ALL-IN meta-analysis

Schure, J.A. ter

5 | Accumulation Bias: How to handle it as a Bayesian

Blog post

This chapter appeared as a blog post and gives more context to the claims in Chapter 3 on accumulation bias. These claims are paradoxical, after all: how can we possibly encounter enormous bias in our meta-analysis estimates and still do valid Bayesian inference? This blog post tries to give some intuition by introducing a very extreme and simple version of accumulation bias and showing by simulation code and plots in R what counteracts the bias in a Bayesian analysis.[1]

An estimated 85% of global health research investment is wasted (Chalmers and Glasziou, 2009); a total of one hundred billion US dollars in the year 2009 when it was estimated. The movement to reduce this research waste recommends that previous study results be taken into account when prioritizing, designing, and interpreting new research (Chalmers et al., 2014; Lund et al., 2016). Yet any recommendation to increase efficiency this way requires that researchers evaluate whether the studies already available are sufficient to complete the research effort; whether a new study is necessary or wasteful. These decisions are essentially stopping rules – or rather noisy accumulation processes, when no rules are enforced – and unaccounted for in standard meta-analysis. Hence reducing waste invalidates the assumptions underlying many typical statistical procedures.

Chapter 3 details all the possible ways in which the size of a study series up for meta-analysis, or the timing of the meta-analysis, might be driven by the results within those studies. Any such dependency introduces *accumulation bias*. Unfortunately, it is often impossible to fully characterize the processes at play in retrospective meta-analysis; the bias cannot be accounted for.

[1] The introduction to this blog post is the same as in Chapter 4 as they describe the same example accumulation bias but a different approach to counteracting it.

This is the second blog post about this type of bias and how to handle it. The first blog post (Chapter 4) detailed how it can be that ALL-IN meta-analysis handles accumulation bias. This second blog post deals with the Bayesian approach. We revisit the same example accumulation bias process, which can be one of many influencing a single meta-analysis, and use it to illustrate the following key points:

- Standard meta-analysis does not take into account that researchers decide on new studies based on other study results already available. These decisions introduce accumulation bias because the analysis assumes that the size of the study series is unrelated to the studies within; it essentially conditions on the number of studies available.
- A Bayesian analysis also conditions on the number of studies available, but can still handle accumulation bias well because it compares the biased sampling distribution under the null hypothesis to those under the alternative hypothesis.
- A Bayesian analysis can have error control under accumulation bias, but this crucially depends on the ratio of null and alternative hypotheses: the prior odds. No Bayesian analysis can handle accumulation bias when the prior odds cannot be specified.
- Specifying prior odds might be difficult for a meta-analysis in retrospect: if information from the study results included in the meta-analysis seeps into the prior odds, they become invalid.
- The e-values that follow from ALL-IN meta-analysis can also be combined with prior odds in a Bayesian analysis. They combine into pseudo-Bayes posterior odds that allow Bayesian error control. By using e-values rather than standard Bayes factors we can avoid specifying prior densities on the parameters within the null and the alternative; but prior odds on H_0 and H_1 are still needed and have to be trusted.
- If trustworthy prior odds can be specified, pseudo-Bayes posterior odds allow for continuous monitoring of the evidence as new studies arrive, even as new interim results arrive. Any decision to start, stop or expand studies is possible while keeping valid inference and Bayesian error control intact. Such decisions can be strategic: increasing the value of new studies, and reducing research waste.

5.1 Our example: extreme *Gold Rush* accumulation bias

We imagine a world in which a series of studies is meta-analyzed as soon as three studies become available. Many topics deserve a first initial study, but the research field is very selective with its replications. Nevertheless, for significant results in the right direction, a replication is warranted. We call this the *Gold Rush* scenario because after each finding of a positive significant result – the gold in science – some research group rushes into a replication, but as soon as a study disappoints, the research effort is terminated and no one bothers to ever try again. This scenario was first proposed by Ellis and Stewart (2009) and formulated in detail and under this name in Chapter 3. Here we consider the most extreme version of the *Gold Rush* where finding a significant positive result not

only makes replication more probable but even inevitable: the dependency of occurring replications on their predecessor's result is deterministic.

The first blog post gave a precise definition of this extreme *Gold Rush accumulation bias* and showed by simulation that the sampling distribution under the null hypothesis is affected by such a process or stopping rule. This is shown in Figure 5.1 for the fixed-effect meta-analysis $z^{(3)}$-scores for a three-study series. The theoretical sampling process, in the pink histogram, is centered around zero and the blue histogram, under accumulation bias process $A(t)$, does not behave like this theoretical distribution at all. It has a smaller variance and is shifted to the right – representing the bias. Here $A(3) = 1$ indicates that we accumulate and analyze 3 studies under the *Gold Rush* process. (For a precise definition, please refer to the blog post Accumulation Bias: How to handle it ALL-IN in Chapter 4.)

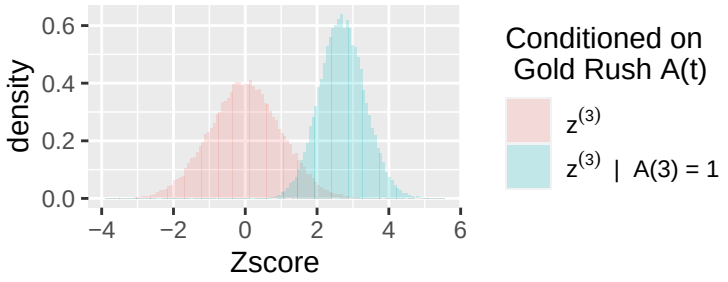

Figure 5.1. *Sampling distributions under the null hypothesis of fixed-effects meta-analysis Z-scores $Z^{(3)}$ of three studies with and without extreme* Gold Rush *accumulation bias $A(t)$, under the assumption of equal study sample size and variance.*

Bayesians claim that they can deal with any such stopping rules. So how can this be when the sampling distribution in Figure 5.1 is so much affected?

5.2 Likelihood ratios

We first turn our attention from the meta-analysis $Z^{(3)}$ statistic for three studies, to a likelihood ratio statistic $\mathbf{LR}^{(3)}$ for three studies. We summarize the results of individual studies into a single per-study Z-score (z_1 for the first study, z_2 for the second, etc), where we follow the same procedure that generated Figure 5.1, but calculate for each sample a likelihood ratio **LR** of two standard normal distributions, one with unit variance and mean 1 (ϕ_1) and one with unit variance and mean 0 (ϕ_0):

$$\mathbf{LR}^{(3)} = \frac{\phi_1(z_1, z_2, z_3)}{\phi_0(z_1, z_2, z_3)} = \prod_{i=1}^{3} \frac{\phi_1(z_i)}{\phi_0(z_i)}.$$

Assume that we are in the scenario that only true null effects are studied in our *Gold Rush* world, such that any new study builds on a false-positive result. How large would the

bias be in our likelihood ratio statistic $\mathbf{LR}^{(3)}$ if we analyze at the three-study series? We illustrate this by simulating this *Gold Rush* world using the R code below.

```
# numSim.study = number of simulated first studies
# you need 1/(0.025*0.025) = 1600 first studies for each series starting with two significant studies
# 50000 series, so 80 milion studies for smooth plot (takes ~4 minutes for simulation + plotting)
numSim.study <- 80000000

Z1 <- rnorm(numSim.study)
Z2 <- rnorm(numSim.study)
Z3 <- rnorm(numSim.study)

# selection based on Gold Rush accumulation bias A(3) = 1
A3 <- which((Z1 > 1.96) & (Z2 > 1.96))

calcLRmeta <- function(Zs) {
  prod(dnorm(Zs, mean = 1)/dnorm(Zs, mean = 0))
}

# meta LRscores for a random sample of 3-study series (you don't need all for a smooth plot)
LRmeta3 <- sapply(sample(1:numSim.study, size = 50000), function(i) calcLRmeta(c(Z1[i], Z2[i], Z3[i])))

# meta LRscores for a biased sample of 3-study series, biased by GoldRush A(3) = 1
LRmeta3.A3 <- sapply(A3, function(i) calcLRmeta(c(Z1[i], Z2[i], Z3[i])))

ggplot(rbind(data.frame(LR = LRmeta3,    GoldRush = ""),
             data.frame(LR = LRmeta3.A3, GoldRush = "A3"))) +
  geom_histogram(aes(x = LR,
                     y = ..density..,  # ..density normalizes by group with/without A(3) GoldRush
                     fill = GoldRush), # each with their own fill
                 bins = 120, position = "identity") +
  scale_x_continuous(trans = 'log10') +
  scale_fill_manual(values = hcl(15, 100, 65, alpha = c(0.8, 0.3)),
                    name = "Conditioned on \n Gold Rush A(t)",
                    labels = c(bquote(LR^(3)), bquote(LR^(3) ~ " |  A(3) = 1")))
```

Figure 5.2. Code to create Figure 5.3

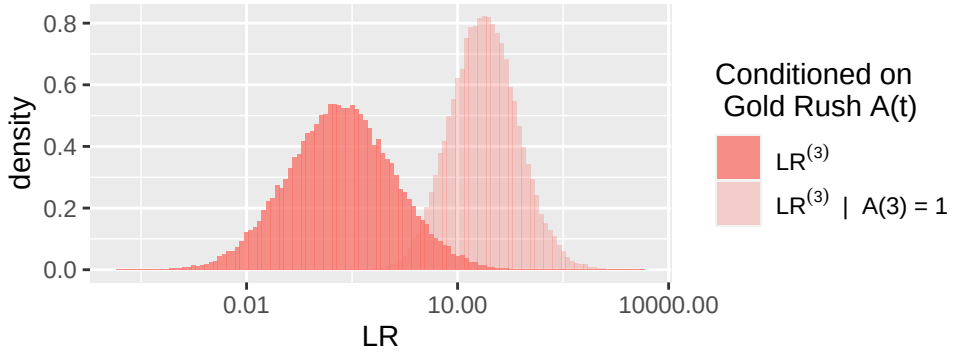

Figure 5.3. Sampling distributions under the null hypothesis of likelihood ratios $\mathbf{LR}^{(3)} = \prod_{i=1}^{3} \phi_1(Z_i)/\phi_0(Z_i)$ *of three studies with and without extreme* Gold Rush *accumulation bias* $A(t)$*. Note that the x-axis is on a log scale.*

Theoretical sampling process: A log-likelihood ratio of standard normal data has a normal sampling distribution. The R code in Figure 5.2 illustrates this sampling process:

First, a large population is simulated of possible first (`Z1`), second (`Z2`) and third (`Z3`) studies from a standard normal distribution. In the line of code for `LRmeta3`, each index `i` represents a possible study series, such that `c(Z1[i], Z2[i], Z3[i])` samples an unbiased study series and `calcLRmeta` calculates its likelihood ratio $\mathbf{LR}^{(3)}$. So the large number of Z-scores in `LRmeta3` captures the unbiased sampling distribution of the likelihood ratios.

***Gold Rush* sampling process:** In contrast, the code resulting in `A3` selects only those study series for which $A(3) = 1$ under extreme *Gold Rush* accumulation bias. So the large number of **LR**-scores in `LRmeta3.A3` capture a biased sampling distribution for $\mathbf{LR}^{(3)} \mid A(3) = 1$.

Likelihood ratios under *Gold Rush* accumulation bias: The final lines of code in Figure 5.2 plot two histograms of $\mathbf{LR}^{(3)}$ samples, one without and one with the *Gold Rush* $A(t)$ accumulation bias process, based on `LRmeta3` and `LRmeta3.A3` respectively. Each is given on the log-scale such that their normal sampling distributions become apparent. Figure 5.3 gives the result.

Here the likelihood ratio is just another statistic, with a sampling distribution that is affected by the *Gold Rush* decision making. The sampling distributions for $\mathbf{LR}^{(3)}$ on a log-scale (so $\log \mathbf{LR}^{(3)}$) in Figure 5.3 look very similar to those for $Z^{(3)}$ in Figure 5.1.

5.3 Two simple hypotheses

A Bayesian does not only care about the sampling distribution under the null hypothesis in Figure 5.3 but also about the sampling distribution under a competing alternative hypothesis. For simplicity, we first assume that we have two simple hypotheses, one representing the null (H_o) and one representing the alternative (H_1). Two simple hypothesis means that each can be represented by a single sampling distribution. We again summarize the results of individual studies into a single per-study Z-score (z_1 for the first study, z_2 for the second, etc). Under the null hypothesis, these Z-scores are generated by a normal distribution ϕ_0 with unit variance and mean 0; under the alternative hypothesis, these Z-scores are generated by an alternative distribution ϕ_1 with unit variance and mean 1.

The code in Figure 5.4 follows the same steps as the code in Figure 5.2 but it repeats each step for both both H_0 and H_1 in the `lapply` statements. We observe in Figure 5.5 that the same bias appears for the alternative hypothesis that we observe for the null hypothesis sampling distribution if we condition on arriving at our meta-analysis under extreme *Gold Rush* accumulation bias ($A(3) = 1$).

As a Bayesian, we simply do not care that our estimates are biased, as long as our posteriors are calibrated. We will first explain what calibration means for a Bayesian before we show that calibration stays intact under accumulation bias.

```
Hs <- c("H0" = "H0", "H1" = "H1")
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/10)  # sample ten times more H0 than H1
mean <- c("H0" = 0, "H1" = 1)

Z1 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z2 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z3 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))

# selection based on Gold Rush accumulation bias A(3) = 1
A3 <- lapply(Hs, function(H) which((Z1[[H]] > 1.96) & (Z2[[H]] > 1.96)))

# meta LRscores for a random sample of 3-study series (you don't need all for a smooth plot)
LRmeta3 <- lapply(Hs, function(H)
                  sapply(sample(1:numSim.study[H], size = numSim.study[H]/1000),
                         function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

# meta LRscores for a biased sample of 3-study series, biased by GoldRush A(3) = 1
LRmeta3.A3 <- lapply(Hs, function(H)
                     sapply(A3[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

ggplot(rbind(data.frame(LR = LRmeta3[["H0"]],    H0H1 = "H0", GoldRush = ""),
             data.frame(LR = LRmeta3[["H1"]],    H0H1 = "H1", GoldRush = ""),
             data.frame(LR = LRmeta3.A3[["H0"]], H0H1 = "H0", GoldRush = "A3"),
             data.frame(LR = LRmeta3.A3[["H1"]], H0H1 = "H1", GoldRush = "A3"))) +
  geom_histogram(aes(x = LR,
                     y = ..density.., # ..density.. normalizes by group H0 or H1 with/without A(3) GoldRush
                     fill = interaction(H0H1, GoldRush)), # each with their own fill
                 bins = 120, position = "identity") +
  scale_x_continuous(trans = 'log10') +
  scale_fill_manual(values = hcl(c(15, 195, 15, 195), 100, 65, alpha = c(0.8, 0.8, 0.3, 0.3)),
                    name = bquote(H[0] ~ "or" ~ H[1] ~ "\n Conditioned on \n Gold Rush A(t)"),
                    labels = c(bquote(H[0] ~ ": " ~ LR^(3)),
                               bquote(H[1] ~ ": " ~ LR^(3)),
                               bquote(H[0] ~ ": " ~ LR^(3) ~ " |  A(3) = 1"),
                               bquote(H[1] ~ ": " ~ LR^(3) ~ " |  A(3) = 1")))
```

Figure 5.4. *Code to create Figure 5.5*

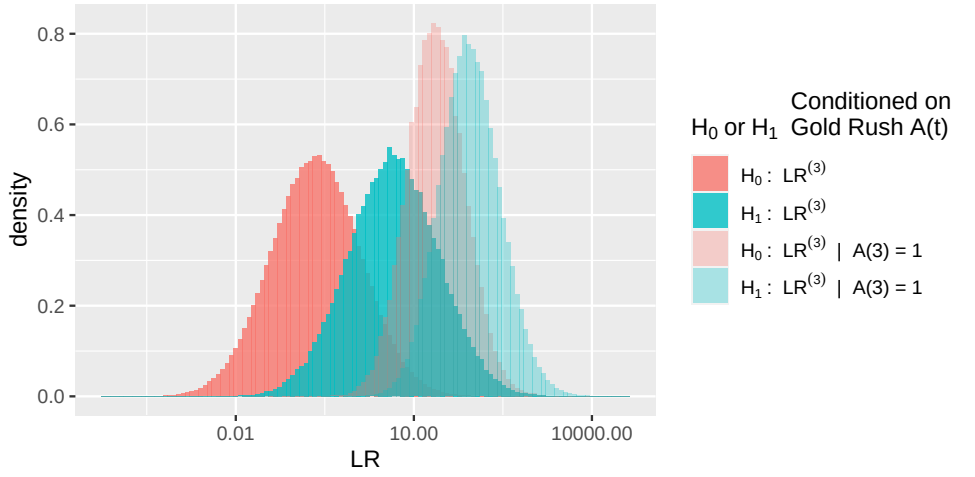

Figure 5.5. *Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of likelihood ratios* $\mathbf{LR}^{(3)} = \prod_{i=1}^{3} \phi_0(Z_i)/\phi_0(Z_i)$ *of three studies with and without extreme* Gold Rush *accumulation bias* $A(t)$. *Note that the x-axis is on a log scale.*

Bayesian calibration of posterior odds and Bayesian error control

No accumulation bias To introduce the notion of Bayesian calibration of the posterior odds and Bayesian error control, we first turn to a situation without accumulation bias. Here we consider the posterior odds, but our discussion is closely related to the literature on Bayes factor calibration (De Heide and Grünwald, 2021; Hendriksen et al., 2020). We obtain the posterior odds by multiplying the likelihood ratio $\mathbf{LR}^{(3)}$ (`LRmeta3`) with a prior odds $\pi(H_1)/\pi(H_0)$ as follows:

$$\frac{\pi(H_1 \mid z_1, z_2, z_3)}{\pi(H_0 \mid z_1, z_2, z_3)} = \frac{\mathbf{P}(z_1, z_2, z_3 \mid H_1) \cdot \pi(H_1)}{\mathbf{P}(z_1, z_2, z_3 \mid H_0) \cdot \pi(H_0)} = \frac{\phi_1(z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot \pi(H_0)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

```
gg.density <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                           data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..density.., # ..density.. normalizes by group H0 or H1
                     fill = H0H1),
                 alpha = 0.8, bins = 120, position = "identity", show.legend = c(fill = FALSE)) +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))))
gg.count <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                         data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..count.., # ..count.. does not normalize by group H0 or H1
                     fill = H0H1), alpha = 0.8, bins = 120, position = "identity") +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))) +
  scale_fill_discrete(name = "",
                      labels = c(bquote(H[0]), bquote(H[1])))
wrap_plots(list(gg.density, gg.count), ncol = 2, nrow = 1)
```

Figure 5.6. Code to create Figure 5.7

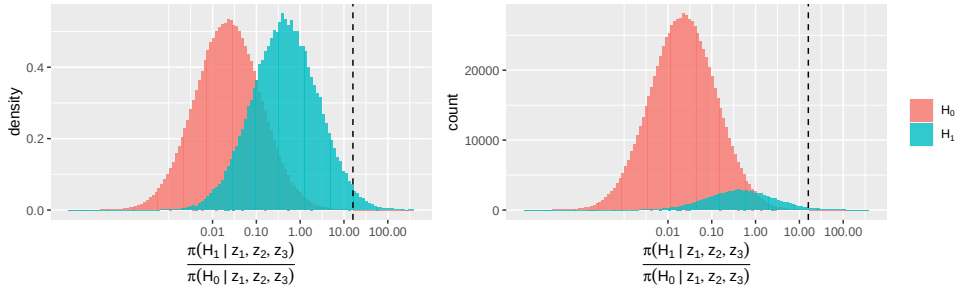

Figure 5.7. Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of $\frac{\pi(H_1 \mid z_1, z_2, z_3)}{\pi(H_0 \mid z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot (1/10)$ with accumulation bias. The vertical line indicates the threshold for the posterior odds at $r = 16$. Note that the x-axis is on a log scale.

We take the unbiased sample of likelihood ratios in `LRmeta3` from the code Figure 5.4 and obtain the posterior odds (`postOdds`) in the code in Figure 5.6 for each likelihood ratio by multiplication with a prior odds odds of (1/10):

$$\frac{\pi(H_1 \,|\, z_1, z_2, z_3)}{\pi(H_0 \,|\, z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = \mathbf{LR}^{(3)} \cdot (1/10).$$

The result of this code is given by Figure 5.7 in two histograms for our sampled posterior odds. One using the statement `..density..` and one using `..count...` The first normalizes the histogram bars such that they add up to one. This is the same plot as Figure 5.5, just with the x-axis scaled by (1/10) because we show posterior odds instead of the likelihood ratio. The second histogram does something different: it just counts the samples and so the histogram bars scale with the number of samples we take in the `sample` statement in calculating `LRmeta3` in Figure 5.4.

Bayesian calibration With calibration of the Bayes posterior odds we mean that if we sample from both H_1 (which is ϕ_1 in our example) and H_0 (which is ϕ_0 in our example) and look at a posterior odds with value o_{post}, observing this value for the posterior odds makes (H_1) our alternative hypothesis o_{post} times more probable than (H_0) our null hypothesis. In other words: the posterior odds of obtaining posterior odds of o_{post} are o_{post}.

$$\frac{\mathbf{P}\left(H_1 \,\middle|\, \frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}}\right)}{\mathbf{P}\left(H_0 \,\middle|\, \frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}}\right)} = o_{\text{post}}$$

$$\text{because} \quad \frac{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = o_{\text{post}}.$$

We can observe Bayesian calibration in the count plot in Figure 5.7, for example by looking at a posterior odds of 1.00 that has exactly the same count in both the histogram generated by H_1 and the one by H_0, which means that the ratio of counts is 1.00. This ratio of counts is calibrated, while the ratio of densities is not. The reason is that the ratio of densities does not take into account the prior odds: we take ten times as many samples from H_0 as from H_1 in the code in Figure 5.4. This agrees with the prior odds of (1/10) that we assume in calculating the posterior odds.

The ratio of densities gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \,|\, Z_1, Z_2, Z_3)}{\pi(H_0 \,|\, Z_1, Z_2, Z_3)} = o_{\text{post}} \,\middle|\, H_0\right)},$$

while the ratio of counts gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad = \quad o_{\text{post}} \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad = \quad o_{\text{post}} \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

Because we look at ratios, as we do if we look at odds, the scale of the counts in the figure does not matter. For simplicity, we are abusing notation a little bit and referring with probabilities **P** to densities, because our histograms of sampling distributions discretize our statistics in small intervals to give a probability instead of a density.

Bayesian error control From the calibration of the Bayes posterior odds we can obtain a notion of Bayesian error control as follows:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad \geq \quad r \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \quad \geq \quad r \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \quad \geq \quad r.$$

This Bayesian calibration is indeed the case for counts of $\mathbf{LR}^{(3)}$ (`LRmeta3`) in Figure 5.7 with the vertical dashed line $r = 16$. Figure 5.8 gives the calculation.

```
> sum(LRmeta3[["H1"]]*(1/10) > 16)/sum(LRmeta3[["H0"]]*(1/10) > 16)
[1] 31.59184
```

Figure 5.8. Code to show Bayesian error control for a threshold of $r = 16$ for the Bayes posterior odds.

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \geq r \,\middle|\, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \geq r \,\middle|\, H_0\right)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = 31.59 \geq r = 16.$$

If we use r as a threshold to decide that we believe H_1 is true and H_0 is false, the probability that we make an error is r smaller than the probability that we are right. In other words: if we use r as a threshold for the posterior odds, the odds for a correct decision are at least r.

5.4 Bayesian error control under extreme *Gold Rush* accumulation bias

We can make the same plots under our scenario of extreme *Gold Rush* accumulation bias and observe calibration. In the count plot in Figure 5.7, the posterior odds of 1.0, for example, has exactly the same count in the histogram for H_0 as it has for H_1, which means that the ratio of counts is 1.0.

```
gg.density <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                           data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..density.., # ..density.. normalizes by group H0 or H1
                     fill = H0H1),
                 alpha = 0.8, bins = 120, position = "identity", show.legend = c(fill = FALSE)) +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))))
gg.count <- ggplot(rbind(data.frame(postOdds = LRmeta3[["H0"]]*(1/10), H0H1 = "H0"),
                         data.frame(postOdds = LRmeta3[["H1"]]*(1/10), H0H1 = "H1"))) +
  geom_histogram(aes(x = postOdds,
                     y = ..count.., # ..count.. does not normalize by group H0 or H1
                     fill = H0H1), alpha = 0.8, bins = 120, position = "identity") +
  geom_vline(aes(xintercept = 16), linetype = "dashed") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(0.01, 0.1, 1, 10, 100),
                     name = bquote(frac(pi(H[1]~"|"~z[1], z[2], z[3]),
                                        pi(H[0]~"|"~z[1], z[2], z[3])))) +
  scale_fill_discrete(name = "",
                      labels = c(bquote(H[0]), bquote(H[1])))
wrap_plots(list(gg.density, gg.count), ncol = 2, nrow = 1)
```

Figure 5.9. *Code to create Figure 5.10*

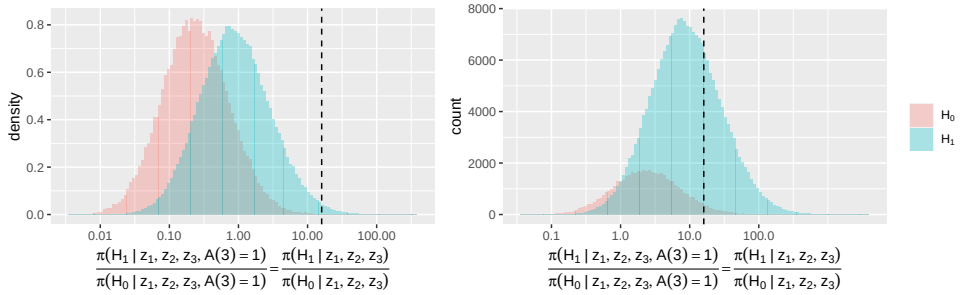

Figure 5.10. *Sampling distributions under the null (H_0) and alternative (H_1) hypothesis of* $\frac{\pi(H_1 \mid z_1,z_2,z_3)}{\pi(H_0 \mid z_1,z_2,z_3)} = \mathbf{LR}^{(3)} \cdot (1/10)$ *with accumulation bias. The vertical line indicates the threshold for the posterior odds at* $r = 16$*. Note that the x-axis is on a log scale.*

Now the ratio of densities gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_0\right)},$$

while the ratio of counts gives:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \quad = \quad o_{\text{post}} \,\middle|\, A(3)=1,H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

What counteracts the accumulation bias is illustrated in Figure 5.12. The posterior odds conditions on accumulating and analyzing three studies, $A(3) = 1$, which means that we are in a very biased sample. But because this situation occurs much more often under H_1 than under H_0

$$\mathbf{P}(A(3)=1 \mid H_1) >> \mathbf{P}(A(3)=1 \mid H_0),$$

our biased sample statistic $\mathbf{LR}^{(3)} \mid A(3) = 3$ can still achieve calibration if we take into account our prior odds.

In the ratio of counts, we also still have Bayesian error control under extreme *Gold Rush* accumulation bias:

```
> sum(LRmeta3.A3[["H1"]]*(1/10) > 16)/sum(LRmeta3.A3[["H0"]]*(1/10) > 16)
[1] 29.88396
```

Figure 5.11. *Code to show Bayesian error control for a threshold for the Bayes posterior odds of* $r = 16$.

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;\geq\; r \,\middle|\, A(3)=1,H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;\geq\; r \,\middle|\, A(3)=1,H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = 29.88 \geq r = 16.$$

We don't have to know the accumulation bias

How can it be that we never have to include anything about $A(3)$ in our calculations? The x-axis label of Figure 5.10 states that

$$\frac{\pi(H_1 \mid z_1,z_2,z_3,A(3)=1)}{\pi(H_0 \mid z_1,z_2,z_3,A(3)=1)} \quad = \quad \frac{\pi(H_1 \mid z_1,z_2,z_3)}{\pi(H_0 \mid z_1,z_2,z_3)},$$

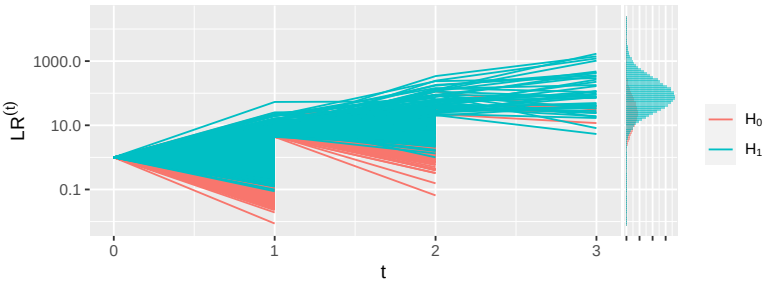

Figure 5.12. *Likelihood ratios* $\mathbf{LR}^{(1)}, \mathbf{LR}^{(2)}, \mathbf{LR}^{(3)}$ *when studies accumulate from 1 to 3 under the extreme* Gold Rush *accumulation bias process. Data simulated under prior odds* $H_1 : H_0 = 1 : 10$*. Note that the y-axis is logarithmic.*

```
numSim.study <- 5000  # these are not for the histogram, so a smaller simulation will do

Hs <- c("H0" = "H0", "H1" = "H1")
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/1)   # sample ten times more H0 than H1
mean <- c("H0" = 0, "H1" = 1)

Z1 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z2 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))
Z3 <- lapply(Hs, function(H) rnorm(numSim.study[H], mean = mean[H]))

A1 <- lapply(Hs, function(H) 1:numSim.study[[H]])
A2 <- lapply(Hs, function(H) which(Z1[[H]] > 1.96))
A3 <- lapply(Hs, function(H) c("A3" = which((Z1[[H]] > 1.96) & (Z2[[H]] > 1.96))))

LR1 <- lapply(Hs, function(H) sapply(A1[[H]], function(i) calcLRmeta(Z1[[H]][i])))
LRmeta2.A2 <- lapply(Hs, function(H) sapply(A2[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i]))))

# this is a smaller sample than LRmeta3.A3, which we use below for the histogram
LRmeta3.A3s <- lapply(Hs, function(H) sapply(A3[[H]], function(i) calcLRmeta(c(Z1[[H]][i], Z2[[H]][i], Z3[[H]][i]))))

ggplot(rbind(data.frame(t = 0, LR = 1,                    H0H1 = "H0", series = paste("H0", A1[["H0"]])),
             data.frame(t = 0, LR = 1,                    H0H1 = "H1", series = paste("H1", A1[["H1"]])),
             data.frame(t = 1, LR = LR1[["H0"]],          H0H1 = "H0", series = paste("H0", A1[["H0"]])),
             data.frame(t = 1, LR = LR1[["H1"]],          H0H1 = "H1", series = paste("H1", A1[["H1"]])),
             data.frame(t = 2, LR = LRmeta2.A2[["H0"]],   H0H1 = "H0", series = paste("H0", A2[["H0"]])),
             data.frame(t = 2, LR = LRmeta2.A2[["H1"]],   H0H1 = "H1", series = paste("H1", A2[["H1"]])),
             data.frame(t = 3, LR = LRmeta3.A3s[["H0"]],  H0H1 = "H0", series = paste("H0", A3[["H0"]])),
             data.frame(t = 3, LR = LRmeta3.A3s[["H1"]],  H0H1 = "H1", series = paste("H1", A3[["H1"]]))))) +
  geom_line(aes(x = t, y = LR, colour = H0H1, group = series)) +
  geom_ysidehistogram(aes(y = LR, x = ..count.., # ..count.. does not normalize by group H0 or H1
                          fill = H0H1),
                      data = rbind(data.frame(LR = LRmeta3.A3[["H0"]], H0H1 = "H0"),
                                   data.frame(LR = LRmeta3.A3[["H1"]], H0H1 = "H1")),
                      alpha = 0.8, bins = 100, position = "identity") +
  scale_y_continuous(trans = 'log10', name = bquote(LR^(t))) +
  scale_ysidex_continuous(labels = NULL) +
  scale_color_discrete(name = "", labels = c(bquote(H[0]), bquote(H[1])))
```

Figure 5.13. *Code to create Figure 5.12*

which follows because

$$\begin{aligned}
\frac{\pi(H_1 \,|\, z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \,|\, z_1, z_2, z_3, A(3) = 1)} &= \frac{\mathbf{P}(z_1, z_2, z_3, A(3) = 1 \,|\, H_1) \cdot \pi(H_1)}{\mathbf{P}(z_1, z_2, z_3, A(3) = 1 \,|\, H_0) \cdot \pi(H_0)} \\
&= \frac{\phi_1(z_1, z_2, z_3) \cdot A(3 \,|\, z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot A(3 \,|\, z_1, z_2, z_3) \cdot \pi(H_0)} \\
&= \frac{\phi_1(z_1, z_2, z_3) \cdot \pi(H_1)}{\phi_0(z_1, z_2, z_3) \cdot \pi(H_0)} \\
&= \frac{\pi(H_1 \,|\, z_1, z_2, z_3)}{\pi(H_0 \,|\, z_1, z_2, z_3)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}.
\end{aligned}$$

This is the reason that given the sample values z_1, z_2, z_3, the only calculations we performed were to get $\mathbf{LR}^{(3)}$ in the `LRmeta3.A3` statement in Figure 5.2. We obtained our posterior odds in Figure 5.11 by simply multiplying $\mathbf{LR}^{(3)}$ with the prior odds $(1/10)$.

Given that we know the data, the probability of accumulating our studies is the same under the null and the alternative hypothesis and drops out of the ratio. We do not need to know what these are to calculate our posterior odds. What matters is how often we are in the null and the alternative situation: our prior odds. This is known as *stopping rule independence* (Hendriksen et al., 2020; Berger and Berry, 1988). If we know our prior odds, we do not need to know the accumulation bias process under which our study results z_1, z_2, z_3 were obtained (the probability $A(3 \mid z_1, z_2, z_3)$, see Chapter 3). We can just calculate our posterior odds in the usual way and decide on a threshold r on that posterior odds.

Prior odds

How often we reach $A(3) = 1$ under H_0 in comparison to under H_1 needs a statement of the relative occurrences of H_0 and H_1: a prior odds $\pi(H_1)/\pi(H_0)$. In the code in Figure 5.4 and Figure 5.13 we sample ten times as many null effects as alternative effects, so we assume that for every clinical trial that studies an effective treatment $\pi(H_1)$, we have $\pi(H_0)/\pi(H_1) = 10$ clinical trials that study an ineffective treatment, so $\pi(H_1)/\pi(H_0) = 1/10$. In Figure 5.12 we show that even if ten times as many studies observe data from the null hypothesis, still a lot more from the alternative hypothesis make it to a three-study-series under extreme *Gold Rush* accumulation bias.

5.5 The prior odds are crucial

The Bayesian calibration is driven by the fact that accumulation bias processes like the extreme *Gold Rush* make it much more likely for study series generated by H_1 to reach the meta-analysis than for study series generated by H_0. How much more depends on how many times either of them can try. As a Bayesian meta-analyst, we can think of this as a property of the research field that we might know and include in the analysis. How many initial studies are measuring a true effect from H_1 for each one that measures a null effect from H_0?

Bayesian calibration does rely crucially on getting the prior odds right. If we set our prior odds to a default $1:1$ and there is extreme *Gold Rush* accumulation bias in our field, we are actually assuming that hardly any series of clinical trials studying a null effect will accumulate three studies. Figure 5.14 shows what we are assuming in this case. For these plots we have set the following in the code in Figure 5.4 and Figure 5.13

```
numSim.study <- c("H0" = numSim.study, "H1" = numSim.study/1)
```

and we calculate the posterior odds based on our assumed $1:1$ prior odds:

$$\frac{\pi(H_1 \mid z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \mid z_1, z_2, z_3, A(3) = 1)} = \mathbf{LR}^{(3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = \mathbf{LR}^{(3)} \cdot \frac{1}{1} = \mathbf{LR}^{(3)}. \tag{5.1}$$

Figure 5.14 shows that assuming 1:1 prior odds in the calculations based on our data,

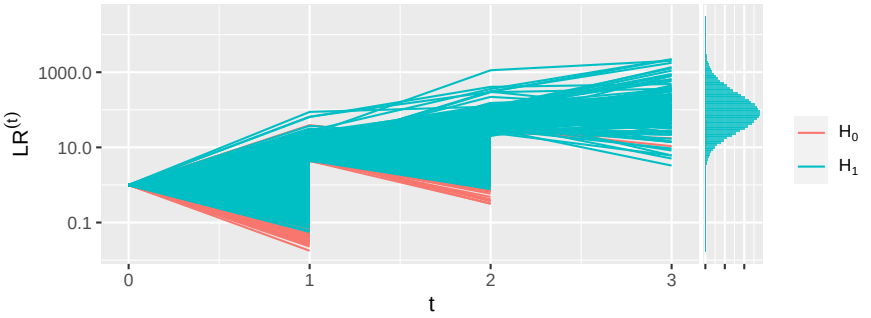

Figure 5.14. *Likelihood ratios* $\mathbf{LR}^{(1)}, \mathbf{LR}^{(2)}, \mathbf{LR}^{(3)}$ *when studies accumulate from 1 to 3 under the extreme* Gold Rush *accumulation bias process. Data simulated under prior odds* $H_1 : H_0 = 1 : 1$. *Note that the y-axis is logarithmic.*

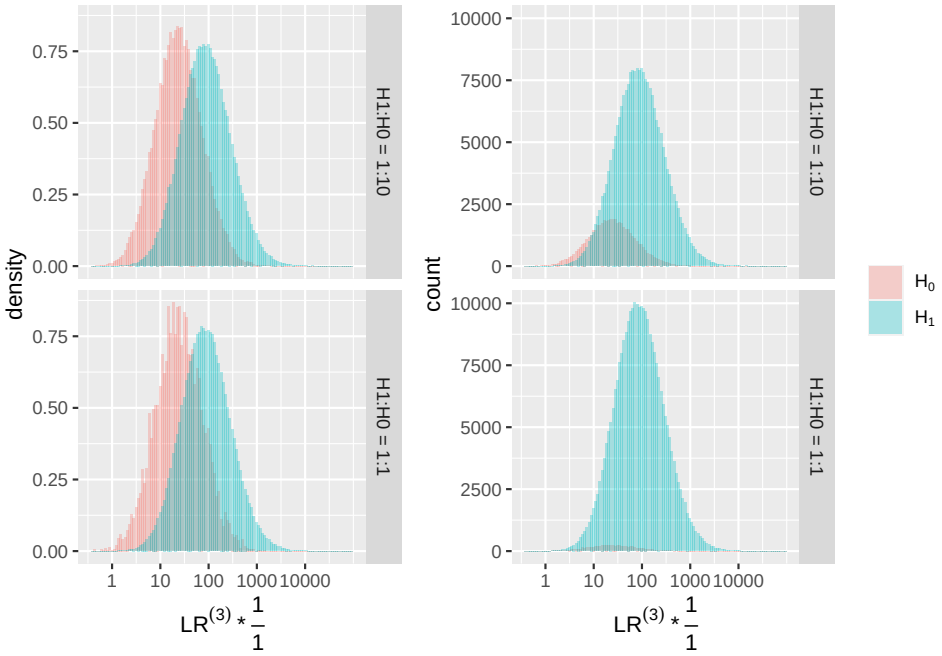

Figure 5.15. *Sampling distributions under the null* (H_0) *and alternative* (H_1) *hypothesis of* $\mathbf{LR}^{(3)} \cdot (1/1)$ *under extreme* Gold Rush *accumulation bias. The upper panels are sampled using* $H_1 : H_0 = 1 : 10$ *and the lower panels using* $H_1 : H_0 = 1 : 1$. *The upper right panel shows that mistakenly assuming* $1 : 1$ *in the posterior odds* $\mathbf{LR}^{(3)} \cdot (1/1)$ *does not give calibration under* $H_1 : H_0 = 1 : 10$, *e.g.* $\mathbf{LR}^{(3)} = 10$ *does not happen ten times as often under* H_0 *than under* H_1. *Note that the x-axis is on a log scale.*

when the true number of null hypotheses studied is ten times larger – the ratio in our research field is 1:10 – breaks the calibration of our posterior odds. For example, for an incorrectly calculated posterior odds $\mathbf{LR}^{(3)} \cdot (1/1) = 10$, we are in a situation that happens just as often under the null as under the alternative (the top-right panel of Figure 5.15) which should not give much evidence in favor of the alternative. As a result, also Bayesian error control breaks.

Setting prior odds is not that easy

We do want to stress that in the setting of retrospective meta-analysis, where the results of individual trials can be known to the meta-analyst before performing the analysis, it might be very difficult to establish prior odds that are not influenced by the data. In such scenarios, relying upon field-specific priors, e.g. established by prediction markets involving many peers (Potthoff, 2007; Dreber et al., 2015), might achieve more reliable prior odds.

What is more, these prior odds need to represent the ratio of alternative to null *initial studies* and not the ratio in meta-analyses. Reaching enough studies – e.g. $t = 3$ under extreme *Gold Rush* – and doing the meta-analysis is part of the data in the likelihood, not part of the prior. We encounter trouble with Bayesian calibration when we use different priors for individual studies than we use for a meta-analysis.

Doing so is appealing, though, since meta-analyses seem to be wrong less often than individual studies. The famous paper "Why Most Published Research Findings Are False" (Ioannidis, 2005b), for example, specifies different prior odds for a clinical trial analysis in comparison to a meta-analysis of clinical trials. This was the paper that introduced the concept of field-specific prior odds to a large audience as "Ratio of True to Not-True Relationships (R)". The different types of prior odds include one for "Adequately powered RCT with little bias" and one for "Confirmatory meta-analysis of good-quality RCTs". The first is set to an R of 1:1 and the second R to 2:1. This means that information seeped into the prior odds about what type of RCTs end up in meta-analyses; getting to the meta-analysis stage is assumed to be more likely under the alternative than the null otherwise the two prior odds would be the same. The meta-analysis prior odds that Ioannidis (2005b) specifies are essentially $\frac{\mathbf{P}(A(t)=1\,|\,H_1)}{\mathbf{P}(A(t)=1\,|\,H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$ with $\mathbf{P}(A(1) = 1\,|\,H_1) = 1$ and $\mathbf{P}(A(1) = 1\,|\,H_0) = 1$ for primary studies. This invalidates the stopping rule principle. Including information about the accumulation process into the prior biases the Bayes posterior and requires that the same information is included in the likelihood as well for Bayesian calibration. We need to know the accumulation bias process $A(t)$ in that case, which is usually impossible.

5.6 Beyond simple hypotheses

The situation with two simple hypotheses that we discussed so far, where each trial is either collecting data from ϕ_0 or ϕ_1, is not very realistic. More generally, we like to vary the parameter μ of our normal distribution ϕ_μ and allow for all possible normal distributions.

We assume we are given a *minimum relevant effect size* μ_{min} as well as a $\mu_0 < \mu_{\text{min}}$, which

respectively define the alternative hypothesis H_1 and the null hypothesis H_0:

$$H_0 = \{\phi_\mu : \mu \leq \mu_0\}, \quad H_1 = \{\phi_\mu : \mu \geq \mu_{\min}\}.$$

We can distinguish two types of prior probabilities: $\pi(H_1)$ and $\pi(H_0)$ for the hypotheses H_1, H_0, and $\pi_1(\mu)$ and $\pi_0(\mu)$ for $\{\mu : \mu \geq \mu_{\min}\}$ and $\{\mu : \mu \leq \mu_0\}$ respectively. Instead of a likelihood ratio of two simple hypotheses, we specify a Bayes Factor of two Bayes marginal distributions, using the priors on μ:

$$\mathbf{BF}(z_1, \ldots, z_t) = \frac{\bar{\phi}_1(z_1, \ldots, z_t)}{\bar{\phi}_0(z_1, \ldots, z_t)},$$

$$\text{with} \qquad \bar{\phi}_1(z) = \int \phi_\mu(z)\pi_1(\mu)dz \quad \text{and} \quad \bar{\phi}_0(z) = \int \phi_\mu(z)\pi_0(\mu)dz;$$

$$\bar{\phi}_1(z_1, \ldots, z_t) = \prod_{i=1}^{t} \bar{\phi}_1(z) \quad \text{and} \quad \bar{\phi}_0(z_1, \ldots, z_t) = \prod_{i=1}^{t} \bar{\phi}_0(z).$$

If π_j puts all its mass on a particular element μ^*, then $\bar{\phi}_j(z) = \phi_{\mu^*}(z)$.

Combining the Bayes Factor with the prior odds gives us the posterior odds, that just like the earlier posterior odds for two simple hypotheses, does not depend on the accumulation bias process for reaching e.g. $A(3) = 3$.

$$\frac{\pi(H_1 \mid z_1, z_2, z_3, A(3) = 1)}{\pi(H_0 \mid z_1, z_2, z_3, A(3) = 1)} = \frac{\bar{\phi}_1(z_1, z_2, z_3) \cdot A(3 \mid z_1, z_2, z_3)}{\bar{\phi}_0(z_1, z_2, z_3) \cdot A(3 \mid z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{5.2}$$

$$= \frac{\bar{\phi}_1(z_1, z_2, z_3)}{\bar{\phi}_0(z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)} \tag{5.3}$$

$$= \mathbf{BF}(z_1, z_2, z_3) \cdot \frac{\pi(H_1)}{\pi(H_0)}. \tag{5.4}$$

The pseudo-Bayes posterior odds

What if we cannot come up with a good prior on the μs? In that case we may want to 'represent' the set of distributions H_0 and H_1 by their 'least extreme elements' respectively, i.e. ϕ_{μ_0} and $\phi_{\mu_{\min}}$. This gives the *pseudo-Bayes posterior odds*,

$$\frac{\pi^{\text{ps}}(H_1 \mid z_1, z_2, z_3)}{\pi^{\text{ps}}(H_0 \mid z_1, z_2, z_3)} = \frac{\phi_{\mu_{\min}}(z_1, z_2, z_3)}{\phi_{\mu_0}(z_1, z_2, z_3)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

$$= \text{BF}^{\text{ps}}(z_1, z_2, z_3) \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

which is just the 'real' posterior odds (5.2) that we would get if we had put all our prior mass on $\mu_{\min}$ and μ_0 respectively.

We can use the 'pseudo-Bayes posterior odds' when (with a Bayesian mindset) we have no good idea about 'good' priors on the μs or (with a frequentist mindset) about what

value of μ may be true if H_1 is true, or what value of μ may be true if H_0 is true. Note in particular that in the pseudo-Bayes posterior odds, we use the *same* priors on H_0 and H_1 as in the 'real' posterior, but different, degenerate priors on the μs.

The GROW e-values that we calculate in ALL-IN meta-analysis (Chapter 1) are pseudo-Bayes factors BF^{ps}. So in ALL-IN meta-analysis, we can very simply extend our conclusions with Bayesian statements by combining our e-values with prior odds to obtain pseudo-Bayes posterior odds. Moreover, with these pseudo-Bayes posterior odds, we can also obtain Bayesian error control.

5.7 Pseudo-Bayesian error control

Throughout this blog post we have shown that if we get the prior odds right, the posterior odds is calibrated under accumulation bias, i.e.:

$$\frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3,A(3)=1)}{\pi(H_0 \mid Z_1,Z_2,Z_3,A(3)=1)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)}$$

$$= \frac{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_1\right)}{\mathbf{P}\left(\frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}} \,\middle|\, A(3)=1, H_0\right)} \cdot \frac{\mathbf{P}(A(3)=1 \mid H_1)}{\mathbf{P}(A(3)=1 \mid H_0)} \cdot \frac{\pi(H_1)}{\pi(H_0)} = o_{\text{post}}$$

$$\text{such that} \qquad \frac{\mathbf{P}\left(H_1 \,\middle|\, \frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}}, \quad A(3)=1\right)}{\mathbf{P}\left(H_0 \,\middle|\, \frac{\pi(H_1 \mid Z_1,Z_2,Z_3)}{\pi(H_0 \mid Z_1,Z_2,Z_3)} \;=\; o_{\text{post}}, \quad A(3)=1\right)} = o_{\text{post}}.$$

The outer probabilities combine a prior odds with a likelihood ratio of the observing the posterior odds of o_{post}, which is a statistic of our data, with the likelihood ratio of observing a three study series ($A(3)=1$), also part of the data. The likelihood ratio of the data combined with prior odds forms posterior odds for the hypotheses conditioned on the data.

We can use this fact of calibration to specify the Bayesian error control further for the pseudo-Bayes posterior odds. We define a threshold on the pseudo-Bayes posterior odds r and decide to reject the null hypothesis and believe the alternative if

$$\frac{\pi^{\text{ps}}(H_1 \mid z_1,\ldots,z_t)}{\pi^{\text{ps}}(H_0 \mid z_1,\ldots,z_t)} \geq r.$$

We can set a threshold such that if we cross it, we reject the null hypothesis and denote so by $\text{REJECT}[A(t)=1, r]$ based on crossing the threshold with our pseudo-Bayes posterior odds conditioned on accumulating t studies.

For a subset of all accumulation bias processes $A(t)$ which includes the extreme *Gold rush* and variations of it, we have the following: for all $t = 1,2,\ldots$ $r > 1$: the *true* Bayes

posterior odds – so not only the pseudo-Bayes posterior odds! – of H_0 satisfies:

$$\frac{\pi(H_1 \mid \text{REJECT}[A(t) = 1, r])}{\pi(H_0 \mid \text{REJECT}[A(t) = 1, r])} \geq r. \tag{5.5}$$

This expresses that, as long as the priors on H_0 and H_1 are chosen correctly, we have Bayesian error control for the pseudo-Bayes posterior odds: a Bayesian's real posterior odds of an incorrect decision can be no larger than the odds to make an error according to the pseudo-Bayes posterior odds on $\{H_0, H_1\}$, even though that the priors on the μs are, according to that same Bayesian, incorrect. Note that this is merely a 'one-sided' calibration, but the inequalities go the right (i.e. practically useful) way. The result holds not just for the normal location family but for a general class of models including all 1-dimensional exponential families and the 1-sample t-test setting. This general result is stated and proved in Appendix Section 5.A and Section 5.B.

5.8 Conclusion

In our imaginary world of extreme *Gold Rush* accumulation bias, the sampling distribution of the meta-analysis Z-score behaves very different from the sampling distribution assumed to calculate p-values and confidence intervals. A meta-analysis sampling distribution conditions on the available number of studies, which means that we are in a situation that is influenced by a selection effect: only some series get there, not others. Bayesian analysis also conditions on the number of studies and therefore also likelihood ratios and Bayes factors have biased sampling distributions. For Bayesian error control, however, we do not need unbiased sampling; we need *calibration*.

Some accumulation bias is at play in almost any retrospective meta-analysis. Bayesian calibration holds no matter the accumulation bias process, as long as the prior odds are trustworthy. The e-values from ALL-IN meta-analysis can also be used to combine with prior odds and obtain pseudo-Bayes posterior odds to obtain Bayesian error control through calibration, although now calibration may only hold for a subset of all accumulation bias processes – which however include the extreme *Gold Rush* scenario. This also allows for continuous monitoring; multiple testing is no problem, as long as the prior odds are correct. Setting trustworthy prior odds in meta-analysis is not easy, however. As long as the prior odds can be trusted, a Bayesian perspective on meta-analysis will reduce research waste by allowing efficient data-driven decisions – not letting them invalidate the inference – and still analyze the posterior odds for any particular meta-analysis, conditioned on arriving at the number of studies so far.

Code availability

This blogpost's R code is available on `https://osf.io/p2rtw/` (Ter Schure, 2021a).

Appendices

5.A Pseudo-Bayes posterior odds for exponential families and beyond

5.A.1 Exponential families

Let $\mathscr{M} = \{P_\delta : \delta \in \Delta\}$ with Δ a (possibly unbounded) open interval in $\mathbb{R}$ represent a 1-dimensional exponential family of probability distributions for some random variable Y, given in its mean-value parameterization. While this already includes important models such as the normal location family (z-test), the Bernoulli model, the Poisson model, and so on, we will later, in Theorem 5.B.1, extend our result to some multi-dimensional families that are not of exponential form. Each P_δ has a density (for continuous-valued Y) or mass function (for discrete-valued Y) p_δ. P_δ and p_δ are extended to i.i.d. sequences by taking product distributions.

We assume we are given two parameter values δ^- and δ^+ in Δ with $\delta^- < \delta^+$ which respectively define the *alternative* hypothesis H_1 and the null hypothesis H_0:

$$H_0 = \{P_\delta : \delta \in \Delta_0\},\ \Delta_0 = \{\delta \in \Delta : \delta \leq \delta^-\},\ H_1 = \{P_\delta : \delta \in \Delta_1\},\ \Delta_1 = \{\delta \in \Delta : \delta \geq \delta^+\}.$$

In the case treated in the main text, $\mathscr{M}$ denotes the normal location family, $\delta^+ = \mu_{\min}$ is the minimum clinically relevant effect size, and $\delta^- = \mu_0$. The fact that the treatment below is entirely symmetric in δ^+ and δ^- (if we swap δ^+ and δ^- and take the reciprocal of all Bayes factors and posterior odds, we get the same result) motivates the switch of notation. Still, in practice we will often have δ^+ interpretable as minimal effect size, $\Delta = \mathbb{R}_0^+$ and $\delta^- = 0$. We need the concept of a *stopping time*. We define this in a standard way in terms of a *randomized stopping rule*. This is any function f from outcome sequences of arbitrary length to $[0, 1]$. The interpretation is that for any actually generated initial sequence of data $y^n = y_1, y_2, \ldots, y_n$, we toss an independent coin with bias $f(y^n)$ after having observed y^n. If the coin lands tails, we stop. If not, we generate y_{n+1} and repeat the procedure for $n + 1$, etc. The stopping time τ is then the random variable set equal to the smallest n at which the coin has landed tails. *Gold Rush* accumulation bias (Chapter 3) defines such a stopping rule. The extreme version of *Gold Rush* accumulation bias presented in this blog post is a non-randomized version of this stopping rule.

5.A.2 The Bayes and the pseudo-Bayes posterior

We can distinguish two types of prior probabilities: $\pi(H_1)$ and $\pi(H_0)$ for the hypotheses H_0, H_1, and Π_1 and Π_0 for the parameter spaces $\Delta_1 = \{\delta : \delta \geq \delta^+\}$ and $\Delta_0 = \{\delta : \delta \leq \delta^-\}$. Π_j can be interpreted as the prior of δ conditioned on it lying in Δ_j; in general, we will allow priors that do not restrict δ to lie in Δ_j (i.e. we may have $\Pi(\delta \notin \Delta_0 \cup \Delta_1) = \pi' > 0$ and then $\pi(H_1) + \pi(H_0) + \pi' = 1$). We can calculate conditional and posterior probabilities and odds in the standard way using Bayes' theorem: illustrating a particular case we need later on, for general measurable events $\mathscr{E}_1$,

$$\frac{\pi(H_1 \mid \mathscr{E}_1)}{\pi(H_0 \mid \mathscr{E}_1)} = \frac{\bar{P}_1(\mathscr{E}_1)}{\bar{P}_0(\mathscr{E}_1)} \cdot \frac{\pi(H_1)}{\pi(H_0)}, \tag{5.A.1}$$

where $\bar{P}_j$ is the Bayes marginal distribution based on the prior Π_j. If Π_j has density π_j, then $\bar{P}_j(\mathscr{E}) = \int P_\delta(\mathscr{E})\pi_j(\delta)$. If Π_j puts all its mass on a particular element δ^*, then $\bar{P}_j(\mathscr{E}) = P_{\delta^*}(\mathscr{E})$.

In our setting, we observe a sequence $y_1, \ldots, y_n$, where n is itself the value that the stopping time τ (whose general underlying definition in terms of some stopping rule f may be unknown to us) takes; so we really observe $Y^n = y^n; \tau = n$. Because of the well-known fact that the Bayes posterior does not depend on the definition of the stopping time as long as it is defined in the (standard) way above (Hendriksen et al., 2020), we have for all n that $\pi(H_1 \mid Y^n = y^n, \tau = n) = \pi(H_1 \mid Y^n = y^n)$. i.e. if a variable stopping time is used and we happen to stop at $\tau = n$, the posterior is the same as if the sample size had been fixed in advance to n. This allows to express the *Bayes factor* $\mathbf{BF}(Y^\tau)$ and the posterior odds $\pi(H_1 \mid Y^\tau)/\pi(H_0 \mid Y^\tau)$ compactly as follows:

$$\bar{p}_j(Y^\tau) = \int_{\delta \in \Delta^+} p_\delta(Y^\tau) d\pi_j(\delta) \ , \text{ for all } n, \text{ for } j \in \{0, 1\} \ .$$
$$\mathbf{BF}(Y^\tau) = \frac{\bar{p}_1(Y^\tau)}{\bar{p}_0(Y^\tau)} \ , \ \frac{\pi(H_1 \mid Y^\tau)}{\pi(H_0 \mid Y^\tau)} = \mathbf{BF}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)}.$$

where we again assume that Π_j has density π_j; again, if Π_j puts all its mass on a particular element δ^*, then $\bar{p}_j(Y^\tau) = p_{\delta^*}(Y^\tau)$.

The pseudo-Bayes posterior odds What if we cannot come up with a good prior on Δ_0 and/or Δ_1? In that case we may want to 'represent' the set of distributions H_0 and H_1 by their 'least extreme elements' respectively, i.e. P_{δ^-} and P_{δ^+}. This gives the *pseudo-Bayes posterior odds* (given $\mathscr{E}_1, \mathscr{E}_2$ based on a prior that was already conditioned on $\mathscr{E}_1$)

$$\frac{\pi^{\text{ps}}(H_1 \mid \mathscr{E}_1, \mathscr{E}_2)}{\pi^{\text{ps}}(H_0 \mid \mathscr{E}_1, \mathscr{E}_2)} = \frac{P_{\delta^+}(\mathscr{E}_1 \mid \mathscr{E}_2)}{\bar{P}_{\delta^-}(\mathscr{E}_1 \mid \mathscr{E}_2)} \cdot \frac{\pi(H_1 \mid \mathscr{E}_2)}{\pi(H_0 \mid \mathscr{E}_2)}, \tag{5.A.2}$$

which is just the 'real' posterior odds that we would get if we had put all our prior mass on δ^+ and δ^- respectively. Similarly we get the *pseudo-Bayes factor*

$$\text{BF}^{\text{ps}}(Y^\tau) = \frac{p_{\delta^+}(Y^\tau)}{p_{\delta^-}(Y^\tau)}.$$

We can use the 'pseudo-Bayes factor' when (with a Bayesian mindset) we have no good idea about what might be a 'good' prior conditioned on $\delta \in \Delta^+$ and/or conditioned on $\delta \in \Delta^-$ or (with a frequentist mindset) about what value of δ in Δ^+ may be true if H_1 is true, or what value of δ in Δ^- may be 'true' if H_0 is true. Based on the pseudo-Bayes factor, we can also calculate the *pseudo-Bayes posterior odds* as if the Bayes factor were correct:

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^\tau)}{\pi^{\text{ps}}(H_0 \mid Y^\tau)} = \text{BF}^{\text{ps}}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)}. \tag{5.A.3}$$

In case $\pi(H_0)+\pi(H_1)=1$, the pseudo-Bayes posterior probability of H_0 is given by:

$$\pi^{\text{ps}}(H_0 \mid Y^\tau) = \frac{p_{\delta^-}(Y^\tau)\cdot\pi(H_0)}{p_{\delta^+}(Y^\tau)\cdot\pi(H_1)+p_{\delta^-}(Y^\tau)\cdot\pi(H_0)} = \frac{1}{\text{BF}^{\text{ps}}(Y^\tau)\cdot(\pi(H_1)/\pi(H_0))+1}.$$

Note in particular that in the pseudo-Bayes posterior, we use the *same* priors on H_0 and H_1 as in the 'real' posterior, but different, degenerate priors on Δ_0 and Δ_1.

5.A.3 The Result

Fix a *significance threshold* $r>1$ and let τ be an arbitrary stopping time. We will *reject* H_0 if $\pi^{\text{ps}}(H_1 \mid Y^\tau)/\pi^{\text{ps}}(H_1 \mid Y^\tau) \geq r$, and *accept* H_0 otherwise. Let $\text{REJECT}_{\tau,r}$ be the event that we reject at level r when using stopping time τ (importantly, in practice it may be unknowable what stopping rule τ is actually being used; to calculate posterior probabilities we only need to know the observed data Y^τ and the sample size of the observed data, and not the general definition of τ, i.e. we do not need to know if we would have stopped at the same n if the data had been different).

Theorem 5.A.1. *Let $\{P_\delta : \delta\in\Delta\}$ represent a 1-dimensional exponential family as above. Fix some $\delta^-<\delta^+$ and define H_1, H_0 and BF^{ps} correspondingly; also fix some arbitrary priors $\pi(H_0), \Pi_0$ on Δ_0, Π_1 on Δ_1. We have the following: for each n and each $r>1$ and each stopping time τ such that*

$$\frac{\bar{P}_1(\tau=n)}{P_{\delta^+}(\tau=n)}\cdot\frac{P_{\delta^-}(\tau=n)}{\bar{P}_0(\tau=n)} \geq 1, \tag{5.A.4}$$

we have: the posterior odds given rejection at time n are well-defined and satisfy

$$\frac{\pi\left(H_1 \mid \text{REJECT}_{\tau,r}, \tau=n\right)}{\pi\left(H_0 \mid \text{REJECT}_{\tau,r}, \tau=n\right)} \geq r. \tag{5.A.5}$$

The theorem implies the statement (5.5) in the main text for the Gaussian location family. The t there corresponds to the n here, and the statement $A(t)=1$ to $\tau=n$; the process $A(t)$ defines the stopping rule τ. The required condition (5.A.4) is easily seen to hold for the *Gold rush* scenario in which we evaluate invariably at $t(=n)=3$: inspecting the definition of $A(t)$, we see that the probability of reaching time 3 (i.e. $A(3)=1$) under P_δ increases monotonically with δ if δ represents the mean of a normal distribution. $\bar{P}_1$ being a mixture of P_δ's with $\delta\geq\delta^+$ and $\bar{P}_0$ being a mixture of P_δ's with $\delta\leq\delta^0$, (5.A.4) then follows.

In case $\pi(H_0)+\pi(H_1)=1$ (we rule out that δ does not lie in $\Delta_0\cup\Delta_1$), we can alternatively work on the scale of probabilities rather than probability ratios and fix a significance level $0<\alpha<1/2$ and reject H_0 if $\pi^{\text{ps}}(H_0 \mid Y^\tau)\leq\alpha$. This is equivalent to the event $\text{reject}_{\tau,r_\alpha}$ with $r_\alpha=(1-\alpha)/\alpha$. (5.A.5) then expresses that for each $0<\alpha<1/2$

$$\pi\left(H_0 \mid \text{REJECT}_{\tau,r_\alpha}\right) \leq \alpha. \tag{5.A.6}$$

The theorem expresses that, as long as the priors on H_0 and H_1 are chosen correctly, the error probabilities of decisions on the pseudo-Bayes posterior are *calibrated*: a Bayesian's real posterior odds of the decision 'reject' being correct (given by conditioning the true prior on the observed stopping time and the fact that at that stopping time, we rejected) can be no smaller than the posterior odds that this decision is correct according to the pseudo-Bayes posterior distribution on $\{H_0, H_1\}$, even though that distribution is, according to that same Bayesian, incorrect. Note that this is merely a 'one-sided' calibration, but the inequalities go the right (i.e. practically useful) way.

In a more frequentist interpretation, we may think of $\pi(H_0)$ as the 'population frequency' that the null is true in the particular field of science that we are working in. Whenever in a study H_0 is true, a particular $\delta_0 \in \Delta_0$ will be 'true' and generate the data , and whenever in a study H_1 is true, a particular $\delta_1 \in \Delta_1$ will be 'true' and generate the data. Theorem 5.A.1 expresses that our *conditional* error probability for rejecting/accepting H_0 is smaller than α, even though we do not know the true δ_0 and δ_1's.

From both a Bayesian and a frequentist stance, the result says that as long as *our prior* $\pi(H_0)$ *on* H_0 *reflects what happens in the real world* and we use if for reject/accept decisions of the kind above, it is o.k. to use the pseudo-Bayes posterior, and we can get away with not having a 'correct' or 'better' prior on the parameters in Δ^+ and Δ^-.

5.B Extension and Proof of Theorem 5.A.1

Our result holds more generally than for i.i.d. exponential families. Namely, we can more generally let $\mathcal{M} = \{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ with Δ a (possibly unbounded) interval in $\mathbb{R}$ denote a family of distributions for some random process $U_1, U_2, \ldots$. Again, δ denotes the 1-dimensional parameter of interest (e.g. an effect size) and now γ denotes potential nuisance parameters. We assume again a δ^+ and a $\delta^- < \delta^+$ are given, defining the null and alternative hypotheses

$$H_0 = \{P_{\delta,\gamma} : \delta \leq \delta^-, \gamma \in \Gamma\} \quad H_1 = \{P_{\delta,\gamma} : \delta \geq \delta^+, \gamma \in \Gamma\}.$$

Our result is valid for general families of this form, if furthermore the following holds: there exists a sequence of random vectors $Y_1, Y_2, \ldots$ such that Y_n is determined (can be written as a function of) $U^n = (U_1, \ldots, U_n)$ and the following two properties hold:

Irrelevance of γ and Full Support The distribution $P_\delta^{(n)}$ of Y^n under process $P_{\delta,\gamma}$ is the same for all γ (hence we can omit it from the notation in $P_\delta^{(n)}$). It has a density $p_\delta^{(n)}$ relative to some fixed underlying measure, and this density has the same support for all $\delta \in \Delta$. That is, we require for all $y^n \in \mathbb{R}^n$ that if for some $\delta \in \Delta$, $p_\delta^{(n)}(y^n) > 0$, then for all $\delta \in \Delta$, $p_\delta^{(n)}(y^n) > 0$. As a consequence, for any stopping rule τ, for every n, if for some $\delta \in \Delta$ we have $P_\delta(\tau = n) > 0$ then for all $\delta \in \Delta$ we have $P_\delta(\tau = n) > 0$. We call the set of n with $P_\delta(\tau = n) > 0$ the *support* of τ.

Monotone likelihood ratio (MLR) Property There exists a function s_n on $\mathbb{R}^n$ such that for each $\delta_0 < \delta_1$ with $\delta_0, \delta_1 \in \Delta$, the likelihood ratio $\frac{p_{\delta_1}^{(n)}(Y^n)}{p_{\delta_0}^{(n)}(Y^n)}$ is an increasing func-

tion of random variable $S_n := s_n(Y^n)$.

Note that both properties automatically hold for 1-dimensional i.i.d. exponential families as above – then we can set Γ to be a singleton, then γ plays no role, we can take $Y_n = U_n$ and $S_n = s_n(Y^n)$ to be the sufficient statistic for n outcomes (if Y_1 is the sufficient statistic for one outcome, then $S_n = \sum_{i=1}^n Y_i$) and then both properties are easily verified (Lehmann, 1986). But they also hold in the t-test setting, where $P_{\delta,\gamma}$ states that the underlying data U_i are i.i.d. normally distributed with variance γ and effect size δ (i.e. mean $\mu = \delta\gamma$). We can then take $Y_i := U_i/|U_1|$ to be the so-called 'maximal invariant statistic' (Hendriksen et al., 2020) and S_n to be the t-statistic based on U^n, which can be written as a function of Y^n. It is a well-known fact that S_n has a non-central t-distribution and that this satisfies the MLR property (Lehmann, 1986). We note that the set of allowed stopping rules/times remains unchanged in this more general set-up. Thus, in the t-test setting, and more generally in settings with $U_i \neq Y_i$, the stopping rule $f(Y^n)$ at time n must be writeable as a function of the Y^n which is a coarsening of (contains less information than) the data U^n. Since the condition above implies that the likelihood ratio can be written as a function of Y^n, and we usually use stopping rules that depend on the likelihood ratio observed so far and possibly some additional data that is independent of the observed data, but nothing else, this poses no great restriction in practice.

We now formulate and prove the theorem for this more general setup. Generalizing (5.A.3), the pseudo-Bayes posterior odds are now defined as:

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^\tau)}{\pi^{\text{ps}}(H_0 \mid Y^\tau)} = \text{BF}^{\text{ps}}(Y^\tau) \cdot \frac{\pi(H_1)}{\pi(H_0)} \text{ with } \text{BF}^{\text{ps}}(Y^\tau) = \frac{p^{(\tau)}_{\delta^+}(Y^\tau)}{p^{(\tau)}_{\delta^-}(Y^\tau)}. \tag{5.B.1}$$

Let again $\text{REJECT}_{\tau,r}$ be the event that $\pi^{\text{ps}}(H_1 \mid Y^\tau)/\pi^{\text{ps}}(H_1 \mid Y^\tau) \geq r$.

Theorem 5.B.1. *Let $\{P_{\delta,\gamma} : \delta \in \Delta, \gamma \in \Gamma\}$ represent a family that satisfies the two properties above for all n. Fix some $\delta^- < \delta^+$ and define H_1, H_0 and BF^{ps} correspondingly; also fix some arbitrary priors $\pi(H_0), \pi(H_1)$ and Π_0 on Δ_0, Π_1 on Δ_1. We have the following for each $r > 1$ and each stopping time τ and each n in the support of τ: the* true *posterior odds of H_1 vs. H_0 satisfy:*

$$\frac{\pi\left(H_1 \mid \text{REJECT}_{\tau,r}, \tau = n\right)}{\pi\left(H_0 \mid \text{REJECT}_{\tau,r}, \tau = n\right)} \geq r \cdot \frac{\bar{P}_1(\tau = n)}{\bar{P}_0(\tau = n)} \cdot \frac{P_{\delta^-}(\tau = n)}{P_{\delta^+}(\tau = n)}. \tag{5.B.2}$$

The earlier Theorem 5.A.1 is immediately seen to be a special case.

Remark The fact that we can go beyond exponential families raises the question of how general the result is. In this respect, we note that our conditions imply that the sequence of pseudo-Bayes factors $\text{BF}^{\text{ps}}(Y^1), \text{BF}^{\text{ps}}(Y^2), \ldots$ in (5.B.1) define a *test martingale* or equivalently, a product of conditional *E-values* under H_0 (Grünwald et al., 2019). Interestingly, *unconditional* frequentist error control under arbitrary stopping times can be given for arbitrary test martingales. All Bayes factors satisfying the conditions of the general version of the theorem below define *two-sided* test martingales: by this, we mean

that $1/\text{BF}^{\text{ps}}(Y^1), 1/\text{BF}^{\text{ps}}(Y^2), \ldots$ defines a test martingale under H_1. One might therefore suspect that our result continues to hold whenever we set our pseudo-Bayes factor equal to a two-sided test martingale, even if the MLR property does not hold. But it is not clear whether this really is the case. An example is the safe logrank test of Ter Schure et al. (2020b) (Chapter 2). The pseudo-Bayes factor we develop there is a ratio of partial likelihoods, and it defines a two-sided test martingale. Nevertheless, it is easily seen that due to the data not being i.i.d. the MLR property does *not* hold, and this property seems crucial for the argument used in the proof. Whether or not a (perhaps slightly weakened, i.e. $\geq r$ in (5.A.6) replaced by $\geq cr$ for some $c < 1$) version of the theorem holds for general pseudo-Bayes factors given by general two-sided test martingales is an interesting topic for future research.

Proof of Theorem 5.B.1

Fix $n \in \mathbb{N}$ in the support of τ. The proof makes crucial use of Lemma 5.B.2, which we state and prove first. We prove the theorem and the lemma only for the discrete case (with each Y_i taking values in a countable set $\mathcal{Y}_i \subset \mathbb{R}$), for which all densities become probability mass functions. It is straightforward to extend the results to the general case by replacing all probability mass functions with appropriate densities and sums by integrals.

Lemma 5.B.2. *Suppose that the MLR Property holds for some given n in the support of τ relative to some S_n as above. Then*

1. *The MLR Property holds for the set of distributions $\{P^{(n)}_\delta(\cdot \mid \tau = n) : \delta \in \Delta\}$ relative to S_n. That is, for each $\delta_0 < \delta_1$ with $\delta_0, \delta_1 \in \Delta$, $p^{(n)}_{\delta_1}(y^n \mid \tau = n)/p^{(n)}_{\delta_0}(y^n \mid \tau = n)$ is an increasing function of $s_n(y^n)$, on the set of all y^n with $p^{(n)}_\delta(y^{(n)} \mid \tau = n) > 0$ for some $\delta \in \Delta$.*
2. *As a consequence, for all $a > 0$,*

$$P_\delta\left(\frac{p^{(n)}_{\delta^+}(Y^n \mid \tau = n)}{p^{(n)}_{\delta^-}(Y^n \mid \tau = n)} \geq a \mid \tau = n\right) \tag{5.B.3}$$

   is increasing in δ for all a.

Proof. For the first part, note that for each y^n as above, we have:

$$\frac{p^{(n)}_{\delta_1}(y^n \mid \tau = n)}{p^{(n)}_{\delta_0}(y^n \mid \tau = n)} = \frac{p^{(n)}_{\delta_1}(y^n)}{p^{(n)}_{\delta_0}(y^n)} \cdot \frac{P^{(n)}_{\delta_1}(\tau = n \mid y^n)}{P^{(n)}_{\delta_0}(\tau = n \mid y^n)} \cdot \frac{P_{\delta_0}(\tau = n)}{P_{\delta_1}(\tau = n)} = \frac{p^{(n)}_{\delta_1}(y^n)}{p^{(n)}_{\delta_0}(y^n)} \cdot \frac{P_{\delta_0}(\tau = n)}{P_{\delta_1}(\tau = n)},$$

where the first equality is Bayes' theorem and the second equality follows, because for the type of stopping rule we employ, conditioned on the sequence y^n, the probability of

stopping exactly after having seen outcomes is independent of δ. But the rightmost expression shows that the likelihood ratio for the densities conditioned on $\tau = n$ must be an increasing function of $s_n(y^n)$ since, by assumption, the original, unconditional likelihood ratio is as well.

The second part follows immediately from the well-known connection (Lehmann, 1986) between monotone likelihood ratios and stochastic dominance; see
`https://math.stackexchange.com/questions/733291/why-mlr-monotone-likelihood-ratio-implies-stochastic-increasing` for a very short, simple, yet correct proof. □

In the remainder of the proof, we write $p_\delta(\cdot \mid \tau = n)$ instead of $p^{(n)}(\cdot \mid \tau = n)$ for brevity. Let $\mathscr{E}_{r,n}$ be the event that $\pi^{\text{ps}}(H_1 \mid Y^n)/\pi^{\text{ps}}(H_0 \mid Y^n) \geq r$. Since by the irrelevance of the stopping rule we have

$$\frac{\pi^{\text{ps}}(H_1 \mid Y^n)}{\pi^{\text{ps}}(H_0 \mid Y^n)} = \frac{\pi^{\text{ps}}(H_1 \mid \tau = n, Y^n)}{\pi^{\text{ps}}(H_0 \mid \tau = n, Y^n)} = \frac{\pi^{\text{ps}}(H_1 \mid \tau = n)}{\pi^{\text{ps}}(H_0 \mid \tau = n)} \cdot \frac{p_{\delta^+}(Y^n \mid \tau = n)}{p_{\delta^-}(Y^n \mid \tau = n)}$$

we have that $\mathscr{E}_{r,n}$ is equivalent to the event that $\frac{p_{\delta^+}(Y^n|\tau=n)}{p_{\delta^-}(Y^n|\tau=n)} \geq r \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)}$. We then have:

$$\begin{aligned}
&\frac{\pi\left(H_1 \mid \mathscr{E}_{r,n}, \tau = n\right)}{\pi\left(H_0 \mid \mathscr{E}_{r,n}, \tau = n\right)} \overset{(a)}{=} \frac{\bar{P}_1\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_1 \mid \tau = n)}{\bar{P}_0\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_0 \mid \tau = n)} \overset{(b)}{\geq} \frac{P_{\delta^+}\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_1|\tau = n)}{P_{\delta^-}\left(\mathscr{E}_{r,n} \mid \tau = n\right)\pi(H_0|\tau = n)} \\
&\overset{(c)}{=} \frac{\pi(H_1 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^+}(y^n \mid \tau = n) \cdot \mathbf{1}_{\frac{p_{\delta^+}(y^n|\tau=n)}{p_{\delta^-}(y^n|\tau=n)} \geq \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)} \cdot r}}{\pi(H_0 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\frac{p_{\delta^+}(y^n|\tau=n)}{p_{\delta^-}(y^n|\tau=n)} \geq \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)} \cdot r}} \overset{(d)}{\geq} \\
&\frac{\pi(H_1 \mid \tau = n) \cdot \left(r \cdot \frac{\pi^{\text{ps}}(H_0|\tau=n)}{\pi^{\text{ps}}(H_1|\tau=n)}\right) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\mathscr{E}_{r,n}}}{\pi(H_0 \mid \tau = n) \cdot \sum_{y^n} p_{\delta^-}(y^n \mid \tau = n) \cdot \mathbf{1}_{\mathscr{E}_{r,n}}} \overset{(e)}{=} r \cdot \frac{\bar{P}_1(\tau = n)}{\bar{P}_0(\tau = n)} \cdot \frac{P_{\delta^-}(\tau = n)}{P_{\delta^+}(\tau = n)}.
\end{aligned}$$

Here (a) is an instance of (5.A.1). We note that this inequality still holds in our generalized set-up as long as the probability of the set $\mathscr{E}_1$ under $P_{\delta,\gamma}$ does not depend on γ. (b) follows by first applying Lemma 5.B.2. (5.B.3) gives, using that $\bar{P}_1$ is a mixture of P_δ with $\delta \geq \delta^+$, that $\bar{P}_1(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) \geq r \mid \tau = n) \geq P_{\delta^+}(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) > r \mid \tau = n)$. Similarly it gives that $\bar{P}_0(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) \geq r \mid \tau = n) \leq P_{\delta^-}(p_{\delta^+}(y^n)/p_{\delta^-}(y^n) > r \mid \tau = n)$, and then (b) follows. (c) is merely writing out the definition, (d) follows by applying the inequality in the event in the indicator function and for (e) we used Bayes' theorem again.

This chain of inequalities gives (5.B.2), thus finishing the proof for the discrete case.