False positive rates of Bayesian A/B tests with optional stopping #227

New issue

Open **fraser** opened this issue on Sep 14, 2014 · 9 comments

fraser commented on Sep 14, 2014

In Chapter 2 > Example: Bayesian A/B testing, it says

> If this probability is too high for comfortable decision-making, we can perform more trials on site B (as site B has less samples to begin with, each additional data point for site B contributes more inferential "power" than each additional data point for site A).

This implies that we don't stop the experiment at a pre-determined sample size, and can stop the experiment whenever we'd like without affecting the results. I've seen this in several discussions of Bayesian experiments and implemented in Bayesian A/B testing software like Google Analytics' Content Experiments, but I think optional stopping has adverse effects on the properties of the experiment. In particular, I show that optional stopping increases the false positive rate.

In frequentist A/B tests, sample size of the experiment is typically determined beforehand based on effect size, false positive rate,and false negative rate, and the experiment cannot be stopped early without changing the properties of the experiment.
(There are ways of adjusting for multiple peeks at the data and early stopping, but fixed sample sizes are typical.)

The example in Chapter 2 covers the case that versions A & B are truly different. Here, I'll consider the case that they are identical - i.e., this is an A/A test. The below is a simulation in R which shows the properties of Bayesian A/B tests are also affected by early stopping. Specifically, the false positive rate increases with the number of peeks at the experiment.

(I use R because I'm not yet familiar enough with PyMC to be confident in the results. Hopefully the code is clear - if not, I can try reproducing the results using PyMC.)

The stopping rule is: stop when a version appears better than the alternative with 95% probability.
This rule is unbiased between the two versions: it is looking for a result, but not a specific result.

The below uses a uniform beta(1, 1) prior, which leads to a beta(successes+1, failures+1) posterior. It determines P(A < B) and P(B < A) by simulating from the posteriors.

```
n.trials = 10000

p.A = 0.05
p.B = 0.05

cat("unbiased stopping rule (stop on any positive)\n")
for (optional.stops in 0:10) {
    cnt.positives = 0
    for (idx.trial in 1:n.trials) {
        n.A = 1500
        n.B = 750

        obs.A = rbinom(1,n.A,p.A)
        obs.B = rbinom(1,n.B,p.B)
        p.A.samples = rbeta(10000,obs.A+1,n.A-obs.A+1) # simulate from the posterior distribution of
        p.B.samples = rbeta(10000,obs.B+1,n.B-obs.B+1)
        delta.samples = p.A.samples - p.B.samples

        ## Stop the experiment if either version appears better with > 95% probability.
        positive.result = .95 < mean(delta.samples < 0) || .95 < mean(0 < delta.samples)

        if (0 < optional.stops) {
```

Assignees
No one assigned

Labels
None yet

Projects
None yet

Milestone
No milestone

3 participants

```
            ## If optional stopping is enabled and we haven't reached a conclusion yet,
            ## collect another 100 samples for A & B and decide again if we should stop the experime
            for (i in 1:optional.stops) {
                if (positive.result)
                    break

                obs.A = obs.A + rbinom(1,100,p.A)
                n.A = n.A + 100
                obs.B = obs.B + rbinom(1,100,p.B)
                n.B = n.B + 100

                p.A.samples = rbeta(10000,obs.A+1,n.A-obs.A+1)
                p.B.samples = rbeta(10000,obs.B+1,n.B-obs.B+1)
                delta.samples = p.A.samples - p.B.samples

                positive.result = .95 < mean(delta.samples < 0) || .95 < mean(0 < delta.samples)
            }
        }

        if (positive.result)
            cnt.positives = cnt.positives + 1
    }
    cat(sprintf("%d optional stops: positive rate: %d/%d = %f\n",
                optional.stops,cnt.positives,n.trials,cnt.positives/n.trials))
}
```

which produces output

> unbiased stopping rule (stop on any positive)
> 0 optional stops: positive rate: 978/10000 = 0.097800
> 1 optional stops: positive rate: 1211/10000 = 0.121100
> 2 optional stops: positive rate: 1472/10000 = 0.147200
> 3 optional stops: positive rate: 1599/10000 = 0.159900
> 4 optional stops: positive rate: 1773/10000 = 0.177300
> 5 optional stops: positive rate: 1748/10000 = 0.174800
> 6 optional stops: positive rate: 1978/10000 = 0.197800
> 7 optional stops: positive rate: 2121/10000 = 0.212100
> 8 optional stops: positive rate: 2129/10000 = 0.212900
> 9 optional stops: positive rate: 2307/10000 = 0.230700
> 10 optional stops: positive rate: 2376/10000 = 0.237600

So the false positive rate is as expected without optional stops (~10%, which makes sense given our stopping rule), but increases substantially with the number of optional stopping points.

What am I missing here? Is increased false positive rate viewed as okay in Bayesian A/B tests? Or am I misinterpreting why Bayesian experiments view optional stopping as okay in general?

CamDavidsonPilon commented on Sep 14, 2014 Owner

Ooo good discussion! Thanks for the R simulation, it's very valuable to see these numbers.

False positives and peeking are still an issue in Bayesian testing. The point I wanted to make was that we are given an extra parameter in Bayesian A/B testing. Whereas with frequentist AB testing, we are returned a boolean (is the p-value below the threshold? Yes or No?), in Bayesian testing we get back the probability of one group being larger than the other, call this *p*. This gives our directors additional degrees of freedom to make a decision: is p high enough to conclude the experiment?

As you showed, false positives will occur if we keep peeking. In practice, what I like to do is wait a few days after the experiment shows significance, to confirm the conclusion. If it is a true conclusion, then it should increase the *p*, or at least not drop too much. If *p* does fall, then what's needed is more either a) more data to make a conclusion, or b) close the experiment and choose whichever the director prefers.

On the subject of A/A tests, there was a great discussion on Cross Validated that you will find very interesting: https://stats.stackexchange.com/questions/97381/why-is-this-distribution-uniform

CamDavidsonPilon commented on Sep 14, 2014 Owner

W.r.t the book, either I can update the text with more clarity, or you can send a pull request too if you'd like.

fraser commented on Sep 15, 2014

Thanks for the response, Cam!

I definitely agree that it is easier to make intuitive sense of `P(A < B)` than a p-value. I find p-values okay to work with, but they are not very useful for communicating results to non-statisticians. :)

In your example, you say you collect more data to confirm the result upon hitting a positive result. But, doesn't any stopping rule that favours positive results (e.g., stop-at-95%) - even if more data appears to confirm the result - bias the experiment away from negative results? I suspect that using more data to confirm the result helps correct for this, but you may see data supporting the conclusion by chance. In practice, maybe the effect is pretty small (and I'm not meaning to say this approach is wrong) - but at the core, I'm trying to understand what the theoretical properties of Bayesian A/B tests are (specifically, false positive and false negative rate) and how they relate to the exact stopping rule chosen.

For example, in a frequentist A/B test, we can say: use this sample size to get a 5% false positive rate and a 20% false negative rate when attempting to detect an effect size of 3%. What is the equivalent for a Bayesian A/B test?

Once I understand this, I will be very happy to make a pull request - it is something I've been trying to track down for quite some time, and it'd be great to share that with others. :)

And, thanks for the link to stackoverflow - that was very interesting!

CamDavidsonPilon commented on Sep 15, 2014 Owner

Here are some of my thoughts, mostly derived from practice. In reality, A & B are always truly different (except in artificial A/A tests). That is *p_A != p_B*. This should be clear - it's highly unlikely two groups, given even a minor change, will respond with exactly equal conversion rates.

Compare this to the null hypothesis test's *H_0* that states that they are indeed equal. The NHST tries to disprove equality. In practice, I don't really want this: I want to know which group has a larger conversion rate. I know they are *not* equal, just tell me which one is larger. For this reason, if my stopping rule was *wait for N samples, stop and look at the results*, then 20% of my tests would, IMO, fail. (Because the test lacked sufficient statistical power to give me the answer I want: which is bigger). For this reason, the *wait for N samples* is not my favourite stopping rule and not the one I use in practice. In truth, I don't have a rigorous stopping rule (I mentioned my heuristic of waiting a few days after significance).

This means some false positive will fall through. That's okay. In the book, I mention that if the experiment is really important, I'll be more strict with how much data I want to collect and stress to the directors to keep waiting to achieve higher *p*. On the other hand, if the experiment is less important, we can be looser with stopping rules. In fact, because the false positive rate is a decreasing function of the delta between the conversion rates, if a false positive does occur, likely the difference between the groups is small, so choosing the wrong one is not a huge disaster.

Does this help? I'm sorta just writing down unwritten thoughts in my head =)

fraser commented on Sep 16, 2014

Hi Cam - that does help, thanks! I follow your reasoning, and it makes sense to adjust the experiment according to the cost of false positives.

Re: A & B are always truly different: this makes sense to me. Then, I got curious about how the error rates changed with optional stopping depending on how similar `p_A` and `p_B` were and performed another simulation below. What follows is mostly for your interest rather than really requiring discussion. At a high level, I think I'm starting to wrap my head around this. Optional stopping does affect the experiment properties (e.g., false positive rates), but it doesn't affect the inference based on the data when using Bayesian methods. I think... I'm still having a bit of trouble putting it all together in my head. (There is a good discussion along similar lines here: http://doingbayesiandataanalysis.blogspot.ca/2013/11/optional-stopping-in-data-collection-p.html .)

In the following, `p_A > p_B`, but by amounts from 0.001 to 0.02.

This code is getting a bit long, but it does mostly the same thing as the previous simulation. The main difference is it loops over cases where there's truly a difference between A&B and now also collect effect sizes.

```
library(ggplot2)

n.trials = 10000

p.trials = list(
    list(p.A = 0.05, p.B = 0.049),
```

```
    list(p.A = 0.05, p.B = 0.04),
    list(p.A = 0.05, p.B = 0.03)
)
for (p.trial in p.trials) {
    p.A = p.trial$p.A
    p.B = p.trial$p.B
    data = c()
    cat(sprintf("\n\ntrue difference exists: %f vs %f\n",p.A,p.B))
    cat("unbiased stopping rule\n")
    for (optional.stops in c(0,1,10,20)) {
        cnt.positives = 0
        cnt.positives.A = 0
        cnt.positives.B = 0
        for (idx.trial in 1:n.trials) {
            n.A = 1500
            n.B = 750

            obs.A = rbinom(1,n.A,p.A)
            obs.B = rbinom(1,n.B,p.B)
            p.A.samples = rbeta(10000,obs.A+1,n.A-obs.A+1)
            p.B.samples = rbeta(10000,obs.B+1,n.B-obs.B+1)
            delta.samples = p.A.samples - p.B.samples

            ## Stop the experiment if either version appears better with > 95% probability.
            positive.A = .95 < mean(0 < delta.samples)
            positive.B = .95 < mean(delta.samples < 0)
            positive.result = positive.A || positive.B

            if (0 < optional.stops) {
                ## If optional stopping is enabled and we haven't reached a conclusion yet,
                ## collect another 100 samples for A & B and decide again if we should stop the expe
                for (i in 1:optional.stops) {
                    if (positive.result)
                        break

                    obs.A = obs.A + rbinom(1,100,p.A)
                    n.A = n.A + 100
                    obs.B = obs.B + rbinom(1,100,p.B)
                    n.B = n.B + 100

                    p.A.samples = rbeta(10000,obs.A+1,n.A-obs.A+1)
                    p.B.samples = rbeta(10000,obs.B+1,n.B-obs.B+1)
                    delta.samples = p.A.samples - p.B.samples

                    positive.A = .95 < mean(0 < delta.samples)
                    positive.B = .95 < mean(delta.samples < 0)
                    positive.result = positive.A || positive.B
                }
            }

            effect.size = mean(delta.samples)
            data = rbind(data,c(optional.stops,effect.size))

            if (positive.A)
                cnt.positives.A = cnt.positives.A + 1
            if (positive.B)
                cnt.positives.B = cnt.positives.B + 1
            if (positive.result)
                cnt.positives = cnt.positives + 1

        }
        cat(sprintf("%d optional stops: positive rate: %d/%d = %f, A: %d/%d = %f, B: %d/%d = %f\n",
                    optional.stops,
                    cnt.positives,n.trials,cnt.positives/n.trials,
                    cnt.positives.A,n.trials,cnt.positives.A/n.trials,
                    cnt.positives.B,n.trials,cnt.positives.B/n.trials))
    }

    colnames(data) = c('optional.stops','effect.size')
    data = as.data.frame(data)
    mean.effect.size.data = aggregate(effect.size ~ optional.stops, data = data, mean)
    cat(sprintf("mean effect size after %d trials\n", n.trials))
    print(mean.effect.size.data)
    g = ggplot(data,aes(x=effect.size)) + geom_histogram() + facet_grid(optional.stops ~ .) + ggtitl
    g = g + geom_vline(aes(xintercept = p.A-p.B), linetype='dashed', colour='blue')
    g = g + geom_vline(aes(xintercept = effect.size), linetype='dashed', colour='red', data = mean.e
    suppressMessages(print(g)) # silence the binwidth warnings
}
```

which produces output

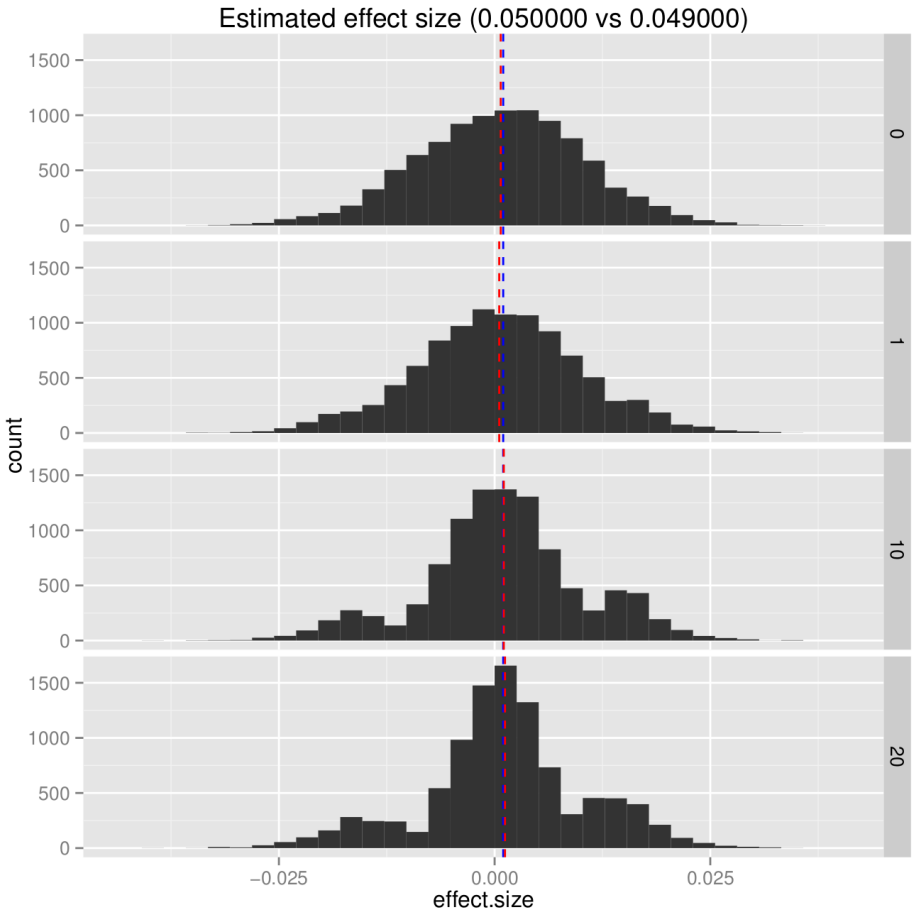
```
true difference exists: 0.050000 vs 0.049000
unbiased stopping rule
0 optional stops: positive rate: 969/10000 = 0.096900, A: 578/10000 = 0.057800, B: 391/10000 = 0.039
```

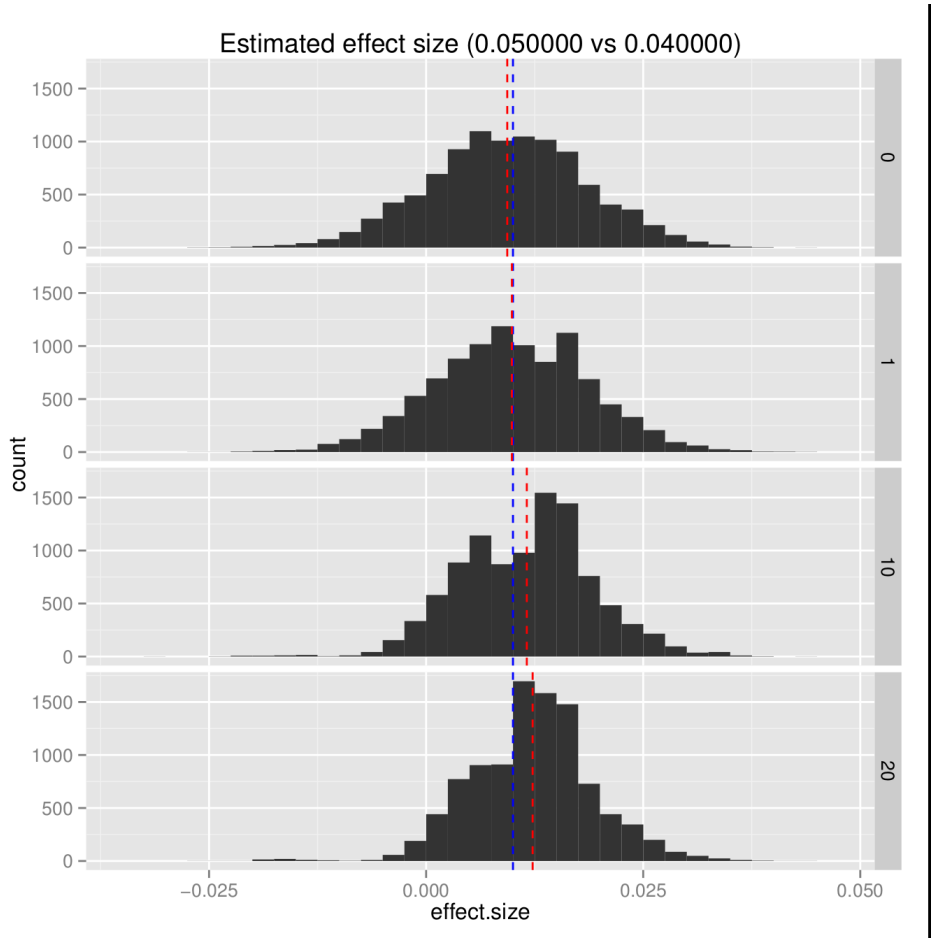
```
1 optional stops: positive rate: 1266/10000 = 0.126600, A: 735/10000 = 0.073500, B: 531/10000 = 0.05
10 optional stops: positive rate: 2417/10000 = 0.241700, A: 1468/10000 = 0.146800, B: 949/10000 = 0.
20 optional stops: positive rate: 3071/10000 = 0.307100, A: 1879/10000 = 0.187900, B: 1192/10000 = 0
mean effect size after 10000 trials
  optional.stops  effect.size
1              0 0.0007094838
2              1 0.0005368143
3             10 0.0010638554
4             20 0.0012055663


true difference exists: 0.050000 vs 0.040000
unbiased stopping rule
0 optional stops: positive rate: 2723/10000 = 0.272300, A: 2687/10000 = 0.268700, B: 36/10000 = 0.00
1 optional stops: positive rate: 3327/10000 = 0.332700, A: 3292/10000 = 0.329200, B: 35/10000 = 0.00
10 optional stops: positive rate: 5848/10000 = 0.584800, A: 5801/10000 = 0.580100, B: 47/10000 = 0.0
20 optional stops: positive rate: 7361/10000 = 0.736100, A: 7303/10000 = 0.730300, B: 58/10000 = 0.0
mean effect size after 10000 trials
  optional.stops effect.size
1              0 0.009344514
2              1 0.009873705
3             10 0.011595168
4             20 0.012267085


true difference exists: 0.050000 vs 0.030000
unbiased stopping rule
0 optional stops: positive rate: 7245/10000 = 0.724500, A: 7245/10000 = 0.724500, B: 0/10000 = 0.000
1 optional stops: positive rate: 7979/10000 = 0.797900, A: 7979/10000 = 0.797900, B: 0/10000 = 0.000
10 optional stops: positive rate: 9687/10000 = 0.968700, A: 9685/10000 = 0.968500, B: 2/10000 = 0.00
20 optional stops: positive rate: 9949/10000 = 0.994900, A: 9949/10000 = 0.994900, B: 0/10000 = 0.00
mean effect size after 10000 trials
  optional.stops effect.size
1              0  0.01934856
2              1  0.01972796
3             10  0.02039392
4             20  0.02050319
```

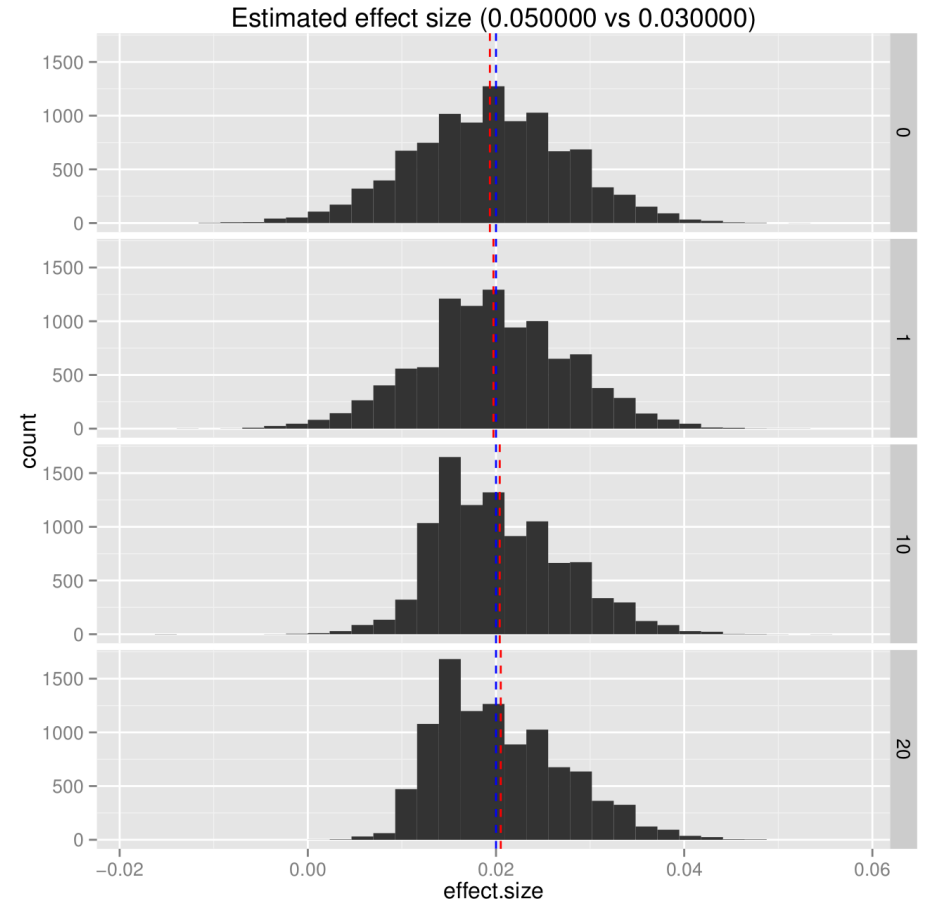
A few observations:

- Optional stopping (collecting more data if there isn't a conclusion) increases the correct positive rate (A) in all cases.
- As you said, the incorrect positive rate decreases as the true difference increases, and the incorrect positive rate doesn't seem as affected by optional stopping with higher differences.
- The effect size does appear to be overestimated with optional stopping. This makes sense, as the stopping rule prefers detecting an effect. (Graphs follow.)
- I didn't expect this, but the effect size appears underestimated without optional stopping. I'm not sure what's happening here - I'm still pondering.

Graphs of the mean effect size in all three cases follow. It appears the distribution shifts away from effects ~0 with optional stopping (speaking very loosely), which is roughly what I expected. The blue line indicates the true effect, and the red line indicates the mean effect size across all 10,000 trials.


Estimated effect size (0.050000 vs 0.049000)
count
effect.size
0
1
10
20
1500
1000
500
0
−0.025
0.000
0.025


Estimated effect size (0.050000 vs 0.040000)
count
effect.size
0
1
10
20
1500
1000
500
0
−0.025
0.000
0.025
0.050



CamDavidsonPilon commented on Sep 16, 2014 Owner

This is great, **@fraser**. Thank you for sharing this. Another influential post I've read was by Gelman, available here. This further explains your correct point

> The effect size does appear to be overestimated with optional stopping. This makes sense, as the stopping rule prefers detecting an effect.

As an interesting side note, if you work out the formula, you will see that the expected number of samples needed to detect a difference between A and B is proportional to:

$$\frac{p_A(1-p_A)+p_B(1-p_B)}{(p_A-p_B)^2}$$

Note the inverse square law here: the smaller the difference, orders of magnitude more data are required! Also, the closer *p_A* and *p_B* are to 0 or 1, the *less* data needed.

CamDavidsonPilon commented on Sep 16, 2014 Owner

You should write a blog post or article: people are really enjoy this stuff!

fraser commented on Sep 16, 2014

Thanks for the links to Gelman! That has further helped clarify things in my mind, especially around multiple comparisons. I like the focus on Type S and Type M errors over I & II.

I am still solidifying things for myself and expect to for a few days. Would you like me to leave this issue open for the moment? I am hopeful that I can come up with a concise addition to the book that touches on these points at a high level and can attach as a pull request.

Thanks for the suggestion about a blog post - I expect I'll put one together within about a month and will share a link when it's ready. :)

lemonlaug commented on Apr 29, 2015

Hello, very late to the discussion, but I found it extremely edifying and wanted to add another source which I found useful in sorting this all out myself:

It's a paper by Rouder, called "Optional stopping: No problem for Bayesians". I found it at:

http://pcl.missouri.edu/sites/default/files/Rouder-PBR-2014.pdf