To fill in the editor, I'll spell out my history with this paper. The author (Chris Hartgerink) first wrote it as a reply to our original paper in PLoS Biology, and I was a reviewer (or rather, I co-wrote a review with Megan Head, whom I shared an office with at the time, and who is first author on the Head et al paper – I am a co- author on the Head paper too). PLoS Biology rejected the reply, and at the time (i.e. >1.5 years ago) I told Hartgerink pretty much what I will say again in this review.

After this rejection, in May 2015, Hartgerink published an earlier version (essentially identical to this one in terms of data and methodology) of this manuscript on Authorea (https://www.authorea.com/users/2013/articles/31568/_show_article). Since the paper is very critical of our work and sort-of implies that we are either dishonest or incompetent, I posted some responses as comments to that article (click the little speech bubble on that site) in which I outlined again why Hartgerink's approach is incorrect.

The fact that Hartgerink is resubmitting this manuscript to another peer-reviewed journal suggests that he is not convinced by my arguments (or perhaps does not understand me), so in this review I have tried very hard to explain to him why I think his test is mistaken.

The crux of my argument is this: in Head et al, we noticed that many of the p-values we collected were rounded off to the nearest 2 decimal places, like p = 0.01, 0.02, 0.03 etc. See Hartgerink's Figure 1, which shows Head et al's data. We reasoned that researchers might be more or less likely to round off in different parts of the p-curve, which could add important biases to our analysis.

For example, one might reasonably assume that researchers are less likely to round off p values in the vicinity of the "magic threshold", p = 0.05, because of the importance attached to it by researchers and reviewers. For example, rounding p = 0.0451 up to p = 0.05 would be a bad idea for researchers, because they will fall afoul of reviewers that think their result is "only just" significant, when actually p = 0.0451 is (widely but wrongly) considered to be good evidence against the null. By contrast, many scientists would not think twice about rounding p = 0.0351 up to p = 0.04, because it looks significant either way. Thus, the rounding would be biased, and the p=0.04 spike would be bigger than the p=0.05 spike (as we observe).

Additionally, rounding p = 0.054 down to p = 0.05 might be seen as cheating (nobody wants to get caught out doing selective rounding, and reviewers would likely ask for more details for any results written as p = 0.05). Conversely it might be considered more acceptable to round p = 0.044 down to p = 0.04. Again this makes the p=0.04 spike bigger than the p=0.05 one, as observed.

For these 2 reasons and many others that one might dream up, it seems there is plenty of reason to assume that the p-values written to 2 decimal places are dodgy, and should not be used in the analysis.

To illustrate that biased rounding practices are expected to make it much harder to detect p-hacking when using both statistical tests that Hartgerink and Head et al used, I made a small simulation (the R code is provided at the end of my review). Basically the simulation does the following:

1. Simulate some fake p-values by drawing random numbers from an exponential distribution, in a fashion that produces a similar p-curve to the one that is shown in Hartgerink's Figure 1.
2. Round off a random subset of these p-values to 2 decimal places. Basically all the p-values <= 0.045 have a probability 'P1' of being rounded and a probability 1-P1 of being left un-rounded. The p-values >=0.045 have a probability 'P2' of being rounded and a probability 1-P2 of being left un-rounded. So, if P1 > P2, this means that we are modeling a scenario where researchers are less likely to round off p-values in vicinity of p=0.05, which is the region of relevance to Hartgerink and Head et al's tests for p-hacking.
3. Plot the p-curve, and calculate the ratio of the number of p-values in the two bins used in Hartgerink's test, namely 0.03875–0.04 and 0.04875–0.05 inclusive (this is shown in the 'ratio' part of my figure's titles). A big number here means that it is harder to detect p-hacking.

So in short, my simulation asks the question "If researchers round off p-values less often in the region near the significance threshold, is Hartgerink's method biased?" The answer is a resounding yes (plot on next page):

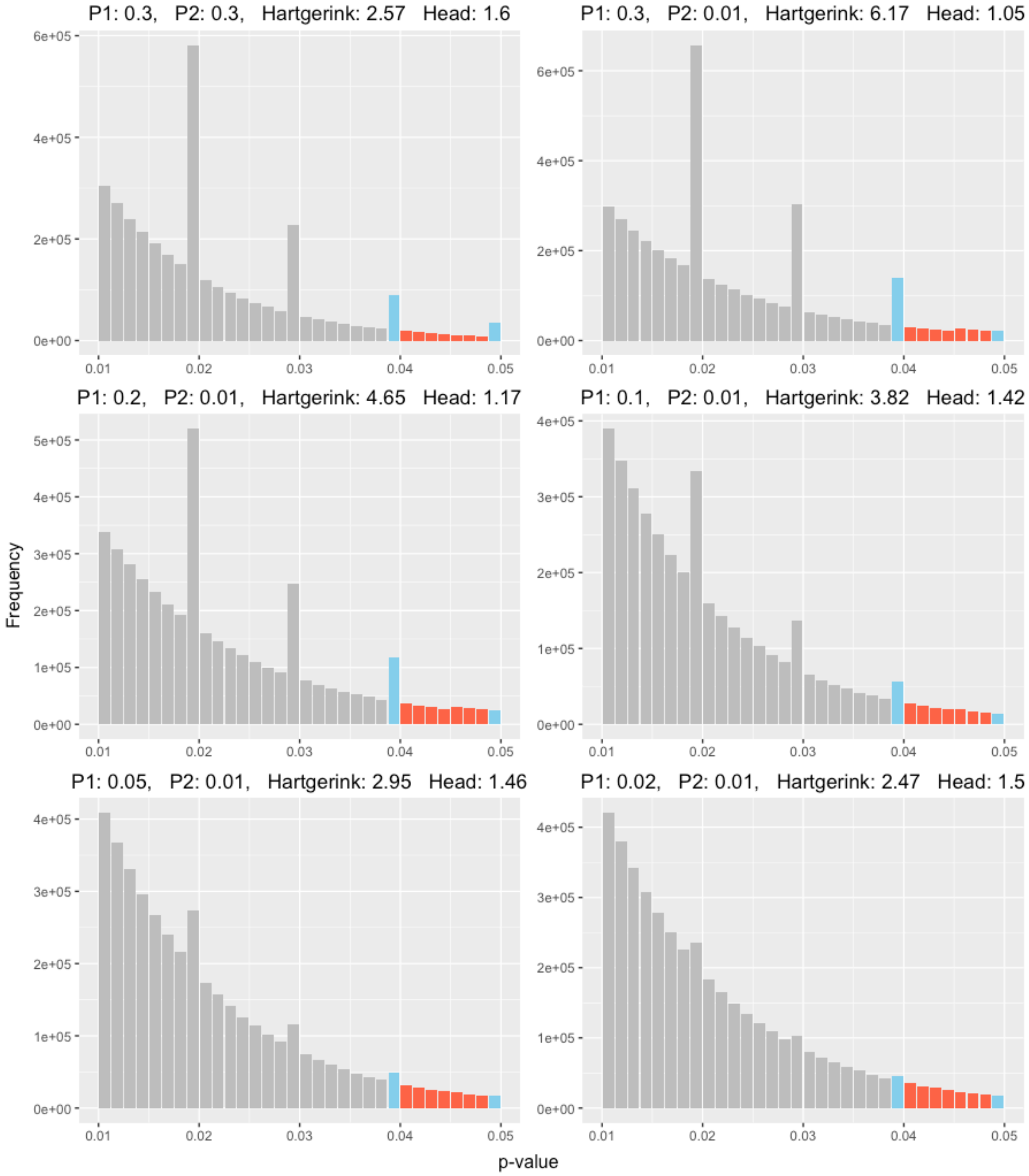

Figure legend: The titles show the values of P1 and P2 used when simulating the fake p-values. Large values of P1 and P2 mean that lots of the p-values get rounded. If P2 < P1, then researchers are less likely to round off p-values in the region p = 0.045 - 0.05 than elsewhere. The blue bins are the ones used in Hartgerink's preferred analysis, and the red zone is (roughly) the zone used in Head et al. The 'Hartgerink' and 'Head' numbers in the title give the number of p-values in the upper bin divided by the number in the lower bin. Higher numbers mean the test is less powerful to detect p-hacking, so low numbers are "good". For clarity, I have not plotted the region p = 0 – 0.01, which makes the plot zoom in on the interesting regions.

Here are the main results from my simulation:

1. Adding lots of rounding makes the dataset more spikey, obviously. Compare the plots with high and low values of P1 and P2.
2. When P2 < P1, the spikes are larger for p = 0.01, 0.02, 0.03 and 0.04 than they are for p = 0.05. This difference starts to get very pronounced when P2 is 10 to 20x smaller than P1 – I think this is a reasonable figure for the real world situation (though the difference between P1 and P2 is pretty much unmeasurable in reality).
3. When there is a big difference between P1 and P2, Hartgerink's measure is very strongly biased against finding p-hacking. Compare the top 2 graphs for example – when P1 = 0.3 and P2 = 0.01, the lower bin is 6 times bigger than the lower bin, so p-hacking would need to increase the size of the upper bin by greater than 6 times in order to stand a chance of being detected.
4. Hartgerink's method is overall less sensitive to p-hacking than Head's method. My simulation assumes the p-curve is right skewed, so the lower bin is naturally larger than the upper bin - this is what the 'Hartgerink' and 'Head' numbers in the title measure. For all assumptions, even P1=P2, Head's measure is more sensitive to detecting p-hacking.
5. The reason Hartgerink's method is less sensitive is because rounding piles even more extra values into p=0.04 than it does into p=0.05 when the p-curve is right skewed. Detecting p-hacking is hard because we are looking for left skew in the region near p=0.05, but this left skew gets overwhelmed by the overall right skew of the curve – we want a method that deals with this as well as possible, and it's clear that the Head method is more sensitive in this regard (though Head's test is very conservative too – this makes our conclusion of p-hacking 'robust', in my opinion).
6. A second reason Hartgerink's method is less sensitive is because the lower bin is substantially lower. The curve slopes up to the top left, and so the further you get from p=0.05, the more p-values there are. Using a really low lower bin makes it harder to detect any p-hacking. This is why Head et al. focused on the numbers between 0.04 and 0.05, rather than including lower ones as Hartgerink does.

So in conclusion, Hartgerink's test includes data that are probably tainted by rounding biases. We (Head et al) intentionally left these data out after discussing issues like the ones I raise here, and I think our main fault was not explaining our rationale at greater length in our paper. Hartgerink's reanalysis includes these tainted data, and it is unsurprising that his test does not detect p-hacking, because his test is set up to fail in multiple ways.

If Hartgerink disagrees with my logic here, I would urge him to engage with my criticism this time around. Previously he has seemed reluctant to debate my points (see my comment thread on his other PeerJ paper: https://peerj.com/preprints/1642/#feedback - I asked at least 3 times, and have still never received a reply to the comments on the PeerJ or Authorea manuscripts). Given that I have told him my objections so many times before, I am surprised (and unimpressed) that he did not try to pre-empt them in this revised manuscript. He also appears to have told his collaborator Marcel van Assen that he replied to me "more than once" (https://peerj.com/preprints/1642/#feedback), but if that's the case I did receive his replies - I am always keen to known if I am wrong, though in this instance I'm pretty sure that I'm not.

So assuming I'm correct: I think this reply adds nothing useful. The main failure of the Head et al paper was not spelling out precisely why we chose the analysis that we did, and I think that has led to the present confusion. Hartgerink does not explain why his analysis is better than Head et al's in a satisfactory way, and it's crucial to do that before this reply can be published.

Here are my specific comments on this new manuscript. In short, I think it's flawed even if the major problems were absent. In particular, the author would benefit from considering readers who have never read anything on p-hacking before.

Line 7: "from their analytic perspective" – not good English. More conventional phrasing: "according to their interpretation" or something.

Line 9-10: Explain what is "left skew p-hacking". I work on p-hacking and I do not recall seeing this term before (I can guess your meaning, but still). At any rate, it's bad practice to use undefined technical jargon right at the start of an abstract.

Line 11-12: The reason for the spikes is obviously rounding, as you are aware – maybe mention this in the abstract so it's clear to readers? Like,

> "Theoretically the distribution of p-values collected from the literature should be smooth, but since many researchers round off their p-values to two decimal places, the distribution shows large numbers of p-values in 'spikes' at 0.03, 0.04, 0.05, etc. To avoid various biases introduced by from confounding their statistical tests, Head et al. removed these rounded p-values".

This is less misleading than what you have written here, which might be taken to mean that we removed those p-values for no good reason (or because we were manipulating the data, i.e. p-hacking in a paper about p-hacking!)

Line 16-17: "when we take into account a second-decimal reporting tendency."

In what sense does your method "take them into account"? You simply include these obviously biased p-values, in an otherwise very similar statistical test to ours. In my opinion this line should say:

> "if I include values which are potentially biased because they have been rounded off (as shown by the spikes in the p-value distribution), then Head et al.'s result disappears."

The onus is the author to explain why one should include the values 0.04 and 0.05, when we have very good reason to suspect that these values are not to be trusted due to rounding errors.

Line 18-19: "Moreover, given the weight of the findings by Head et al. (2015b), it is important that these findings are robust to choices that can be debated" Again, this is not very clear English. Maybe you mean "Given that these results are important and have far-reaching implications…" when you say "weight". "Choices that can be debated" is also not the best phrasing

Abstract: Throughout, you totally ignore the findings from the non-text mining part of Head et al's paper, which also found evidence of p-hacking and are not challenged by your reanalysis (even if your reanalysis were correct, which it is not, in my opinion). Thus, your conclusion should be "I challenge one part of the Head et al study but not the other part, which also found evidence of p-hacking…"

Line 27-30: This line makes no sense. Also the University of Melbourne does not have a subscription to that psychology journal with the Simonsohn et al paper, so I can't figure out what you're trying to say here. The title of that paper implies it is a critique of a different paper, not Head et al, so check you cited it correctly, and maybe spell out what exactly the criticism is in the text instead of just alluding.

Intro: Again, you don't define what left skew p-hacking is. I assume you mean "types of p-hacking that are theoretically predicted to generate a left skewed bump under 0.05"

Intro:

"Their mining procedure included all reported p-values, including those that were reported without an accompanying test statistic. For example, the p-value from the result t(59) = 1.75, p > .05 was included, but also a lone p < .05."

Firstly, it's "text mining" not "mining".

Secondly and more importantly, this description of our methods is not correct. We only used p-values given exactly (i.e. with an equals sign, not less-than sign). This should be obvious to you; if the p value was written "p<0.05", how could you have plotted it in your Figure 1?

"Data analytic strategy" is not good English

"The binwidth of .005 and the 72 bins .04 < p < .045 and .045 < p < .05 were chosen by Head et al. (2015b) because they expected the 73 signal of this form of p-hacking to be strongest in this part of the distribution" Explain to the reader why the p-hacking should be most obvious in this part of the p-curve (hint: look at Figure 1 – you can see that the massive right skew would overwhelm the comparatively tiny right skew from the p-hacking in the region close to p=0). Incidentally, this is another reason your tests are less sensitive than ours. Your tests use data from regions of the p-curve that are closer to zero, which means the null expectation is that the lower bin is much bigger than the upper bin. Because our binomial tests (what you call Caliper tests) specify the conservative null of equal probability, using values below 0.04 as you do makes the tests less sensitive.

Line 91-95: You write:

"Moreover, the analytic strategy by Head et al. (2015b) eliminates p = .045 without justification and p = .05 based on a potentially invalid assumption of when researchers regard results as statistically significant. P = .045 is not included in the bins selected (.04 < p < .045 versus .045 < p < .05), while seriously affecting the results. If p = .045 is included, no evidence of a bump below .05 is found (the left black bin in Figure 1 is then included; frequency .04 < p ≤ .045 = 20114 versus .045 < p < .05 = 18132)."

Firstly, the reason we eliminated p=0.05 is because of the potential for rounding errors to bias our data (see my introduction and model). As you yourself point out, "For example, p = .041 might be correctly rounded to p = .04", which would explain the spikes in our data. People might perform rounding in a biased way – for example, they might be more likely to round down to p=0.04 than they are to round up to p=0.05. This is a clear reason not to include these potentially biased p-values in our statistical tests.

Secondly, we did not "eliminate p = .045 without justification". We eliminated p=.045 so that the test has symmetry – if we removed p=0.05 but left in p=0.045, then we give an "unfair advantage" to the lower bin (i.e. .04 < p ≤ .045) over the upper bin (i.e. 0.045 < p < 0.05). For this 'Caliper test' to make sense, it is imperative that the two bins are the same size; imagine if you tested p=0-0.04 against 0.4-0.5! We did not "provide justification" for why we left out 0.045 because this seems completely obvious to me.

Thirdly, you then make the mistake I just explained: your test compares the bin ".04 < p ≤ .045" with the bin ".045 < p < .05". You have purposely given more numbers to the lower bin while excluding 0.05 from the upper bin, so you have "set up the test to fail". Look at your Figure 1 – if you were to compare .04 < p ≤ .045 with .045 < p ≤.05, you'd find massive evidence for p-hacking. However that test would clearly be wrong too, because more people might round off to p=0.05 than p=0.045 (And they clearly do – look at your Figure). Thus, I think our original test remains the most logical choice.

Line 96-99: You write “Moreover, upon inspecting the original code to test for a bump below .05 (Head et al., 2015a), the inclusion or exclusion of the endpoints of the bins is not consistent. The endpoints are excluded when comparing 98 .04 < p < .045 versus .045 < p < .05, but the lower end is included when comparing .03 ≤ p < .04 99 versus .04 ≤ p < .05”

It’s not impossible there are typos in our analysis – I haven’t checked. Did you find that the typo you mentioned here changed the results? I would guess not. The latter test you mention (i.e. .03 ≤ p < .04 99 versus .04 ≤ p < .05) did not find any evidence for p-hacking, and I am certain that the results will be qualitatively identical if you fix the typo. Did you check?

Given that this possible typo has no bearing on our paper’s main results, maybe you should clearly say “This typo has no bearing on the results, I just wanted to point it out” – or just remove this passage? It incorrectly gives the reader the impression that we made a typo in the test for the main result of our paper.

Line 99: “P = .05 was consistently excluded because Head et al. (2015b) assumed researchers did not interpret this as statistically significant. Researchers interpret p = .05 as statistically significant more frequently than they thought: 94% of 236 cases investigated by Nuijten et al. (2015) interpreted p = .05 as statistically significant, indicating this assumption might not be valid.”

No – we excluded p = 0.05 because of the massive spikes and rounding errors shown in your Figure 1. I assume I have told you that many times before - check my Authorea and PeerJ comments. I do agree that an ideal test would include p=0.05, but we cannot include because of the rounding bias issue.

Line 103-104: Here you write “Given that systematically more p-values are reported to two decimal places and the disputable selection of the bins .04 < p < .045 versus .045 < p < .05, I did not exclude p = .045 and p = .05”

I would translate this as: “Even though it’s obvious that 0.04 and 0.05 are stuffed with rounded numbers, and we know that people almost certainly round off their numbers in a dubious way around p=0.05, I decided to include this tainted data anyway…”

This is the crux of why your re-analysis is wrong. You later say that “This altered bin selection takes such a reporting tendency into account and consequently includes the information available in these data.” I would argue that your test does not take anything into account – you just throw the tainted data into the test, and let them confound the results. If I am wrong, you need to explain what is ‘being taken into account’ at least once in your paper.

108-109: “the data show systematically more p-values reported to two decimal places, which might indicate a reporting tendency” What is a “reporting tendency”? Why might it be important? Are you just saying “lots of p-values were reported to 2 decimal places, which means people tend to report p-values to 2 decimal places”? Because that seems redundant. Spell out what you mean.

Line 111-114: Can we please stop calling it “the Caliper test”? The test is extremely simple and doesn’t need a fancy and confusing name. It is just a binomial test comparing the number of p-values in the two bins against the null hypothesis that they each have the same number of p-values in them. Note that our test is very conservative, because we expect the bin closest to zero to have more p-values in it because of the evidential value in the data (See your Figure 1). The odds are thus stacked against finding a significant result in a test for p-hacking. When you find one (as we did, in both our text mining and literature review), it implies p-hacking is pretty darn common and/or strong.

Additionally, there is little or no added value in your Bayesian binomial/Caliper test, as far as I can. You’re just testing the same data with two near-identical tests, right?

Line 127-128: You should set a different prior for the Bayesian test. As I just said (see my Figure and your Figure), we expect there to be more p-values in the lower bin (the one closest to zero) even if there is no p-hacking, so you are intentionally assuming a prior that we know is incorrect (i.e. that they are equal). A better prior would be to assume that the lower bin is a bit bigger.

Line 130-131: “when we take into account a second-decimal reporting tendency” Again, here what you actually mean is “when I choose to include rounded-off p-values that are probably biased.”

Best regards,

Luke Holman

P.S. here’s the R code for my simulation, with annotations. You should be able to just paste it into R and it will run (it takes about 8 seconds to simulate the data and make the graph – no big deal). If it fails, un-comment the first two lines and run them first to install those 2 packages. Email me if it doesn’t work.

```
# install.packages(ggplot2) # Run this line if you haven't
 already installed it
# install.packages(gridExtra) # Run this line if you haven't
 already installed it

library(ggplot2) # load this package that makes graphs
library(gridExtra)


############
# Global parameters:
```

```
#############

# Sample size for the simulation:
n <- 10000000

# Evidential value parameter - bigger values make the skew
 more extreme, as in a dataset where most studies were of real
 effects
# I chose a number that makes the curve look quite a lot like
 Figure 1
skew.parameter <- 0.2

#############
# The simulation:
#############

make.a.plot <- function(prob.rounding.outside.magic.range,
 prob.rounding.inside.magic.range){

  # Make some fake data with an exponential distribution that
 looks a lot like our p-curve
  df <- data.frame(x = rexp(n, rate = 1))
  df$x <- (df$x / max(df$x)) * skew.parameter
  df$rounded.x <- df$x

  # Make a version of the p-values with rounding applied. The
 simulation randomly rounds off some of the p-values arccording
 to the probabilities given above

  random.numbers1 <- rbinom(n, 1,
 prob.rounding.outside.magic.range) # roll numbers to see which
 ones get rounded
  random.numbers2 <- rbinom(n, 1,
 prob.rounding.inside.magic.range)

  df$rounded.x[random.numbers1 == 1 & df$x < 0.045] <-
 round(df$x[random.numbers1 == 1 & df$x < 0.045], 2)
  df$rounded.x[random.numbers2 == 1 & df$x >= 0.045] <-
 round(df$x[random.numbers2 == 1 & df$x >= 0.045], 2)

  # Set up the data to make the plot
  plotting.data <- data.frame(x =
 seq(0.000625,0.049375,length=40), counts =
 hist(df$rounded.x[df$rounded.x <=0.05], breaks =
 seq(0,0.05,length=41), plot=F)$counts)
  plotting.data$bar.colour <-  rep("a", 40)
  plotting.data$bar.colour[plotting.data$x %in% c(0.039375,
0.049375)] <- "b"
  plotting.data$bar.colour[33:39] <- "c"
  plotting.data$bar.colour <- factor(plotting.data$bar.colour)

  # The property "Hartgerink" is the ratio of the left and
right blue bins
  hartgerink.ratio <- plotting.data$counts[plotting.data$x ==
0.039375] / plotting.data$counts[plotting.data$x == 0.049375]

  # The property "Head" is the ratio of the left and right
blue bins
  head.ratio <- sum(df$rounded.x > 0.04 & df$rounded.x <
0.045) / sum(df$rounded.x > 0.045 & df$rounded.x < 0.05)

  title <- paste("P1: ", prob.rounding.outside.magic.range, ",
P2: ", prob.rounding.inside.magic.range, ",   Hartgerink: ",
round(hartgerink.ratio, 2), "   Head: ", round(head.ratio, 2),
sep="")


  ggplot(plotting.data[plotting.data$x > 0.01, ], aes(x=x,
y=counts)) + geom_bar(stat="identity", aes(fill = bar.colour))
+ xlab(NULL) + ylab(NULL) + ggtitle(title) +
theme(legend.position = "none") + scale_fill_manual(values =
c("grey", "skyblue", "tomato"))
}

#############
# Make the plot:
#############

p1 <- make.a.plot(prob.rounding.outside.magic.range = 0.3,
 prob.rounding.inside.magic.range = 0.3)
p2 <- make.a.plot(prob.rounding.outside.magic.range = 0.3,
 prob.rounding.inside.magic.range = 0.01)
p3 <- make.a.plot(prob.rounding.outside.magic.range = 0.2,
 prob.rounding.inside.magic.range = 0.01)
p4 <- make.a.plot(prob.rounding.outside.magic.range = 0.1,
 prob.rounding.inside.magic.range = 0.01)
p5 <- make.a.plot(prob.rounding.outside.magic.range = 0.05,
 prob.rounding.inside.magic.range = 0.01)
p6 <- make.a.plot(prob.rounding.outside.magic.range = 0.02,
 prob.rounding.inside.magic.range = 0.01)

grid.arrange(p1, p2, p3, p4, p5, p6, bottom = "p-value", left
 = "Frequency")
```