How should we choose a prior distribution? Several schools of thought answer this question differently:

- Subjective Bayesianism: The prior should reflect in as much detail as possible the researcher's prior knowledge of and uncertainties about the problem. These should be determined through *prior elicitiation*.

- Objective Bayesianism: The prior should incorporate as little information as possible. Priors with this property are known as *non-informative*.

- Robust Bayesianism: Reasonable people may hold different priors, and it is difficult to precisely express even one person's prior; we should therefore consider the *sensitivity* of our inferences to changes in the prior.

A Bayesian analysis will often incorporate more than one of these ideas.

The simplest kind of non-informative prior places a uniform distribution on θ. When the range of θ is bounded, this prior gives a valid PDF, since it integrates to 1.

It is also possible to assign a uniform prior when the range of θ is not bounded. For example, suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$. We could take $f(\theta) \propto 1$. This prior is called "improper," since $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$.

However, we can still apply the Bayesian machinery to get

$$\begin{aligned} f(\theta|x^n) &\propto f(x^n|\theta)f(\theta) \\ &\propto \exp\left\{-\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n]\right\} \end{aligned}$$

which is the kernel of a $N(\bar{X}_n, 1/n)$ distribution for θ. Therefore we still have a "proper posterior."

One property we might want a noninformative prior to possess is that it be transformation invariant. For example, if θ represents a distance, our inference shouldn't depend on whether θ is expressed in miles or kilometers.

That is, if instead of expressing the likelihood given θ, we express it given $\phi = g(\theta)$, we want a rule for choosing the priors f_θ and f_ϕ such that

$$f_\phi(\phi) = f_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|$$

The Jeffreys prior does just this. For 1-D θ, we have $f(\theta) \propto I(\theta)^{1/2}$. The multivariate case is

$$f(\theta) \propto det(I(\theta))^{1/2}$$

Example:

1. Find the Jeffreys prior for λ when $X_1, \ldots, X_n \overset{iid}{\sim} Pois(\lambda)$.

2. Is the Jeffreys prior proper?

3. Find the implied prior distribution for $\phi = \log \lambda$.

4. Show the prior in 3 is the same as the Jeffreys prior for ϕ.

In a Bayesian analysis, hypotheses, like parameters, can be described using probability distributions.

The simplest case is when the hypotheses describe regions into which θ can fall, and these all have positive prior probability. If $H_0 : \theta \in \Theta_0$, then

$$\begin{aligned} \text{Prior probability: } P(H_0) &= \int_{\Theta_0} f(\theta) d\theta \\ \text{Posterior probability: } P(H_0|x^n) &= \int_{\Theta_0} f(\theta|x^n) d\theta \end{aligned}$$

Example: Consider the earthquake example in your homework. Suppose the null hypothesis is "earthquakes of 3+ magnitude occur within a 10 mile radius of Berkeley more than twice per year". λ represents the average waiting time between earthquakes, so $H_0 : \lambda < 365/2$. Now calculate the prior and posterior probability of this using `pbeta` in `R`.

Suppose $H_0, \ldots, H_{K-1}$ are K hypotheses under consideration. (Typically $K = 2$, but in theory we can have more.) Suppose that under H_k, $\theta \sim f(\theta|H_k)$. θ may mean different things under the various hypotheses.

Note that

$$P(H_k|x^n) = \frac{f(x^n|H_k)P(H_k)}{\sum_{k=1}^{K} f(x^n|H_k)P(H_k)}$$

Therefore, the posterior odds of H_i relative to H_j equals

$$\frac{P(H_i|x^n)}{P(H_j|x^n)} = \frac{f(x^n|H_i)}{f(x^n|H_j)} \times \frac{P(H_i)}{P(H_j)}$$

The term $f(x^n|H_i)/f(x^n|H_j)$ is called the Bayes Factor for comparing H_i to H_j. I'll denote it BF_{ij}.

Computing the Bayes Factor

When H_i and H_j represent regions of the parameter space, it's easier to calculate the prior and posterior odds, and from this compute the Bayes Factor.

If $H_i : \theta = \theta_i$ and $H_j : \theta = \theta_j$, then the Bayes Factor is just the ratio of likelihoods under the two values.

More generally,

$$f(x^n|H_i) = \int_\Theta f(x^n|\theta, H_i) f(\theta|H_i) d\theta,$$

which is called the marginal likelihood. If $f(\theta|H_i)$ is conjugate, it can be calculated in closed form. Otherwise, we use sampling to approximate it. For example, we could use MC integration, sampling from $f(\theta|H_i)$.

Example: Albert Pujols (St. Louis Cardinals) was recently voted the "most feared hitter in baseball." However, Ichiro Suzuki (Seattle Mariners) has a very similar batting average. Here are their career statistics since 2001, when they both started playing major league baseball.

Pujols: 5146 at bats; 1717 hits
Suzuki: 6099 at bats; 2030 hits

If we consider that each player has a "true" batting average p, around which their actual batting average fluctuates, we might be interested in looking at evidence for/against the hypothesis $p_{Pujols} = p_{Suzuki}$.

Suppose $X|p_1 \sim Bin(n, p_1)$ and $Y|p_2 \sim (m, p_2)$. Under $H_0 : p_1 = p_2$, we assign prior distribution $p_1 \sim Unif(0, 1)$ (and $p_2 = p_1$) and under $H_1 : p_1 \neq p_2$, we assign independent priors $p_1 \sim Unif(0, 1)$ and $p_2 \sim Unif(0, 1)$.

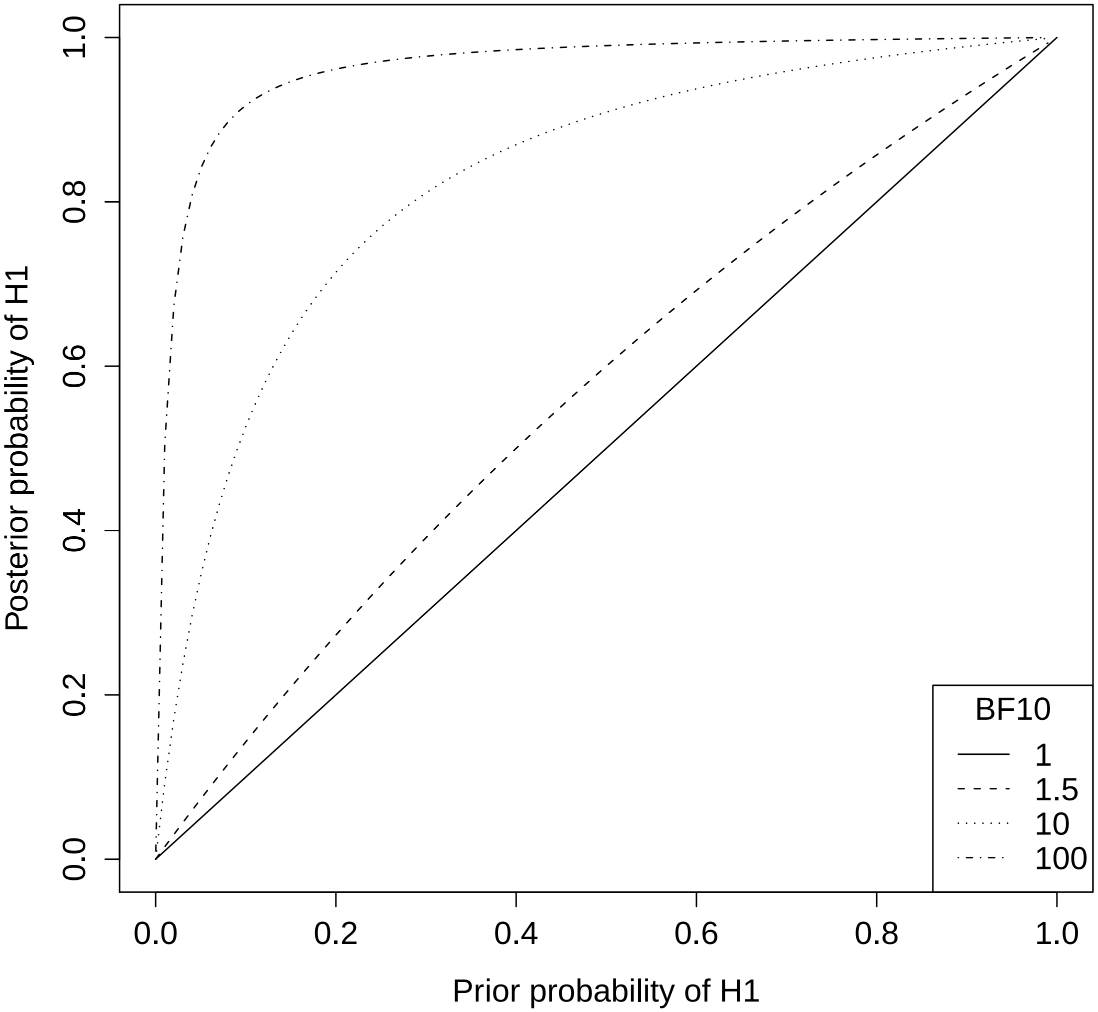
Calculate $f(X|H_1)$. (The rest of the problem is in the homework.)

Let BF_{10} be the Bayes Factor for comparing H_1 to H_0. We might classify BF_{10} as a measure of evidence against H_0 and in favor of H_1 as follows

$log_{10}(B_{10})$	B_{10}	Evidence
$0 - 1/2$	$1 - 3/2$	Weak
$1/2 - 1$	$3.2 - 10$	Moderate
$1 - 2$	$10 - 100$	Strong
> 2	> 100	Decisive

The key to this interpretation is to note that if $p = P(H_1)$ and $p^* = P(H_1|Data)$, then

$$p^* = \frac{\frac{p}{1-p}BF_{10}}{1 + \frac{p}{1-p}BF_{10}}$$


Posterior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
Prior probability of H1
0.0
0.2
0.4
0.6
0.8
1.0
BF10
1
1.5
10
100

Decision Theory

Statistical decision theory is concerned with making decisions under uncertainty. We express our uncertainties around the problem in terms of an unknown quantity or "state of nature" θ.

The particular decision made is also referred to as an "action," and we'll denote it by a, with the collection of all possible actions denoted by $\mathcal{A}$.

A loss function

$$L(\theta, a) : (\Theta \times \mathcal{A}) \to [0, \infty)$$

describes the consequences of taking action a when the true state of nature is θ. In reality, we never know the true value of the loss (at least not at the time of the decision).

Example: A drug company has developed a new pain reliever. They are trying to determine how much of the drug to produce, but they are uncertain about the proportion of the market the drug will capture (θ).

Suppose a is an estimate of θ. The company plans to produce an amount proportional to a. One possible loss function is

$$L(\theta, a) = \begin{cases} K(\theta - a) & a - \theta < 0 \\ 2K(a - \theta) & a - \theta \geq 0 \end{cases}$$

for some constant K. This loss function implies that an overestimate of demand (leading to overproduction of the drug) is considered twice as costly as an underestimate. The loss is also taken to be linear, which may be reasonable if the total cost is proportional to the number of units produced.

Many results are based on the following “standard” loss functions. These are expressed in generic “units of utility.”

- Squared error loss: $L(\theta, a) = (\theta - a)^2$
- Linear loss: $L(\theta, a) = \begin{cases} K_1(\theta - a) & a - \theta < 0 \\ K_2(a - \theta) & a - \theta \geq 0 \end{cases}$
- Absolute error loss: $L(\theta, a) = |\theta - a|$ (linear loss with $K_1 = K_2$)
- L^p loss: $L(\theta, a) = |\theta - a|^p$
- Zero-one loss: $L(\theta, a) = \begin{cases} 0 & a = \theta \\ 1 & a \neq \theta \end{cases}$

Since we don't know the actual loss, we may consider an "expected loss" and then choose an "optimal" decision with respect to this. This "expected loss" is known as risk. However, there are several ways of thinking about the expectation; hence, several different risks.

Note: in what follows, we'll consider estimation problems only, that is, actions $a = \hat{\theta}(x)$.

1. The posterior risk

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta$$

averages over uncertainty in θ after conditioning on observations x. We may think of it as a function of x, as well as the particular form of $\hat{\theta}$. Another way to think of this is that, conditional on the x we observed, we just get a single number for each estimator $\hat{\theta}$ we might consider.

2. The frequentist risk (or sometimes just "risk")

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

averages over different possible realizations x of the random variable, given that the true "state of nature" is θ. It is a function of θ, as well as the particular form of $\hat{\theta}$.

Consider two estimators, $\hat{\theta}$ and $\hat{\theta}'$. We say $\hat{\theta}'$ dominates $\hat{\theta}$ if

$$\begin{aligned} R(\theta, \hat{\theta}') &\leq R(\theta, \hat{\theta}) \text{ for all } \theta \\ R(\theta, \hat{\theta}') &< R(\theta, \hat{\theta}) \text{ for at least one } \theta \end{aligned}$$

The estimator $\hat{\theta}$ is called inadmissible if there is at least one other estimator $\hat{\theta}'$ that dominates it. Otherwise it is called admissible.

Example: Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate the risk in terms of c and θ.

- Calculate the risk when $c = 1$.

- Show that $\hat{\theta}_c$ is inadmissible when $c > 1$.

- Make a plot comparing the risk when $c = 1/2$ and $c = 1$.

3. The Bayes risk:

$$\begin{aligned}
r(f,\hat{\theta}) &= \int R(\theta,\hat{\theta}) f(\theta) d\theta \\
&= \int \left[\int L(\theta,\hat{\theta}(x)) f(x|\theta) dx \right] f(\theta) d\theta \\
&= \int \int L(\theta,\hat{\theta}(x)) f(x,\theta) dx\, d\theta \\
&= \int \left[\int L(\theta,\hat{\theta}(x)) f(\theta|x) d\theta \right] f(x) dx \\
&= \int r(\hat{\theta}|x) f(x) dx
\end{aligned}$$

averages over both θ and X. It depends on the particular form of $\hat{\theta}$.

A decision rule that minimizes the Bayes risk is called a Bayes rule. The estimator $\hat{\theta}$ is a Bayes rule, or Bayes estimator (under a particular model and loss function) if

$$r(f,\hat{\theta}) = \inf_{\tilde{\theta}} r(f,\tilde{\theta})$$

We can also find the Bayes estimator using the posterior risk. For each x, let $\hat{\theta}(x)$ be the value of $\hat{\theta}$ that minimizes $r(\hat{\theta}|x)$. (Recall that for each x, $r(\hat{\theta}|x)$ returns a single number for each $\hat{\theta}$.) The estimator defined in this way is the Bayes estimator. This is because

$$r(f,\hat{\theta}) = \int f(\hat{\theta}|x)f(x)dx$$

and we have defined this $\hat{\theta}$ to minimize the quantity being integrated for each x; hence we've also minimized the whole integral.

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate the Bayes risk using the prior $\theta \sim N(0, \tau^2)$.

- Find the Bayes rule among estimators $\hat{\theta}_c$.

- Find the Bayes risk of this estimator.

We can calculate the Bayes rule explicitly for several standard loss functions.

- Squared error loss: posterior mean

- Absolute error loss: posterior median

- Zero-one loss: posterior mode

Another way of summarizing the frequentist risk function is to look at its maximum. Define

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta})$$

A decision rule that minimizes the maximum frequentist risk is called a minimax rule. The estimator $\hat{\theta}$ is a minimax rule (under a particular loss function) if

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate $\sup_{\theta} R(\theta, \hat{\theta}_c)$.

- Use this to determine the minimax estimator of θ.