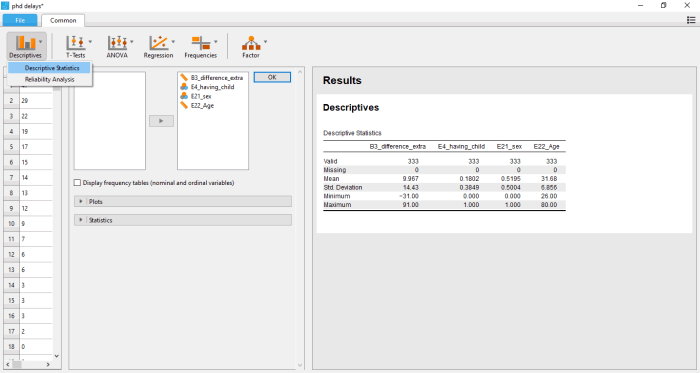
C:\Users\lionb\AppData\Local\Microsoft\Windows\INetCache\Content.Word\RvdS-logo-zwartig-1.png

**Stats Training  
JASP: how to get started – Answers and Explanations**

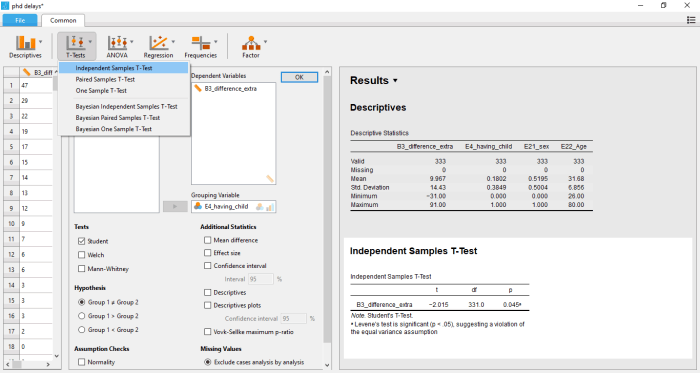
**Frequentist T-Test**

**Step 1.** Start by downloading the dataset *'phd delays.csv'*and loading this datafile into JASP. Do so by clicking on File -> Open. Have a look at the descriptive statistics. Has all data been loaded in correctly?



| **Descriptive Statistics** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **B3\_difference\_extra** | | **E4\_having\_child** | | **E21\_sex** | | **E22\_Age** | |
| **Valid** |  | 333 |  | 333 |  | 333 |  | 333 |  |
| **Missing** |  | 0 |  | 0 |  | 0 |  | 0 |  |
| **Mean** |  | 9.967 |  | 0.1802 |  | 0.5195 |  | 31.68 |  |
| **Std. Deviation** |  | 14.43 |  | 0.3849 |  | 0.5004 |  | 6.856 |  |
| **Minimum** |  | -31.00 |  | 0.000 |  | 0.000 |  | 26.00 |  |
| **Maximum** |  | 91.00 |  | 1.000 |  | 1.000 |  | 80.00 |  |
|  | | | | | | | | | |

**Step 2.** We will first do a frequentist T-test before we move to the Bayesian alternative including Bayes Factors. Perform an independent samples T-test and interpret the output. Using a significance criterion of 0.05, is there a significant difference between the two groups? Recently, a group of 72 notable statisticians proposed to shift the significance threshold to 0.005 ([Benjamin et al. 2017](https://osf.io/preprints/psyarxiv/mky9j)). How does your conclusion change if you follow this advice?



| **Independent Samples T-Test** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **t** | | **df** | | **p** | |
| B3\_difference\_extra |  | -2.015 |  | 331.0 |  | 0.045 | ᵃ |
|  | | | | | | | |
| Note.  Student's T-Test. | | | | | | | |
| ᵃ Levene's test is significant (p < .05), suggesting a violation of the equal variance assumption | | | | | | | |

Using the conventional significance criterion of 0.05, Student’s T-test suggests that the difference in Ph.D. delay between those who had at least one child up to 18 years old and those who didn’t is indeed significantly different from 0. Given a population where the difference is exactly 0, a T-statistic that’s equal to or more extreme than +/- 2.015 would appear in 4.5% of the cases. Following the advice of [Benjamin et al. (2017)](https://psyarxiv.com/mky9j), there would not be enough evidence to reach this conclusion. In this line of reasoning, one would not reject the null hypothesis that the difference in Ph.D. delay between the two groups is exactly 0.

Note that the assumption of equality of variances (in the population) is likely to be violated here. As the group-specific descriptives document, standard deviations of the two groups obtain the values 13.91 and 16.30 in our specific sample. To correct for this, one could choose to conduct a Welch’s test, which drops the equality of variance assumption.

**Step 3.** In the window that opens, you can easily request additional statistics and group-specific descriptives. Request discriptives, the mean difference between the two groups, a 95% confidence interval and the effect size (*Cohen's d*). Interpret your newly obtained output. Which group took longer to complete their Ph.D. trajectory? What about the size of the effect?

Requesting descriptive statistics within the ‘T-tests’ box (under ‘Additional Statistics’), one can see that those who had at least one child had a mean delay of 13.350 months, while those who didn’t were on average 9.223 months delayed. The mean difference in Ph.D. delays between the two groups is 4.127 months. If repeatedly sampled, an interval between 0.097 months and 8.156 months would include the true population value 95% of times. Both through substantial effect size interpretation as through interpreting Cohen’s d, one could treat this as a small effect.

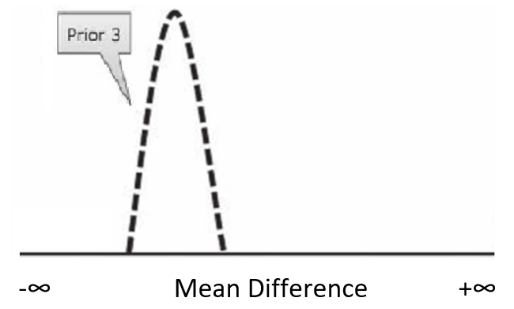
| **Group Descriptives** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Group** | | **N** | | **Mean** | | **SD** | | **SE** | |
| B3\_difference\_extra |  | 0 |  | 273 |  | 9.223 |  | 13.91 |  | 0.842 |  |
|  |  | 1 |  | 60 |  | 13.350 |  | 16.30 |  | 2.104 |  |
|  | | | | | | | | | | | |

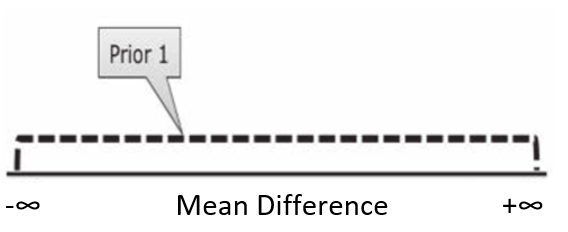
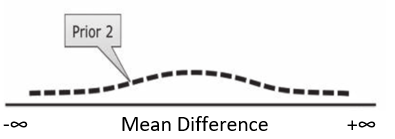
| **Independent Samples T-Test** | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | **95% Confidence interval** | | | |
|  | | **t** | | **df** | | **p** | | **Mean Difference** | | **SE Difference** | | **Cohen's d** | | **Lower** | | **Upper** | |
| B3\_difference\_extra |  | -2.015 |  | 331.0 |  | 0.045 | ᵃ | -4.127 |  | 2.048 |  | -0.287 |  | -8.156 |  | -0.097 |  |
|  | | | | | | | | | | | | | | | | | |
| *Note.*  Student's T-Test. | | | | | | | | | | | | | | | | | |
| ᵃ Levene's test is significant (p < .05), suggesting a violation of the equal variance assumption | | | | | | | | | | | | | | | | | |

**Bayesian T-Test**

**Step 5.** Do you have any prior knowledge on how parenthood affects the achievement of individual mid-term life goals (like a Ph.D.)? Think about a reasonable prior distribution. How would your prior distribution for the effect of parenthood look like? Give reasons for your choice.

There are different ways of how to reach a prior distribution for parenthood. For example, one could rely on past studies that investigated the same or similar research questions. Also, a prior distribution can be formulated with reference to theoretical literature. If no prior knowledge is acquired, one can specify a flat distribution won’t give some parameter values more weight than others and won’t influence the posterior distribution (Prior 1). Such flat distributions are often called *‘uninformative’* distributions (although it can be disputed if this term is adequate or not). Also, one could argue that (e.g. based on empirical studies or theoretical reasoning) one is unsure about the effect of parenthood, but since the overwhelming number of effects in Psychology should be small, the a priori probability of the effect of parenthood to be small as well is quite large. Thus, one could formulate a symmetrical prior distribution that peaks at 0 (Prior 2). Lastly, one could aquire prior knowledge of an effect in either direction in thus formulate a prior distribution like displayed in Prior 3.



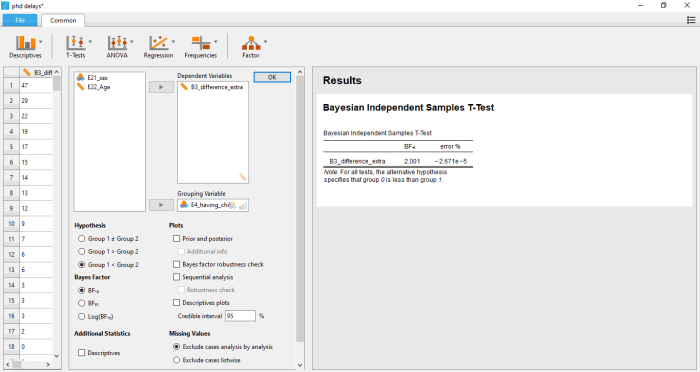


**Step 6.** Now, perform your Bayesian T-test in JASP. Note that a crucial feature of the JASP software is that users *cannot specify individual prior distributions*. Instead, JASP uses default priors, which peak at 0 (no effect). In this tutorial, you are strongly advised not to use default prior settings in your own research. However, since this is the only option available in JASP, proceed by performing a Bayesian independent sample T-test. To do so, set your dependent and independent variable in the respective boxes and choose a hypothesis to test (right below).

Interpret your output. What does the Bayes Factor mean? How does its interpretation differ from the classical p-value above? What are potential pitfalls to the interpretation of a Bayes Factor? After having collected your own ideas, have a look at [Konijn et al. (2015)](https://dspace.library.uu.nl/handle/1874/329896) for further reasoning.

Like with the frequentist T-test, different hypothesis can be tested in the Bayesian counterpart. Since we already collected some evidence that parenthood does indeed lead to larger Ph.D. delay, here the hypothesis Group 1 < Group 2 is tested. Instead of a p-value, the Bayesian T-test uses a Bayes Factor in order to reach conclusions. A Bayes Factor of 2.001 emerges. Thus it is two times as likely that the collected data comes from a population where the mean of Group 1 is indeed smaller than the mean of Group 2 than the other way around. Simultaneously, the Bayes Factor for the hypothesis Group 1 > Group 2 is 0.054. Note that this interpretation is fundamentally different from the one of the p-value, since in Bayesian hypothesis testing, evidence can actually quantified *in favor* of hypotheses. In frequentist statistics, tests are exclusively designed to quantify evidence *against* a pre-specified null hypothesis that is sought be rejected (but often unrealistic or non-interesting anyways).

A Bayes Factor of 2 is often referred to as anecdotal evidence. Many people set the threshold for substantial evidence at 3. A potential pitfall of the Bayes Factor is to treat the number 3 as a fixed, objective criterion that separates findings from non-findings in a similar manner as the p-value does in frequentist statistics. Rather, it should be interpreted *context-specific* and substantial reasoning should argue in favor or against a strong effect. Consult [Konijn et al. (2015)](https://dspace.library.uu.nl/handle/1874/329896) for more information about this topic.

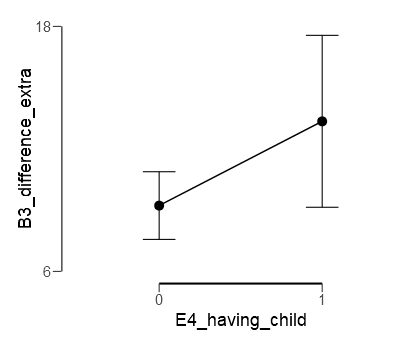


**Step 7.** As a last step, get a better understanding of the evidence and its robustness by ticking 'Descriptives' under 'Additional Statistics' and all available boxes under 'Plots'. By doing this, a large amount of output is created.

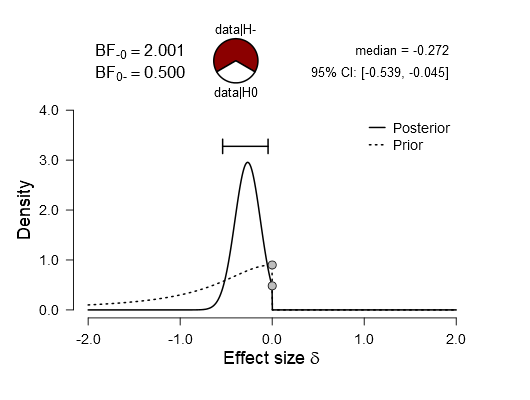
Work through the entire output and try to understand every single diagram and number. If you have difficulties with some interpretation, consult the solutions file under 'Answers and Explanation' on the top of this exercise. How would you evaluate the effect of parenthood based on your new insights? How robust is this effect?

| **Group Descriptives** | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | **95% Credible Interval** | | | |
|  | | **Group** | | **N** | | **Mean** | | **SD** | | **SE** | | **Lower** | | **Upper** | |
| B3\_difference\_extra |  | 0 |  | 273 |  | 9.223 |  | 13.91 |  | 0.842 |  | 7.566 |  | 10.88 |  |
|  |  | 1 |  | 60 |  | 13.350 |  | 16.30 |  | 2.104 |  | 9.139 |  | 17.56 |  |
|  | | | | | | | | | | | | | | | |

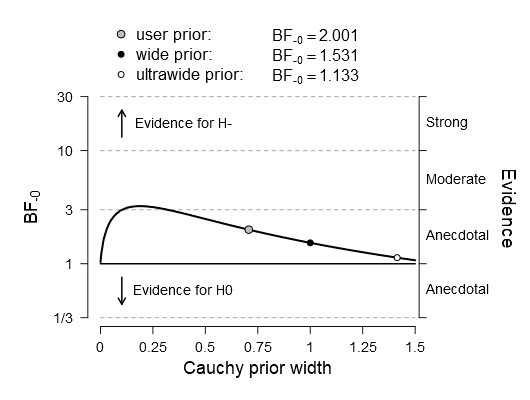
The descriptive plot (below) displays both group means and their 95% credible interval. Note that the credible intervals overlap quite substantially. This speaks in favor of a small or non-relevant difference.



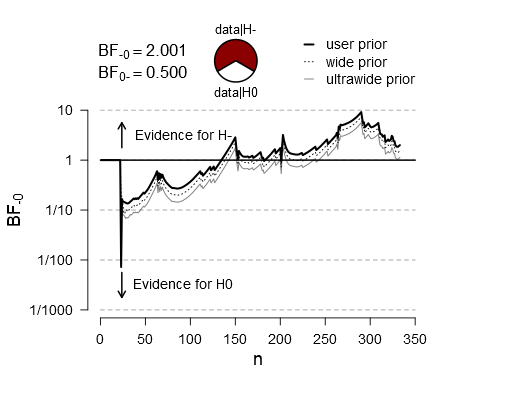
The first inferential plot gives you a better understanding of your data and entails a lot of information. At first, you see the Bayes Factors again that we already evaluated in Step 6 (top left) as well as a graphical representation of both Bayes Factors in the form of a pie chart. Also graphically, you can see that the data is twice as likely under the hypothesis of Group 1 < Group 2 rather than under the hypothesis of Group 1 > Group 2. Below, you see a plot of your prior and posterior distribution expressed in effect size 𝛿. Here you can see which prior distribution JASP was automatically using behind the scenes. Based on the data, it chose a Cauchy prior distribution that gave weight to negative effect sizes and peaked at 0. The obtained posterior distribution is thus restricted to negative values. In the top right corner, some of its summary statistics are displayed. The posterior distribution obtains a median of -0.272 and a 95% credible interval between -0.539 and -0.045. Thus, there is a 95% probability that the effect in the population is in this interval. You are 95% certain that the effect size 𝛿 lies between -0.045 and -0.539.



The summary of the Bayes Factor and the relevant distributions is followed by a Bayes Factor robustness check. It shows you how your Bayes Factor (y-axis) would change if the Cauchy prior width (x-axis) is changed. As you can see, a default width of slightly less than 0.75 was used. A smaller width gives effect sizes close to 0 much more weight and is thus way more peaked at 0. A larger width will distribute effect size weights wider and you will see less of a peak. Note that if increasing prior width, evidence would shrink. Simultaneously, a width of 0 would result in a Bayes Factor of 1. In general, you can see that (reasonable) changes in prior width would not affect the substantial conclusion drawn out of your Bayes Factor.



Lastly, JASP provides you with a sequential analysis. You can see how your conclusions would change through the data collection process if you would re-analyze your data after adding every participant. As you can see, up to the first 100 participants, data seemed to be in favor of Group 1 > Group 2. Analyzing the first 300 participants, a Bayes Factor of almost 10 would have emerged. Using the complete sample, the value 2.001 was achieved.



**References**

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. (2017, July 22). [Redefine statistical significance.](https://psyarxiv.com/mky9j) Retrieved from psyarxiv.com/mky9j

JASP Team (2017). [JASP (Version 0.8.2)](https://jasp-stats.org/)[Computer software].

Konijn, Elly A., van de Schoot, Rens, Winter, Sonja D., Ferguson, Christopher J. (2015). [Possible Solution to Publication Bias Through Bayesian Statistics, Including Proper Null Hypothesis Testing.](https://dspace.library.uu.nl/handle/1874/329896) Communication Methods and Measures, 9(4): 280 - 302.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J. and van Aken, M. A.G. (2014). [A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research](http://onlinelibrary.wiley.com/doi/10.1111/cdev.12169/abstract). Child Dev, 85: 842–860. doi:10.1111/cdev.12169

van de Schoot R, Yerkes MA, Mouw JM, Sonneveld H (2013). [What Took Them So Long? Explaining PhD Delays among Doctoral Candidates.](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068839) PLoS ONE 8(7): e68839. https://doi.org/10.1371/journal.pone.0068839