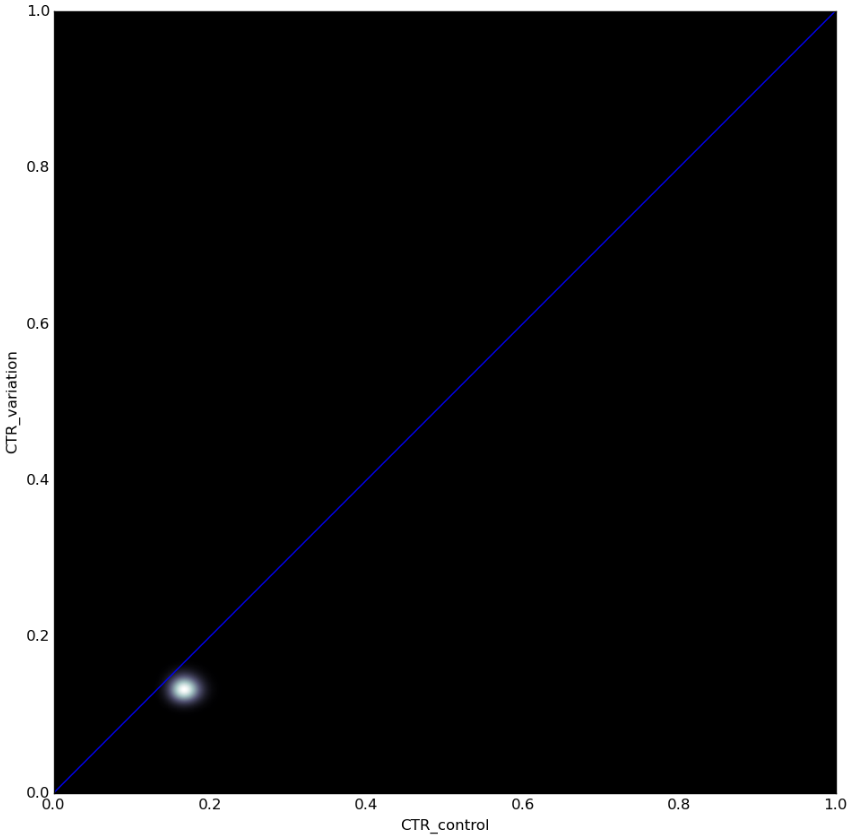

Figure 7: Joint posterior at the end of the test. The posterior has narrowed to the point where it lives almost entirely on one side of the blue line, and we can conclude that control is superior to the variation.

6.1 Chance to beat control

Suppose we have some evidence, and then we choose to display variant A. What is the probability we made a mistake?

Definition 6.1 *The* error probability *is the probability we made a mistake:*

$$E[\mathcal{I}](A) = \int_0^1 \int_0^{\lambda_A} P(\lambda_A, \lambda_B) d\lambda_B d\lambda_A \tag{6.2a}$$

I.e., this is the probability that $\lambda_B > \lambda_A$. Similarly:

$$E[\mathcal{I}](B) = \int_0^1 \int_{\lambda_A}^1 P(\lambda_A, \lambda_B) d\lambda_B d\lambda_A \tag{6.2b}$$

This definition is intuitively the obvious choice to determine whether it is worthwhile to continue the test.

However, this metric is flawed in an important way for making decisions - it treats all errors as equally bad.

6.2 The Loss Function

The loss function corrects the error function in an important way. It treats small errors as less bad than big ones.

Definition 6.2 *The* loss function *is the amount of uplift that one can expect to be lost by choosing a given variant, given particular values of λ_A and λ_B:*

$$\mathcal{L}(\lambda_A, \lambda_B, A) = \max\left(\lambda_B - \lambda_A, 0\right) \tag{6.3a}$$

$$\mathcal{L}(\lambda_A, \lambda_B, B) = \max\left(\lambda_A - \lambda_B, 0\right) \tag{6.3b}$$

As an example, suppose we choose to display variant A. Suppose the conversion rate for A is known to be $\lambda_A = 0.1$ and the conversion rate for B is $\lambda_B = 0.15$. Then the loss $\max(0.15 - 0.1, 0) = 0.05$. In contrast,if we chose B, the loss would be $\max(0.1 - 0.15, 0) = 0.0$.

Definition 6.3 *The* expected loss *given a joint posterior is the expected value of the loss function:*

$$E[\mathcal{L}](?) = \int_0^1 \int_0^1 \mathcal{L}(\lambda_A, \lambda_B, ?) P(\lambda_A, \lambda_B) d\lambda_B d\lambda_A \tag{6.4}$$

Here, the ? symbol can take the value of either A or B, depending on which version we choose to display.

6.3 Computing the loss and error functions

It is fairly straightforward to compute the error or loss function computationally. Computationally, suppose we have two posteriors:

Algorithm 6.4 *Computing the Joint Posterior.*

```
posteriorA = ...do work here...
posteriorB = ...do work here...

joint_posterior = zeros( shape=(100, 100) ) #The joint posterior is a 2d array

for i in range(100):
    for j in range(100):
        joint_posterior[i,j] = posteriorA[i] * posteriorB[j]
```

The error function can be computed as follows:

Algorithm 6.5 *Computing the Error Function.*

```
errorFunctionA = 0.0
for i in range(100):
    for j in range(i, 100):
        errorFunctionA += joint_posterior[i,j]
```

This code is equivalent to (6.2a)*. To compute* (6.2b)*, one would do this:*

```
errorFunctionB = 0.0
for i in range(100):
    for j in range(0, i):
        errorFunctionB += joint_posterior[i,j]
```

6.3.1 Computing the loss function

To compute the loss function, one would do:

Algorithm 6.6 *Computing the loss function.*

```
def loss(i, j, var):
    if var == 'A':
        return max(j*0.01 - i*0.01, 0.0)
    if var == 'B':
        return max(i*0.01 - j*0.01, 0.0)

lossFunction = 0.0
for i in range(100):
    for j in range(100):
        lossFunction += joint_posterior[i,j] * loss(i,j,'A')
```

Remark 6.7 It is also possible, in the case of measuring conversion rates and two variables, to compute things with a closed form solution. This is described here: `https://www.chrisstucchio.com/blog/2014/bayesian_ab_decision_rule.html` and is based on a formula derived by Evan Miller here: `http://www.evanmiller.org/bayesian-ab-testing.html`. The code I wrote does not use this method, since that would only work in the two-variable case.

7 Running a Bayesian A/B test

Before moving on to 3+ variants, lets now explain how an A/B test works. The basic idea is to choose a desired error tolerance, which we denote by ε, a threshold of loss which is considered acceptable. Then the A/B test is run until the expected loss is below this specified tolerance.

The interpretation of ε is as follows - ε is a percentage. It represents how much lift one would expect to lose by making a particular choice *given that the choice is wrong*. It should be set to a number so low that one does not care if an error is made by this amount.

Example 7.1 *Suppose we are testing two button colors, and we are interested in measuring lift of 10%. Conversely, if the lift we get from this test changes negatively by 0.2% or less, this change is so small that we don't care. In that case, we can choose* $\varepsilon = 0.002$*.*

Algorithm 7.2 *Running a Bayesian A/B test.*

1. *Run an experiment, displaying variants A and B to a random selection of users.*

2. *Periodically, compute the aggregate statistics* n_A, c_A, n_B *and* c_B*. (Recall that* n_A *is the number of times variant A was displayed, and* c_A *is the number of times variant A converted.)*

3. *Compute* $E[\mathcal{L}](A)$ *and* $E[\mathcal{L}](B)$*. Define the set* $\mathfrak{A} = \{x : E[\mathcal{L}](x) < \varepsilon\}$*. The computation of* $E[\mathcal{L}](A)$ *can be performed as described in section 6.3.1.*

4. *If* $\mathfrak{A}$ *is empty, go to step 1. Otherwise stop the test, choose an element of* $\mathfrak{A}$ *and declare it to be the winner.*

8 Running a test with more than two variants

For more variants, the loss function is straightforward to define:

$$\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, A) = \max\left(\lambda_B - \lambda_A, \lambda_C - \lambda_A, 0\right) \tag{8.1a}$$

To compute $\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, B)$, one does:

$$\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, B) = \max\left(\lambda_A - \lambda_B, \lambda_C - \lambda_B, 0\right) \tag{8.1b}$$

and similarly for C, etc.

If you had 4 variants, the loss function would be:

$$\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, \lambda_D, A) = \max\left(\lambda_B - \lambda_A, \lambda_C - \lambda_A, \lambda_D - \lambda_A, 0\right) \tag{8.2}$$

One would define the loss function similarly for 5 or more variants.

In principle, we could run the multivariant version of Algorithm 7.2 to determine whether to finish or continue the test:

Algorithm 8.1 *Running a Bayesian A/B/C/etc test.*

1. *Run an experiment, displaying variants $A, B, C, \ldots$ to a random selection of users.*
2. *Periodically, compute the aggregate statistics $n_A, c_A, n_B, c_B, \ldots$. (Recall that n_A is the number of times variant A was displayed, and c_A is the number of times variant A converted.)*
3. *Compute $E[\mathcal{L}](A), E[\mathcal{L}](B), \ldots$. Define the set $\mathfrak{A} = \{x : E[\mathcal{L}](x) < \varepsilon\}$.*
4. *If $\mathfrak{A}$ is empty, go to step 1. Otherwise stop the test, choose an element of $\mathfrak{A}$ and declare it to be the winner.*

Unfortunately, computing the expected loss is not so straightforward using the methods we've described previously. If we have 3 variants, the numerical integration technique described above would involve a triple for loop:

```
def loss(i, j, var):
    if var == 'A':
        return max(j*0.01 - i*0.01, k*0.01-i*0.01, 0.0)
    if var == 'B':
        return max(i*0.01 - j*0.01, k*0.01-j*0.01, 0.0)
    if var == 'C':
        return max(i*0.01 - k*0.01, j*0.01-k*0.01, 0.0)

lossFunction = 0.0
for i in range(100):
    for j in range(100):
        for k in range(100):
            lossFunction += joint_posterior[i,j,k] * loss(i,j,k,'A')
```

This requires summing $100^3 = 1,000,000$ numbers, which is not terribly unreasonable. If we have 4 variants, it will require summing $100^4 = 100,000,000$ numbers. This will rapidly become intractable.

Enter Monte Carlo methods, which allow us to approximate the integral via sampling.

8.1 Chance to beat Control, Chance to beat All

Given the posterior, we can compute these two useful quantities.

The first is the *chance to beat control.* This is defined as the probability that a given variation (say B) has a higher conversion rate than the control.

$$\mathcal{CTBC}(B) = \int \ldots \int_{\lambda_B > \lambda_A} P(\lambda_A, \lambda_B, \ldots, \lambda_k) d\lambda_A d\lambda_B \ldots d\lambda_k \tag{8.3}$$

We can also define the *chance to beat all* as the probability that a given variation (say B) has a higher conversion rate than *every other variation.*

$$\mathcal{CTBA}(B) = \int \ldots \int_{(\lambda_B > \lambda_A) \wedge (\lambda_B > \lambda_C) \wedge \ldots} P(\lambda_A, \lambda_B, \ldots, \lambda_k) d\lambda_A d\lambda_B \ldots d\lambda_k \tag{8.4}$$

8.2 Monte Carlo - how to compute the integrals

Monte Carlo sampling relies on the law of large numbers, which states the following. Suppose that $({\lambda_A}^i, {\lambda_B}^i, {\lambda_C}^i, \ldots)$ is, for each i, a randomly selected sample from the joint probability distribution. Suppose we have drawn $N \gg 0$ such samples. Then:

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}({\lambda_A}^i, {\lambda_B}^i, {\lambda_C}^i, A) \approx E[\mathcal{L}](A) \tag{8.5}$$

(and similarly for choice $B, C, \ldots$.)

Remark 8.2 Note that if necessary, this calculation can be performed in parallel - for example, if $N = 2 \cdot 10^6$, one CPU can compute the sum for $n = 1 \ldots 10^6$ while a second CPU can compute the sum for $n = 10^6 + 1 \ldots 2 \cdot 10^6$.

An important question to ask - how accurate is (8.5)?

Theorem 8.3 *Take δ as given. Suppose we wish to ensure that:*

$$P\left(\left| \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A) - E[\mathcal{L}](A) \right| < \epsilon \right) < \delta \tag{8.6}$$

Then we must make N at least as large as $-\ln(\delta)/(2\epsilon^2)$.

Remark 8.4 Note the logarithmic growth of N with δ. Since it is unlikely that VWO will ever run 10^{10} tests, we can choose $\delta = 10^{-10}$ resulting in $-\ln(10^{-10}) = 23.03$.

Proof. Note that $\mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A) \leq 1$. We can then apply Hoeffding's inequality:

$$P\left(\left| \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A) - E[\mathcal{L}](A) \right| < \epsilon \right) < e^{-2N\epsilon^2} \tag{8.7}$$

Setting $\ln(\delta) = -2N\epsilon^2$ or $N = -\ln(\delta)/(2\epsilon^2)$ yields the result we seek. □

Theorem 8.5 *Suppose N is large. Then we have the approximate value:*

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A) - E[\mathcal{L}](A) \to N(0, S^2/N) \tag{8.8}$$

Here S^2 satisfies the bound:

$$S^2 \leq VAR[\lambda_B - \lambda_A] + VAR[\lambda_C - \lambda_A] + \ldots \\ = VAR[\lambda_B] + VAR[\lambda_C] + \ldots + M\, VAR[\lambda_A]$$

where M is the number of variations.

The convergence is convergence in distribution - see ***http://en.wikipedia.org/wiki/Convergence_in_distribution#Convergence_in_distribution.***

This theorem is proved in Section A of the Appendix.

Thus, if we wish to make the error from the Monte Carlo approximation smaller than ϵ with probability τ, we must choose:

$$N = \left[(S/\epsilon) \mathrm{cdf}^{-1}(\tau/2) \right]^2 \tag{8.9}$$

where $\mathrm{cdf}(z)$ is the cumulative distribution function of the normal distribution, and $\mathrm{cdf}^{-1}(z)$ is the inverse of that function.

Remark 8.6 Suppose λ_A is given by a Beta distribution $B(a, b)$. Then:

$$\mathrm{VAR}[\lambda_A] = \frac{ab}{(a+b)^2(a+b+1)} \tag{8.10}$$

Thus, if you use Section 5 to compute posteriors, the bound on S^2 can be computed exactly.

So now consider the case where each variation had a fixed prior, (α, β) and an experiment has run. Then:

$$S \leq \sum_{x=B,C,\ldots} \frac{(\alpha + c_x)(\beta + n_x - c_x)}{(\alpha + \beta + n_x)^2(\alpha + \beta + n_x + 1)} \\ + M\left(\frac{(\alpha + c_A)(\beta + n_A - c_A)}{(\alpha + \beta + n_A)^2(\alpha + \beta + n_A + 1)} \right) \tag{8.11}$$

An important observation here is the asymptotic complexity - suppose each variation has approximately the same number of trials, i.e. $n_A \approx n_B \approx n$. Then:

$$S \leq O\left(\frac{M}{n^2}\right)$$

Substituting this into (8.9) yields:

$$N = O\left(\left[\frac{M}{n^2 \epsilon} \mathrm{cdf}^{-1}(\tau/2)\right]^2\right)$$

Thus, bringing the error below ϵ grows quadratically with ϵ^{-1} and shrinks quartically as the number of trials increases.

In practice, typical values of N will be on the order of 1-50 million, corresponding to $\tau = 10^{-5}$ and $\epsilon = 0.0001$.

9 Approximate worst case for the Bernoulli test

The following calculations are *approximate only.*

9.1 Two variations

We might ask the question, what is the worst case scenario for the Bernoulli test to finish? The test will finish when $E[\mathcal{L}](A) \leq \varepsilon$, or:

$$E[\mathcal{L}](A) = \int \int \max(\lambda_A - \lambda_B, 0) P(\lambda_A, \lambda_B) d\lambda_A d\lambda_B \leq \varepsilon$$

The worst case occurs when both variants are identically equal.

To compute a bound, first note that:

$$\int \int \max(\lambda_A - \lambda_B, 0) P(\lambda_A, \lambda_B) d\lambda_A d\lambda_B \leq \int \int |\lambda_A - \lambda_B| \, P(\lambda_A, \lambda_B) d\lambda_A d\lambda_B \tag{9.1}$$

Second, note that a Beta distribution is (in the limit of large $a = N\zeta, b = N\zeta$) *approximated* by a normal distribution of variance $\sigma^2 = \zeta(1-\zeta)/N$. Thus:

$$(9.1) \approx \int \int |\lambda_A - \lambda_B| \, C \exp\left(-\frac{(\lambda_A - \zeta)^2 + (\lambda_B - \zeta)^2}{2\sigma^2}\right) d\lambda_A d\lambda_B \tag{9.2}$$

To evaluate (9.2), we change variables to $u = (\lambda_A - \lambda_B)/\sqrt{2}$, $v = (\lambda_A + \lambda_B)/\sqrt{2}$.

This yields:

$$
\begin{aligned}
(9.2) \approx \int\int 2^{-1/2}\,|u|\,C\exp\left(-\frac{u^2+(v-\zeta)^2}{2\sigma^2}\right)dudv \\
= 2^{1/2}\int\int |u|\,C\exp\left(-\frac{u^2}{2\sigma^2}\right)\exp\left(\frac{(v-\zeta)^2}{2\sigma^2}\right)dudv \\
= 2^{1/2}\left[\int \bar{C}\exp\left(\frac{(v-\zeta)^2}{2\sigma^2}\right)dv\right]\left[\int |u|\,\tilde{C}\exp\left(-\frac{u^2}{2\sigma^2}\right)du\right] \\
= 2^{1/2}\left[\int |u|\,\tilde{C}\exp\left(-\frac{u^2}{2\sigma^2}\right)du\right] = 2^{1/2}\sigma\sqrt{2/\pi} = 2\frac{\sigma}{\sqrt{\pi}} \\
= 2\sqrt{\frac{\zeta(1-\zeta)}{\pi N}} \quad (9.3)
\end{aligned}
$$

The value for $E[|u|] = \sigma\sqrt{2/\pi}$ is given on wikipedia: `http://en.wikipedia.org/wiki/Normal_distribution#Moments`.

Thus, we find that in the worst case, we must wait until:

$$2\sqrt{\frac{\zeta(1-\zeta)}{\pi N}} \leq \varepsilon \tag{9.4}$$

or equivalently

$$N \geq 4\frac{\zeta(1-\zeta)}{\pi\varepsilon^2} \tag{9.5a}$$

Note that (9.5a) is dependent on the fact that the beta distribution approximates the normal distribution, and this approximation is only valid when:

$$0 \leq p \pm 3\sqrt{\zeta(1-\zeta)/N} \leq 1 \tag{9.5b}$$

Thus, the constraints (9.5a) and (9.5b) together are both necessary.

9.2 Many Variations

If we want to extend this analysis to K variations, we do the following. First note that:

$$\max\{\lambda_B - \lambda_A, \lambda_C - \lambda_A, \ldots\} \leq \sqrt{\sum_{?=B,C,\ldots} |\lambda_? - \lambda_A|^2} \tag{9.6}$$

Plugging this into the equation for expected loss, we find:

$$
\begin{aligned}
\int \ldots \int \max(\lambda_B - \lambda_A, \lambda_C - \lambda_A, \ldots, 0) P(\lambda_A, \lambda_B, \ldots) d\lambda_A d\lambda_B \ldots d\lambda \\
\leq \int \ldots \int \sqrt{\sum_{?=B,C,\ldots} |\lambda_? - \lambda_A|^2} P(\lambda_A, \lambda_B, \ldots) d\lambda_A d\lambda_B \ldots d\lambda \quad (9.7)
\end{aligned}
$$

If we change variables to $(\lambda_A, \vec{u})$ with $\vec{u} = [\lambda_B - \lambda_A, \lambda_C - \lambda_A, \ldots]^T$, we obtain:

$$(9.7) = \int\int |\vec{u}|\, P(\lambda_A, \vec{u}) d\vec{u} d\lambda_A \tag{9.8}$$

We again use the normal approximation to the beta distribution, yielding:

$$\begin{aligned}(9.8) &\approx \int\int |\vec{u}|\, \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} \frac{1}{(\sigma\sqrt{2\pi})^{K-1}} e^{-|\vec{u}|^2/2\sigma^2} d\vec{u} d\lambda_A \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} \int r \frac{1}{(\sigma\sqrt{2\pi})^{K-1}} e^{-r^2/2\sigma^2} r^{K-2} V_{K-1} dr d\lambda_A \qquad (9.9)\end{aligned}$$

where in the second line we shifted to polar coordinates with $r = |\vec{u}|$. Here $V_{K-1} = (K-1)\pi^{(K-1)/2}/\Gamma((K-1)/2+1)$ is the area of the sphere in $K-1$ dimensions. Now set $s = r/\sigma$, then $dr = \sigma ds$ and:

$$\begin{aligned}(9.9) &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} \int s\sigma \frac{1}{(2\pi)^{(K-1)/2}\sigma} e^{-s^2/2} s^{K-2} V_{K-1} \sigma ds d\lambda_A \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} (2\pi)^{-(K-1)/2} V_{K-1}\sigma \int s^{K-1} e^{-s^2/2} ds d\lambda_A \qquad (9.10)\end{aligned}$$

The inner integral can be computed via Proposition C.3:

$$\begin{aligned}(9.10) &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} \sigma (2\pi)^{-(K-1)/2} V_{K-1} 2^{(K-1)/2}\Gamma(K/2) d\lambda_A \\ &= \sigma(2\pi)^{-(K-1)/2} V_{K-1} 2^{(K-1)/2}\Gamma(K/2) \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(\lambda_A-\zeta)^2/2\sigma^2} d\lambda_A \\ &= \sigma \frac{(K-1)\Gamma(K/2)}{\Gamma((K+1)/2)} = \sqrt{\frac{\zeta(1-\zeta)}{N}} \frac{(K-1)\Gamma(K/2)}{\Gamma((K+1)/2)} \qquad (9.11)\end{aligned}$$

We want to make (9.11) $\leq \varepsilon$. To do that, we require

$$\sqrt{\zeta(1-\zeta)} \frac{(K-1)\Gamma(K/2)}{\Gamma((K+1)/2)} \frac{1}{\varepsilon} \leq \sqrt{N}$$

or

$$N \geq \zeta(1-\zeta) \frac{1}{\varepsilon^2} \left(\frac{(K-1)\Gamma(K/2)}{\Gamma((K+1)/2)} \right)^2 \tag{9.12}$$

Remark 9.1 In the case of $K = 2$, this reduces to (9.5a):

$$N \geq \zeta(1-\zeta) \frac{1}{\varepsilon^2} \left(\frac{\Gamma(2/2)}{\Gamma((2+1)/2)} \right)^2 = \frac{\zeta(1-\zeta)}{\varepsilon^2} \frac{4}{\pi}$$

10 Modeling revenue

To model revenue we choose the following generative model. The revenue generated by an individual user i is given by:

$$\alpha_i \leftarrow \text{Bernoulli}(\lambda) \tag{10.1a}$$

$$r_i \leftarrow \text{Expon}(\theta) \tag{10.1b}$$

$$v_i \leftarrow \alpha_i \cdot r_i \tag{10.1c}$$

In this model, the variable $\alpha_i \in \{0, 1\}$ represents whether or not the user bought anything. The variable r_i represents the size of their purchase if they bought anything - otherwise this variable is meaningless and unobservable (since if $\alpha_i = 0$, then $\alpha_i \cdot r_i = 0 \cdot r_i = 0$ regardless of r_i). Finally the variable v_i represents the actual revenue generated by a visitor.

Note that the average revenue per sale is given by θ^{-1} and the average revenue per visitor is $\lambda\theta^{-1}$.

We wish to fit this model to an existing data set with the goal of computing a posterior on c_i, r_i and v_i.

10.1 Notation

In a given A/B test, let n_A represent the number of visitors in variation A, c_A the numer of sales in variation A, and s_A the *empirical* revenue per sale for variation A. Similarly let $n_B, \ldots$ represent the same variables for variation B.

Sometimes we will need to discuss individual sale variables. Toward that end, let $\mathbf{s}_A^k$ represent the size of the k-th customer's sale in variation A. Thus:

$$s_A = \frac{1}{c_A} \sum_{k=1}^{c_A} \mathbf{s}_A^k \tag{10.2}$$

When we are discussing only a single variation (mostly during the calculation of posteriors), we will sometimes drop the subscripts and discuss n, c, s instead.

11 Posterior distribution for revenue/sale

11.1 Computing a posterior on sale size

Define the $\Gamma(k, \theta)$-distribution as being the probability distribution having pdf $\gamma(k, \theta, x)$:

$$\gamma(k, \theta, x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \tag{11.1}$$

We can prove the following result:

Lemma 11.1 *Consider a set of variables* $\mathbf{s}^i, i = 1 \ldots c$ *drawn from an exponential distribution with decay rate* θ*. Suppose* θ *has a prior* $\Gamma(k, \Theta)$*. Then the posterior is given by:*

$$P(\theta|\mathbf{s}) \sim \Gamma\left(k + c, \frac{\Theta}{1 + \Theta c s}\right) \tag{11.2}$$

Proof. To compute the posterior $P(\theta|\mathbf{s})$, we use Bayes rule. In this calculation, we let $C = C(\Theta, k, \mathbf{s})$ be a constant which varies from line to line, but does not vary with θ.

$$\begin{aligned} P(\theta|\mathbf{s}) = \frac{P(\mathbf{s}|\theta)\gamma(k, \Theta, \theta)}{C} &= C\left(\prod_{i=1}^{c} \theta e^{-\theta \mathbf{s}^i}\right) \frac{1}{\Gamma(k)\Theta^k} \theta^{k-1} e^{-\theta/\Theta} \\ &= C\theta^{c+k-1} \exp\left[-\theta\left(\frac{1}{\Theta} + \sum_{i=1}^{c} \mathbf{s}^i\right)\right] \\ = \gamma\left(k + c, \frac{1}{1/\Theta + \sum_{i=1}^{c} \mathbf{s}^i}, \theta\right) &= \gamma\left(k + c, \frac{\Theta}{1 + \Theta \sum_{i=1}^{c} \mathbf{s}^i}, \theta\right) \end{aligned} \tag{11.3}$$

Determining that the last line is a gamma distribution follows from the second to last line by noting that the θ-variance of the expression fits the form of a Γ-distribution, and since the expression is normalized, there is no other possibility for it besides being a gamma distribution.

Equation (11.2) follows from (11.3) by using (5.5). □

11.2 The total posterior

Given the above, we can now compute a posterior on (λ, θ) jointly. Let us assume a prior $f(\lambda; a, b)$ on λ and $\gamma(k, \Theta, \theta)$ on θ. Then:

$$P(\lambda, \theta|n, c, s) = f(\lambda; a + c, b + n - c)\gamma\left(k + c, \frac{\Theta}{1 + \Theta c s}, \theta\right) \tag{11.4}$$

Facts about this distribution:

$$\mathrm{E}[\lambda] = \frac{a + c}{a + b + n} \tag{11.5a}$$

$$\mathrm{Median}[\lambda] = I_{1/2}^{-1}(a + c, b + n - c) \tag{11.5b}$$

Here $I_z^{-1}(\alpha, \beta)$ represents the inverse (w.r.t z) of the incomplete beta function $I_z(\alpha, \beta)$. This must be computed numerically, and quantiles can be computed similarly.

$$\mathrm{VAR}[\lambda] = \frac{(a + c)(b + n - c)}{(a + b + n)^2(a + b + n + 1)} \tag{11.5c}$$

$$\mathrm{E}[\theta] = \frac{\Theta(k+c)}{1+\Theta cs} \tag{11.5d}$$

Note that for large c this approaches s^{-1} as expected. There is no simple closed form for the median of θ, but it can be computed via numerical inversion of the CDF (for which a closed form does exist).

$$\mathrm{VAR}[\theta] = \frac{\Theta^2(k+c)}{(1+\Theta cs)^2} \tag{11.5e}$$

As expected, the variance decreases like $O(c^{-1})$.

It's also useful to know the variance of the *average sale size*, which can be computed via Proposition B.2 on page 28:

$$\mathrm{VAR}[\theta^{-1}] = \frac{(\Theta^{-1}+cs)^2}{(k+c-1)^2(k+c-2)} \tag{11.5f}$$

As expected this behaves like $O(c^{-1})$ as $c \to \infty$.

12 Chance to beat all for Revenue

The total gain from any variation is λ/θ - i.e. the probability of a sale times the average value of a sale.

Suppose we've computed a posterior for λ_A, θ_A and λ_B, θ_B. The chance to beat is given by:

$$\begin{aligned} &P(\lambda_B/\theta_B > \lambda_A/\theta_A) \\ &\qquad = \int\int\int\int \mathcal{H}\left(\lambda_B/\theta_B - \lambda_A/\theta_A\right) \\ &\qquad \cdot f(\lambda; a+c_A, b+n_A-c_A)\gamma\left(k+c_A, \frac{\Theta}{1+\Theta c_A s_A}, \theta_A\right) \\ &\qquad \cdot f(\lambda_B; a+c_B, b+n_B-c_B)\gamma\left(k+c_B, \frac{\Theta}{1+\Theta c_B s_B}, \theta_B\right) d\lambda_A d\lambda_B d\theta_A d\theta_B \end{aligned} \tag{12.1}$$

Here the function $\mathcal{H}(x)$ is the Heaviside step function defined by $\mathcal{H}(x) = 1$ for $x >= 0$ and $\mathcal{H}(x) = 0$ for $x < 0$.

Because of the $\mathcal{H}(\lambda_B/\theta_B - \lambda_A/\theta_A)$ term, this integral is not separable. Because it is 4 dimensional, using a standard numerical quadrature will be extremely computationally intensive. However, this integral can be computed via Monte Carlo sampling:

1. Draw samples ${\lambda_A}^i, {\lambda_B}^i, {\theta_A}^i, {\theta_B}^i$ for $i = 1 \ldots M$ from the relevant distributions.
2. Count the number of samples for which ${\lambda_B}^i/{\theta_B}^i > {\lambda_A}^i/{\theta_A}^i$ and divide by M.

13 Making decisions based on Revenue

We now define our loss function as the difference in expected revenue (assuming a mistake) between the chosen variation and the max of the others:

$$\mathcal{L}(\lambda_A, \lambda_B, \theta_A, \theta_B, A) = \max(\lambda_B/\theta_B - \lambda_A/\theta_A, 0) \tag{13.1a}$$

$$\mathcal{L}(\lambda_A, \lambda_B, \theta_A, \theta_B, B) = \max(\lambda_A/\theta_A - \lambda_B/\theta_B, 0) \tag{13.1b}$$

With more variants, it would be defined as:

$$\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, \theta_A, \theta_B, \theta_C, A) = \max\left(\frac{\lambda_B}{\theta_B} - \frac{\lambda_A}{\theta_A}, \frac{\lambda_C}{\theta_C} - \frac{\lambda_A}{\theta_A}, 0\right) \tag{13.2a}$$

$$\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, \theta_A, \theta_B, \theta_B, B) = \max\left(\frac{\lambda_A}{\theta_A} - \frac{\lambda_B}{\theta_B}, \frac{\lambda_C}{\theta_C} - \frac{\lambda_B}{\theta_B}, 0\right) \tag{13.2b}$$

We can then define the expected losses:

$$E[\mathcal{L}](A) = E[\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, \theta_A, \theta_B, \theta_C, A)] \tag{13.3a}$$

$$E[\mathcal{L}](B) = E[\mathcal{L}(\lambda_A, \lambda_B, \lambda_C, \theta_A, \theta_B, \theta_C, B)] \tag{13.3b}$$

Remark 13.1 For the Monte Carlo simulations above, I'm using the formula $M = (-\ln\delta)/\epsilon^2$. Here δ is a desired probability of error (I typically choose 10^{-10}) and ϵ is a desired magnitude of error (I typically choose 10^{-3}). This error bound is not strictly accurate - it's the same bound as in Theorem 8.3. Proving a similar bound for a distribution with infinite support (such as the Gamma distribution) is an open problem.

As in the Bernoulli case, we allow a decision to be made when the expected loss drops below a *threshold of caring* ε.

The decision procedure consists of finding the set of variations who's loss is below our threshold of caring:

$$\{X \in \{A, B, ...\} : E[\mathcal{L}](X) \leq \varepsilon\} \tag{13.4}$$

The test can be stopped when this set is nonempty, and any variation in this set is an acceptable choice.

14 Revenue Empirics

To measure the efficacy of this scheme according to frequentist criteria I ran several A/A and A/B tests.

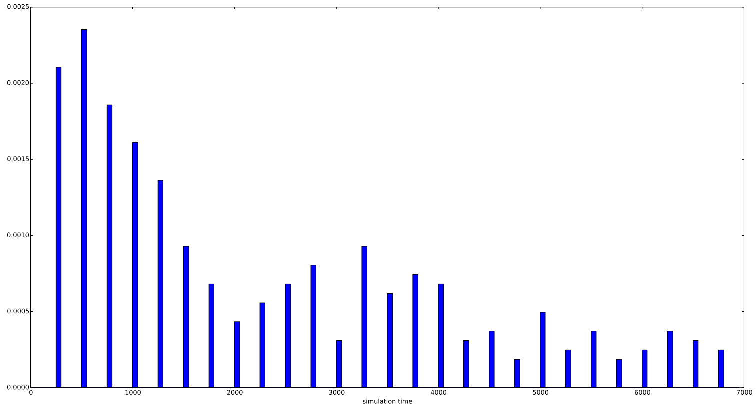

Figure 8: Illustration of the termination time of the second A/B test.

14.1 Test - A/A test

To begin I ran an A/A test comparing a sales process with a 5% conversion rate and a mean of 25$ revenue/sale. In this example, the standard deviation of the data is $5.6 (compared to a mean revenue/visitor of $1.25).

According to Evan Miller's t-test calculator at `http://www.evanmiller.org/ab-testing/t-test.html`, this test will require 4,250 data points (per sample) to resolve a 20% lift.

In the simulation, the Bayesian test finished with an average of about 3,292 data points (10th/90th percentile 250/7,750). The distribution is plottedin Figure 8.

14.2 Test - varying conversion rates

We now consider a comparison between a sales process with a conversion rate of 5% and 6% in the variant. The revenue/sale is the same (25$) in both variants.

I ran a set of 400 A/B simulations. The threshold of caring ε was chosen to be 2% of the mean visitor value of $1.25, or $0.025. I.e., a $0.025 loss was considered acceptable.

In the A/B simulations I ran, the Bayesian test finished with an average of about 3,000 data points. Out of 400 simulations, 95.75% returned the correct result. The 10'th percentile of test durations was 250 samples, the 50'th percentile was 1,500 samples and the 95'th sample was 4,750. See Figure 9 for an illustration.

This approximately corresponds to the *balanced mode* in the tool.

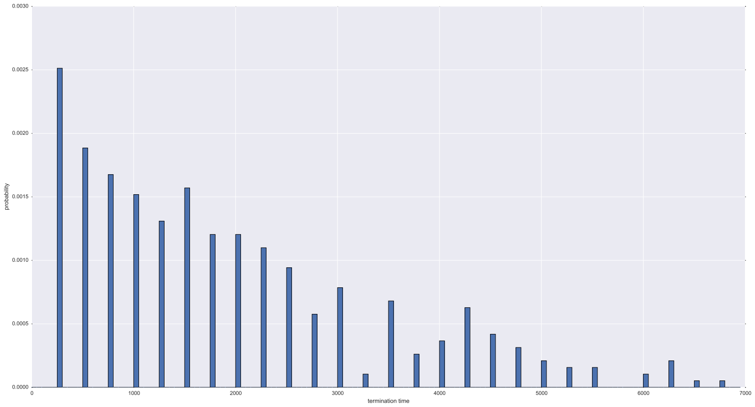

Figure 9: Illustration of the termination time of the first A/B test.

14.3 Test - varying average sale price

The second comparison considers a sales process with a conversion rate of 5% in both variants, but one variant has an average sale price of $30 rather than $20. The results were quite similar, and an error occurred in only 6% of cases. The results are plotted in Figure 9.

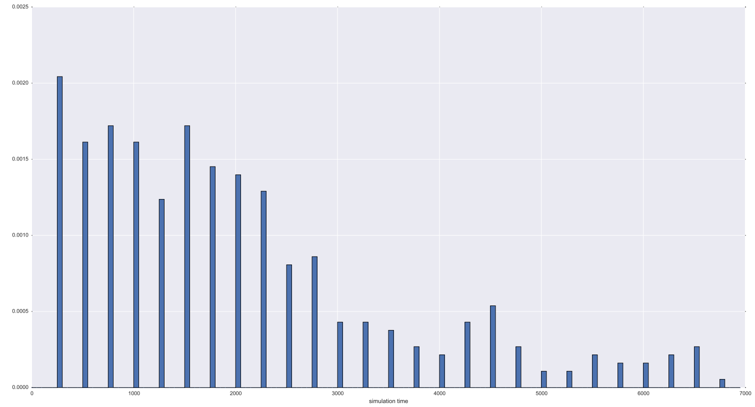

Figure 10: Illustration of the termination time of the second A/B test.

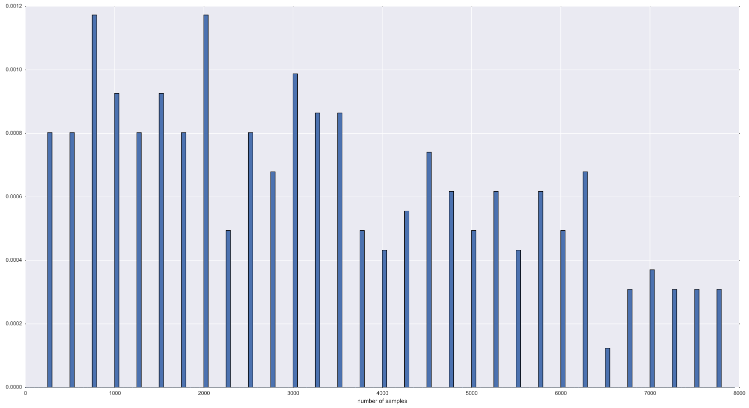

Figure 11: Illustration of the termination time of the non-exponential revenue distribution.

14.4 Test - revenue not from an exponential distribution

The same test as in section 14.2 was run, but this time the distribution of sale prices was chosen uniformly from [\$49,\$129,\$259]. The conversion rate was 5% in variation A and 6% in variation B. This time the A/B test yielded the correct answer 81% of the time. This time the mean number of samples required was 3,290, and the 95'th percentile was 7,000.

19% of the A/B tests did not reach convergence before 8,000 samples. These were coded as returning variation A - as a result, the accuracy of the non-exponential distribution was only 81%.

14.5 Why does Bayesian testing require fewer samples?

I attribute the drastically reduced number of samples primarily to the exact modelling of the relevant probability distributions, i.e. modelling the data via (10.1) rather than simply unknown random variables via the central limit theorem. This is likely a major source of win - the CLT model allows for a large probability of *negative revenue*, since the standard deviation of \$5.6 is much larger than the mean of \$1.25.

It is my belief that if we wished, we could develop a frequentist model of conversions based on the exact probability distribution with similar accuracy.

15 Approximate worst case for the Revenue test

The following calculations are *approximate only*, and roughly correspond to those in Section 9.

15.1 Worst case for 2 variants

We might ask the question, what is the worst case scenario for the Bernoulli test to finish? The test will finish when $E[\mathcal{L}](A) \leq \varepsilon$, or:

$$E[\mathcal{L}](A) = \int\int\int\int \max(\lambda_A/\theta_A - \lambda_B/\theta_B, 0) P(\lambda_A, \lambda_B, \theta_A, \theta_B) d\lambda_A d\lambda_B d\theta_A d\theta_B \\ \leq \varepsilon \quad (15.1)$$

The worst case occurs when both variants are identically equal. So let us suppose that $n_A = n_B = N$, $c_A = c_B = \zeta N$ and $s_A = s_B = \nu$.

From section 9, we recall that for large N, the Beta distribution approximates a normal distribution of mean ζ and standard deviation $\sqrt{\zeta(1-\zeta)/N}$. In a similar way, the distribution of $1/\theta_A$ when drawn from Gamma$(k+c, \Theta/(1+\Theta cs))$ approximates a normal distribution with mean

$$\mu_r = \frac{1 + \Theta cs}{\Theta(k+c)} = \frac{1 + \Theta N\zeta s}{\Theta(k + N\zeta)} \quad (15.2a)$$

and standard deviation

$$\sigma_r = \frac{(1 + \Theta N\zeta s)}{\Theta(k + N\zeta - 1)\sqrt{k + N\zeta - 2}} \quad (15.2b)$$

In what follows, let $g(x; \zeta, \sigma)$ be the pdf of a normal distribution with mean ζ and standard deviation σ.

Definition 15.1 *Define the function $\mathfrak{h}(z)$ by:*

$$\mathfrak{h}(z) = \int |x| \frac{e^{-(x-z)^2/2}}{\sqrt{2\pi}} dx \quad (15.3)$$

Note that $\mathfrak{h}(z) \approx z$ for large z.

Remark 15.2 Note that:

$$\int |x| \frac{e^{-(x-z)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} dx = \sigma\mathfrak{h}(z/\sigma) \quad (15.4)$$

We can thus approximate:

$$\begin{aligned}\int\int\int\int &\max(\lambda_A/\theta_A - \lambda_B/\theta_B, 0) P(\lambda_A, \lambda_B, \theta_A, \theta_B) d\lambda_A d\lambda_B d\theta_A d\theta_B \\ &\leq \int\int\int\int |\lambda_A/\theta_A - \lambda_B/\theta_B| \, P(\lambda_A, \lambda_B, \theta_A, \theta_B) d\lambda_A d\lambda_B d\theta_A d\theta_B \\ &\approx \int\int\int\int |\lambda_A/\theta_A - \lambda_B/\theta_B| \, T d\lambda_A d\lambda_B d\theta_A d\theta_B\end{aligned} \quad (15.5)$$

where

$$T = g(\lambda_A; \zeta, \sqrt{\zeta(1-\zeta)}/\sqrt{N}) g(\lambda_B; \zeta, \sqrt{\zeta(1-\zeta)}/\sqrt{N}) \\ \cdot g(s_A; \mu_r \sigma_r) g(s_B; \mu_r \sigma_r)$$

We can apply Lemma C.2 to (15.5) to show that:

$$(15.5) \leq \frac{2}{\sqrt{\pi}} \left[\mathfrak{h} \left(\zeta \sqrt{\frac{N}{\zeta(1-\zeta)}} \right) + \mathfrak{h} \left(\mu_r / \sigma_r \right) \right] \sigma_r \sqrt{\frac{\zeta(1-\zeta)}{N}}$$

Thus, to compute N we must choose N sufficiently large so that:

$$\frac{2}{\sqrt{\pi}} \left[\mathfrak{h} \left(\zeta \sqrt{\frac{N}{\zeta(1-\zeta)}} \right) + \mathfrak{h} \left(\mu_r / \sigma_r \right) \right] \sigma_r \sqrt{\frac{\zeta(1-\zeta)}{N}} \leq \varepsilon \tag{15.6}$$

A Proof of Theorem 8.5

A proof is provided here since I don't know of an external source to cite.

Lemma A.1 *The function $\mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A)$ is convex in $(\lambda_A, \lambda_B, \ldots)$.*

Proof. The functions $\lambda_B - \lambda_A$, $\lambda_C - \lambda_A$ and 0 are all convex. The maxima of a set of convex functions is convex. Note that $\mathcal{L}({\lambda_A}^i, {\lambda_B}^i, \ldots, A)$ is defined as the maxima of the aforementioned linear functions, and hence is convex. □

To prove Theorem 8.5, we must apply the Lindeberg-Levy Central Limit Theorem - see `http://en.wikipedia.org/wiki/Central_limit_theorem#Classical_CLT`.

Specifically, we must provide a bound on the standard deviation of $\mathcal{L}(\lambda_A, \lambda_B, \ldots, A)$.

Consider the random variable:

$$\lambda_B - \lambda_A = \max(\lambda_B - \lambda_A, 0) + \min(\lambda_B - \lambda_A, 0) \tag{A.1}$$

We can compute the variance of both sides:

$$\mathrm{VAR}[\lambda_B - \lambda_A] = \mathrm{VAR}[\max(\lambda_B - \lambda_A, 0)] + \mathrm{VAR}[\min(\lambda_B - \lambda_A, 0)] \\ + \text{covariance}\left[\max(\lambda_B - \lambda_A, 0), \min(\lambda_B - \lambda_A, 0)\right]$$

The latter term is zero since the two variables inside the covariance have disjoint support. Thus:

$$\mathrm{VAR}[\lambda_B - \lambda_A] = \mathrm{VAR}[\max(\lambda_B - \lambda_A, 0)] + \mathrm{VAR}[\min(\lambda_B - \lambda_A, 0)]$$

or

$$\mathrm{VAR}[\max(\lambda_B - \lambda_A, 0)] = \mathrm{VAR}[\lambda_B - \lambda_A] - \mathrm{VAR}[\min(\lambda_B - \lambda_A, 0)] \\ \leq \mathrm{VAR}[\lambda_B - \lambda_A] \tag{A.2}$$

Similarly for λ_C we have VAR[$\max(\lambda_C - \lambda_A, 0)$] $=\leq$ VAR[$\lambda_C - \lambda_A$], etc. We now observe that:

$$\begin{aligned}&\mathrm{VAR}[\max(\lambda_B - \lambda_A, \lambda_C - \lambda_A, \ldots, 0)] \leq \\ &\qquad \mathrm{VAR}[\max(\lambda_B - \lambda_A, \ldots, 0)] + \mathrm{VAR}[\max(\lambda_C - \lambda_A, \ldots, 0)] + \ldots \\ &\qquad\qquad \leq \mathrm{VAR}[\lambda_B - \lambda_A] + \mathrm{VAR}[\lambda_C - \lambda_A] + \ldots \end{aligned} \tag{A.3}$$

Thus, the variance of $\mathcal{L}(\lambda_A, \lambda_B, \ldots, A)$ satisfies the bound:

$$\mathrm{VAR}[\mathcal{L}(\lambda_A, \lambda_B, \ldots, A)] \leq \mathrm{VAR}[\lambda_B - \lambda_A] + \mathrm{VAR}[\lambda_C - \lambda_A] + \ldots \tag{A.4}$$

B Numerical results for Gamma variables

This section merely computes a few simple facts about the $\Gamma(k, \theta)$ distribution. Let x be a random variable distributed according to $\Gamma(k, \theta)$.

Proposition B.1 *Let $k > 2$. Then:*

$$E[x^{-1}] = (k\theta)^{-1} \tag{B.1}$$

Proof. A computation:

$$\begin{aligned}&E[x^{-1}] = \int_0^\infty x^{-1} \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} dx = \\ &\quad \frac{\Gamma(k-1)}{\Gamma(k)\theta} \int_0^\infty \frac{1}{\Gamma(k-1)\theta^{(k-1)-1}} x^{(k-1)-1} e^{-x/\theta} = \frac{\Gamma(k-1)}{\Gamma(k)\theta} = \frac{1}{\theta(k-1)}\end{aligned}$$

□

Proposition B.2 *Let $k > 3$. Then:*

$$VAR[x^{-1}] = \theta^{-2}(k-1)^{-2}(k-2)^{-1} \tag{B.2}$$

Proof. A computation. First compute $E[x^{-2}]$:

$$\begin{aligned}&E[x^{-2}] = \int_0^\infty x^{-2} \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} dx = \\ &\frac{\Gamma(k-2)}{\Gamma(k)\theta} \int_0^\infty \frac{1}{\Gamma(k-2)\theta^{(k-2)-1}} x^{(k-2)-1} e^{-x/\theta} = \frac{\Gamma(k-2)}{\Gamma(k)\theta^2} = \frac{1}{\theta^2(k-1)(k-2)}\end{aligned}$$

Now we compute:

$$\begin{aligned}\mathrm{VAR}[x^{-1}] = E[x^{-2}] - (E[x^{-1}])^2 &= \frac{1}{\theta^2(k-1)(k-2)} - \frac{1}{\theta^2(k-1)^2} \\ &= \frac{1}{\theta^2} \frac{(k-1)-(k-2)}{(k-1)^2(k-2)} = \frac{1}{\theta^2} \frac{1}{(k-1)^2(k-2)}\end{aligned}$$

□

C Facts about gaussians

Lemma C.1

$$\int\int |x-y|\, g(x;\zeta,\sigma) g(y;\zeta,\sigma) dx dy = \frac{2\sigma}{\sqrt{\pi}} \tag{C.1}$$

Proof. Change variables to $u = (x-y)/\sqrt{2}$, $v = (x+y)/\sqrt{2}$. The integral becomes:

$$\begin{aligned}
&\int\int |x-y|\, g(x;\zeta,\sigma) g(y;\zeta,\sigma) dx dy \\
&\qquad = \int\int \sqrt{2}\, |u|\, g(u;\zeta,\sigma) g(v - \sqrt{2}\zeta;\zeta,\sigma) du dv \\
&\qquad = \left(\int g(v-\sqrt{2}\zeta;\zeta,\sigma) dv\right)\left(\int \sqrt{2}\, |u|\, g(u;\zeta,\sigma) du\right) \\
&\qquad = 1 \cdot \left(\int \sqrt{2}\, |u|\, g(u;\zeta,\sigma) du\right) = \frac{2\sigma}{\sqrt{\pi}}
\end{aligned} \tag{C.2}$$

The last equality is given on wikipedia: `http://en.wikipedia.org/wiki/Normal_distribution#Moments` □

Lemma C.2 *Let $x_1, x_2 \sim N(\zeta, \sigma_x)$ and let $y_1, y_2 \sim N(\xi, \sigma_y)$. Then:*

$$\begin{aligned}
E[|x_1 y_1 - x_2 y_2|] &\leq \frac{2}{\sqrt{\pi}} \left(\sigma_x \mathfrak{h}(\zeta/\sigma_x)\sigma_y + \sigma_y \mathfrak{h}(\xi/\sigma_y)\sigma_x\right) \\
&= \frac{2\sigma_x\sigma_y}{\sqrt{\pi}} \left(\mathfrak{h}(\zeta/\sigma_x) + \mathfrak{h}(\xi/\sigma_y)\right)
\end{aligned} \tag{C.3}$$

Proof. First note that:

$$\begin{aligned}
|x_1 y_1 - x_2 y_2| &= |x_1 y_1 - x_1 y_2 + x_1 y_2 - x_2 y_2| \\
&\leq |x_1 y_1 - x_1 y_2| + |x_1 y_2 - x_2 y_2| \\
&= |x_1|\, |y_1 - y_2| + |y_1|\, |x_1 - x_2|
\end{aligned} \tag{C.4}$$

Using this, we find that:

$$\begin{aligned}
&\int\int\int\int |x_1 y_1 - x_2 y_2|\, g(x_1;\zeta,\sigma_x) g(x_2;\zeta,\sigma_x) g(y_1;\zeta,\sigma_y) g(y_2;\zeta,\sigma_y) dx_1 dx_2 dy_1 dy_2 \\
&\qquad \leq \int\int\int\int |x_1|\, |y_1 - y_2|\, T dx_1 dx_2 dy_1 dy_2 \\
&\qquad\qquad + \int\int\int\int |y_1|\, |x_1 - x_2|\, T dx_1 dx_2 dy_1 dy_2
\end{aligned} \tag{C.5}$$

where $T = g(x_1;\zeta,\sigma_x) g(x_2;\zeta,\sigma_x) g(y_1;\zeta,\sigma_y) g(y_2;\zeta,\sigma_y)$.

The first integral can be simplified as follows:

$$\begin{aligned}
&\int\int\int\int |x_1|\,|y_1-y_2|\,T dx_1dx_2dy_1dy_2 \\
&\qquad = \int\int\int |x_1|\,|y_1-y_2|\,g(x_1;\zeta,\sigma_x)g(y_1;\zeta,\sigma_y)g(y_2;\zeta,\sigma_y)dx_1dy_1dy_2 \\
&\qquad = \int |x_1|\,g(x_1;\zeta,\sigma_x)dx_1\int\int |y_1-y_2|\,g(y_1;\zeta,\sigma_y)g(y_2;\zeta,\sigma_y)dy_1dy_2 \\
&\qquad\qquad = \sigma_x\mathfrak{h}(\zeta/\sigma_x)\frac{2\sigma_y}{\sqrt{\pi}} \qquad \text{(C.6)}
\end{aligned}$$

The same calculation can be applied to show that the second integral is equal to $\sigma_y\mathfrak{h}(\xi/\sigma_y)2\sigma_x/\sqrt{\pi}$. This yields the result we seek. □

Proposition C.3 *We have the identity:*

$$\int_0^\infty x^t e^{-x^2/2}dx = 2^{(t-1)/2}\Gamma([t+1]/2) \qquad \text{(C.7)}$$

Proof. A calculation. Let $u = x^2/2$, then $x = \sqrt{2u}$ and $dx = (2u)^{-1/2}du$.

$$\begin{aligned}
\int_0^\infty x^t e^{-x^2/2}dx &= \int_0^\infty (2u)^{t/2}e^{-u}\frac{du}{\sqrt{2u}} \\
&= 2^{(t-1)/2}\int_0^\infty u^{t/2-1/2}e^{-u}du = 2^{(t-1)/2}\Gamma([t+1]/2) \qquad \text{(C.8)}
\end{aligned}$$

□

D Facts about inverse gamma distributions

Consider an inverse gamma distribution, having pdf

$$\mathfrak{g}(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x} \qquad \text{(D.1a)}$$

and with cdf

$$\mathfrak{G}(x;\alpha,\beta) = \frac{\Gamma(\alpha,\beta/x)}{\Gamma(\alpha)}. \qquad \text{(D.1b)}$$

Proposition D.1 *We have the following results about expected values of loss functions:*

$$E[\max\{x-z,0\}] = \frac{\beta}{\alpha-1}[1-\mathfrak{G}(\zeta;\alpha-1,\beta)] - z[1-\mathfrak{G}(\zeta;\alpha,\beta)] \qquad \text{(D.2a)}$$

$$E[\max\{z-x,0\}] = z\mathfrak{G}(\zeta;\alpha,\beta) - \frac{\beta}{\alpha-1}\mathfrak{G}(\zeta;\alpha-1,\beta) \qquad \text{(D.2b)}$$

Proof. First note that

$$x\mathfrak{g}(x;\alpha,\beta) == \left(\frac{\beta}{\alpha-1}\right)\frac{\beta^{\alpha-1}}{\Gamma(\alpha-1)}x^{-\alpha-2}e^{-\beta/x} = \frac{\beta}{\alpha-1}\mathfrak{g}(x;\alpha-1,\beta).$$

$$\begin{aligned}\int_z^\infty (x-z)\mathfrak{g}(x;\alpha,\beta)dx &= \frac{\beta}{\alpha-1}\int_z^\infty \mathfrak{g}(x;\alpha-1,\beta)dx - z\int_z^\infty \mathfrak{g}(x;\alpha,\beta)dx \\ &= \frac{\beta}{\alpha-1}[1-\mathfrak{G}(\zeta;\alpha-1,\beta)] - z[1-\mathfrak{G}(\zeta;\alpha,\beta)] \quad \text{(D.3)}\end{aligned}$$

Similarly:

$$\begin{aligned}\int_0^z (z-x)\mathfrak{g}(x;\alpha,\beta)dx &= \frac{\beta}{\alpha-1}\int_z^\infty \mathfrak{g}(x;\alpha-1,\beta)dx - z\int_z^\infty \mathfrak{g}(x;\alpha,\beta)dx \\ &= z\mathfrak{G}(\zeta;\alpha,\beta) - \frac{\beta}{\alpha-1}\mathfrak{G}(\zeta;\alpha-1,\beta) \quad \text{(D.4)}\end{aligned}$$

□

Proposition D.2 *Let x be distributed according to an inverse gamma distribution with params α,β. Then:*

$$E[|x-z|] = z\left[2\mathfrak{G}(z;\alpha,\beta)-1)\right] + \frac{\beta}{\alpha-1}\left[2\mathfrak{G}(z;\alpha-1,\beta)-1\right] \quad \text{(D.5)}$$

Proof. Apply proposition D.1 twice. □

Proposition D.3 *Let $u, w > 0$. Then we can compute:*

$$\int\int |wx-uy|\,\mathfrak{g}(x;\alpha_1,\beta_1)\mathfrak{g}(y;\alpha_2,\beta_2)dxdy \quad \text{(D.6)}$$

Proof.

$$\begin{aligned}&\int\int |wx-uy|\,\mathfrak{g}(x;\alpha_1,\beta_1)\mathfrak{g}(y;\alpha_2,\beta_2)dxdy \\ &\qquad = w\int\left[\int |x-uy/w|\,\mathfrak{g}(x;\alpha_1,\beta_1)dx\right]\mathfrak{g}(y;\alpha_2,\beta_2)dy \\ &= \int\left(uy\left[2\mathfrak{G}(uy/w;\alpha_1,\beta_1)-1\right] + \frac{\beta_1}{(\alpha_1-1)w}\left[2\mathfrak{G}(uy/w;\alpha_1-1,\beta_1)-1\right]\right)\mathfrak{g}(y;\alpha_2,\beta_2)dy \\ &\qquad = \int u\frac{\beta_2}{(\alpha_2-1)w}\left[2\mathfrak{G}(uy/w;\alpha_1,\beta_1)-1\right]\mathfrak{g}(y;\alpha_2-1,\beta_2)dy \\ &\qquad\quad + \int \frac{\beta_1}{(\alpha_1-1)w}\left[2\mathfrak{G}(uy/w;\alpha_1-1,\beta_1)-1\right]\mathfrak{g}(y;\alpha_2,\beta_2)dy \quad \text{(D.7)}\end{aligned}$$

□