p-values and Bayes factors under sequential and block sampling procedures

Berenice López-Casal & Rocío Alcalá-Quintana[1]
Universidad Complutense de Madrid

Bayes Factors (BFs) are sometimes used as a means to putting hypotheses to test instead of the more widespread Null Hypothesis Significance Testing (NHST) procedure. Despite the many conceptual differences between BFs and *p*-values, García-Pérez (2016) showed that, under certain conditions, they have a one-to-one correspondence that make them equivalent statistics. Whether this equivalence holds upon failure to meet the assumptions of either method is up for analysis.

While it is widely known that the validity of the NHST procedure depends on whether its sampling assumptions are met (i.e., the sample size should be fixed before collecting the data), it has been claimed that BFs make no assumptions regarding the process of data collection and are thus immune to changes in the sampling procedure (Etz et al., 2016). This app illustrates the relation between *p*-values and Bayes Factors for various statistical tests under different sampling procedures, showing that both methods are similarly affected by sequential and block sampling.

The texts displayed on the information section of each of the the application's tabs are presented below.

t-test for a single mean

Sequential sampling
This program simulates data and performs **two-sided Student's t-tests** within sequential sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes. Bayes Factors are computed as described by Gronau & Wagenmakers (2017).

Usage

Choose the method used to define the stopping criterion for the sampling process (either *p*-values or BFs) and the stopping threshold for rejecting H_0 (*p*-values) or for gathering enough evidence in favor of H_1 (BFs), as well as the minimum and maximum sample size allowed. Whenever the minimum and maximum sample sizes are set equal, the procedure corresponds to a fixed sampling design. The application only requires one stopping threshold (b_1), but it gives the option to add an extra upper threshold (b_2), which yields the sampling procedure described below. Also, a reference value for the method not being used to define the stopping rule can be entered, which will be shown as a mere benchmark in the output plots.

In the frequentist framework, if $p < b_1$, the sampling stops and the null hypothesis is rejected; if $p > b_2$, the sampling stops and H_0 is not rejected; if $b_1 \leq p \leq b_2$ the sampling

[1] To whom correspondence should be addressed (ralcala@psi.ucm.es)

continues. When b_1 = 0.01 and b_2 = 0.36 this procedure corresponds to Frick's (1998) Composite Adaptive Sequential Test (COAST), which sets the rate of Type I error rate to roughly 0.05. In the Bayesian framework, the sampling procedure with two thresholds is analogous to the one just described (i.e., $BF_{01} < b'_1$ leads to acceptance of H_1; $BF_{01} > b'_2$ leads to acceptance of H_0; and if $b'_1 < BF_{01} < b'_2$ further data are gathered). Also, the user can change the specifications for the Normal prior distribution over effect sizes (Cohen's *d*), which are set to a standard Normal *N(0,1)* by default (the null and alternative hypotheses can be expressed in terms of effect sizes).

Output

The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process). When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size. When H_0 is true and there is a mismatch between H0 and the mean for the prior distribution, the relation between log BF_{01} and log *p*-values turns into a two-limbed curve. The lower limb corresponds to those simulations where the evidence (i.e., data) gathered lies to the left of H_0, whilst the upper limb corresponds to evidence that lies to the right of such point. In the frequentist framework, it can be seen that the empirical type I error rate differs from the nominal rate under sequential sampling (both, with and without an extra stopping threshold).

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq b'_1$, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

Two additional plots can be accessed by pressing the button between the information and warning icons. On the left, a plot with the **prior distribution** for the Bayesian framework is displayed; on the right, either the *p*-value or the BF_{01} as a function of sample sizes for each simulation are plotted.

Block sampling

This program simulates data and performs **two-sided Student's *t*-tests** within block sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes. Bayes Factors are computed as described by Gronau & Wagenmakers (2017).

Usage

Choose the method used to define the stopping criterion for the sampling process (either *p*-values or BFs) and the stopping threshold for rejecting H_0 (*p*-values) or for gathering enough evidence in favor of H_1 (BFs), as well as the minimum sample size, the size of the sample blocks, and the maximum number of blocks allowed. In both the frequentist and Bayesian frameworks, data are gathered in blocks of the specified size until the stopping criterion is met. Inflated type I error rates due to multiple comparisons are inherent to sequential sampling procedures under the frequentist approach. Those are

controlled by default with the (overly conservative) Bonferroni correction, but the user is allowed to uncheck this option and not correct for multiple comparisons.

Whenever the size of the blocks is set to 1, the procedure corresponds to a sequential sampling design. Also, the user can change the specifications for the Normal prior distribution over effect sizes (Cohen's *d*), which are set to a standard Normal *N(0,1)* by default (the null and alternative hypotheses can be expressed in terms of effect sizes).

Output

The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process). When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size. When H_0 is true and there is a mismatch between H0 and the mean for the prior distribution, the relation between log BF_{01} and log *p*-values turns into a two-limbed curve. The lower limb corresponds to those simulations where the evidence (i.e., data) gathered lies to the left of H_0, whilst the upper limb corresponds to evidence that lies to the right of such point. In the frequentist framework, it can be seen that the empirical type I error rate differs from the nominal rate under sequential sampling (both, with and without an extra stopping threshold).

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq b'_1$, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

Test for a single Pearson's correlation

Sequential sampling

This program simulates data and performs **two-sided** tests for **Pearson's correlation** with H_0: $\rho = 0$ within sequential sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes. Bayes Factors are computed as described by Wetzels & Wagenmakers (2012).

Usage

Choose the method used to define the stopping criterion for the sampling process (either *p*-values or BFs) and the stopping threshold for rejecting H_0 (*p*-values) or for gathering enough evidence in favor of H_1 (BFs), as well as the minimum and maximum sample size allowed. Whenever the minimum and maximum sample sizes are set equal, the procedure corresponds to a fixed sampling design. The application only requires one stopping threshold (b_1), but it gives the option to add an extra upper threshold (b_2), which yields the sampling procedure described below. Also, a reference value for the method not being used to define the stopping rule can be entered, which will be shown as a mere benchmark in the output plots. In the frequentist framework, if $p < b_1$, the sampling stops and the null hypothesis is rejected; if $p > b_2$, the sampling stops and H_0 is not rejected;

if $b_1 \leq p \leq b_2$ the sampling continues. In the Bayesian framework, the sampling procedure with two thresholds is analogous to the one just described (i.e., $BF_{01} < b'_1$ leads to acceptance of H_1; $BF_{01} > b'_2$ leads to acceptance of H_0; and if $b'_1 < BF_{01} < b'_2$ further data are gathered). The null hypothesis is fixed to $\rho = 0$.

Output

The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process). When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size.

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq b'_1$, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

Two additional plots can be accessed by pressing the button between the information and warning icons. On the left, a plot with the **prior distribution** for the Bayesian framework is displayed; on the right, either the *p*-value or the BF_{01} as a function of sample sizes for each simulation are plotted.

Block sampling

This program simulates data and performs **two-sided** tests for **Pearson's correlation** within block sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes. Bayes Factors are computed as described by Wetzels & Wagenmakers (2012).

Usage

Choose the method used to define the stopping criterion for the sampling process (either *p*-values or Bayes Factors) and the stopping threshold for rejecting H_0 or gathering enough evidence in favor of H_1 respectively, as well as the minimum sample size, the size of the sample blocks, and the maximum number of blocks allowed. In both the frequentist and Bayesian frameworks, data are gathered in blocks of the specified size until the stopping criterion is met. Inflated type I error rates due to multiple comparisons are inherent to sequential sampling procedures under the frequentist approach. Those are controlled by default with the (overly conservative) Bonferroni correction, but the user is allowed to uncheck this option and not correct for multiple comparisons. Whenever the size of the blocks is set to *1*, the procedure corresponds to a sequential sampling design. The null hypothesis is fixed to $\rho = 0$.

Output

The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those

that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process). When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size.

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq b'_1$, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

Binomial test for a single proportion

Sequential sampling
This program simulates data and performs **two-sided** tests for a single proportion in a (finite) **Bernoulli process** within sequential sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes. Bayes Factors are computed as described by Wagenmakers et. al. (2010).

Usage
Choose the method used to define the stopping criterion for the sampling process (either *p*-values or BFs) and the stopping threshold for rejecting H_0 (*p*-values) or for gathering enough evidence in favor of H_1 (BFs), as well as the minimum and maximum sample size allowed. Whenever the minimum and maximum sample sizes are set equal, the procedure corresponds to a fixed sampling design. The application only requires one stopping threshold (b_1), but it gives the option to add an extra upper threshold (b_2), which yields the sampling procedure described below. Also, a reference value for the method not being used to define the stopping rule can be entered, which will be shown as a mere benchmark in the output plots. In the frequentist framework, if $p < b_1$, the sampling stops and the null hypothesis is rejected; if $p > b_2$, the sampling stops and H_0 is not rejected; if $b_1 \leq p \leq b_2$ the sampling continues. In the Bayesian framework, the sampling procedure with two thresholds is analogous to the one just described (i.e., $BF_{01} < b'_1$ leads to acceptance of H_1; $BF_{01} > b'_2$ leads to acceptance of H_0; and if $b'_1 < BF_{01} < b'_2$ further data are gathered). Also, the user can change the specifications for the Beta prior distribution over the probability of success, which are set at *Beta(1,1)* by default.

Output
The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process).. When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size. In addition, specifying a Beta prior that concentrates mass relatively far from the true probability of success when H_0 holds turns the scatter plot into a two-limbed curve. The lower limb corresponds to those simulations where the evidence (i.e., data)

gathered lies to the left of H_0, whilst the upper limb corresponds to evidence that lies to the right of such point.

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq b'_1$, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

Two additional plots can be accessed by pressing the button between the information and warning icons. On the left, a plot with the **prior distribution** for the Bayesian framework is displayed; on the right, either the *p*-value or the BF_{01} as a function of sample sizes for each simulation are plotted.

Block sampling
This program simulates data and performs **two-sided** tests for a single proportion in a (finite) **Bernoulli process** within block sampling designs. It reports the distributions of the corresponding *p*-values and Bayes Factors for the null (BF_{01}), in addition to a histogram of final sample sizes.

Usage

Choose the method used to define the stopping criterion for the sampling process (either *p*-values or Bayes Factors) and the stopping threshold for rejecting H_0 or gathering enough evidence in favor of H_1 respectively, as well as the minimum sample size, the size of the sample blocks, and the maximum number of blocks allowed. In both the frequentist and Bayesian frameworks, data are gathered in blocks of the specified size until the stopping criterion is met. Inflated type I error rates due to multiple comparisons are inherent to sequential sampling procedures under the frequentist approach. Those are controlled by default with the (overly conservative) Bonferroni correction, but the user is allowed to uncheck this option and not correct for multiple comparisons. Whenever the size of the blocks is set to *1*, the procedure corresponds to a sequential sampling design. Also, the user can change the specifications for the Beta prior distribution over the probability of success, which are set at *Beta(1,1)* by default.

Output

The **scatter plot on the left** displays log BF_{01} against log p-values, including the proportion of p-values (bottom right) and BF_{01} (top left) smaller than their corresponding boundaries (either the stopping threshold or the reference value, as selected). The green dots show simulations which stopped with the minimum sample size, the orange dots represent those that ended with the maximum sample size, and the purple dots correspond to simulations with other sample sizes (i.e., the stopping threshold was reached at some point during the sampling process). When all sample sizes are the same, all dots are colored green. Note that the offset of the relation between log *p*-values and log Bayes Factors depends on the sample size. In addition, specifying a Beta prior that concentrates mass relatively far from the true probability of success when H_0 holds turns the scatter plot into a two-limbed curve. The lower limb corresponds to those simulations where the evidence (i.e., data) gathered lies to the left of H_0, whilst the upper limb corresponds to evidence that lies to the right of such point.

The blue and red **histograms on the right** show the distribution of log BF_{01} (top) and *p*-values (bottom), with tags indicating the proportion of $p < b_1$, $p \geq b_1$, $BF_{01} < b'_1$ and $BF_{01} \geq$

b'_1, as appropriate. The extra threshold used for data collection (if set) is not shown in the histograms. The green histogram displays the distribution of final sample sizes.

References

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P., & Baribault, B. (2016). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, 30(4), 690-697.

García-Pérez, M. A. (2016). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*, 0013164416668232.

Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2017). Informed Bayesian *t*-tests. arXiv preprint arXiv:1704.02479.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology*, 60(3), 158-189.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, 19(6), 1057-1064.