8 Comparing Gaussian means

with Ruud Wetzels

Popular theories are difficult to overthrow. Consider the following hypothetical sequence of events. First, Dr John proposes a Seasonal Memory Model (SMM). The model is intuitively attractive and quickly gains in popularity. Dr Smith, however, remains unconvinced and decides to put one of SMM's predictions to the test. Specifically, SMM predicts that a glucose-driven increase in recall performance is more pronounced in summer than in winter. Dr Smith conducts the relevant experiment using a within-subjects design and finds the opposite result: as shown in the fictitious data in Table 8.1, the increase in recall performance is *smaller* in summer than in winter, although this difference is not significant. With $n = 41$ and a t value of 0.79, the corresponding two-sided p-value equals 0.44.

Table 8.1 Glucose-driven increase in recall performance in summer and winter.

Season	N	Mean	SD
Winter	41	0.11	0.15
Summer	41	0.07	0.23

Clearly, Dr Smith's data do not support SMM's prediction that the glucose-driven increase in performance is larger in summer than in winter. Instead, the data seem to suggest that the null hypothesis is plausible, and that no difference between summer and winter is evident. Dr Smith submits his findings to the *Journal of Experimental Psychology: Learning, Memory, and the Seasons*. Three months later, Dr Smith receives the reviews. Inevitably, one of the reviews is from Dr John, and it includes the following comment:

> From a null result, we cannot conclude that no difference exists, merely that we cannot reject the null hypothesis. Although some have argued that with enough data we can argue for the null hypothesis, most agree that this is only a reasonable thing to do in the face of a sizeable number of data that have been collected over many experiments that control for all concerns. These conditions are not met here. Thus, the empirical contribution here does not enable readers to conclude very much, and so is quite weak.

Formally, Dr John's first statement is completely correct, since p-values cannot be used to quantify the support in favor of the null hypothesis. A p-value of 0.44 could indicate that the data support $\mathcal{H}_0$, but it could also indicate that the data

are too few in number to result in a rejection of $\mathcal{H}_0$. In this chapter we show how this ambiguity can be overcome, and how Dr Smith and other researchers can use the Bayes factor to quantify evidence in favor of $\mathcal{H}_0$. As explained in Chapter 7, the Bayes factor measures the change from prior model odds to posterior model odds brought about by the data. This means that, in contrast to the p-value, the Bayes factor is able to quantify evidence both in favor of $\mathcal{H}_0$ and in favor of $\mathcal{H}_1$.

The sections below highlight some properties of the Bayes factor in the context of the popular t-test (Rouder et al., 2009). We show how to specify $\mathcal{H}_0$ and $\mathcal{H}_1$, and then use the Savage–Dickey density ratio to calculate the Bayes factor.[1] This then allows us to address the key point of contention between Dr John and Dr Smith. To what extent, if at all, do the observed data contradict the prediction from the Seasonal Memory Model?

8.1 One-sample comparison

When we use the one-sample t-test, we assume that the data follow a Gaussian distribution with unknown mean μ and unknown variance σ^2. This is a natural assumption for a within-subjects experimental design, like that undertaken by Dr Smith. The data consist of one sample of standardized difference scores (i.e., "winter scores − summer scores"). The null hypothesis states that the mean of the difference scores is equal to zero, that is, $\mathcal{H}_0 : \mu = 0$. The alternative hypothesis states that the mean is not equal to zero, that is, $\mathcal{H}_1 : \mu \neq 0$.

We follow Rouder et al. (2009) and use a Cauchy$(0, 1)$ prior for effect size δ. The advantage of defining a prior on effect size, instead of on the mean, is that it is very general. The same prior can be used across many experiments, dependent variables, and measurement scales. The Cauchy distribution used for this prior is a t-distribution with 1 degree of freedom, and resembles a Gaussian distribution with fatter tails. The choice for the Cauchy is theoretically motivated, and details are provided by Jeffreys (1961), Liang et al. (2008), and Zellner and Siow (1980). For the standard deviation we use a half-Cauchy distribution, so that $\sigma \sim \text{Cauchy}(0, 1)_{\mathcal{I}(0,\infty)}$. This is a Cauchy$(0, 1)$ distribution that is defined only for positive numbers (Gelman & Hill, 2007).

The graphical model for the one-sample comparison of means is shown in Figure 8.1. In the graphical model, x represents the observed data that follow a Gaussian distribution with mean μ and a variance σ^2. The effect size δ is defined as $\delta = \mu/\sigma$, and so μ is given by $\mu = \delta\sigma$. The null hypothesis puts all prior mass for δ on a single point, that is, $\mathcal{H}_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ has a Cauchy distribution, with $\mathcal{H}_1 : \delta \sim \text{Cauchy}(0, 1)$.

The script `OneSample.txt` implements the graphical model in WinBUGS:

[1] More information can be found on the website of Ruud Wetzels, `www.ruudwetzels.com`, and the website of Jeff Rouder, `pcl.missouri.edu`.

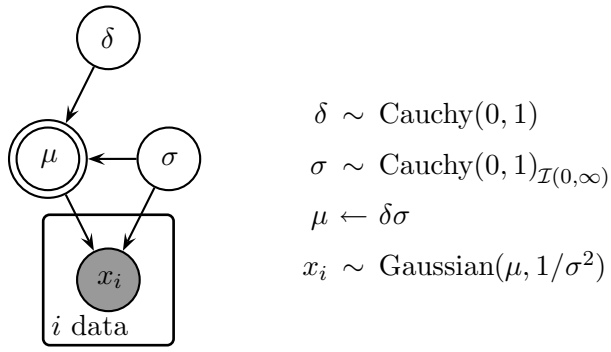


Fig. 8.1 Graphical model for the one-sample within-subjects comparison of means.

```
# One-Sample Comparison of Means
model{
  # Data
  for (i in 1:ndata){
    x[i] ~ dnorm(mu,lambda)
  }
  mu <- delta*sigma
  lambda <- pow(sigma,-2)
  # delta and sigma Come From (Half) Cauchy Distributions
  lambdadelta ~ dchisqr(1)
  delta ~ dnorm(0,lambdadelta)
  lambdasigma ~ dchisqr(1)
  sigmatmp ~ dnorm(0,lambdasigma)
  sigma <- abs(sigmatmp)
  # Sampling from Prior Distribution for Delta
  deltaprior ~ dnorm(0,lambdadeltaprior)
  lambdadeltaprior ~ dchisqr(1)
}
```

Note that the Cauchy distribution is not directly available in WinBUGS. This can be addressed by assigning $\delta \sim \text{Gaussian}\big(0, \lambda_\delta\big)$, where the precision λ_δ has a chi-square distribution with one degree of freedom, $\lambda_\delta \sim \chi^2\big(1\big)$. This two-step assignment procedure corresponds to $\delta \sim \text{Cauchy}\big(0, 1\big)$. Note also that the WinBUGS script generates prior as well as posterior samples of the effect size δ.

The code `OneSample.m` or `OneSample.R` applies the model to the data in Table 8.1, plots the prior and the posterior distributions for δ, and applies the Savage–Dickey density ratio test to the posterior samples of δ to compute the Bayes factor for $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_1 : \delta \sim \text{Cauchy}\big(0, 1\big)$.

Figure 8.2 shows the results. The posterior distribution is peaked near zero, with a little more density given to positive, rather than negative, effect sizes. The critical point $\delta = 0$ is about 5 times more likely in the posterior distribution than it is in the prior distribution. This means that the Bayes factor is about 5:1 in favor of the null hypothesis $\mathcal{H}_0$.

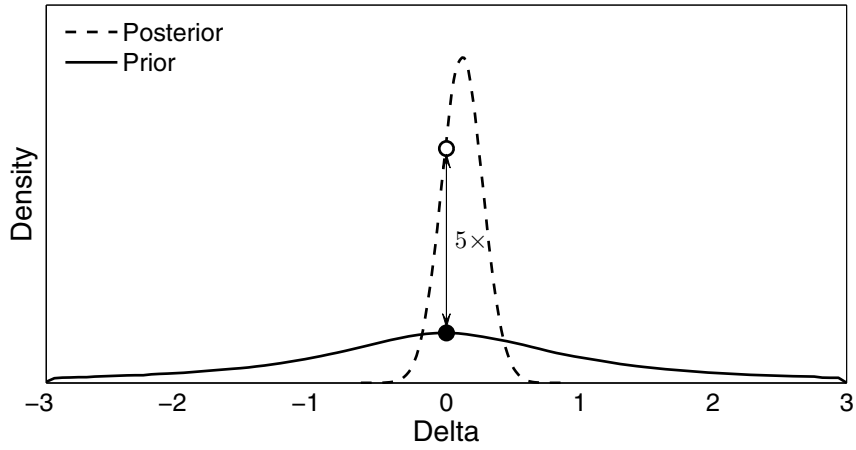


Fig. 8.2 Prior and posterior distributions on effect size δ for the summer and winter data. Markers show the height of the prior and posterior distributions at $\delta = 0$ needed to estimate the Bayes factor between $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_1 : \delta \sim \text{Cauchy}\big(0, 1\big)$ using the Savage–Dickey method.

Exercises

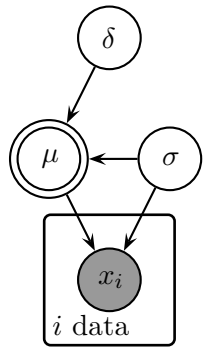
Exercise 8.1.1 Here we assumed a half-Cauchy prior distribution on the standard deviation `sigma`. Other choices are possible and reasonable. Can you think of a few?

Exercise 8.1.2 Do you think the different priors on `sigma` will lead to substantially different conclusions? Why or why not? Convince yourself by implementing a different prior and studying the result.

Exercise 8.1.3 We also assumed a Cauchy prior distribution on effect size `delta`. Other choices are possible and reasonable. One such choice is the standard Gaussian distribution. Do you think this prior will lead to substantially different conclusions? Why or why not? Convince yourself by implementing the standard Gaussian prior and studying the result.

8.2 Order-restricted one-sample comparison

The Bayes factor computed in the previous section quantified the strength of evidence in favor of $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_1 : \delta \sim \text{Cauchy}\big(0, 1\big)$. However, this particular $\mathcal{H}_1$ was not the SMM hypothesis that Dr Smith set out to test. The SMM hypothesis specifically stated that δ should be *negative*. Hence, a more appropriate alternative hypothesis incorporates the constraint $\delta < 0$. This corresponds to a half-Cauchy prior distribution that is defined for negative numbers only,



$$\begin{aligned}\delta &\sim \text{Cauchy}(0,1)_{\mathcal{I}(-\infty,0)}\\ \sigma &\sim \text{Cauchy}(0,1)_{\mathcal{I}(0,\infty)}\\ \mu &\leftarrow \delta\sigma\\ x_i &\sim \text{Gaussian}(\mu, 1/\sigma^2)\end{aligned}$$

Fig. 8.3 Graphical model for the order-restricted one-sample within-subjects comparison of means.

$\mathcal{H}_2 : \text{Cauchy}\big(0,1\big)_{\mathcal{I}(-\infty,0)}$. This alternative hypothesis is called an order-restricted or one-sided hypothesis.

The graphical model for this analysis is shown in Figure 8.3, and it only changes the prior on effect size. The script `OneSampleOrderRestricted.txt` implements the graphical model in WinBUGS:.

```
# One-Sample Order Restricted Comparison of Means
model{
  # Data
  for (i in 1:ndata){
    x[i] ~ dnorm(mu,lambda)
  }
  mu <- delta*sigma
  lambda <- pow(sigma,-2)
  # delta and sigma Come From (Half) Cauchy Distributions
  lambdadelta ~ dchisqr(1)
  delta ~ dnorm(0,lambdadelta)I(,0)
  lambdasigma ~ dchisqr(1)
  sigmatmp ~ dnorm(0,lambdasigma)
  sigma <- abs(sigmatmp)
  # Sampling from Prior Distribution for Delta
  deltaprior ~ dnorm(0,lambdadeltaprior)I(,0)
  lambdadeltaprior ~ dchisqr(1)
}
```

The code `OneSampleOrderRestricted.m` or `OneSampleOrderRestricted.R` again applies the model to the data in Table 8.1. Figure 8.4 plots the prior and the posterior distributions for δ, and shows the key densities for the Savage–Dickey density ratio test at $\delta = 0$ to compute the Bayes factor for $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_2 : \text{Cauchy}\big(0,1\big)_{\mathcal{I}(-\infty,0)}$. The data are now about 10 times more likely under $\mathcal{H}_0$ than they are under the order-restricted $\mathcal{H}_2$ associated with SMM. According to the classification scheme proposed by Jeffreys (1961), as presented in Table 7.1, this could be considered "strong evidence" for the null hypothesis.

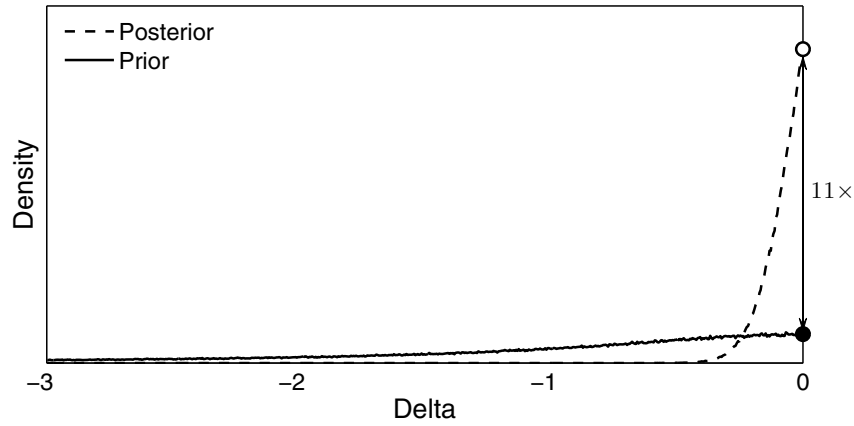


Fig. 8.4 Prior and posterior distributions on effect size δ for the Summer and Winter data. Markers show the height of the prior and posterior distributions at $\delta = 0$ needed to estimate the Bayes factor between $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_2 : \text{Cauchy}(0, 1)_{\mathcal{I}(-\infty,0)}$ using the Savage–Dickey method.

Exercises

Exercise 8.2.1 For completeness, estimate the Bayes factor for the summer and winter data between $\mathcal{H}_0 : \delta = 0$ versus $\mathcal{H}_3 : \text{Cauchy}(0, 1)_{\mathcal{I}(0,\infty,)}$, involving the order-restricted alternative hypothesis that assumes the effect is positive.

Exercise 8.2.2 In this example, it matters whether the alternative hypothesis is unrestricted, order-restricted to negative values for δ, or order-restricted to positive values for δ. Why is this perfectly reasonable? Can you think of a situation where the three versions of the alternative hypothesis yield exactly the same Bayes factor?

Exercise 8.2.3 From a practical standpoint, we do not need a new graphical model and WinBUGS script to compute the Bayes factor for $\mathcal{H}_0$ versus the order-restricted $\mathcal{H}_2$. Instead, we can use the original graphical model in Figure 8.1 that implements the unrestricted Cauchy distribution and discard those prior and posterior MCMC samples that are inconsistent with the $\delta < 0$ order-restriction. The Savage–Dicky density ratio test still involves the height of the prior and posterior distributions at $\delta = 0$, but now the samples from these distributions are truncated, respecting the order-restriction, such that they range only from $\delta = -\infty$ to $\delta = 0$. Implement this method in Matlab or R, and check that the same conclusions are drawn from the analysis.

Exercise 8.2.4 Wagenmakers and Morey (2013) describe yet another method to obtain the Bayes factor for order-restricted model comparisons. This method is perhaps the most reliable because it avoids the numerical complications associated with having to estimate the posterior density at a boundary. Go to `http://www.ejwagenmakers.com/papers.html`, download the Wagenmakers

Box 8.1 Estimating densities from samples

Estimating Bayes factors using the Savage–Dickey approach requires estimating the height of the prior and posterior distributions at a specific value of a parameter. Often the height of the prior distribution can be obtained analytically. It is always possible to estimate the required prior and posterior densities from MCMC samples. The simplest approach is by binning. A more advanced approach is to use a non-parametric density estimator (e.g., Stone et al., 1997). In R, one such estimator is included in the `polspline` package. This can be installed by starting R and selecting the `Install Package(s)` option in the `Packages` menu. Once you choose your preferred CRAN mirror, select `polspline` in the `Packages` window and click on `OK`. In Matlab, the Statistics toolbox provides the `ksdensity` function, or you may try the non-parametric density estimators available in the free package developed by Zdravko Botev, `www.mathworks.com/matlabcentral/fileexchange/14034`.

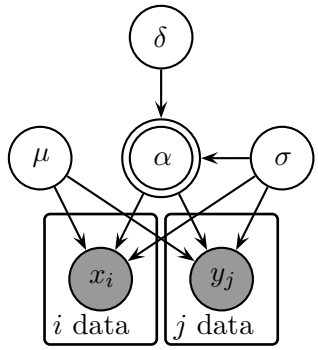
Both binning and density estimation approaches depend on tuning parameters—like the width of the bins—and so some experimentation and tests of robustness are usually required. For example, the Savage–Dickey analysis in Figure 8.2 can give Bayes factors between (at least) about 4.7 and 6.1 using different reasonable density estimation methods. Since the goal is to estimate the Bayes factor, what is important is that the substantive conclusions are trusted, rather than obtaining the exact number. The interpretive framework provided by Table 7.1 is very helpful in this regard.

and Morey paper, and read the introduction with a focus on Equation 1. Implement their suggested method and compare the results to those obtained earlier.

8.3 Two-sample comparison

Often in cognitive science, the comparison of means is based on data from two independent groups, rather than a single group. The two-sample t-test is commonly used as a standard frequentist approach for these sorts of between-subjects designs.

Most textbooks on research methods have their own introductory example, and we consider the one presented by Evans and Rooney (2011, pp. 279–283). Their example involves a between-subjects experiment to test the effect of drinking plain versus oxygenated water. The raw data are presented in Evans and Rooney (2011,



$$
\begin{aligned}
\delta &\sim \text{Cauchy}(0,1) \\
\mu &\sim \text{Cauchy}(0,1) \\
\sigma &\sim \text{Cauchy}(0,1)_{\mathcal{I}(0,\infty)} \\
\alpha &\leftarrow \delta\sigma \\
x_i &\sim \text{Gaussian}(\mu + \tfrac{\alpha}{2}, 1/\sigma^2) \\
y_j &\sim \text{Gaussian}(\mu - \tfrac{\alpha}{2}, 1/\sigma^2)
\end{aligned}
$$

Fig. 8.5 Graphical model for the two-sample between-subjects comparison of means.

Table 13.3), and consists of a set of memory scores for 20 subjects who took plain water, and a set of memory scores for 20 subjects who took oxygenated water. For the plain water (control) group, the mean score is 68.35 with standard deviation 6.38, while for the oxygenated water (treatment or experimental) group, the mean score is 76.65 with standard deviation 4.06. A two-sample t-test gives $t\,(38) = 4.47$, $p < 0.01$.

In our Bayesian approach of making inferences about the means, we rescale the data so that one group has mean 0 and standard deviation 1. This rescaling procedure ensures that the prior distributions for the parameters hold regardless of the scale of measurement. Therefore it does not matter whether, say, response times are measured in seconds or in milliseconds.

The graphical model for the two-sample comparison is shown in Figure 8.5. The variables x and y represent the experimental and control data, respectively. Both x and y follow Gaussian distributions with shared variance σ^2. The mean of x is given by $\mu + \alpha/2$, and the mean of y is given by $\mu - \alpha/2$, so that α is the difference in the means.

Because $\delta = \alpha/\sigma$, α is given by $\alpha = \delta\sigma$. As for the one-sample scenario, the null hypothesis puts all prior mass for δ on a single point, so that $\mathcal{H}_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ follows a Cauchy distribution, so that $\mathcal{H}_1 : \delta \sim \text{Cauchy}\big(0, 1\big)$.

The script `TwoSample.txt` implements the graphical model in WinBUGS:

```
# Two-sample Comparison of Means
model{
  # Data
  for (i in 1:n1){
    x[i] ~ dnorm(mux,lambda)
  }
  for (j in 1:n2){
    y[j] ~ dnorm(muy,lambda)
  }
  # Means and precision
  alpha <- delta*sigma
  mux <- mu+alpha/2
```

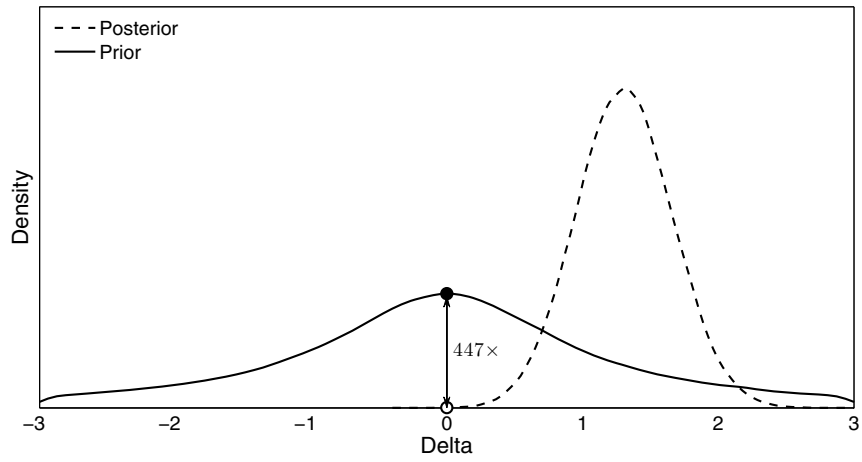


Fig. 8.6 Prior and posterior distributions on effect size δ for the Evans and Rooney (2011) data using the model for a two-sample comparison of means. Markers show the height of the prior and posterior distributions at $\delta = 0$ needed to estimate the Bayes factor between $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_1 : \text{Cauchy}\big(0, 1\big)$ using the Savage–Dickey method.

```
  muy <- mu-alpha/2
  lambda <- pow(sigma,-2)
  # delta, mu, and sigma Come From (Half) Cauchy Distributions
  lambdadelta ~ dchisqr(1)
  delta ~ dnorm(0,lambdadelta)
  lambdamu ~ dchisqr(1)
  mu ~ dnorm(0,lambdamu)
  lambdasigma ~ dchisqr(1)
  sigmatmp ~ dnorm(0,lambdasigma)
  sigma <- abs(sigmatmp)
  # Sampling from Prior Distribution for Delta
  lambdadeltaprior ~ dchisqr(1)
  deltaprior ~ dnorm(0,lambdadeltaprior)
}
```

The code `TwoSample.m` or `TwoSample.R` applies the model to the data in Evans and Rooney (2011, Table 13.3). Figure 8.6 plots the prior and the posterior distributions for δ, and shows the key densities for the Savage–Dickey density ratio test at $\delta = 0$ to compute the Bayes factor. It is clear that there is a large effect of oxygenated versus plain water on memory performance. The data are now more than 400 times more likely under $\mathcal{H}_1$ than they are under $\mathcal{H}_0$, and this large Bayes factor can be interpreted as providing decisive evidence.

Exercise

Exercise 8.3.1 The two-sample comparison of means outlined above assumes that the two groups have equal variance. How can you extend the model when this assumption is not reasonable?