Equivalence Testing

Any science that wants to be taken seriously needs to be able to provide support for the null-hypothesis. A limitation of null-hypothesis significance testing is that the null-hypothesis can be rejected, but not accepted. When you perform a statistical test, and the outcome is a *p*-value larger than the alpha level (α), the only formally correct conclusion is that the data are not surprising, assuming the null hypothesis is true. It is not possible to conclude there is *no* effect – our test might simply have lacked the statistical power to detect it.

In this assignment, we will examine how to provide support for the lack of a meaningful effect. We will do this using both Frequentist and Bayesian statistics. In Bayesian statistics, a Bayes Factor expresses the relative evidence for the alternative hypothesis compared to the null hypothesis. In Frequentist statistics, a traditional *t*-test does not allow you to make a statement about the absence of a meaningful effect. But with a simple twist, you can use *t*-tests or confidence intervals to test for the lack of a meaningful effect. This is known as *equivalence testing*.

Equivalence Hypothesis Testing

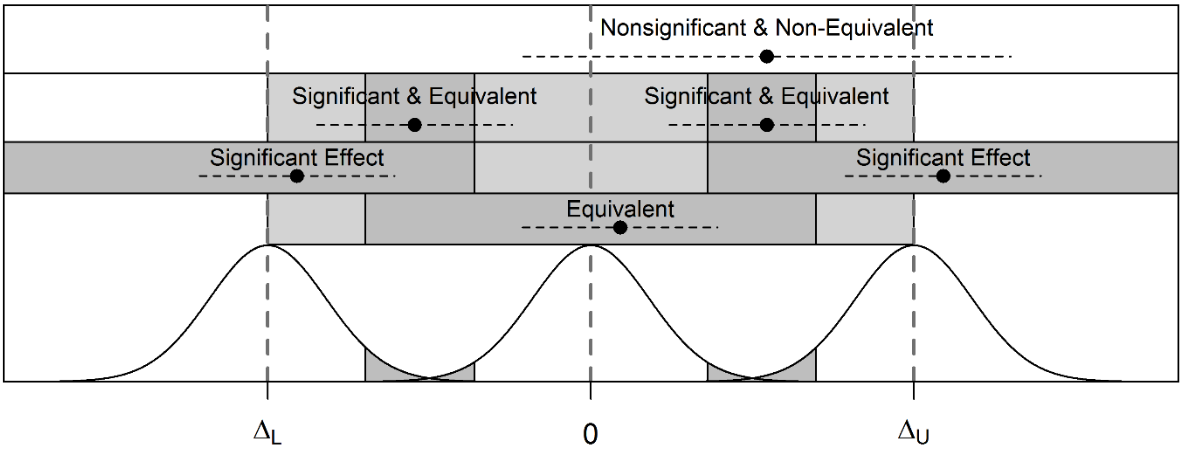
When researchers want to argue for the absence of an effect that is worthwhile to examine, they should test for *equivalence* (Wellek, 2010). In equivalence testing, an upper and lower equivalence bound is specified of the smallest effect size of interest. When effect sizes larger than these equivalence bounds can be statistically rejected based on the collected data, researchers can conclude the effect is not large enough to be worthwhile.

Equivalence testing originates from the field of pharmacokinetics where researchers might want to show that a new, cheaper, drug works just as well as an existing drug. A very simple equivalence testing approach is the 'two-one-sided t-tests' (TOST) procedure (Schuirmann, 1987), which tests whether the observed effect is larger than the lower equivalence bound, and smaller than the upper equivalence bound.

The basic idea of the TOST procedure is to reverse the null hypothesis and the alternative hypothesis. In Equivalence Hypothesis Testing the null hypothesis is that there *is* a true effect larger than a smallest effect size of interest (SESOI), and smaller than the largest effect size of interest. That's right – the null-hypothesis is now that there *is* an effect, and we are going to try to reject it (with a $p < 0.05$). The alternative hypothesis is that the effect is anywhere in the *equivalence range*. The equivalence range is a range of effect sizes that are deemed equivalent to the absence of an effect that is worthwhile to examine (e.g., Δ_L

= -0.3 to ΔU = 0.3, where Δ is a difference that can be defined by either standardized differences such as δ, or raw differences such as 1 scale point or 50 milliseconds). If the *p*-value for both tests indicates the observed data is surprising, assuming $-\Delta_L$ or Δ_U are true, we can follow a Neyman-Pearson approach to statistical inferences and reject effect sizes larger than the equivalence bounds.

When NHST and equivalence tests are used, there are four possible outcomes of a study: The effect can be significant (statistically different from zero), equivalent (statistically larger than Δ_L and smaller than Δ_U), significant and equivalent, or undetermined (neither statistically significant, nor statistically equivalent). In Figure 1, three hypothetical effect size distributions are visualized, one assuming the true effect size is 0, one assuming the true effect is Δ_L, and one assuming the true effect is Δ_U. To conclude equivalence, significant *p*-values should indicate the effect is both surprisingly larger than Δ_L (indicated by the darker grey tail area in right tail of the left distribution), and surprisingly smaller than Δ_U (indicated by the darker grey tail area in left tail of the right distribution). For symmetric equivalence bounds around zero, the 90% confidence interval around the observed effect size should exclude the Δ_L and Δ_U values (indicated by the grey area). A 90% confidence interval (1-2α) is used instead of a 95% confidence interval (1-α) because two one-sided tests (each with an alpha of 5%) are performed.



The traditional two-sided null hypothesis significance test rejects the null of the observed effect size is extreme enough to be surprising, assuming the null hypothesis is true. Effect sizes (indicated by black dots) should be more extreme than the upper or lower critical thresholds (2.5% of the middle distribution, in the left or right tail, indicated by dark-grey areas). The confidence interval around the effect size should not include 0 (and completely fall within light-grey areas). Effects can be significant *and* equivalent, for example when an effect is significantly smaller than the upper bound Δ_U, but also significantly larger than 0. Finally, an effect can be undetermined, or non-significant and non-equivalent, when the effect is not statistically different from 0, nor from the upper or lower bound (Δ_L or Δ_U).

The TOST procedure entails performing two one-sided tests to examine whether the observed data is surprisingly larger than a lower equivalence boundary (Δ_L), or surprisingly smaller than an upper equivalence boundary (Δ_L).

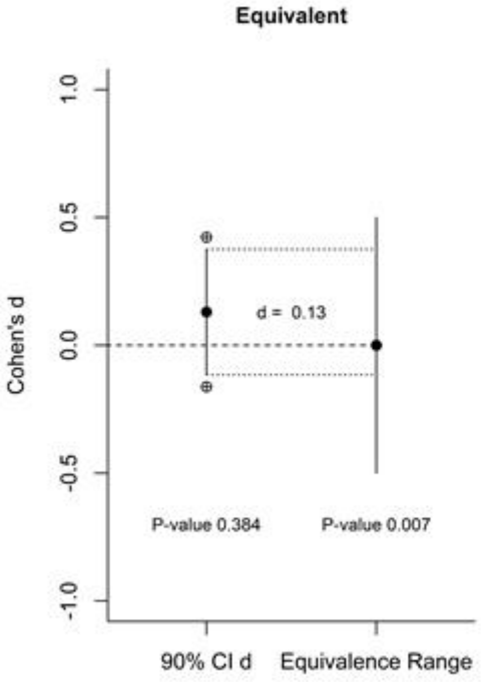
$$t_L = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_L}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ and } t_U = \frac{\bar{M}_1 - M_2 - \Delta_U}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad (1)$$

Where *M* indicates the means of each sample, *n* is the sample size, and σ is the pooled standard deviation:

$$\sigma = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}} \qquad (2)$$

The formulas are highly similar to the normal formula for the *t*-statistic. The difference between a NHST *t*-test and the TOST procedure is that the lower equivalence boundary Δ_L and the upper equivalence boundary Δ_U are subtracted from the mean difference between groups.

The TOST procedure for a t-test is identical to calculating a 90% CI around the effect, and check whether this 90% CI falls completely within the equivalence range. You can see this in the picture below: The 90% CI around the Cohen's d of 0.13 (the left vertical line) falls within the equivalence range (from d = -0.5 to d = 0.5) illustrated by the right vertical line.



Q1) When the 90% CI around an effect size falls within the equivalence range, the observed effect is statistically smaller than the smallest effect size of interest. Based on your knowledge about confidence intervals, and looking at the picture above, when you lower the equivalence range from -0.5 – 0.5 to -0.3 – 0.3, what is needed for the equivalence test to be significant (assuming the effect size estimate remains the same)?

A) A larger effect size

B) A lower alpha level

C) A larger sample size

D) Lower statistical power

To answer question 2, open the spreadsheet TOST t-test.xlsx. The spreadsheet will allow you to perform the TOST procedure for independent and dependent *t*-tests by filling in the means, standard deviations, and sample sizes for each group, and the equivalence range. It also provides the t-value, degrees of freedom, and *p*-value for the normal two-sided NHST.

Q2) Researchers often manipulate something they are interested in. To ensure their manipulation does not inadvertently alter participants' moods, they assess positive and negative emotions using the PANAS. Let's assume in one specific experiment, positive mood in one condition (M_1 = 4.55, SD_1 = 1.05, n_1 = 15) did not differ from the mood in the other condition (M_2 = 4.87, SD_2 = 1.11, n_2 = 15). The researchers conclude: "*Mood did not differ between conditions, t = 0.81, p =.42*'. Let's assume we consider any effect larger than *d*=-0.5 and smaller than *d* = 0.5 equivalent (even though *d* = 0.5 is actually a medium effect size!). Use the spreadsheet for the independent *t*-test, and fill in the 8 numbers (note: you need to fill in -0.5, not 0.5, as the lower equivalence bound). Were the authors correct in concluding mood did not differ between conditions, given the equivalence range of -0.5 to 0.5?

A) Yes

B) No

Q3) If we increase the sample size in the above example to 150 participants in each condition, and assuming the means and standard deviations remain the same, which conclusion would we draw?

A) Equivalent: The difference in mood is not statistically significant, and it is statistically equivalent.

B) Undetermined: The difference in mood is not statistically significant, and it is not statistically equivalent.

C) Not zero, and not meaningful: The difference in mood is statistically significant, and it is statistically equivalent.

D) Not zero, and meaningful: The difference in mood is statistically significant, and it is not statistically equivalent.

Deciding upon a smallest effect size of interest

When should you consider an effect too small to be meaningful? This is clearly a subjective choice, and what you consider meaningful can change over time. Ideally, you can determine the equivalence range either based on practical considerations or theoretical considerations. For example, if an advertisement campaign increases sales, but not enough to earn back the cost of the advertising campaign, you might decide an advertising campaign has an effect smaller than your smallest effect size of interest. Alternatively, you might believe based on available theories that a delay in response selection in a Stroop experiment where people respond verbally is larger than 20 milliseconds.

The choice for a smallest effect size of interest is also relevant when performing a power analysis. Because it is often unclear which effect size you can expect (after all, if you already knew the effect size, you would not need to do the experiment!) researchers sometimes make sure they have sufficient power for the smallest effect size they find interesting. Although it might be difficult to decide upon a SESOI, you might find it much easier to decide upon a maximum sample size you are willing to collect.

Open the spreadsheet TOST t-test.xlsx. The spreadsheet also allows you to perform a power analysis for the TOST procedure (based on an approximation, which is very close to the formally correct value).

Q4) We used an equivalence range of d=-0.5 to d = 0.5 in the previous question, even with a sample size of 15 participants in each condition. One might wonder if that gave us sufficient statistical power to detect equivalence. In the TOST power analysis section of the independent *t*-test, use an alpha of 0.05, a desired power of 0.8, an equivalence range from -0.5 to 0.5, and assume the true effect size is *d* = 0. How many participants in each independent group would you need to have 80% power to detect equivalence with this equivalence range?

Fill in the right number.

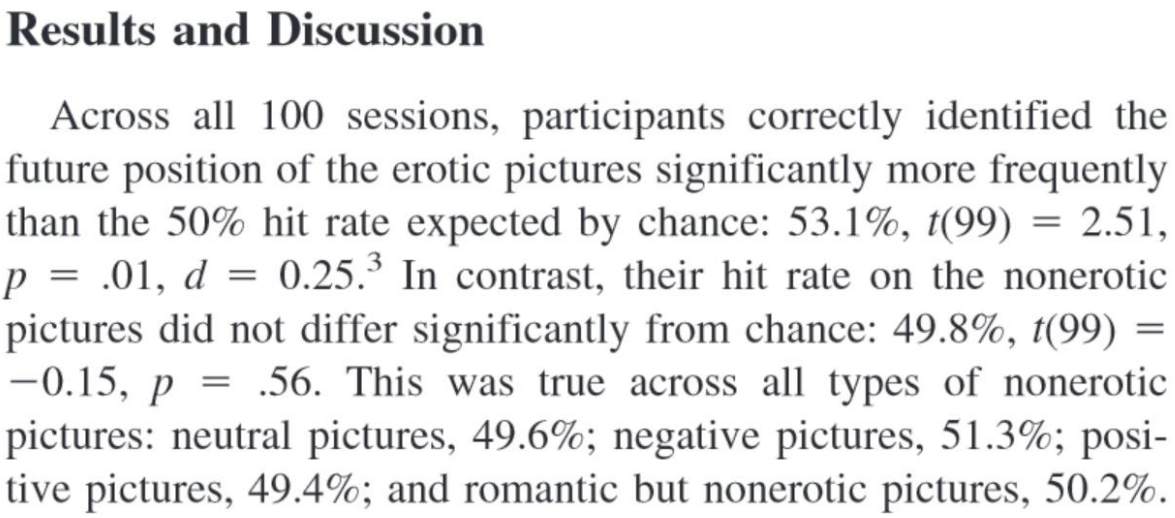
Q5) Change the equivalence range to -0.1 and 0.1. To be able to reject effect outside such a very narrow equivalence range, you'll need a large sample size. With an alpha of 0.05, and a desired power of 0.9 (or 90%), and assuming a true effect size of 0, how many participants would you need in each group?

Fill in the right number.

Supporting the null with Bayes Factors

Bayesian statistics, through its reliance on likelihoods, allows you to express the relative support for one hypothesis over the other hypothesis. You can conclude both that the alternative hypothesis is more plausible than the null model, or vice versa. Here we will use r code by Jeff Rouder to decide upon priors, calculate the posterior based on observed data, and calculate the Bayes Factor for a one-sample *t*-test.

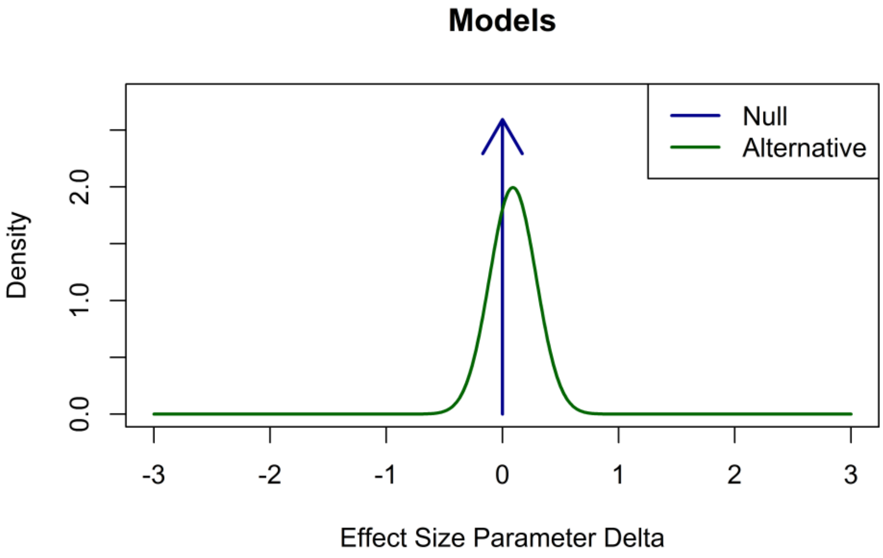
Let's take a look at the pre-cognition experiments by Daryl Bem (2011). In the first study, 100 participants classified the future position of pictures, before the computer had randomly chosen where the pictures would appear. For erotic pictures, this difference was statistically significant (see the result section below).



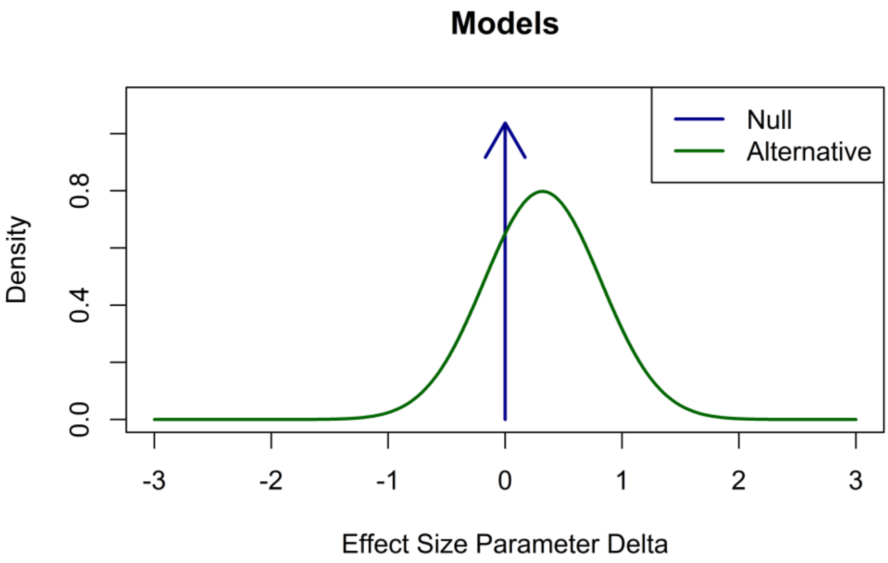
Results and Discussion

Across all 100 sessions, participants correctly identified the future position of the erotic pictures significantly more frequently than the 50% hit rate expected by chance: 53.1%, $t(99) = 2.51$, $p = .01$, $d = 0.25$.[3] In contrast, their hit rate on the nonerotic pictures did not differ significantly from chance: 49.8%, $t(99) = -0.15$, $p = .56$. This was true across all types of nonerotic pictures: neutral pictures, 49.6%; negative pictures, 51.3%; positive pictures, 49.4%; and romantic but nonerotic pictures, 50.2%.

We will calculate a Bayes Factor for this experiment. To do so, we need to choose a prior for the null-model, and a prior for the alternative. The null-prior we will use is known as a point prior of $d_z = 0$. Because Bayesian priors are normally distributions, and a $d_z = 0$ point prior means the density at this point goes up to infinity (which is illustrated by an arrow in the figures below). For the alternative hypothesis, we'll use a normal distribution for the effect size, which means we need to specify a mean and standard deviation that reflects our prior. We don't know which prior Daryl Bem had for the expected effect size, but with $N = 100$, an experiment has 90% power to observe a d_z of 0.32. So let's assume a prior belief of an effect size around d_z of 0.32. We also need to specify the standard deviation of the normal distribution – the higher the standard deviation, the wider the distribution. With a standard deviation of 0.2, the prior looks like:



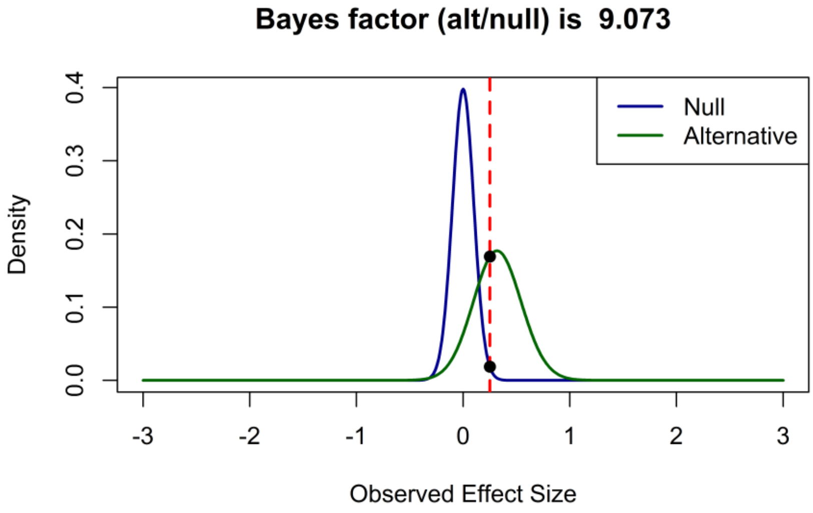
With a standard deviation of 0.5, the prior looks like the graph below:



Because pre-cognition effects are very unlikely to be so large as to include $d_z = 1$, and are much more likely small, let's use a $sd = 0.2$ for the prior here (but remember: this is your prior, so there are no right or wrong answers, as long as the prior correctly represents your belief!).

Open the 'Bayes Factor One-Sided T-test.R' script. Daryl Bem observed an effect size of $d_z = 0.25$ with N = 100. In line number 4 (see the black line numbers on the left of the source window) you can enter the sample size, and in line 5 you can fill in the **observed** effect size d_z. In lines 7 and 8 you can fill in the **expected** effect size for the prior (0.32), and the **expected** standard deviation for the prior (0.2).

Run the code, which will create two plots (remember you can browse through plots with the blue left and right arrows in the Rstudio interface). The first is a plot for the prior (the top Figure on the previous page), and the second plot looks like the figure below.



The vertical red dotted line indicates our observed effect size. The blue line gives the posterior for the null model, and the green line gives the posterior for the alternative model. We can see that given the observed effect size of d_z = 0.25, the alternative model is more likely than the null model, and the Bayes Factor tells us that given this data and our prior, the alternative model has become 9.073 times more likely.

Bayes Factors can also be used to provide support for the null hypothesis. Let's take a look at the reported effect for the remaining stimulus types in the study. Performance on these trials did not differ from guessing average, *t*(99) = -0.15. This corresponds to a Cohen's d_z of -0.015.

Q6) In line 5, change the observed effect size to -0.015. Run the code, using the same prior. What can we conclude? The title of the figure provides the Bayes Factor for the alternative over the null (BF_{10}). Remember that a Bayes Factor of 1 means both models are equally likely – values smaller than BF_{10} = 1 mean the null hypothesis is more likely. You can reverse the BF_{10} to reflect how much more likely the null is compared to the alternative (BF_{01}) by computing 1/BF. When the Bayes factor is smaller than 0.333 or larger than 3, the Bayes Factor can be interpreted as modest support. When the Bayes factor is smaller than 0.1 or larger than 10 the Bayes Factor can be interpreted as strong support. When the Bayes factor falls within the 0.333 to 3 range, it is considered too weak support for either hypothesis to draw a conclusion on the data.

A) The alternative model is more likely than the null model, with a BF_{10} of 0.147

B) The alternative model is more likely than the null model, with a BF_{10} of 14.70

C) The null model is more likely than the alternative model, with a BF_{10} of 0.147

D) The null model is more likely than the alternative model, with a BF_{10} of 14.70

In 2015 Bem and colleagues published a meta-analysis of pre-cognition effects. In this meta-analysis, the authors argue pre-cognition has a meta-analytic effect size of Hedges' *g* = 0.09.

Q7) In line 7, change the prior to an effect size of dz_prior = 0.09. Given this prior, can we still conclude that the Bayes Factor provides support for the null model for the remaining conditions, where the statistical test was *t*(99) = -0.15, d_z = -0.015?

A) With a BF_{10} = 7.365, the data now actually provide support for the alternative model, compared to the null model.

B) With a $BF_{10} = 0.405$, we can no longer conclude the data provide support for the null model, compared to the alternative model.

C) With a $BF_{10} = 0.405$, the data provide support for the null model, compared to the alternative model.

Conclusion

It is important to be able to provide support for the null-hypothesis, if you want to be able to test theories that predict no effect, or when you want to be able to falsify theory that predict an effect. According to Popper, falsifiability is the demarcation criteria between science and pseudoscience, so being able to falsify hypotheses is very important. According to Lakatos, even though we rarely outright reject our hypotheses, we enter a degenerative research line when our alternative hypothesis is rejected. Finally, being able to conclude an effect is too small to be worthwhile to examine enables us to improve our statistical inferences.

You will always have to make assumptions about the alternative hypothesis, either be specifying an equivalence region consisting of a range of effect sizes you find meaningful, or by specifying a prior distribution. You can use either Frequentist or Bayesian approaches to test absence of evidence (or even both!). Don't simply conclude that a $p > 0.05$ means there is no effect – instead, provide quantitative arguments for the conclusion that there is no effect by using equivalence tests or Bayes Factors.