Frequentist vs Bayesian inference

Dr Benjamin T. Vincent

NOTE: This session is an amalgamation of 2 previous lectures that have looked at Hypothesis Testing in detail, and Bayesian inference in detail. This session is an attempt to combine these topics. So this session is not simply a combination of those previous sessions, we cover different examples and have different emphases. So it may be that in working through this material in your own time, going back to refer to these other sessions could be very useful as they go into more depth in certain places.

🚨 TODO's for Ben to further improve this notebook 🚨

- Pimp the Bayes Factor plot.
- Add animated graphic where you can chance the Cauchy prior (mean and variance) to see the effect on the posterior.
- Firm up my wording on the interpretation of Bayes Factors - evidence for H_0 and H_1 vs evidence of H_0 under prior and posterior.

The goal of this session

This session is going to give an introduction to Bayesian statistics. We will start with a motivating example, using one of the built in datasets included in JASP, and highlight the differences in the outputs.

Frequentist output

This is the kind of output you get from JASP when running a traditional Frequentist independent samples T-Test. We can see that we get a few values:

- t: this is the t-statistic value
- df: is the degrees of freedom
- p: is the p-value, indicating the level of statistical significance
- Cohen's d: this is the effect size, which gives us an indication of how meaningful the

differences are between groups.

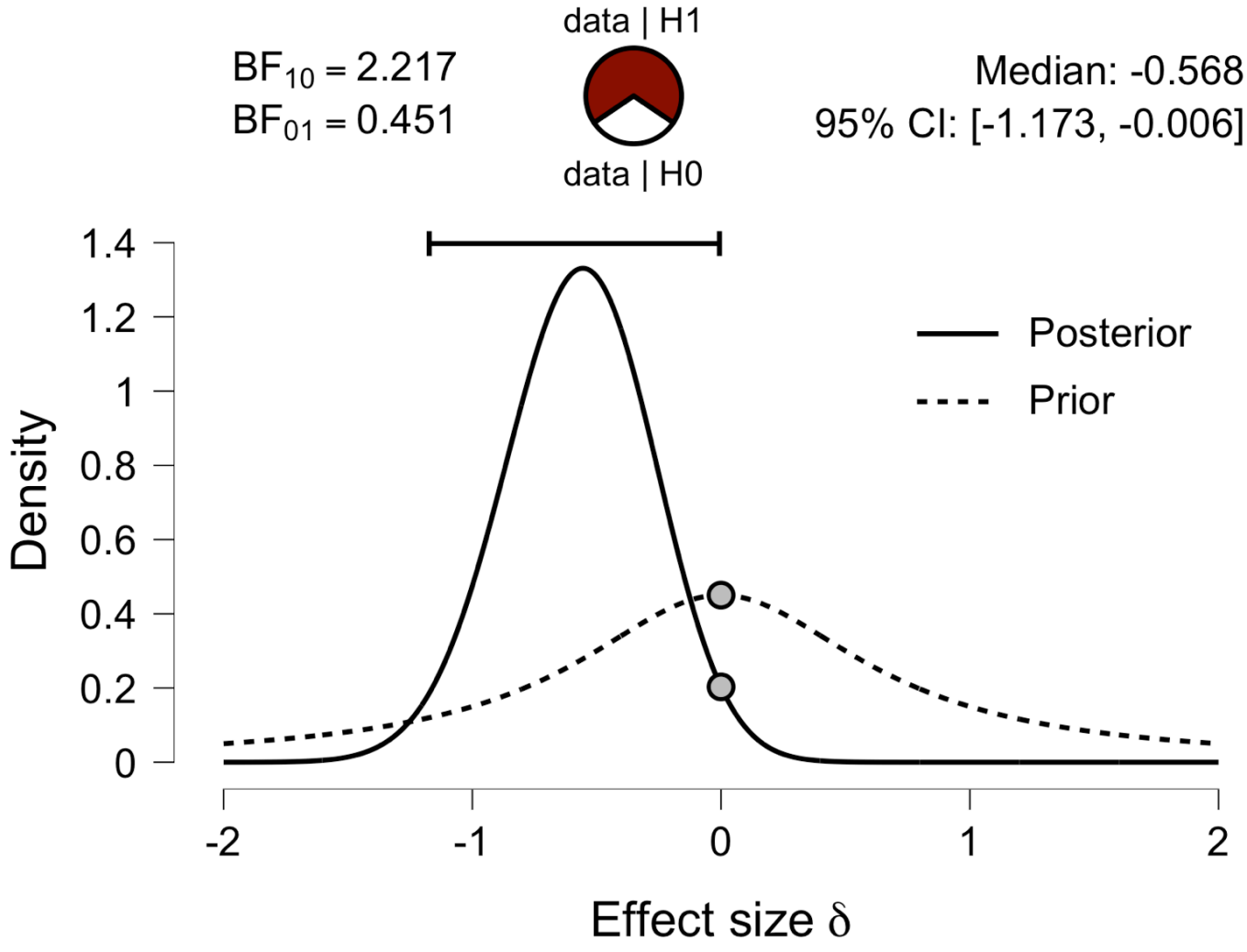
Independent Samples T-Test ▼

	t	df	p	Cohen's d
drp	−2.267	42	0.029	−0.684

Note. Student's t-test.

Bayesian output

This figure below shows the kind of output you get from a Bayesian version of the same statistical test in JASP. The key result here is what is known as a Bayes Factor, $BF_{10} = 2.217$. But we also get a nice graphical output. At the moment this might look all very mysterious, but by the end of this session you should understand it 🙂.



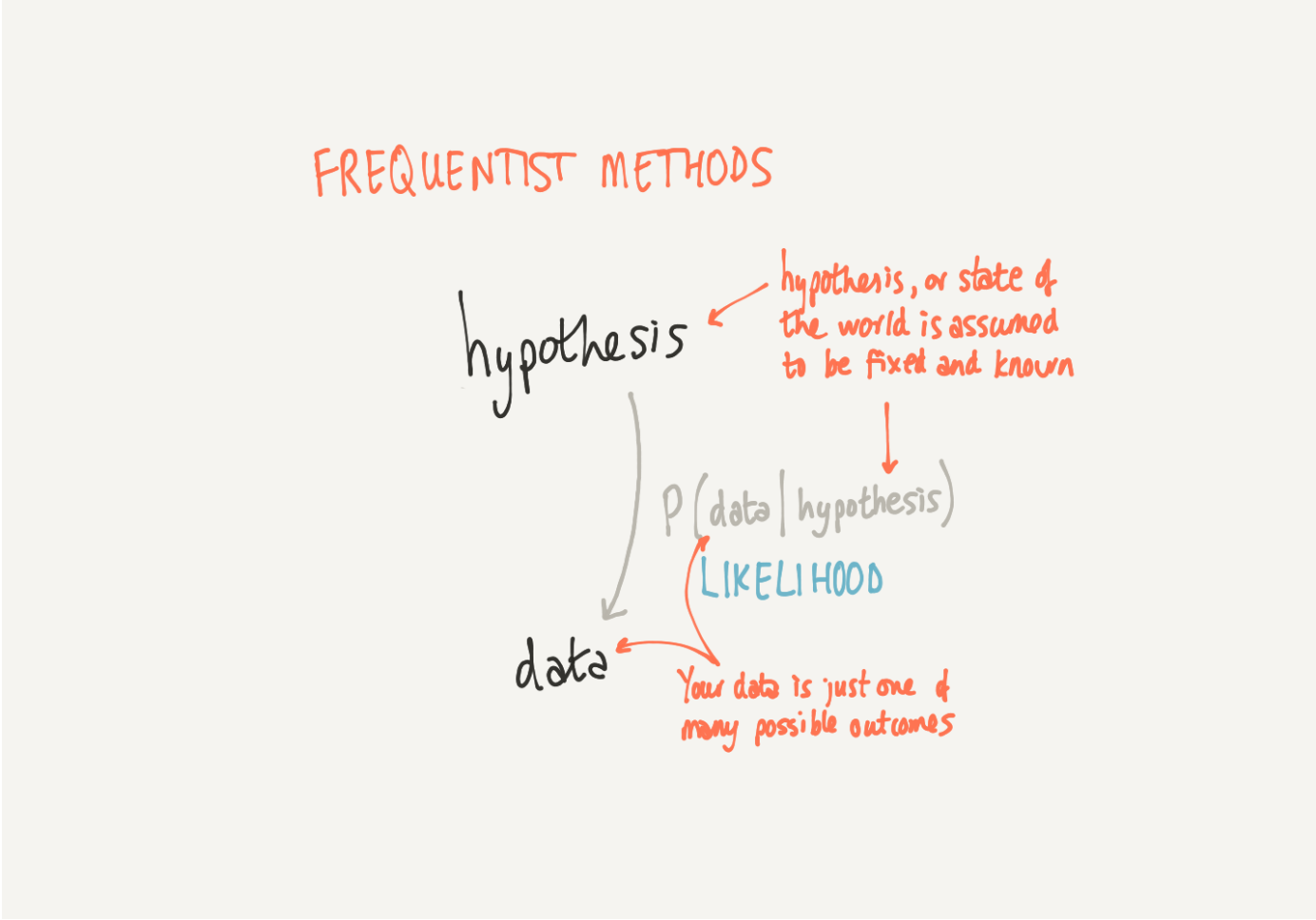
The focus of this session

Having seen this motivating example and had a brief look at the differences in the outputs of a Frequentist and Bayesian output, our focus will be digging into the differences of the Frequentist and the Bayesian approaches.

Part I: Conceptual differences

The Frequentist approach

Under the Frequentist approach, the hypothesis (or state of the world) is assumed to be fixed. The data you collect from your experiment is assumed to be stochastic. That is, a mere instance of one possible dataset that you could have observed. You can imagine theoretically running the same experiment many many times, and each time you would observe slightly different data.



The data generating process

The downward arrow in this diagram symbolicalises what we know about what processes occured that led to our dataset - we can call this the data generating process. For the moment we will keep this somewhat vauge until we look at a concrete example below. Having said that, it is not a difficult idea. It could be as simple as saying: *I believe the data from group 1 and group 2 are both normally distributed around zero*.

Making claims about the data with the likelihood

We can simulate what we would expect to see *if* the null hypothesis is true. You can imagine that this would give us slightly different datasets, and thus slightly different t-statistic values. This basically describes what we would expect to see given the null hypothesis is true. This is what is known as the likelihood, $P(\text{data}|\text{H}_0)$.

What is the likelihood? The most important thing you need to know about the likelihood is that it is a claim about the probability of observing some particular data according to the particular hypothesis under consideration, in this case H_0. This is both a good thing and a bad thing.

The likelihood is very useful in telling us how likely it is that we would have observed something (like a t-statistic) under the null hypothesis.

- If $P(\text{data}|\text{H}_0)$ is a very low number, then we know that the data is not very consistent with H_0
- If $P(\text{data}|\text{H}_0)$ is a high number, then we know that the data is pretty consistent with H_0.

You can see that this is useful. Frequentists use this in order to make claims about how likely or unlikey the data is under the null hypothesis. If the data is sufficiently unlikely to have been observed under H_0, then we could try to make a claim (a statistical inference) about the data.

Questions

But there are some questions that spring to mind here:

- What does 'sufficiently unlikely' mean? How unlikely should the data be under H_0 for us to make a claim?
- What kind of claim can we really make? Can we make claims about our hypothesis, or just about the data?

Why are claims about the data unsatisfactory?

- Why isn't this good enough?
- Why do we have to bother about the Bayesian approach?
- Isn't it sufficient to observe data that are very improbable under H_0 and make scientific claims based on this?

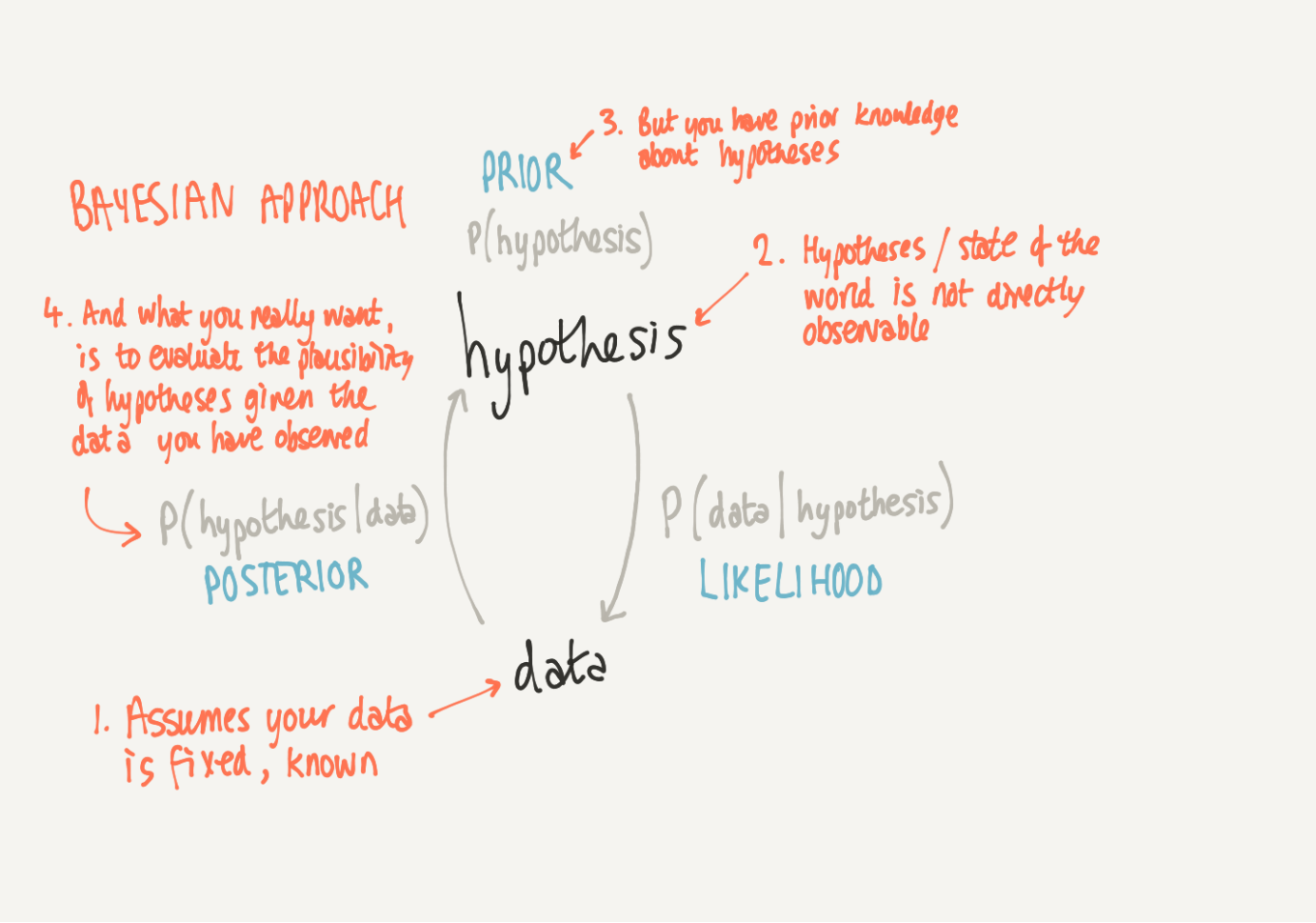
Hint

Some thoughts:

- We often want to make claims about the credibility of hypotheses. We cannot do this under the frequentist approach
- Rejecting H_0 is not a very strong statement, rather than rejecting what is not the case, maybe it would be better if we could make statements about what is the case

The Bayesian approach

The Bayesian approach builds upon what we have looked at before, namely the data generating process. However, Bayesians want to make claims about how credible hypotheses are given the data $P(\text{hypothesis}|\text{data})$. This is known as the **posterior**, and in some ways it is the opposite of the likelihood.



Understanding how the likelihood and prior are different

In a way, it is clear that:

- the likelihood, $P(\text{data}|\text{hypothesis})$, is about how consistent a set of data are with a given hypothesis and that
- the posterior, $P(\text{hypothesis}|\text{data})$, is about how consistent a hypothesis is with some data.

But it is good to dig into this. A great example can be found in Understanding Psychology as a Science by Zoltan Dienes. We can consider the probability that someone will have died within 2 years, given they have had their head bitten off by a shark. It is not often we can be 100% sure of something, but this is a good example where we would be extremely confident:

$$P(\text{dead within 2 years}|\text{head bitten off by shark}) = 1$$

But now we could consider the inverse of this...

Question: What is the probability that someone who died within the last few years had their head bitten off by a shark? Roughly.

Hint

$P(\text{head bitten off by shark}|\text{died within 2 years}) = ?$

So this is a good demonstration that just because we know $P(\text{data}|\text{hypothesis})$, that really doesn't tell us about $P(\text{hypothesis}|\text{data})$.

How do we work out the posterior?

If a Bayesian wants to make claims about hypotheses, not about how consistent data is with a hypothesis, then how do they do that?

Funnily enough, with Baye's Equation.

$$\underbrace{P(\text{hypothesis}|\text{data})}_{posterior} \propto \underbrace{P(\text{data}|\text{hypothesis})}_{likelihood} \cdot \underbrace{P(\text{hypothesis})}_{prior}$$

Let's just recap, because we are getting into equations, but what we are doing is conceptually very simple...

Baye's equation simply says... to work out what we believe about a hypothesis (given some data), all we need to do is to mulitple the probability of seeing that data given the hypothesis is true, mulitplied by the probability that the hypothesis is true.

Baye's equation and the shark example

Let's have a quick look to see how Baye's equation can help us calculate the posterior, $P(\text{head bitten off by shark}|\text{died in last 2 years})$. We will do this using the prior and the likelihood.

The **likelihood** (which we can also call our data generating model) specifies $P(\text{died in last 2 years}|\text{head bitten off by shark}) = 1$. We already discussed how this is equal to 1, because it is certain that you will be dead within 2 years of having your head bitten off by a shark.

```
likelihood = 1.0
```

The **prior** is something that we can change using this slider below, but it will be a number between 0 and 1, representing our prior belief that someone who died within 2 years had their head bitten off by a shark:

Prior: $P(\text{shark bit head off})$ = 0.0

Posterior: $P(\text{shark bit head off}|\text{died in last 2 years})$ = 1 $\times$ 0.0 = 0.0

This is a simple example, but it shows how by using Baye's equation, we can multiply the likelihood by the prior to arrive at our posterior.

Important point

This is pretty cool! We have broken free of the shakles of confinement of making claims about data. Now we can make claims about hypotheses! This is what we are very often interested in doing as scientists.

Now we will go forth and apply what we have learnt to our T-Test case study.

Part II - Frequentist T-Test

So far we have introduced our motivating example of an independent samples T-Test. We saw that the output you get in JASP or Frequentist and Bayesian version of the test are rather different. We then started on our journey of understanding the differences between the approaches on a conceptual level. Of particular relevance for the Frequentist approach was the likelihood term, which can be described as the data generating process.

Here we return to our T-Test example and show, through simulation approaches, exactly what the Frequentist approach does. As I mentioned in the beginning, I have a existing material on Hypothesis testing, there will be some overlap, but you may wish to refer to that for more detail, and a different example.

The 'logic' of frequentist hypothesis testing

There are many different statistical tests. It can be confusing which one you should use and why.

What are all the different statistics and where do they come from? We can think about this in a simpler way:

By thinking about simulating experiments (rather than doing maths), we can see that there is only one core procedure used in the hypothesis testing approach.

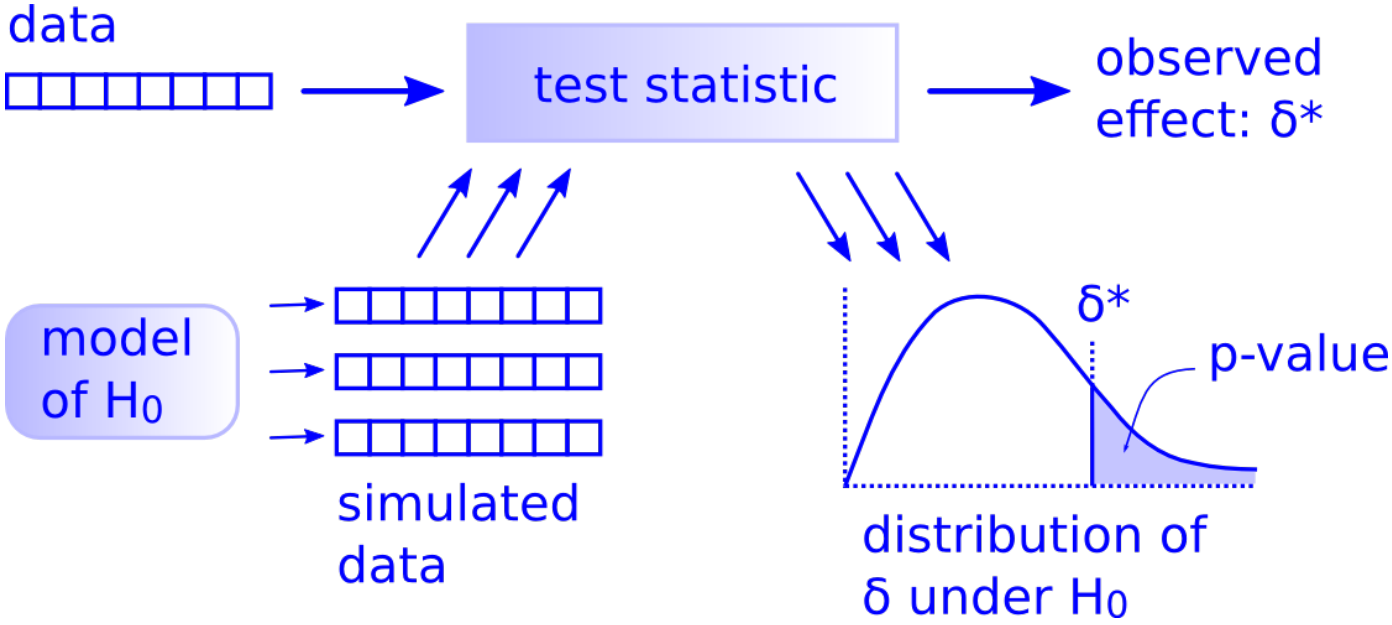


Figure by Allen Downney from this blog post http://allendowney.blogspot.com/2016/06/there-is-still-only-one-test.html

The data is relatively straightforward. It is just a table of data that you have collected. You might have columns for different measures and rows for each observation.

In order to test your hypothesis, you construct a test statistic. This takes your data and reduces it to a single number which captures the size of the effect that you are interested in. Example hypotheses could be:

- The mean of group A is different than group B.
- The median of group A is different than group B. (Might be more robust to outliers).
- The mean of group A is higher than group B. (A directional hypothesis).
- Each individual increased their score from condition A to condition B. (As in a repeated measures context).
- Etc.

We could come up with a whole range of statistics (ways of converting a dataset into an effect size) which test different hypotheses. We will look at an example below.

Then the heart of the hypothesis testing approach comes in... You need to work out what are the chances of seeing an effect (a test statistic), for your particular dataset, as big as this by chance. If

your effect size is low, then it is more consistent with chance as compared to if your effect size is large.

So how do we do this, specifically? We define a null hypothesis, H_0, which states what we would expect to see if there is no effect. This is the 'top down' or deductive approach, where we say, *assuming the null and a certain experiment structure and a certain sample size, then I expect to see this.*

This allows you to work out the consistency (also known as the likelihood) of the data given the null hypothesis, which we could call $P(data|H_0)$.

Frequentist independent T-Test by hand

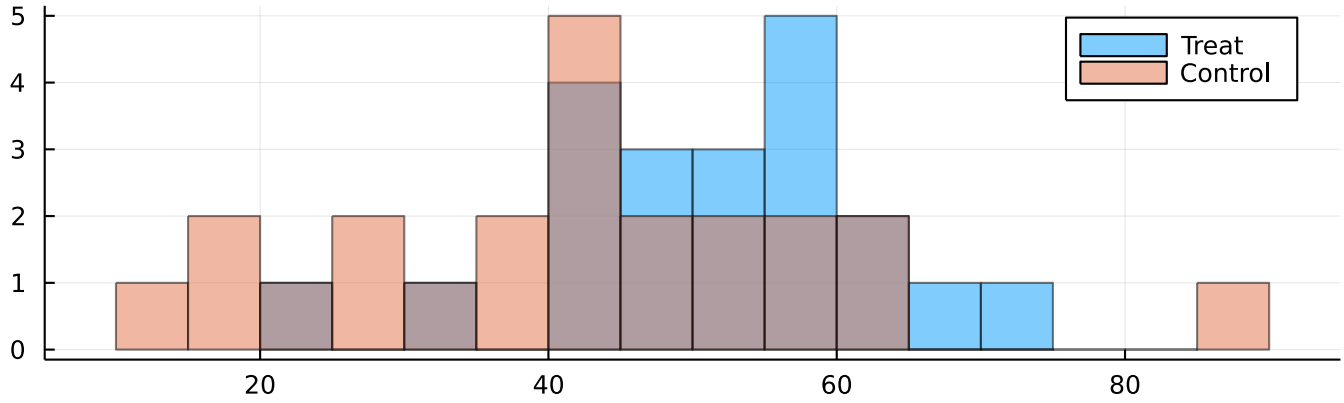
So we are going to follow this approach and simulate what we would expect to see, assuming H_0 . Before doing that, we need to understand a bit about our dataset.

```
using DataFrames
```

```
data =
```

	id	group	g	drp
1	1	"Treat"	0	24
2	2	"Treat"	0	56
3	3	"Treat"	0	43
4	4	"Treat"	0	59
5	5	"Treat"	0	58
6	6	"Treat"	0	52
7	7	"Treat"	0	71
8	8	"Treat"	0	62
9	9	"Treat"	0	43
10	10	"Treat"	0	54
	more			
44	44	"Control"	1	85

```
data = CSV.read("Directed_Reading_Activities.csv", DataFrame)
```



Let's see how many participants we have in the control and treatment groups

```
N_treat = 21
```
```
N_treat = sum(data.group .== "Treat")
```

```
N_control = 23
```
```
N_control = sum(data.group .== "Control")
```

Now let's calculate the t-statistic (equal variances assumed) for our data. As I mentioned, you could do this with a much simpler statistic (e.g. difference in means), but the t-statistic has some more desirable properties in how it is affected by group variance etc, so we will look up the t-statistic equation and use that.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

```
function t_statistic(x₁, x₂)
    return (mean(x₁) - mean(x₂)) /
    (pooled_std(x₁, x₂) * sqrt(1/length(x₁) + 1/length(x₂)))
end;
```

```
function pooled_std(x₁, x₂)
    n₁, n₂  = length(x₁), length(x₂)
    return √(((n₁ - 1) * var(x₁) + (n₂ - 1) * var(x₂)) / (n₁ + n₂ - 2))
end;
```

So let's use that to calculate the t-statistic for our data.

```
t_observed = 2.2665515995859433
```

```
t_observed = t_statistic(data.drp[data.group .== "Treat"],
                         data.drp[data.group .== "Control"])
```

Important point

Our t-statistic for our dataset is 2.2665515995859433. Compare that to the t-statistic calculated through JASP, up at the top of this notebook. It's the same!

So the t-statistic for our dataset is 2.2665515995859433. Nice. But what does this mean? Is it a small number or a large number? How does it compare to what we'd expect to see under the null hypothesis?

In order to work that out, we will follow the simulation approach in the flow diagram above.

Define the generative model

$$\alpha = \delta \cdot \sigma$$

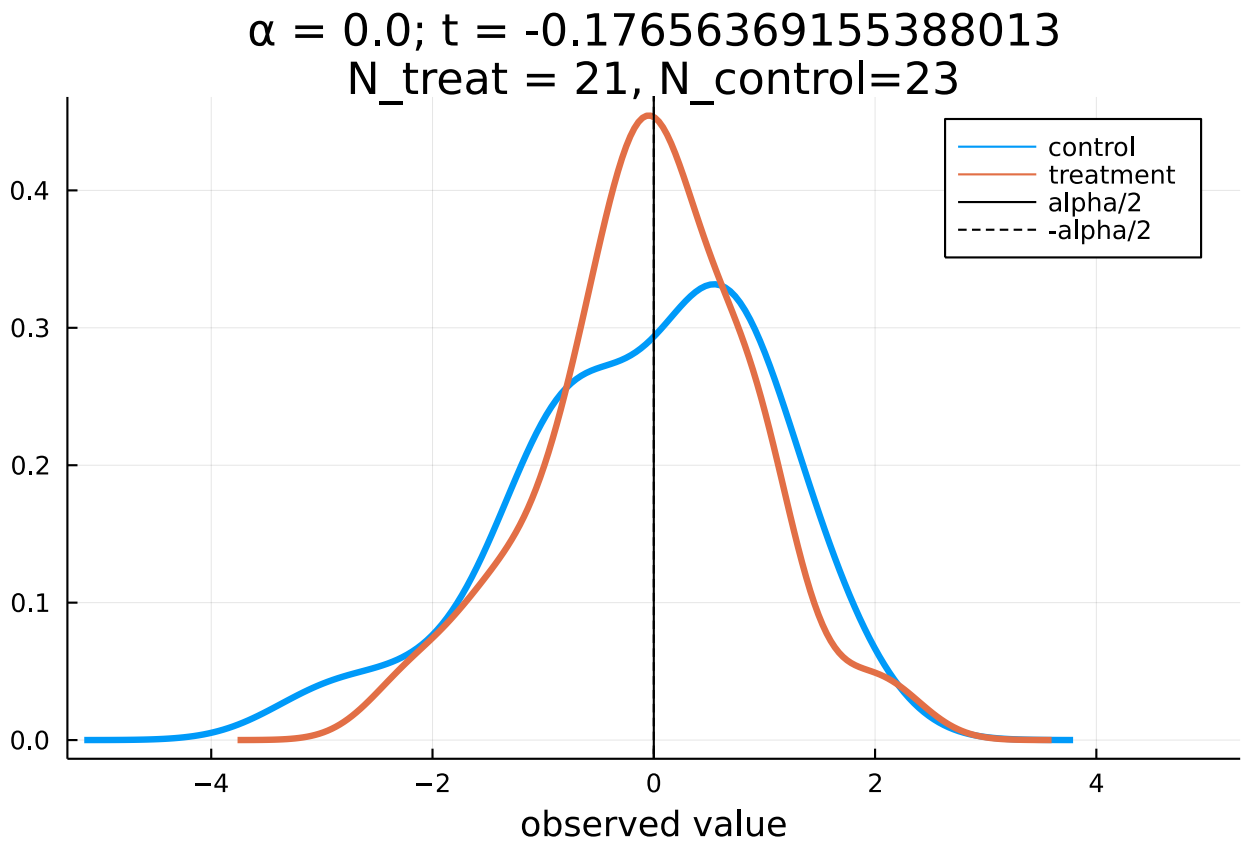
$$x_i \sim \text{Normal}\left(\mu - \frac{\alpha}{2}, \sigma\right), i = 1, \ldots, N_{\text{treat}}$$

$$y_i \sim \text{Normal}\left(\mu + \frac{\alpha}{2}, \sigma\right), i = 1, \ldots, N_{\text{control}}$$

Note that the effect size $\delta = \alpha/\sigma$

```
function generative_process(α, N_treat, N_control)
    μ, σ = 0, 1
    x = rand(Normal(μ + α/2, σ), N_treat)
    y = rand(Normal(μ - α/2, σ), N_control)
    return x, y
end;
```

Choose a value for α. Note that H_0 is that there is no difference between the group, and when H_0 is true, $\alpha = 0$.




```
seriestype density has been moved to StatsPlots.  To use: `Pkg.add("StatsPlots"); using StatsPlots`
```

Now what we want to do is to repeat this simulation many times to calculate a distribution of what t-statistic values we would expect to see under H_0. The steps are:

1. Set $\alpha = 0$ to correspond to H_0
2. Use the generative process to simulate data we'd expect to see.
3. Calculate the t-statistic for that simulated dataset.
4. Repeat steps 2-3 many times, building up a list of t-statistic values.
5. Work out the proportion of the time the observed t-statistic is greater than what we'd expect under H_0.

```
simulate (generic function with 1 method)
```

```
function simulate(n_simulations, N_treat, N_control)
    α = 0
    x, y = generative_process(α, N_treat, N_control)
    return t_statistic(x, y)
end
```

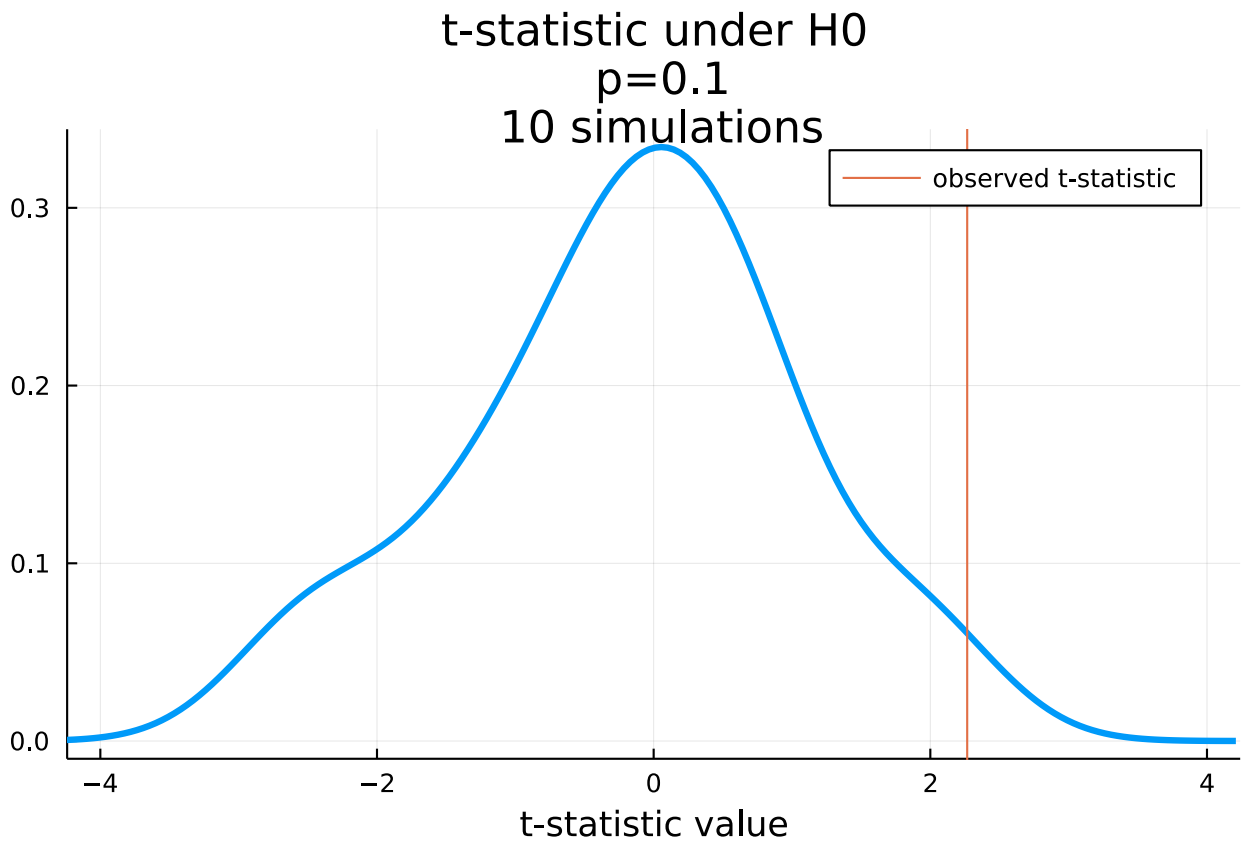
```
# function many_simulations(n_simulations, N_treat, N_control)
#     t_vec = []
#     for n = 1 : n_simulations
#         # Note: we assume alpha = 0
#         x, y = generative_process(0, N_treat, N_control)
#         append!(t_vec, t_statistic(x, y))
#     end
#     return t_vec
# end;

function many_simulations(n_simulations, N_treat, N_control)
    return [simulate(n_simulations, N_treat, N_control) for n = 1:n_simulations]
    # t_vec = []
    # for n = 1 : n_simulations
    #     # Note: we assume alpha = 0
    #     x, y = generative_process(0, N_treat, N_control)
    #     append!(t_vec, t_statistic(x, y))
    # end
    # return t_vec
end;
```

Note that we are using a 2-tailed test in that we are not making a directional hypothesis. Therefore we need compare the distribution of t-statistics to the *absolute value* of the observed t-statistic.

```
function p_two_tailed(observed, distribution)
    sum(abs.(distribution) .> observed) / length(distribution)
end;
```

Choose the number of simulations with the slider:

```
@bind n_simulations Slider(10 : 10 : 1_000_000)
```




```
seriestype density has been moved to StatsPlots.  To use: `Pkg.add("StatsPl
ots"); using StatsPlots`
```

Important point

The two-tailed p-value using our simulation approach is 0.1. We got a $p = 0.029$ from JASP, so that is a match when our estimates are based on a large number of simulated datasets.

Recap

Let's circle back to the output we got from the Frequentist independent samples T-Test:

Independent Samples T-Test ▼

	t	df	p	Cohen's d
drp	−2.267	42	0.029	−0.684

Note. Student's t-test.

We were able to replicate the t-statistic value *and* the p-value using our simulation based approach. Hopefully this all clarifies how Frequentist hypothesis testing works. The most important points are:

- The Frequentist approach uses a data generating process. We assume H_0 is true and then simulate data consistent with this hypothesis.
- By repeatedly simulating possible datasets under H_0, then we can calculate a distribution of t-statistics we would expect to see.
- We can then compare that to the t-statistic value for the actual data that we observed. This allows us to see how likely or unlikely the observed t-statistic value is
- We can then make claims about the plausibility of the data under the null hypothesis, $P(\text{data}|H_0)$, which we know as the likelihood.
- But we must be aware that this does *not* allow us to make any claims about how likely the null hypothesis is to be true or false.

What you *cannot* say:

- I have falsified the null hypothesis.
- There is a 95% chance of the null hypothesis being false

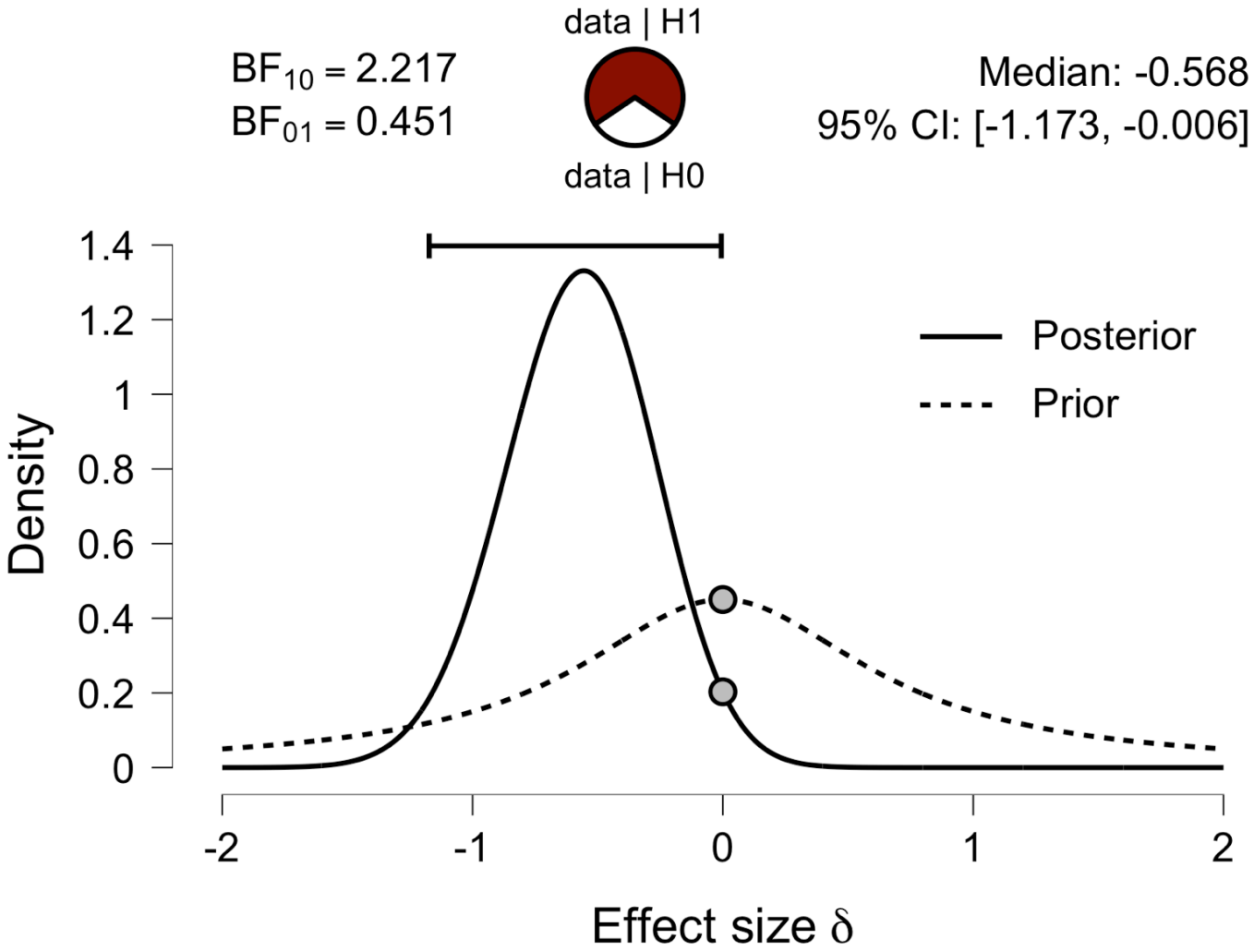
What you *can* say:

- If I theoretically repeated the same experiment a very large number of times, then I would expect to get a statistic value this extreme or more p=0.1 of the time.
- If I conclude that H_0 is false, I would be wrong 10.0% of the time.

I highly recommend checking out my other Hypothesis Testing notebook which goes into this in more depth.

Part III - Bayesian T-Test

It's been a while, so let's recap the kind of output that we get from the Bayesian version of the independent samples T-Test.



The most important points we will cover are:

- How to understand the main result, the Bayes Factor, $BF_{01} = 0.451$
- How to understand the prior and posterior over effect sizes, δ.

To pre-empt our full understanding:

1. The prior curve represents the relative plausibility/credibility in different effect sizes prior to having seen the data. This prior knowledge could be based upon our knowledge of the effect sizes that we observe in psychological experiments.
2. The posterior curve represents the credibility of effect sizes after having observed the data.

This is arrived at using Baye's equation.

3. The Bayes Factor describes whether the credibility of H_0 (effect size is zero) has gone up or down, and by how much, after having seen the data relative to our prior beliefs.

One of the major conceptual leaps here is that we no longer just consider one hypothesis, such as H_0 which states that group differences are zero. We saw in the generative model, this was done by setting $\alpha = 0$.

In the Bayesian approach we actually consider a whole span of hypotheses and consider the credibility for a range of effect sizes consistent with our prior beliefs or our posterior beliefs after having seen the data.

Multiple definitions for PlutoUI.
Combine all definitions into a single reactive cell using a `begin ... end` block.

```
png_joinpathsplit__FILE__1assetsimagepng = let
    import PlutoUI
    PlutoUI.LocalResource(joinpath(split(@__FILE__, '#')[1] * ".assets", "image.png"))
end
```

Define the generative model

This is basically the same generative model as before. The only real difference is that we can define our prior beliefs about the effect size, δ.

$$\delta \sim \text{Cauchy}(0, 0.707)$$

$$\alpha = \delta \cdot \sigma$$

$$x_i \sim \text{Normal}\left(\mu - \frac{\alpha}{2}, \sigma\right), i = 1, \ldots, N_{\text{treat}}$$

$$y_i \sim \text{Normal}\left(\mu + \frac{\alpha}{2}, \sigma\right), i = 1, \ldots, N_{\text{control}}$$

Let's look at these components:

- μ is the mean of the observed data.
- δ is the effect size and is defined as $\delta = \alpha \cdot \sigma$.
- α is the distance between the group means.
- σ is the standard deviation of the data.
- x_i defines the likelihood of the data from the treatment group.
- y_i defines the likelihood of the data from the control group.

Note that the effect size $\delta = \alpha / \sigma$

```
effect_sizes = LinRange(-10, 10, 1000);
```

```
dx = 0.02
```

```
normalize (generic function with 1 method)
```

```
normalize(x, dx) = x./(sum(x)*dx)
```

Calculate the posterior probability for a given effect size δ.

```
function calc_posterior(δ, σ, treat, control)
    # Note we sum log posteriors rather than multiply raw posteriors
    # Mathematically this is the same thing, but numerically it is more stable
    μ = mean([treat; control])
    lp = 0.0
    α = δ * σ
    # prior
    lp += logpdf(Cauchy(0, 0.707), δ)
    # likelihood
    lp += sum(logpdf.(Normal(μ-(α/2), σ), treat))
    lp += sum(logpdf.(Normal(μ+(α/2), σ), control))
    return exp(lp)  # return posterior, not the log posterior
end;
```

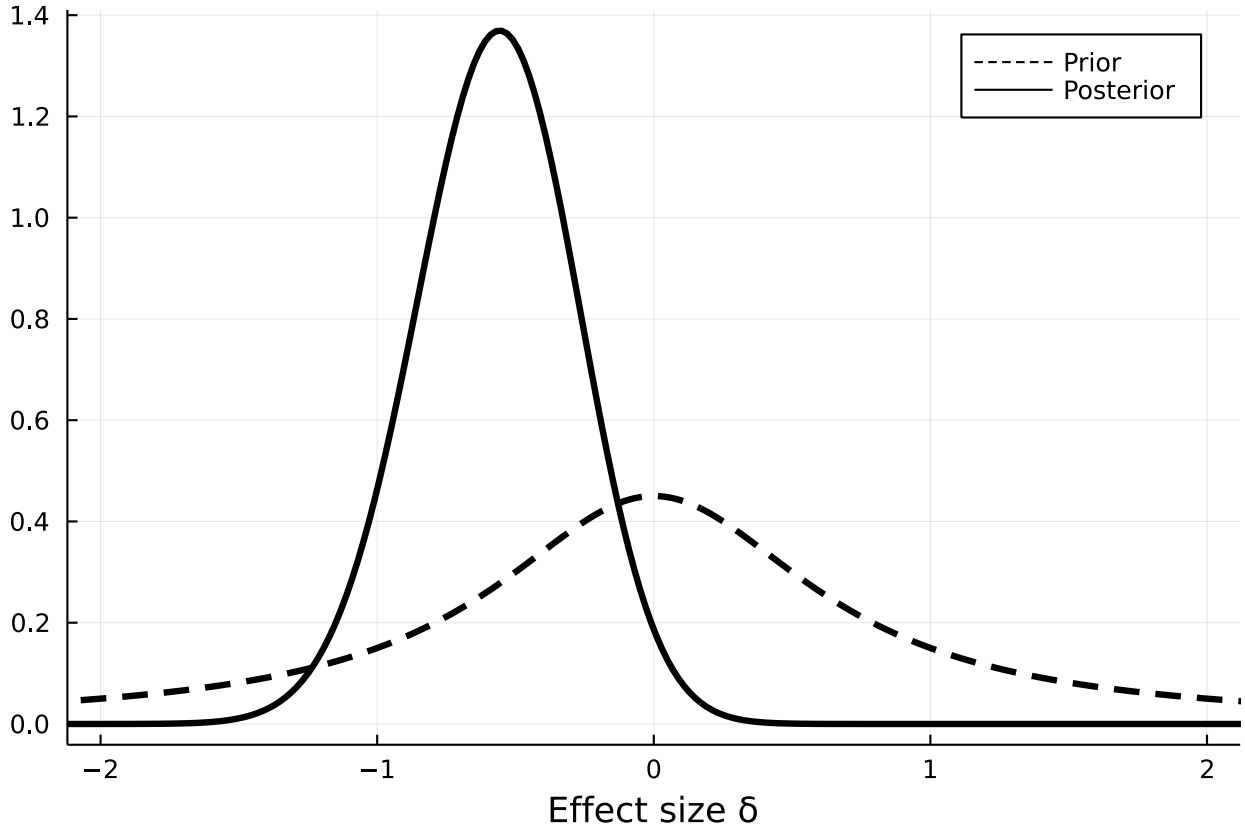
```
function calc_posterior_for_many_effect_sizes(treat, control)
    σ = pooled_std(treat, control)
    post = [calc_posterior(δ, σ, treat, control) for δ in effect_sizes]
    return normalize(post, dx)
end;
```

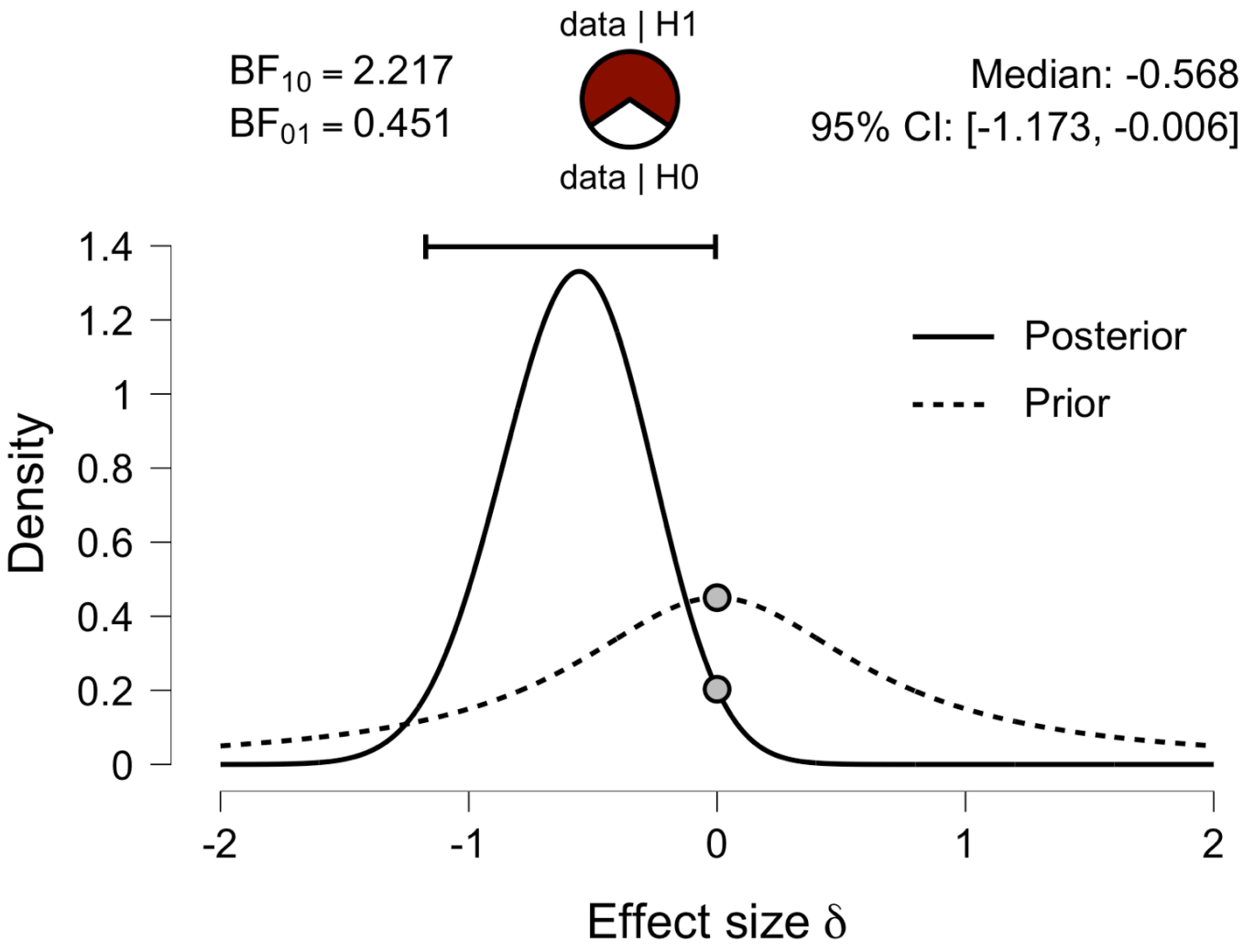
```
posterior = calc_posterior_for_many_effect_sizes(
    data.drp[data.group .== "Treat"],
    data.drp[data.group .== "Control"]);
```

Let's look at the results. We know they only approximate the results that JASP gives.

Here we can see the the credibility that the effect size equals zero ($\delta = 0$) has now gone up slightly after having observed the data. This constrasts to the result given by JASP.

The discrepancy is simply because I have not yet fully understood how they deal with the σ parameter.


Prior
Posterior
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
Effect size δ


data | H1
BF10 = 2.217
BF01 = 0.451
Median: -0.568
95% CI: [-1.173, -0.006]
data | H0
Posterior
Prior
Density
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-2
-1
0
1
2
Effect size δ

But let's interpret the JASP result:

Bayes factor, BF_{01}			Interpretation
	$>$	100	Extreme evidence for H_0 compared to H1
30	–	100	Very Strong evidence for H_0 compared to H1
10	–	30	Strong evidence for H_0 compared to H1
3	–	10	Substantial evidence for H_0 compared to H1
1	–	3	Anecdotal evidence for H_0 compared to H1
	1		No evidence
1/3	–	1	Anecdotal evidence for H_1 compared to H0
1/10	–	1/3	Substantial evidence for H_1 compared to H0
1/30	–	1/10	Strong evidence for H_1 compared to H0
1/100	–	1/30	Very strong evidence for H_1 compared to H0
	$<$	1/100	Extreme evidence for H_1 compared to H0

So for the JASP result, you could say:

- The most likely effect size is $\delta = -0.568$, and the 95% Bayesian credible intervals do not overlap with zero (just).
- Relative to my prior beliefs, the credibility in the null hypothesis ($\delta = 0$) has gone *down* by 2.2 times after having observed the data.
- By convention, a Bayes Factor of 2.2 is classed as *anecdotal evidence* and so I will not seek to draw strong research conclusions either way.
- The 'dart board' part of the plot gives an indication of how surprised you should be if you throw a dart and it hits H_0 or H_1.

Note: there are different schools of thought about the relative importance of Bayes Factors vs Bayesian credible intervals. But the tide seems to be in favour of Bayes Factors when it comes to hypothesis testing.

If you are interested in this particular way of explaining the Bayesian approach, then I would recommend a previous Bayesian inference notebook which focusses on a linear regression example. There is also an accompanying YouTube video.

Summary

We can do hypothesis testing in the Frequentist and Bayesian approaches. There are similarities and differences.

- Similar in having a level of evidence which is compared to a threshold (or thresholds) which is decided by convention.
- But the Frequentist approach only allows you to make claims about data.
- Very often, as scientists, we are interested in making claims about hypotheses.
- The Bayesian approach allows you to do this.
- It is conceptually simple, in terms of incorporating your prior beliefs and information from the data (via the likelihood) to arrive at your final beliefs (the posterior).

There are many other aspects of Frequentism vs Bayesianism which I've not touched upon here, but hopefully this has served as a good foundation to build upon.

Any questions for me?

References

- Dienes, Z. (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave Macmillan.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, 16(2), 225–237. http://doi.org/10.3758/PBR.16.2.225

Code blocks below are just setup stuff to make eveything work. I am hiding them away in order to focus on the concepts rather than mundane practical aspects of the code.

Multiple definitions for PlutoUI.
Combine all definitions into a single reactive cell using a `begin ... end` block.

```
using PlutoUI    , CSV    , Statistics    , Distributions    , StatsPlots
```

```
hint(text) = Markdown.MD(Markdown.Admonition("hint", "Hint", [text]));
```

```
correct(text) = Markdown.MD(Markdown.Admonition("correct", "Important point",
[text]));
```

```
danger(text) = Markdown.MD(Markdown.Admonition("danger", "Warning!", [text]));
```