Bayes Factor Design Analysis

1. BFDA

We conduct a design analysis to determine the necessary sample size for testing whether exposure to a foreign language can reduce the illusion of causality. Our primary hypothesis centers on the difference between two groups: one group using their native language (NL) and the other using a foreign language (FL). We aim to apply Bayesian inference to evaluate evidence for either H_0 (no difference between the two groups) or H_1 (the FL reduces the illusion), and to estimate a credible interval for the effect size (ES) through the posterior distribution.

To achieve this, the Bayes Factor appears to be the most appropriate method. For this reason, our design analysis will be based on the Bayes Factor Design Analysis (BFDA). This approach is well outlined in two key articles by Schönbrodt and Wagenmakers (2016) and Stefan et al. (2017), which provide a comprehensive tutorial on BFDA.

1.1 Bayes Factor

The **Bayes Factor** (Jeffreys, 1935) is a statistical measure that quantifies the strength of evidence for one hypothesis over another. It is mathematically defined as the ratio of the marginal likelihoods of the data under the alternative hypothesis (H_1) and the null hypothesis (H_0):

$$BF_{10} = \frac{p(D|H_1)}{p(D|H_0)}$$

Where:

- $p(D|H_1)$ is the likelihood of the data under the **alternative hypothesis** (H_1).
- $p(D|H_0)$ is the likelihood of the data under the **null hypothesis** (H_0).

If the null hypothesis (H_0) and the alternative hypothesis (H_1) are considered equally probable a priori (i.e., $p(H_1) = p(H_0) = 0.5$), then a BF of $BF_{10} = 6$ would indicate that the data are 6 times more likely under the alternative hypothesis (H_1) than under the null hypothesis (H_0). This would correspond to a posterior probability of 86% for H_1 and 14% for H_0 (Stefan et al., 2017).

The Bayes Factor is particularly useful in hypothesis testing, as it quantifies evidence in favor of both hypotheses (H_0 and H_1) – a feature not provided by traditional Null Hypothesis Significance Testing (NHST). In our context, where we are interested in determining whether the foreign language effect (FLE) influences the illusion of causality, the BF offers a significant advantage over classical testing approaches.

1.1.1 Decision Rules for Bayes Factor

Decision rules for supporting either H_0 or H_1 often involve specifying upper and lower thresholds for the BF. Specifically:

- If the BF exceeds the upper threshold, it is considered evidence for the alternative hypothesis (H_1).
- If the BF falls below the lower threshold, it is considered evidence for the null hypothesis (H_0).
- If the BF lies between these thresholds, the evidence is considered inconclusive.

1.1.2 Interpretation of Bayes Factors

Several researchers have proposed heuristic classification schemes for interpreting BF. Below is a rough scale for interpreting BF_{10}, adapted from Lee and Wagenmakers (2014) :
This classification provides a helpful guide for interpreting the strength of evidence in favor of either H_1 or H_0 based on the observed BF.

1.2 Another Perspective on Power

In the traditional NHST framework, power analysis is used to estimate the probability of correctly rejecting the null hypothesis, based on a point estimate of the effect size under the alternative hypothesis. This analysis is useful for determining the required sample size to detect an effect, assuming that the effect size is known in advance.
In contrast, Bayesian Design Analysis has a different objective. Rather than focusing on the likelihood of rejecting the null hypothesis, the aim is to estimate the precision of the result. Specifically, it quantifies the probability that the BF will be either conclusive or inconclusive.

This approach provides a more nuanced understanding of the evidence in favor of the null or alternative hypotheses, emphasizing the strength of evidence rather than a simple binary decision.

1.3 Prior Distributions for BFDA

In order to perform the BFDA, we need to consider two prior distributions under H_1:

- The **Design Prior**, as defined by Schönbrodt and Wagenmakers (2016), is the prior used before data collection to quantify prior beliefs about the true state of nature and to assist in experimental design. Its purpose is to make compelling evidence likely and to avoid misleading evidence. The design prior incorporates our beliefs about the possible effect sizes (ESs), given the experimental conditions, and helps to guide the sampling process for calculating the BF. In our case, the prior distribution is the calculated distribution of Cohen's d based on our previous simulation.

- The **Analysis Prior** is used for Bayesian statistical analysis after the data are in – that is, to compute the BF. In our case, the analysis prior under H_1 is the personalized function described in the previous simulation:

$$f(x) = \left(\frac{1}{1 + e^{-40(x-0.2)}} + \frac{1}{1 + e^{35(x-0.9)}} - 1 \right) \cdot \frac{10}{7}$$

Our analysis prior under H_0 is that there is no difference between means (i.e., the effect size is equal to 0). The two models have been specified in Stan code to run simulations to compute the necessary power.

```
# STAN model specification under H_1
library(rstan)


Stan_Code_H1 <- '
functions{
  real funzione_delta(real delta){
```

```
    return (10/7)*((1 / (1 + exp(-65 * (delta - 0.2))) + 1 /
    (1 + exp(65 * (delta - 0.9))))-1);
  }
}
data {
  int<lower=1> N1;  // number of observations for the first group
  int<lower=1> N2;  // number of observations for the second group
  vector[N1] y1;  // Dependent variable for the first group
  vector[N2] y2;  // Dependent variable for the second group
}
parameters {
  real <lower=0, upper=100> mu;  // overall mean
  real delta;  // Cohens d
  real<lower=0> sigma_2;
  // standard deviation for both groups (assumed to be the same)
}

model {
  // Defining the priors
  target += log(1/(sigma_2)^2); // prior for the variance
  // (Jeffreys non-informative prior)
  target +=  log(funzione_delta(delta)); // prior for Cohens d

  // Data model
  target += normal_lpdf(y1 | mu + (delta * sigma_2 / 2), sigma_2);
  target += normal_lpdf(y2 | mu - (delta * sigma_2 / 2), sigma_2);
}
```

```
'


# STAN model specification under H_0
Stan_Code_H0 <- '
data {
  int<lower=1> N1;  // number of observations for the first group
  int<lower=1> N2;  // number of observations for the second group
  vector[N1] y1;  // Dependent variable for the first group
  vector[N2] y2;  // Dependent variable for the second group
}
parameters {
  real <lower=0, upper=100> mu;  // overall mean
  real<lower=0> sigma_2;  // standard deviation for both groups
}

model {
  // Prior
  target += log(1/(sigma_2)^2); // prior for the variance
  //(Jeffreys non-informative prior)

  // Data model
  target += normal_lpdf(y1 | mu, sigma_2);
  target += normal_lpdf(y2 | mu, sigma_2);
}
'

stan_M_H1 <- stan_model(model_code = Stan_Code_H1, model_name="stanmodel")
```

```
stan_M_H0 <- stan_model(model_code = Stan_Code_H0, model_name="stanmodel")
```

2. Design Analysis

The following is an adapted version of the algorithmic model outlined by Schönbrodt and Wagenmakers (2016) for performing a design analysis with BF:

1. Define a population (or distributions) reflecting the expected ES under the alternative hypothesis H_1 (i.e., the design prior): In our case, this correspond to the situation where a true difference exists between the two populations (i.e., we are hypothesizing a meaningful difference from 5 to 20 points on a scale from 0 to 100 between the NL and FL groups).

2. Draw a random sample of size n from each population: The sample size n should be fixed and identical for both groups.

3. Compute the BF for simulated data: The BF will be computed using the analysis prior that will also be used in the actual data analysis. We are interested in checking if the BF exceeds a certain threshold y. If $BF > y$, the result is considered a "correct answer." If $\frac{1}{y} < BF \leq y$, the result is considered "inconclusive." If $BF < \frac{1}{y}$, the result is considered to be in the "wrong direction."

4. Repeat the simulation (e.g., 5.000 times): Repeat steps 2 and 3 multiple times to obtain a distribution of BF values under H_1, for given sample sizes ($n_x, n_{x+y}, ..$). This allow to calculate the different types of results that may be yielded from different sample sizes.

5. Simulate data under H_0 for false-positive probability: Perform the same simulation process as in steps 2 and 3, but this time assume the null hypothesis H_0 is true (i.e., no true effect exists between the populations). This allows us to estimate the false-positive rate (i.e., the probability of incorrectly supporting H_1 when H_0 is true).

2.1 Under H_1

We run simulations to compute the BF using different sample sizes to determine which sample size will allow us to achieve 80% power (i.e., a proportion of .8 for the BF in the correct direction). We use a threshold of $y = 3$ and $-y = \frac{1}{3}$ for compelling evidence.

To compute the design prior, we extract the population parameters based on the dataframe we created from the truncated normal distribution file. This dataframe accounts for different combinations of mean differences between 5 and 20 points on a bounded scale from 0 to 100. The NL group mean ranges from 55 to 65 (with $SD = 20$), and the SD for the FL group varies between 20 and 25. We sample rows from this dataframe to run the BFDA considering different hypothesized parameters.

```
#Loading data from previous simulation


load("Effsize.Rda")
library(bridgesampling)


# Function to generate from a truncated normal distribution


tnorm_f <- function(n, mean, sd, a = 0, b = 100) {
  qnorm(runif(n, pnorm(a, mean, sd), pnorm(b, mean, sd)), mean, sd)
}


# Function to compute power under H1


run_simulation <- function(n_sim = 100000, sample_size = 100) {


  BF_set <- rep(NA, n_sim)
```

```
for(i in 1:n_sim) {
  sam <- effect_size_analysisP[sample(nrow(effect_size_analysisP), 1), ]

  # Simulate NL group and FL group
  FL_gr <- round(tnorm_f(n = sample_size, mean = as.numeric(sam$FLmean),
                         sd = sam$FLmean))
  NL_gr <- round(tnorm_f(n = sample_size, mean = as.numeric(sam$NLmean),
                         sd = 20))

  # Perform BF
  stan_Fit_H1 <- sampling(stan_M_H1, data = list(y2 = FL_gr,
                                                 y1 = NL_gr,
                                                 N1 = sample_size,
                                                 N2 = sample_size),
                          iter = 20000, warmup = 500, chains = 4, cores = 1,
                          NL_gr = list(adapt_delta = .99))


  stan_Fit_H0 <- sampling(stan_M_H0, data = list(y2 = FL_gr,
                                                 y1 = NL_gr,
                                                 N1 = sample_size,
                                                 N2 = sample_size),
                          iter = 20000, warmup = 500, chains = 4, cores = 1,
                          NL_gr = list(adapt_delta = .99))
  H0 <- bridge_sampler(stan_Fit_H0, silent = TRUE)
  H1 <- bridge_sampler(stan_Fit_H1, silent = TRUE)
  BF10 <- bf(H1, H0)
```

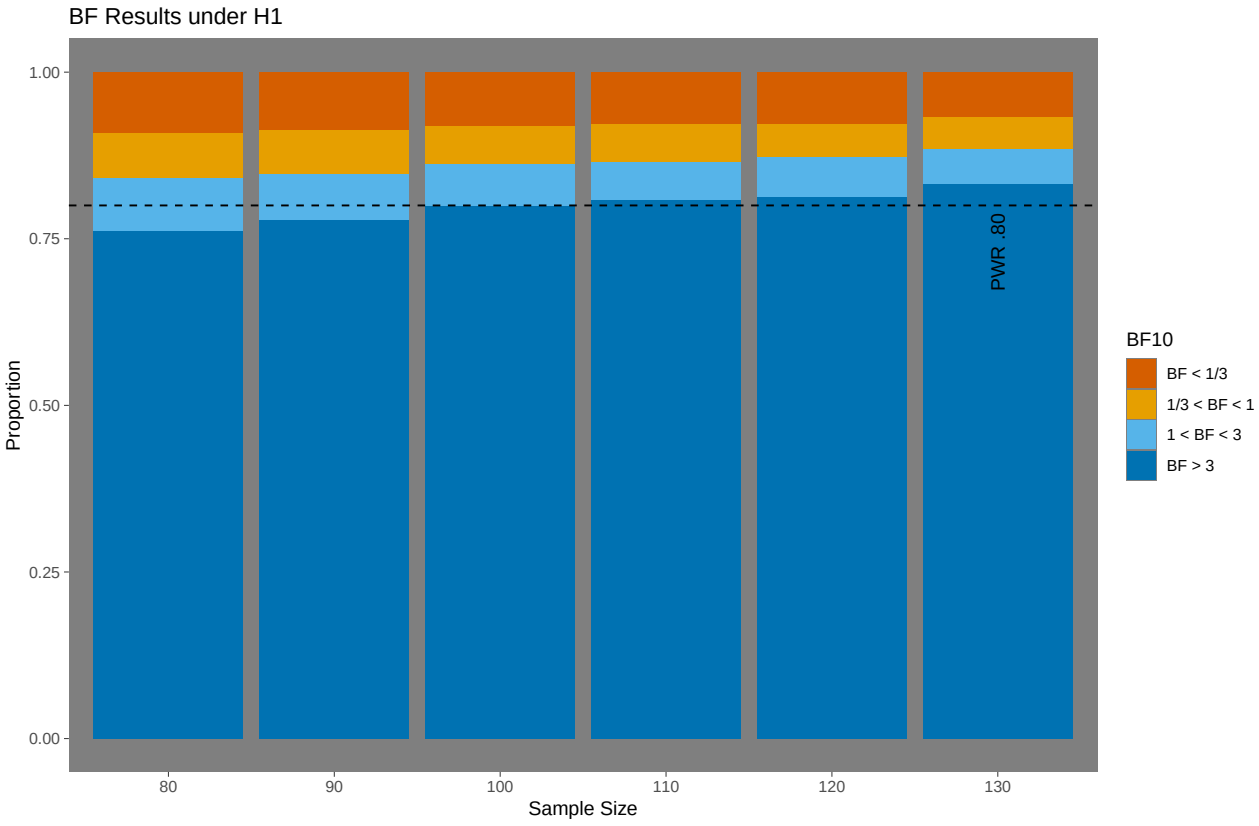
```
    # Store BF
    BF_set[i] <- as.numeric(BF10)[1]
  }

  # Classify the BF
  vector_Waffle <- rep(NA, length(BF_set))
  vector_Waffle[BF_set < 1/3] <- "Wrong Direction"
  vector_Waffle[BF_set >= 1/3 & BF_set <= 1] <- "Inconclusive Wrong"
  vector_Waffle[BF_set >= 1 & BF_set <= 3] <- "Inconclusive Right"
  vector_Waffle[BF_set > 3] <- "Right Direction"
  # Return the results
  return(data.frame(BF = BF_set, direction = as.factor(vector_Waffle)))
}



# Simulations for each sample size (uncomment to reproduce)
sample_sizes <- c(80,80, 90,90,100,100, 110, 110,120,120,130,130)
#simulation_results <- list()

#library(doParallel)
#registerDoParallel(cores = 20) #Change n of cores

# Parallel loop to run simulations
#simulation_results <- foreach(i = sample_sizes,
#.combine = rbind, .packages = c("rstan", "bridgesampling")) %dopar% {
  #sim_result <- run_simulation(n_sim = 2500, sample_size = i)   size
```

```
  #sim_result$sample_size <- i


 #return(sim_result)
#}


#H1 <- do.call(cbind, simulation_results)
```

We plot the results that we already store from this function and that are available under the Files area on the OSF project page.

```
# Plot the results


H1 <- read.csv("H_1.csv")


H1$direction[H1$BF>3] <- "4"
H1$direction[H1$BF<1/3] <- "1"
H1$direction[H1$BF<3 & H1$BF>1] <- "3"
H1$direction[H1$BF<1 & H1$BF>1/3] <- "2"
H1$direction <- as.factor(H1$direction)


library(ggplot2); library(ggokabeito)


PWR80 <- ggplot(H1, aes(x = factor(sample_size),
                        fill = factor(direction))) +
  scale_fill_manual(values = c("#D55E00","#E69F00", "#56B4E9", "#0072B2"),
        labels = c("BF < 1/3", "1/3 < BF < 1", "1 < BF < 3", "BF > 3")) +
  geom_bar(position = "fill") +
  labs(title = "BF Results under H1",
```

```
       x = "Sample Size",
       y = "Proportion",
       fill = "BF10") +
  geom_hline(yintercept = .8, lty = "dashed", col = "black") +
  theme_dark() + theme(panel.grid.major = element_blank(),
                       panel.grid.minor = element_blank()) +
  annotate("text", x = 6, y = 0.73, label = "PWR .80",
           angle = '90', col = "black")


PWR80
```



Our analysis shows that in order to obtain a conclusive BF in favor of H_1, more than 80% of the time, we must collect 110 participants per group. At 110 participants per group, we also have a 7 % probability of finding a result in the wrong direction and a 13% probability of obtaining inconclusive results, under the assumption of H_1.

```r
# Contingency table of sample sizes
H1$direction <- as.character(H1$direction)
H1$direction[H1$BF>3] <- "BF > 3"
H1$direction[H1$BF<1/3] <- "BF < 1/3"
H1$direction[H1$BF<3 & H1$BF>1] <- "1 < BF < 3"
H1$direction[H1$BF<1 & H1$BF>1/3] <- "1/3 < BF < 1"
table(H1$direction, H1$sample_size) /5000
```

```
                  80     90    100    110    120    130
  1 < BF < 3   0.0800 0.0684 0.0630 0.0566 0.0598 0.0524
  1/3 < BF < 1 0.0670 0.0660 0.0570 0.0572 0.0508 0.0476
  BF < 1/3     0.0920 0.0874 0.0812 0.0782 0.0772 0.0674
  BF > 3       0.7610 0.7782 0.7988 0.8080 0.8122 0.8326
```

2.2 Under H_0

We perform the same simulation process as in the previous section, but now we assume H_0 is true. This allows us to compute the false positive rate – the probability of incorrectly supporting H_1 when no true effect exists. We expect the false positive rate to be less than 5% for 110 participants per group.

```r
# Function to compute power under H0

run_simulation <- function(n_sim = 100000, sample_size = 100) {

  BF_set <- rep(NA, n_sim)
  for(i in 1:n_sim) {
    sam <- effect_size_analysisP[sample(nrow(effect_size_analysisP), 1), ]
```

```r
# Simulate NL group and FL group
FL_gr <- round(tnorm_f(n = sample_size,
                       mean = as.numeric(sam$NLmean),
                       sd = 20))
NL_gr <- round(tnorm_f(n = sample_size,
                       mean = as.numeric(sam$NLmean),
                       sd = 20))

# Perform BF
stan_Fit_H1 <- sampling(stan_M_H1,
                        data = list(y2 = FL_gr,y1 = NL_gr,
                                    N1 = sample_size,
                                    N2 = sample_size),
                      iter = 20000, warmup = 500, chains = 4, cores = 1,
                      NL_gr = list(adapt_delta = .99))


stan_Fit_H0 <- sampling(stan_M, data = list(y2 = FL_gr,
                                            y1 = NL_gr,
                                           N1 = sample_size,
                                          N2 = sample_size),
                    iter = 20000, warmup = 500, chains = 4, cores = 1,
                  NL_gr = list(adapt_delta = .99))
H0 <- bridge_sampler(stan_Fit_H0, silent = TRUE)
H1 <- bridge_sampler(stan_Fit_H1, silent = TRUE)
BF01 <- bf(H0, H1)
```

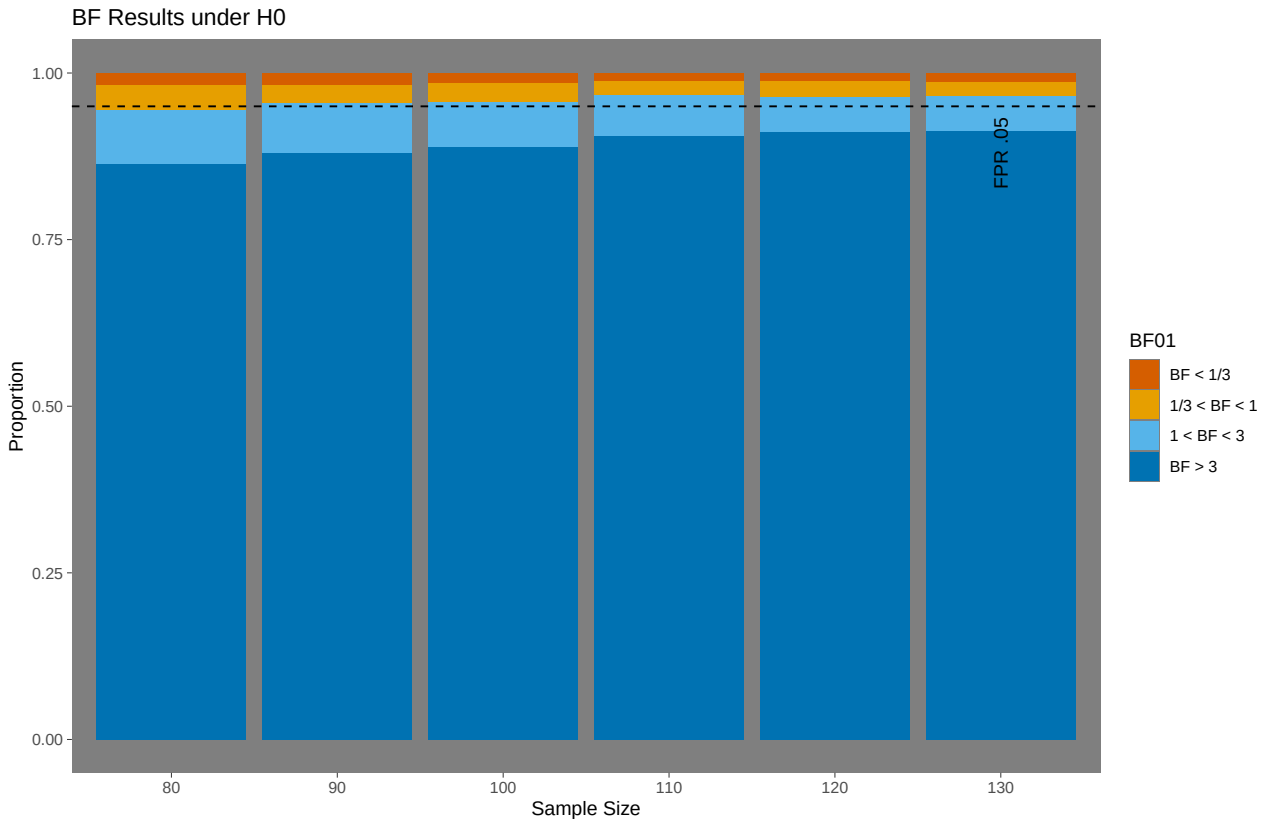
```
    # Store BF
    BF_set[i] <- as.numeric(BF01)[1]
  }

  # Classify the BF
  vector_Waffle <- rep(NA, length(BF_set))
  vector_Waffle[BF_set < 1/3] <- "Wrong Direction"
  vector_Waffle[BF_set >= 1/3 & BF_set <= 1] <- "Inconclusive Wrong"
  vector_Waffle[BF_set >= 1 & BF_set <= 3] <- "Inconclusive Right"
  vector_Waffle[BF_set > 3] <- "Right Direction"
  # Return the results
  return(data.frame(BF = BF_set, direction = as.factor(vector_Waffle)))
}



# Simulations for each sample size (uncomment to reproduce)
sample_sizes <- c(80,80, 90,90,100,100, 110, 110,120,120,130,130)
#simulation_results <- list()


#library(doParallel)
#registerDoParallel(cores = 20)  #Change n of cores


# Parallel loop to run simulations
#simulation_results <- foreach(i = sample_sizes, .combine = rbind,
# .packages = c("rstan", "bridgesampling")) %dopar% {
  #sim_result <- run_simulation(n_sim = 2500, sample_size = i)   size
  #sim_result$sample_size <- i
```

```
 #return(sim_result)
#}


#H0 <- do.call(cbind, simulation_results)
```

We plot the result that we already store from this function that are available under the Files area on the page OSF project page.

```
# Plot the results
H0 <- read.csv("H_0.csv")


H0$direction[H0$BF>3] <- "4"
H0$direction[H0$BF<1/3] <- "1"
H0$direction[H0$BF<3 & H0$BF>1] <- "3"
H0$direction[H0$BF<1 & H0$BF>1/3] <- "2"
H0$direction <- as.factor(H0$direction)



FPR05 <- ggplot(H0, aes(x = factor(sample_size), fill = factor(direction))) +
  scale_fill_manual(values = c("#D55E00","#E69F00", "#56B4E9", "#0072B2"),
labels = c("BF < 1/3", "1/3 < BF < 1", "1 < BF < 3", "BF > 3")) +
  geom_bar(position = "fill") +
  labs(title = "BF Results under H0",
       x = "Sample Size",
       y = "Proportion",
       fill = "BF01") +
  geom_hline(yintercept = .95, lty = "dashed", col = "black") +
```

```
  theme_dark() + theme(panel.grid.major = element_blank(),
                       panel.grid.minor = element_blank()) +
  annotate("text", x = 6, y = 0.88, label = "FPR .05",
           angle = '90', col = "black")


FPR05
```



Our analysis also shows that in order to obtain a conclusive BF in favor of H_0, almost 90% of the time, we must collect 110 participants per group.

```
# Contingency table of sample sizes
H0$direction <- as.character(H0$direction)
H0$direction[H0$BF>3] <- "BF > 3"
H0$direction[H0$BF<1/3] <- "BF < 1/3"
H0$direction[H0$BF<3 & H0$BF>1] <- "1 < BF < 3"
```

```r
H0$direction[H0$BF<1 & H0$BF>1/3] <- "1/3 < BF < 1"


table(H0$direction, H0$sample_size) / 5000
```

```
                  80     90    100    110    120    130
  1 < BF < 3   0.0798 0.0742 0.0674 0.0614 0.0520 0.0518
  1/3 < BF < 1 0.0376 0.0280 0.0280 0.0204 0.0254 0.0216
  BF < 1/3     0.0188 0.0180 0.0152 0.0130 0.0114 0.0140
  BF > 3       0.8638 0.8798 0.8894 0.9052 0.9112 0.9126
```

Barrett, M. (2024). *Ggokabeito: 'Okabe-ito' scales for 'ggplot2' and 'ggraph'*. R package version 0.1.0.9000, commit e28e8b7a0a3301ac40722fb07ed082bde424bb8f.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10), 1–29. https://doi.org/10.18637/jss.v092.i10

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222. https://doi.org/10.1017/S030500410001330X

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Masked Citation. (n.d.). *Masked Title*.

Morey, R., & Rouder, J. (2024). *BayesFactor: Computation of bayes factors for common designs*. R package version 0.9.12-4.7.

Schönbrodt, F. D., & Stefan, A. M. (2019). *BFDA: An r package for bayes factor design analysis* (Version 0.5.0).

Schönbrodt, F. D., & Wagenmakers, E. (2016). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *23*(2), 254–271. https://doi.org/10.3758/s13423-017-1230-y

Stan Development Team. (2024). *RStan: The R interface to Stan*. https://mc-stan.org/

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. (2017). A tutorial on bayes factor design analysis using an informed prior. *Behavior Research Methods*, *49*(2), 413–428. https://doi.org/10.3758/s13428-018-01189-8

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Bayes Factor (BF_{10})	**Evidence Category**
> 100	Extreme evidence for H_1
$30 - 100$	Very strong evidence for H_1
$10 - 30$	Strong evidence for H_1
$3 - 10$	Moderate evidence for H_1
$1 - 3$	Anecdotal evidence for H_1
1	No evidence
$\frac{1}{3} - 1$	Anecdotal evidence for H_0
$\frac{1}{10} - \frac{1}{3}$	Moderate evidence for H_0
$\frac{1}{30} - \frac{1}{10}$	Strong evidence for H_0
$\frac{1}{100} - \frac{1}{30}$	Very strong evidence for H_0
$< \frac{1}{100}$	Extreme evidence for H_0