Sequential Bayes Factors
A flexible and efficient way of optional stopping
• Is the same number of red and blue candies in the bag?
  Your research question: Is it fair, or not?

• Every item you investigate costs 0.10 €.
  Don’t sample more than necessary!

• Think aloud about your choices, hunches, and evaluations of evidence.
A-Priori Power Analysis with G*Power

$g = \text{difference from H}_0 \text{ proportion, e.g. } 0.5 + 0.1 = 0.6$

constant proportion = H$_0$ proportion

$\alpha$ err prob = 0.05
Power (1−$\beta$ err prob) = 0.8
Constant proportion = 0.5

Output parameters:
- Lower critical N = 85,000,000
- Upper critical N = 114
- Total sample size = 199
- Actual power = 0.8037146
- Actual $\alpha$ = 0.0468865

X-Y plot for a range of values
Observations:

• *Nobody* does a priori sample size planning

• *Everybody* samples sequentially, and increases sample size if data feel inconclusive

• Very low evidential threshold for stopping - typically stopping before $\text{BF}_{10} \geq 2$
Evidence accumulation in two-choice decision tasks

Sequential sampling done wrong
Under $H_0$, $p$ values meander infinitely.
Repeated Significance Tests on Accumulating Data

By P. Armitage, C. K. McPherson and B. C. Rowe

Department of Medical Statistics and Epidemiology,
London School of Hygiene and Tropical Medicine

Table 2

The probability of being absorbed at or before the nth observation in sampling from a normal distribution with known variance, with repeated tests at a nominal two-sided significance level $2\alpha$ (i.e. standardized normal deviate $k$)†

<table>
<thead>
<tr>
<th>$2\alpha$</th>
<th>$0.10$</th>
<th>$0.05$</th>
<th>$0.02$</th>
<th>$0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$1.645$</td>
<td>$1.960$</td>
<td>$2.326$</td>
<td>$2.576$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$Q$</th>
<th>$S$</th>
<th>$Q$</th>
<th>$S$</th>
<th>$Q$</th>
<th>$S$</th>
<th>$Q$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.10000</td>
<td>0.0970</td>
<td>0.05000</td>
<td>0.0545</td>
<td>0.02000</td>
<td>0.0230</td>
<td>0.01000</td>
<td>0.0135</td>
</tr>
<tr>
<td>2</td>
<td>0.16015</td>
<td>0.1650</td>
<td>0.08312</td>
<td>0.0885</td>
<td>0.0345</td>
<td>0.0385</td>
<td>0.0456</td>
<td>0.0506</td>
</tr>
<tr>
<td>3</td>
<td>0.20207</td>
<td>0.1980</td>
<td>0.10726</td>
<td>0.1115</td>
<td>0.0545</td>
<td>0.0585</td>
<td>0.0620</td>
<td>0.0660</td>
</tr>
<tr>
<td>4</td>
<td>0.23399</td>
<td>0.2295</td>
<td>0.12617</td>
<td>0.1260</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.25963</td>
<td>0.2590</td>
<td>0.14169</td>
<td>0.1420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>0.63315</td>
<td></td>
<td>0.40829</td>
<td></td>
<td>0.2083</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>0.64301</td>
<td></td>
<td>0.41677</td>
<td></td>
<td>0.2135</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.65165</td>
<td></td>
<td>0.42429</td>
<td></td>
<td>0.2182</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>0.670</td>
<td></td>
<td>0.440</td>
<td></td>
<td>0.228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.720</td>
<td></td>
<td>0.487</td>
<td></td>
<td>0.259</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>0.746</td>
<td></td>
<td>0.513</td>
<td></td>
<td>0.276</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>0.763</td>
<td></td>
<td>0.529</td>
<td></td>
<td>0.288</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

With long enough sampling, it is guaranteed to get a significant result!
Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling IN PSYCHOLOGY

Leslie K. John¹, George Loewenstein², and Drazen Prelec³
¹Marketing Unit, Harvard Business School; ²Department of Social & Decision Sciences, Carnegie Mellon University; and ³Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology
Practical problems with p-values

• The “$p = .08$” problem
• Committing to an effect size in a power analysis
Optional stopping without guilt: Sequential Bayes Factors
Today’s program:

1. Sequential Bayes Factor (SBF) designs
2. Evaluating the properties of a SBF design with prospective (pre-data) design analysis
3. Does sequential sampling bias parameter estimates?
Abraham Wald, 1945: 
Sequential Probability Ratio Test (SPRT)
Compared to NHST with same error rates, average savings in sample size: -50%
Abraham Wald, 1945:
Sequential Probability Ratio Test (SPRT)

Because of the substantial savings in the expected number of observations effected by the sequential probability ratio test, and because of the simplicity of this test procedure in practical applications, the National Defense Research Committee considered these developments sufficiently useful for the war effort to make it desirable to keep the results out of the reach of the enemy, at least for a certain period of time. The author was, therefore, requested to submit his findings in a restricted report [7] which was dated September, 1943. In this report the sequential probability ratio test is devised and its mathematical theory is developed. In July 1944 a second report [8] was issued by the Statistical

Problem: You have to commit yourself to an exact point- $H_1$
Group sequential methods in the design and analysis of clinical trials

BY STUART J. POCOCK
Medical Computing and Statistics Group,
Medical School, University of Edinburgh

SUMMARY

In clinical trials with sequential patient entry, fixed sample size designs are unjustified on ethical grounds and sequential designs are often impracticable. One solution is a group sequential design dividing patient entry into a number of equal-sized groups so that the decision to stop the trial or continue is based on repeated significance tests of the accumulated data after each group is evaluated. Exact results are obtained for a trial with two treatments and a normal response with known variance. The design problem of determining the required size and number of groups is also considered. Simulation shows that these normal results may be adapted to other types of response data. An example shows that group sequential designs can sometimes be statistically superior to standard sequential designs.
Special issue article: Methods and statistics in social psychology: Refinements and new developments

Performing high-powered studies efficiently with sequential analyses

DANIEL LAKENS*
Human Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Running studies with high statistical power, while effect size estimates in psychology are often inaccurate, leads to a practical challenge when designing an experiment. This challenge can be addressed by performing sequential analyses while the data collection is still in progress. At an interim analysis, data collection can be stopped whenever the results are convincing enough to conclude that an effect is present, more data can be collected, or the study can be terminated whenever it is extremely unlikely that the predicted effect will be observed if data collection would be continued. Such interim analyses can be performed while controlling the Type I error rate. Sequential analyses can greatly improve the efficiency with which data are collected. Additional flexibility is provided by adaptive designs where sample sizes are increased on the basis of the observed effect size. The need for pre-registration, ways to prevent experimenter bias, and a comparison between Bayesian approaches and null-hypothesis significance testing (NHST) are discussed. Sequential analyses, which are widely used in large-scale medical trials, provide an efficient way to perform high-powered informa-
“The following rule of experimentation is therefore suggested:

Perform that experiment for which the expected gain in information is the greatest, and continue experimentation until a preassigned amount of information has been attained”

Dennis Lindley (1956)

Also suggested by Berger, 1985; Dienes, 2008; Kass & Raftery, 1995; Lindley, 1957; Wagenmakers et al., 2012, ...
Schönbrodt et al., 2015: Sequential Hypothesis Testing with Bayes Factors

Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences

Felix D. Schönbrodt  
Ludwig-Maximilians-Universität München, Germany

Eric-Jan Wagenmakers  
University of Amsterdam

Michael Zehetleitner  
Ludwig-Maximilians-Universität München, Germany

Marco Perugini  
University of Milan – Bicocca

Unplanned optional stopping rules have been criticized for inflating Type I error rates under the null hypothesis significance testing (NHST) paradigm. Despite these criticisms this research practice is not uncommon, probably as it appeals to researcher’s intuition to collect more data in order to push an indecisive result into a decisive region. In this contribution we investigate the properties of a procedure for Bayesian hypothesis testing that allows optional stopping with unlimited multiple testing, even after each participant. In this procedure, which we call Sequential Bayes Factors (SBF), Bayes factors are computed until an a priori defined level of evidence is reached. This allows flexible sampling plans and is not dependent upon correct effect size guesses in an a priori power analysis. We investigated the long-term rate of misleading evidence, the average expected sample sizes, and the biasedness of effect size estimates when an SBF design is applied to a test of mean differences between two groups. Compared to optimal NHST, the SBF design typically needs 50% to 70% smaller samples to reach a conclusion about the presence of an effect, while having the same or lower long-term rate of wrong inference.

Manuscript accepted for publication in Psychological Methods.
This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

NHST

• Define expected/minimal meaningful ES
• Define long-term error rates $\alpha$ and $\beta$
• Run a priori power analysis to determine fixed sample size
• Do study with that sample size, compute $p$, reject $H_0$ if $p < \alpha$

SBF

• Define desired evidential threshold (e.g., $BF_{10} > 10$, resp. $< 1/10$)
• Define prior distribution of effect sizes under $H_1$
• Run a minimal sample size (e.g. $n_{min} = 20$ per group), then compute BF and increase sample until BF exceeds one of both thresholds
The interpretation of BFAs is independent of stopping rules

The rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience. [...] This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels. (Edwards, Lindman, and Savage 1963, p. 193)

➤ you can apply the weirdest/most creative stopping rules to your data collection, without need for special corrections and adjustments!

➤ NHST, in contrast: Sampling intentions must be incorporated to get a proper $p$-value!

Consistency

Default Bayes factors converge towards (and across) the correct boundary
Under $H_0$, $p$ values meander infinitely

Exemplary $p$ value trajectory for $d = 0$

Do not reject $H_0$

Reject $H_0$
BF always converges to 0 or $\infty$

Exemplary Bayes Factor trajectory for $d = 0$
BF always converges to 0 or $\infty$

Exemplary Bayes Factor trajectory for $d = 0$

Bayes Factor trajectory for $d = 0$

false positive evidence

Very strong $H_1$

Strong $H_1$

moderate $H_1$

Anecdotal $H_1$

Anecdotal $H_0$

moderate $H_0$

Strong $H_0$

Very strong $H_0$
Symmetry of BFs

under $H_1$

under $H_0$
Sequential analysis: Compute BF after each candy

\[ \text{BF}_{\text{final}} = 4.37 \]
Let’s check our own candy bag sequence!

Try the app:

Prospective design analysis for SBF designs

aka

Frequentist properties of a Bayesian procedure


Planning stage (pre-data)

Frequentist properties:
• Simulate many hypothetical studies
• Record long-term probabilities of certain outcomes
• Goal: plan a study in a way that you have desirable properties.

Analysis stage (post-data)

Bayesian analysis:
BF carries all the information we have learned from this specific study.
Planning stage (pre-data)

*Bayesian design analysis:* Assumptions about true effect size: quantify uncertainty of expected effect size under $H_1$ with **design priors**.

Can be informed, non-central priors. Extreme case: point prediction of expected effect size (as in classical power analysis)

Analysis stage (post-data)

*Bayesian inference:* Use **analysis priors** for effect sizes under $H_1$ that convince a skeptical audience.

Typically more uninformative, with more weight on small effect sizes.

No “oracle prior”.

![Graph of plausibility vs. effect size (Cohen's d)](image1)

![Graph of plausibility vs. effect size (Cohen's d)](image2)
Planning stage
(pre-data)

Useful for practical reasons:
• Get an idea about the necessary sample size
• Tune design to make compelling evidence likely and to avoid misleading evidence
• Apply for a certain sample size in a grant

Analysis stage
(post-data)

As soon as the data are in, all pre-data planning is irrelevant!

That’s the real stuff: here the actual knowledge generation is done

Completely subjective: You don’t have to convince anyone about your choices here

Convince a skeptical audience
Probability of misleading evidence

- *Misleading evidence* = strong evidence towards the false hypothesis
Probability of misleading evidence

• *Misleading evidence* = strong evidence towards the false hypothesis

• Probability of misleading evidence: What is the proportion of hypothetical SBF studies that will stop at the wrong boundary (given the assumed true effect size (prior)?)

• Not to be confused with the post-data concept: Probability that this specific evidence is misleading
Expected sample size

Trajectories under $H_1$

$\delta = 0.5$
boundary = 10 (resp. 1/10)

density of stopping-$n$

92% stopped at $H_1$ boundary

8% stopped at $H_0$ boundary

Sample size

Bayes factor ($BF_{10}$)
Scenario: Independent t-test, 2-sided

Table 1
Percentages of wrong inference and average sample number (ASN) for the SBF design.

| r/Effect size | BF = 3 | | BF = 5 | | BF = 6 | | BF = 7 | | BF = 10 |
|---------------|--------|--------|--------|--------|--------|--------|--------|--------|
|               | % err  | ASN    | % err  | ASN    | % err  | ASN    | % err  | ASN    |
| \( \delta = 0 \) (% err relates to false positive evidence) | | | | | | | | |
| \( r = \sqrt{2}/2 \) | 7.5 | 30 | 6.6 | 96 | 6.0 | 146 | 5.5 | 205 |
| \( r = 1 \) | 5.6 | 24 | 5.1 | 50 | 4.7 | 75 | 4.3 | 105 |
| \( r = \sqrt{2} \) | 4.2 | 22 | 3.4 | 30 | 3.3 | 39 | 3.2 | 54 |
| \( \delta > 0 \) (% err relates to false negative evidence) | | | | | | | | |
| \( r = \sqrt{2}/2 \) | | | | | | | | |
| \( \delta = 0.20 \) | 77.9 | 34 | 50.6 | 133 | 35.9 | 203 | 24.1 | 269 |
| \( \delta = 0.30 \) | 60.9 | 36 | 21.7 | 108 | 10.4 | 140 | 4.4 | 162 |
| \( \delta = 0.40 \) | 42.8 | 35 | 7.0 | 79 | 1.9 | 91 | 0.4 | 97 |
| \( \delta = 0.50 \) | 26.8 | 33 | 1.7 | 57 | 0.2 | 61 | 0.0 | 64 |
| \( \delta = 0.60 \) | 15.2 | 30 | 0.3 | 42 | 0.0 | 45 | 0.0 | 46 |
| \( \delta = 0.70 \) | 7.9 | 27 | 0.0 | 34 | 0.0 | 35 | 0.0 | 36 |
| \( \delta = 0.80 \) | 3.7 | 25 | 0.0 | 28 | 0.0 | 29 | 0.0 | 30 |
| \( \delta = 1.00 \) | 0.6 | 22 | 0.0 | 23 | 0.0 | 24 | 0.0 | 24 |

ASN = Average sample number

Efficiency of sequential design, compared to an optimal* NHST design with the same error rates

* optimal = expected ES is exactly the true ES

Efficient on average, but …

<table>
<thead>
<tr>
<th>$r$/Effect size</th>
<th>BF = 3</th>
<th></th>
<th>BF = 5</th>
<th></th>
<th>BF = 20</th>
<th></th>
<th>BF = 30</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% err</td>
<td>ASN</td>
<td>% err</td>
<td>ASN</td>
<td>% err</td>
<td>ASN</td>
<td>% err</td>
<td>ASN</td>
</tr>
<tr>
<td>$\delta = 0$ (%) err relates to false positive evidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = \sqrt{2}/2$</td>
<td>7.5</td>
<td>30</td>
<td>6.6</td>
<td>96</td>
<td>2.4</td>
<td>1825</td>
<td>1.7</td>
<td>4057</td>
</tr>
<tr>
<td>$r = 1$</td>
<td>5.6</td>
<td>24</td>
<td>5.1</td>
<td>50</td>
<td>2.0</td>
<td>927</td>
<td>1.4</td>
<td>2107</td>
</tr>
<tr>
<td>$r = \sqrt{2}$</td>
<td>4.2</td>
<td>22</td>
<td>3.4</td>
<td>30</td>
<td>1.6</td>
<td>472</td>
<td>1.1</td>
<td>1070</td>
</tr>
</tbody>
</table>

→ if effect is zero or very small, it can take thousands of participants to reach compelling evidence for $H_0$ (e.g., $BF_{10} < 1/30$)
The \( SBF+maxN \) design

- Problem of open-ended SBF design: risk of very large samples (in particular, when a true but very small effect exists)
- Granting agencies: demand planning of maximal sample size.
- \( \rightarrow \) modification of SBF design: Sample until the BF exceeds an evidential threshold, or you reach \( n_{\text{max}} \).
SBF+maxN design analysis

Settings for exemplary design analysis:

Quantify uncertainty about true effect under $H_1$ with prior:

\[ \delta \sim N(0.5, \sigma^2 = 0.12) \]

- Thresholds = $BF_{10} > 30$ or $BF_{10} < 1/5$ (asymmetric)
- $n_{\text{min}} = 40$
- $n_{\text{max}} = 100$
- Directional $H_1$
$SBF+maxN$ design analysis

Trajectories under $H_1$

- 68% stopped at $H_1$ boundary
- 7% stopped at $H_0$ boundary

Bayes factor ($BF_{10}$)

Sample size

Very strong $H_1$
Strong $H_1$
Moderate $H_1$
Anecdotal $H_1$
Anecdotal $H_0$
Moderate $H_0$
Strong $H_0$
SBF+maxN design analysis

Settings for exemplary design analysis:

- Quantify uncertainty about true effect under $H_1$ with prior:
  \[ \delta \sim N(0.5, \sigma^2 = 0.12) \]

- Thresholds = $BF_{10} > 30$ or $BF_{10} < 1/5$ (asymmetric)
- $n_{\text{min}} = 40$
- $n_{\text{max}} = 100$
- Directional $H_1$

Results of design analysis:

- $\text{prob(false positive evidence)} = 0.3\%$
- $\text{prob(false negative evidence)} = 6.7\%$
- $\text{prob(true positive evidence)} = 68\%$
- $\text{prob(true negative evidence)} = 90.1\%$
- Median sample size under $H_1$: $n = 62$; under $H_0$: $n = 42$
Target properties of a sequential design

- Expected sample size
- Expected rate of misleading evidence (i.e., false negative evidence, false positive evidence)
- Expected rate of weak evidence
- Expected Bayes factor
Effect size estimation after a sequential procedure: Biased or unbiased?
Biased effect sizes in a sequential procedure?

“we found that truncated RCTs provide biased estimates of effects on the outcome that precipitated early stopping”

In a fixed-n design, this distribution of sample effect size estimates would be a symmetric $t$-distribution around the true effect. The optional stopping reshapes the distribution in strange ways ...

\[ \delta = 0.6, \text{ boundary } = 4 \]

Usually we do not recommend a boundary of 4 (too weak evidence, too high risk of misleading evidence). Here we use it for didactic purpose to pronounce the effects of early stopping.

Under the $H_1$ ...

- Early $H_1$ hits overestimate the true effect
- Late $H_1$ hits underestimate the true effect
- All $H_0$ hits underestimate the true effect

But: in a meta-analysis, the overall bias is negligible. Early terminations (which overestimate) are perfectly balanced by late terminations (which underestimate).

The small downward bias in the plot is not induced by the optional stopping rule, but by the effect size prior that shrinks extreme estimates towards zero.

Effect size estimates after a sequential procedure - summary

When only early terminations are reported and late terminations disappear into the file drawer,

a meta-analytic aggregation will overestimate the real effect.

Effect size estimates after a sequential procedure - summary

When only early terminations are reported and late terminations disappear into the file drawer,

a meta-analytic aggregation will overestimate the real effect.

Effect size estimates after a sequential procedure - summary

• Unsurprising: If one selectively averages larger-than-average effect sizes, the estimate is larger than average. This is true for Bayesian, frequentist, sequential, and non-sequential designs.

• Hence: Publication bias is always a concern. Sequential designs are unconditionally unbiased - but only when all studies are meta-analyzed.

Solution for conditional bias: Bayesian shrinkage

• If the result of a single study should be evaluated:
Overall summary
(Tentative) Recommendations for an SBF design in the two-sided independent $t$-test*

• Assumed we use the default JZS BF from JASP:
• Typical scenario in psychology: ES (Cohen’s $d$) of 0.5
• A plausible choices for early lines of research: boundary of 6 (resp 1/6) and a scale parameter of $r = 1$ (Cauchy distribution as effect size prior under $H_1$)
• ➞ balanced false positive and false negative rates (4.7% and 4.6%); on average 46% smaller samples than optimal NHST-PA with the same error rates.
• Whenever you have prior information about your paradigm: 
  **Use it!**

*for all other scenarios, compute your own tailored Bayes factor design analysis (Schönbrodt & Wagenmakers, submitted)
each week of data collection and adaptively increase the sample size until a predefined Bayes factor has been reached. In particular, skeptics and proponents set out to test at least 20 participants in each of the three eye movement conditions and agreed to stop testing whenever the Bayes factor for the horizontal eye movement versus no eye movement condition comparison reflects “strong” evidence for $H_0$ (i.e., $BF_{01} > 10$) or $H_1$ (i.e., $BF_{01} < 1/10$; see Jeffreys, 1961, for a classification
Practical Advantages

• A very intuitive way of testing

• A fair test: Possible to gain support for $H_0$

• No a-priori commitment to “smallest effect size of interest” necessary
  ➔ if you plan for the minimal interesting effect size, you often have much too large samples. SBF prevents that, as you can stop when the effect is larger than “minimally interesting”

• Flexible: Procedure adapts both too larger-than-expected and smaller-than-expected effects

• Efficient: Typically, sample sizes are on average more than 50% smaller compared to an optimal NHST fixed-$n$ design (while having same error rates)

• Guarantees compelling evidence, 0% risk of inconclusive results (if you sample long enough)

Small disadvantage

• Conditionally biased effect size estimates - be careful about ES estimate from a single small-$n$ study!
Strong inference

Scientists these days tend to keep up a polite fiction that all science is equal. But I think anyone who looks at the matter closely will agree that some fields of science are moving forward very much faster than others, perhaps by an order of magnitude.

**Strong inference:**
“[d]evising a crucial experiment [...], with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses.“

John Platt (1964), p. 347

SBF guarantees strong evidence (unless you run out of resources), which is a necessary ingredient of strong inference.