8 Hypothesis Testing

As pointed out in chapter 4, Bayesian hypothesis testing is straightforward. For a hypothesis of the form \( H_i : \theta \in \Theta_i \), where \( \Theta_i \) is a subset of the parameter space \( \Theta \), we can compute the prior probability
\[
\pi_i = P(H_i) = P(\theta \in \Theta_i)
\]
and the posterior probability
\[
p_i = P(H_i | y) = P(\theta \in \Theta_i | y).
\]

Often, there are only two hypotheses, the “null hypothesis” \( H_0 \) and its logical negation \( H_1 : \theta \not\in \Theta_0 \), called the “alternative hypothesis”. The hypothesis with the highest probability can be chosen as the “best” hypothesis; a more sophisticated choice can be made using Decision Theory (to be discussed in section 14).

Two hypotheses can be compared using odds. The posterior odds in favour of \( H_0 \) against \( H_1 \) given data \( y \) are given by the ratio
\[
\frac{p_0}{p_1} = \frac{P(H_0 | y)}{P(H_1 | y)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(y | H_0)}{P(y | H_1)}.
\]

The number \( B \), called the Bayes factor, tells us how much the data alters our prior belief. In general, the Bayes factor depends on the prior:
\[
B = \frac{P(H_0 | y) / \pi_0}{P(H_1 | y) / \pi_1} = \frac{\int_{\Theta_0} p(\theta | y) d\theta / \pi_0}{\int_{\Theta_1} p(\theta | y) d\theta / \pi_1} = \frac{\int_{\Theta_0} p(y | \theta) p(\theta) / \pi_0 d\theta}{\int_{\Theta_1} p(y | \theta) p(\theta) / \pi_1 d\theta}.
\]

However, when the parameter space has only two elements, the Bayes factor is the likelihood ratio
\[
B = \frac{p(y | \theta_0)}{p(y | \theta_1)}
\]
which does not depend on the choice of the prior. This interpretation applies in the following example.

Example: Transmission of hemophilia  The human X chromosome carries a gene that is essential for normal clotting of the blood. The defect in this gene that is responsible for the blood disease hemophilia is recessive: no disease develops in a woman at least one of whose X chromosomes has a normal gene. However, a man whose X chromosome has the defective gene develops the disease. As a result, hemophilia occurs almost exclusively in males who inherit the gene from non-hemophiliac mothers. Great Britain’s Queen Victoria (pictured here) carried the hemophilia gene, and it was transmitted through her daughters to many of the royal houses of Europe.
**Question.** Alice has a brother with hemophilia, but neither she, her parents, nor her two sons (aged 5 and 8) have the disease. What is the probability that she is carrying the hemophilia gene?

**Solution.** Let $H_0$: Alice does not carry the hemophilia gene, and $H_1$: she does. The X chromosome that Alice inherited from her father does not have the defective gene, because he’s healthy. We know that Alice’s mother has one X chromosome with the defective gene, because Alice’s brother is sick and her mother is healthy. The X chromosome that Alice inherited from her mother could be the good one or the bad one; let’s take $P(H_0) = P(H_1) = \frac{1}{2}$ as our prior, that is, we assume prior odds to be 1 to 1. Let $Y$ denote the fact that Alice’s two sons are healthy. Because the sons are not identical twins, we can assume $P(Y|H_1) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$, $P(Y|H_0) = 1$. The posterior probability is then

$$P(H_1|Y) = \frac{P(Y|H_1)P(H_1)}{P(Y|H_1)P(H_1) + P(Y|H_0)P(H_0)} = \frac{\frac{1}{8} \cdot \frac{1}{2}}{\frac{1}{8} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = \frac{1}{5},$$

and $P(H_0|Y) = 1 - \frac{1}{5} = \frac{4}{5}$. The posterior odds in favour of $H_0$ against $H_1$ are 4 to 1, much improved (Bayes factor $B = 4$) compared to the prior odds.

A one-sided hypothesis for a continuous parameter has a form such as $H_0: \theta \leq \theta_0$, where $\theta_0$ is a given constant. This could represent a statement such as “the new fertilizer doesn’t improve yields”. After you compute the posterior probability

$$p_0 = P(H_0|y) = P(\theta \leq \theta_0|y) = \int_{\theta_0}^{\theta_0} p(\theta|y) \, d\theta,$$

you can make straightforward statements such as “The probability that $\theta \leq \theta_0$ is $p_0$”, or “The probability of that $\theta > \theta_0$ is $1 - p_0$”.

Some practitioners like to mimic Frequentist procedures and choose beforehand a “significance level” $\alpha$ (say, $\alpha = 5\%$), and then if $p_0 < \alpha$ they “reject $H_0$ (and accept $H_1$) at the $\alpha$ level of significance”. This is all rather convoluted, however, and as we shall see in section 14, a systematic approach should be based on Decision Theory. Simply reporting the probability value $p_0$ is more direct and informative, and usually suffices.

A two-sided hypothesis of the form $H_0: \theta = \theta_0$ might be used to model statements such as “the new fertilizer doesn’t change yields.” In Bayesian theory, such a “sharp” hypothesis test with a continuous prior pdf is pointless because it is always false:

$$P(\theta = \theta_0|y) = \int_{\theta_0}^{\theta_0} p(\theta|y) \, d\theta = \int_{\theta_0}^{\theta_0} p(y|\theta)p(\theta) \, d\theta = 0.$$

Thus it would seem that the question is not a sensible one. It nevertheless arises fairly often, for example when trying to decide whether to add or remove terms to a regression model. How, then, can one deal with such a hypothesis?

---

1For simplicity, we neglect the fact that hemophilia can also develop spontaneously as a result of a mutation.
• One could test whether \( \theta_0 \) lies in some credibility interval \( C_\varepsilon \). However, this isn’t a Bayesian hypothesis test.

• A Bayesian hypothesis test consists of assigning a prior probability \( \pi_0 > 0 \) to the hypothesis \( H_0 : \theta = \theta_0 \), yielding a prior that is a mixture of discrete pmf and continuous pdf. We shall discuss this further in section 13.

9 Simple Multiparameter Models

Often, even though one may need many parameters to define a model, one is only interested in a few of them. For example, in a normal model with unknown mean and variance \( y_i | \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2) \), one is usually interested only in the mean \( \mu \). The uninteresting parameters are called nuisance parameters, weighted by the posterior density of the nuisance parameters. Thus, denoting \( \theta = (\theta_1, \theta_2) \), where \( \theta_1 \) is the vector of parameters of interest and \( \theta_2 \) is the vector of ‘nuisance’ parameters, we have

\[
p(\theta_1 | y) = \int p(\theta | y) d\theta_2.
\]

This marginalisation integral can also be written as

\[
p(\theta_1 | y) = \int p(\theta_1 | \theta_2, y)p(\theta_2 | y) d\theta_2,
\]

which expresses \( \theta_1 | y \) as a mixture of conditional posterior distributions given the nuisance parameters, weighted by the posterior density of the nuisance parameters. This explains why the posterior pdf for the parameters of interest is generally more diffuse than \( p(\theta_1 | \theta_2, y) \) for any given \( \theta_2 \).

9.1 Two-parameter normal model

Consider a normal model with unknown mean and variance, that is, \( y_i | \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2) \) with \( y = y_1, \ldots, y_n \) conditionally independent given \( \mu, \sigma^2 \). The likelihood is then

\[
p(y | \mu, \sigma^2) \propto \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2} = \sigma^{-n} e^{-\frac{n(y - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2}{2\sigma^2}} = \sigma^{-n} e^{-\frac{n(y - \bar{y})^2 + (n-1)s^2}{2\sigma^2}},
\]

where \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) is the sample mean and \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \) is the sample variance.

We assume the following prior information:

• \( \mu \) and \( \sigma^2 \) are independent.

• \( p(\mu) \propto 1 \). This improper distribution expresses indifference about the location of the mean, because a translation of the origin \( \mu' = \mu + c \) gives the same prior.

• \( p(\sigma) \propto \frac{1}{\sigma} \). This improper distribution expresses indifference about the scale of the standard deviation, because a scaling \( \sigma' = c \sigma \) gives the same prior \( p(\sigma') = p(\sigma) | d\sigma / d\sigma' \propto \frac{1}{\sigma} \). Equivalently, we can say that this improper distribution expresses indifference about the location of \( \log(\sigma) \), because a flat prior \( p(\log(\sigma)) \propto 1 \) corresponds to \( p(\sigma) = \frac{d\log(\sigma)}{d\sigma} p(\log(\sigma)) \propto \frac{1}{\sigma} \). The corresponding prior distribution for the variance is \( p(\sigma^2) \propto \frac{1}{\sigma^2} \).
With this prior and likelihood, the joint posterior pdf is

\[
p(\mu, \sigma^2 | y) \propto (\sigma^2)^{-\left(\frac{\nu}{2}+1\right)}e^{-\frac{(n-1)s^2 + n(y - \mu)^2}{2\sigma^2}},
\]

(2)

The posterior mode can be found as follows. The mode’s \( \mu \) value is \( \bar{y} \) because of the symmetry about \( \mu = \bar{y} \). Then, denoting \( \nu = \log \sigma \) and \( A = (n-1)s^2 + n(y - \mu)^2 \), we have

\[
\log p(\mu, \sigma^2 | y) = -(n+2)\nu - \frac{1}{2}Ae^{-2\nu}.
\]

Differentiating this with respect to \( \nu \), equating to zero, and solving gives

\[
e^{-2\nu} = \frac{n+2}{A}.
\]

Substituting \( e^\nu = \sigma \) and \( \mu = \bar{y} \) gives

\[
\text{mode}(\mu, \sigma^2 | y) = (\bar{y}, \frac{n-1}{n+2}s^2).
\]

(3)

The marginal posterior pdf of \( \sigma^2 \) is obtained by integrating over \( \mu \):

\[
p(\sigma^2 | y) = \int_{-\infty}^{\infty} p(\mu, \sigma^2 | y) \, d\mu
\]

\[
= (\sigma^2)^{-\left(\frac{n}{2}+1\right)}e^{-\frac{(n-1)s^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{n(y - \mu)^2}{2\sigma^2}} \, d\mu
\]

\[
\propto (\sigma^2)^{-\left(\frac{n}{2}+1\right)}e^{-\frac{(n-1)s^2}{2\sigma^2}},
\]

that is, \( \sigma^2 | y \sim \text{InvGam}\left(\frac{n-1}{2}, \frac{n-1}{2}s^2\right) \), for which

\[
E(\sigma^2 | y) = \frac{n-1}{n-3}s^2, \quad \text{mode}(\sigma^2 | y) = \frac{n-1}{n+1}s^2, \quad V(\sigma^2 | y) = \frac{2(n-1)^2s^4}{(n-3)^2(n-5)}.
\]

Notice that the mode of the marginal posterior is different from (a bit larger than) the joint posterior mode’s \( \sigma^2 \) value given in (3).

The following general result will be useful.

**Lemma 1** If \( x | w \sim \text{Normal}(0, w) \) and \( w \sim \text{InvGam}\left(\frac{m}{2}, \frac{m}{2}S^2\right) \) then \( \frac{1}{w} \sim \tau_m \), a standard Student-t distribution with \( m \) degrees of freedom.

**Proof:** The marginal pdf is

\[
p(x) = \int_0^{\infty} p(x | w) p(w) \, dw \propto \int w^{-\frac{1}{2}}e^{-\frac{x^2}{2w}}w^{-(m+1)}e^{-\frac{mw^2}{2}} \, dw
\]

\[
= \left(\frac{x^2 + mS^2}{2}\right)^{-(m+1)/2} \left[ z^{-(m+3)/2}e^{-1/z} \, dz \right] \quad [ \text{integral of an inverse-gamma pdf}]
\]

\[
\propto \left(\frac{x^2}{mS^2}\right)^{-(m+1)/2} \left[ \int \right]
\]

4
From (2) we have $\mu | \sigma^2, y \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$, so that $\frac{\mu - \bar{y}}{\sqrt{s_i} / \sqrt{n}} | \sigma^2, y \sim \text{Normal}(0, \sigma^2)$. Then by Lemma 1 we have $\mu | y \sim t_{n-1}$, that is, $\mu | y \sim t_{n-1}(\bar{y}, \frac{\sigma^2}{n})$, and

$$E(\mu | y) = \bar{y}, \quad \text{mode}(\mu | y) = \bar{y}, \quad V(\mu | y) = \frac{n - 1}{n - 3} \frac{s^2}{n}.$$  

The Student-t distribution roughly resembles a Normal distribution but has heavier tails. This marginal posterior distribution has the same mean as that of the posterior $\mu | y \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$ that we found in section 5.1 for the one-parameter normal model with known variance $\nu$ and uniform prior $p(\mu) \propto 1$.

Next, we find the posterior predictive distribution. The model is $\bar{y} | \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$ (conditionally independent of $y$ given $\mu, \sigma^2$). Because $\bar{y} - \mu | \sigma^2, y \sim \text{Normal}(0, \sigma^2)$ and $\mu | \sigma^2, y \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$ are independent given $\sigma^2, y$, we have $\bar{y} | \sigma^2, y \sim \text{Normal}(\bar{y}, (1 + \frac{1}{n})\sigma^2)$, that is, $\bar{y} | \sigma^2, y \sim \text{Normal}(\bar{y}, (1 + \frac{1}{n})\sigma^2)$. Then by Lemma 1, we obtain $\frac{\bar{y} - \bar{y}}{(1 + \frac{1}{n})^{1/2} s} \sim t_{n-1}$, that is, $\bar{y} | y \sim t_{n-1}(\bar{y}, (1 + \frac{1}{n})s^2)$, for which

$$E(\bar{y} | y) = \bar{y}, \quad \text{mode}(\bar{y} | y) = \bar{y}, \quad V(\bar{y} | y) = \frac{n - 1}{n - 3} (1 + \frac{1}{n} s^2).$$  

The formulas derived in this section can be generalised by considering families of conjugate prior distributions. However, the formulas and their derivations do not add much insight to the results already derived. We proceed instead to look at how a two-parameter normal model can be analysed using numerical simulation.

**Example: Two-parameter normal model for Cavendish’s data**  We have $n = 23$, $\bar{y} = 5.4848$, and $s^2 = (0.1924)^2 = 0.0370$. With the prior $p(\mu, \sigma^2) = 1/\sigma^2$, the posterior pdf’s are:

\[
\begin{align*}
\text{mode}(\mu, \sigma^2 | y) &= (5.4848, 0.0326), \\
E(\mu | y) &= 5.4848, \quad \text{mode}(\mu | y) = 5.4848, \quad V(\mu | y) = 0.0018, \\
E(\sigma^2 | y) &= 0.0407, \quad \text{mode}(\sigma^2 | y) = 0.0339, \quad V(\sigma^2 | y) = 1.84 \cdot 10^{-4}, \\
E(\bar{y} | y) &= 5.4848, \quad \text{mode}(\bar{y} | y) = 5.4848, \quad V(\bar{y} | y) = 0.0425.
\end{align*}
\]

For a WinBUGS model, we assume $\mu$ and $\sigma^2$ to be independent a priori. As in section 4.2, we choose $\mu \sim \text{Normal}(5, 0.5)$. Our prior for $\sigma^2$ is based on
the judgement that $\sigma^2 \approx 0.04 \pm 0.02$. Assuming $\sigma^2 \sim \text{InvGam}(\alpha, \beta)$ and solving $E(\sigma^2) = 0.04$ and $V(\sigma^2) = 0.02^2$ for $\alpha$ and $\beta$, we obtain $\sigma^2 \sim \text{InvGam}(6, 0.2)$. The corresponding prior distribution for the precision $\tau = 1/\sigma^2$ is $\tau \sim \text{Gamma}(6, 0.2)$.

The WinBUGS model is

```winbugs
model {
  for (i in 1:n) { y[i] ~ dnorm(mu, tau) }
  mu ~ dnorm(5, 2)
  tau ~ dgamma(6, 0.2)
  sigma2 <- 1/tau
  ypred ~ dnorm(mu, tau)
}
```

The double link in the DAG and the “<-” in the code denote the logical function that specifies $\sigma^2$ as a deterministic function of $\tau$. The results after 2000 simulation steps are

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu</td>
<td>5.483</td>
<td>0.04028</td>
<td>5.403</td>
<td>5.483</td>
<td>5.56</td>
</tr>
<tr>
<td>sigma2</td>
<td>0.03767</td>
<td>0.0097</td>
<td>0.023</td>
<td>0.036</td>
<td>0.060</td>
</tr>
<tr>
<td>ypred</td>
<td>5.483</td>
<td>0.195</td>
<td>5.116</td>
<td>5.482</td>
<td>5.869</td>
</tr>
</tbody>
</table>

The results after 2000 simulation steps are

- **mu sample**: 2000
- **sigma2 sample**: 2000
- **ypred sample**: 2000