Hypothesis Testing I

In Bayesian inference, hypotheses are represented by prior distributions. There is nothing special about H_0, and H_0 and H_1 need not be nested.

Let $\pi(\theta|H_0)$, $\theta \in \Theta_0$ be the prior distribution of θ under hypothesis H_0. Here $\pi(\theta|H_0)$ is a pmf/pfd as $\theta|H_0$ is continuous/discrete.

Composite $H_0 : \theta \in \Theta_0$, with $\Theta_0 \subset \Theta$,

$$\pi(\theta|H_0) = \frac{\pi(\theta)}{\pi(\theta \in \Theta_0)}\mathbb{I}(\theta \in \Theta_0)$$

If Θ_0 has more than one element then H_0 is a composite hypothesis.

Simple $H_0 : \theta = \theta_0$, so that $\pi(\theta_0|H_0) = 1$. This is a simple hypothesis.

However, since any statement about the form of the prior amounts to a hypothesis about θ, we are not restricted to statements about set membership (not just simple and composite).

Example 1 In a quality inspection program components are selected at random from a batch and tested. Let θ denote the failure probability. Suppose that we want to test for $H_0 : \theta \leq 0.2$ against $H_1 : \theta > 0.2$ and that the prior is $\theta \sim \text{Beta}(2,5)$ so that

$$\pi(\theta) = 30\theta(1-\theta)^4,\ 0 < \theta < 1.$$

Now if $\pi(H_0) = \pi(\theta \in \Theta_0)$ then $\pi(H_0) = \int_0^{0.2} 30\theta(1-\theta)^4 d\theta$ so that $\pi(H_0) \simeq 0.345$ and $\pi(H_0) \simeq 1 - 0.345$ so

$$\pi(\theta|H_0) = \frac{30\theta(1-\theta)^4}{\pi(H_0)},\ 0 < \theta \leq 0.2$$

and

$$\pi(\theta|H_1) = \frac{30\theta(1-\theta)^4}{\pi(H_1)},\ 0.2 < \theta < 1$$

Marginal Likelihood

$P(x|H_0)$ is called the marginal likelihood (for hypothesis H_0). We can think of a parameter $H \in \{H_0, H_1\}$, with likelihood $P(x|H)$, prior $P(H)$ and posterior $P(H|x)$.

By the partition theorem for probability ($\pi(\theta|H_0)$ a pdf, say)

$$
\begin{aligned}
P(x|H_0) &= \int_{\Theta_0} P(x|\theta, H_0) P(\theta|H_0) d\theta \\
&= \int_{\Theta_0} L(\theta; x) \pi(\theta|H_0) d\theta,
\end{aligned}
$$

since, given θ, x is determined by the observation model, and independent of the process (H_0) that generated θ.

In the discrete case

$$P(x|H_0) = \sum_{\theta \in \Theta_0} L(\theta; x)\pi(\theta|H_0).$$

In the special case that H_0 is a simple hypothesis $\Theta_0 = \{\theta_0\}$, $\pi(\theta_0|H_0) = 1$, and

$$P(x|H_0) = L(\theta_0; x).$$

Example 1 (cont) In the quality inspection program suppose n components are selected for independent testing. The number X that fail is $X \sim$ Binomial(n, θ). Recall $H_0 : \theta \leq 0.2$ with $\theta \sim$ Beta$(2, 5)$ in the prior.

The marginal likelihood for H_0 is

$$
\begin{aligned}
P(x|H_0) &= \int_{\Theta_0} L(\theta; x)\pi(\theta|H_0)d\theta \\
&= \binom{5}{x} \int_0^{0.2} \theta^x (1-\theta)^{n-x} \frac{30\theta(1-\theta)^4}{\pi(H_0)} d\theta
\end{aligned}
$$

For one batch of size $n = 5$, $X = 0$ is observed. Recall that

$\pi(H_0) \simeq 0.345$. Then

$$
\begin{aligned}
P(x|H_0) &= \binom{5}{0} \int_0^{0.2} \frac{30\theta(1-\theta)^9}{\pi(H_0)} d\theta \\
&\simeq 0.185/0.345 = 0.536.
\end{aligned}
$$

Similarly, for $H_1 : \theta > 0.2$

$$
\begin{aligned}
P(x|H_1) &= \binom{5}{0} \int_{0.2}^{1} \frac{30\theta(1-\theta)^9}{\pi(H_1)} d\theta \\
&\simeq 0.134.
\end{aligned}
$$

Notice (skip this slide at first reading) that

(i)

$$P(x|H_0) = \mathbb{E}(L(\vartheta; x)|H_0),$$

that is, the marginal likelihood is the average likelihood given the prior $\pi(\theta|H_0)$, and

(ii) the marginal likelihood is the normalizing constant we often leave off when we write the posterior

$$\pi(\theta|x, H_0) = \frac{L(\theta; x)\pi(x|H_0)}{P(x|H_0)},$$

$$\text{posterior} = \frac{\text{likelihood} \quad \times \quad \text{prior}}{\text{marginal likelihood}},$$

Prior and Posterior Probabilities for Hypotheses

We have a posterior probability for H_0 itself. This is actually where we started with Bayesian inference. In the simple case where we have two hypotheses H_0, H_1, exactly one of which is true,

$$P(H_0 \mid x) = \frac{P(H_0)P(x \mid H_0)}{P(x)},$$

where

$$P(x) = P(H_0)P(x \mid H_0) + P(H_1)P(x \mid H_1)$$

so that $P(H_0 \mid x) + P(H_1 \mid x) = 1$.

When we estimate the value of a discrete parameter $H \in \{H_0, H_1\}$, we are making a Bayesian hypothesis test.

Example 1 (cont) $X \sim$ Binomial$(5, \theta)$ with $\theta \sim$ Beta$(2, 5)$ in the prior and $H_0 : \theta \leq 0.2$ and $H_1 : \theta > 0.2$. The posterior probability for H_0 given we observe $X = 0$ is

$$
\begin{aligned}
P(H_0|x) &= \frac{P(x|H_0)P(H_0)}{P(x)} \\
P(H_0) &= \frac{P(\theta \in \Theta_0)}{(P(\theta \in \Theta_0) + P(\theta \in \Theta_1))} = \pi(H_0) \\
P(x|H_0)\pi(H_0) &\simeq 0.185 \\
P(x|H_1)\pi(H_1) &\simeq 0.088 \\
P(x) &\simeq P(x|H_0)P(H_0) + P(x|H_1)P(H_1) \\
&\simeq 0.273 \\
P(H_0|x) &\simeq 0.185/0.273 \\
&= 0.678 \\
P(H_1|x) &\simeq 0.322
\end{aligned}
$$

Hypothesis Testing II, Bayes factors: Suppose we have two hypotheses H_0, H_1, exactly one of which is true. Data x.

The Prior Odds

$$Q = \frac{P[H_0]}{P[H_1]}$$

These are prior odds since $P[H_1] = 1 - P[H_0]$. Here H_0 *is* Q *times more probable than* H_1*, given the prior model.*

The Posterior Odds

$$Q^* = \frac{P[H_0 \mid x]}{P[H_1 \mid x]}$$

are the posterior odds, so that H_0 *is* Q^* *times more probable than* H_1*, given the data and prior model.*

The posterior odds for H_0 against H_1 can be written

$$Q^* = \frac{P[H_0]}{P[H_1]} \times \frac{P(x \mid H_0)}{P(x \mid H_1)} = Q \times B$$

where Q is the prior odds and

$$B = \frac{P(x \mid H_0)}{P(x \mid H_1)}$$

is the Bayes Factor.

The Bayes Factor is a criterion for model comparison since H_0 *is* B *times more probable than* H_1*, given the data and a prior model which puts equal probability on* H_0 *and* H_1. The Bayes factor tells us how the data shifts the strength of belief (measured as a probability) in H_0 relative to H_1.

Example 1 (cont) $X \sim$ Binomial$(5, \theta)$ with $\theta \sim$ Beta$(2, 5)$ in the prior and $H_0 : \theta \leq 0.2$ and $H_1 : \theta > 0.2$.

The prior odds are

$$\begin{aligned} Q &= P(H_0)/P(H_1) \\ &\simeq 0.345/(1 - 0.345) \simeq 0.527 \end{aligned}$$

The posterior odds are

$$\begin{aligned} Q^* &= P(H_0|x)/P(H_1|x) \\ &\simeq 0.678/(1 - 0.678) \simeq 2.1 \end{aligned}$$

The Bayes factor comparing H_0 and H_1 is

$$\begin{aligned} B &= \frac{P(x|H_0)}{P(x|H_1)} \\ &\simeq 0.536/0.134 = 4 \end{aligned}$$

Explicitly, from the beginning,

$$
\begin{aligned}
B &= \frac{\int_{\Theta_0} L(x;\theta)\pi(\theta|H_0)d\theta}{\int_{\Theta_1} L(x;\theta)\pi(\theta|H_1)d\theta} \\
&= \frac{\int_{\Theta_0} L(x;\theta)\pi(\theta)d\theta}{\int_{\Theta_1} L(x;\theta)\pi(\theta)d\theta} \times \frac{\pi(H_1)}{\pi(H_0)} \\
&= \frac{\binom{5}{0}\int_0^{0.2} 30\theta(1-\theta)^9 d\theta}{\binom{5}{0}\int_{0.2}^{1} 30\theta(1-\theta)^9 d\theta} \frac{\int_{0.2}^{1} 30\theta(1-\theta)^4 d\theta}{\int_0^{0.2} 30\theta(1-\theta)^4 d\theta} \\
&= 6619897/1654272 \simeq 4.002 \qquad \text{(Maple)}.
\end{aligned}
$$

This is 'positive' evidence for $\theta \leq 0.2$. Notice that the Bayes factor is 'more positive' than the posterior odds, as the prior odds were weighted against H_0.

Adrian Raftery gives this table (values are approximate, and adapted from a table due to Jeffreys) interpreting B.

'$P(H_0 \mid x)$'	B	$2\log(B)$	evidence for H_0
< 0.5	$<$ 1	$<$ 0	negative (supports H_1)
0.5 to 0.75	1 to 3	0 to 2	barely worth mentioning
0.75 to 0.92	3 to 12	2 to 5	positive
0.92 to 0.99	12 to 150	5 to 10	strong
> 0.99	$>$ 150	$>$ 10	very strong

I added the leftmost column (posterior for prior odds equal one). We sometimes report $2\log(B)$ because it is on the same scale as the familiar deviance and likelihood ratio test statistic.

Simple-Simple and Simple-Composite hypothesis

If both hypotheses are simple $H_0 : \theta = \theta_0$; $H_1 : \theta = \theta_1$, with priors $P(H_0)$ and $P(H_1)$ for the two hypotheses, the posterior probability for H_0 is

$$
\begin{aligned}
P(H_0|x) &= \frac{P(x|H_0)P(H_0)}{P(x)} \\
&= \frac{L(\theta_0; x)P(H_0)}{L(\theta_0; x)P(H_0) + L(\theta_1; x)P(H_1)},
\end{aligned}
$$

since $P(x|H_0)$ is just $L(\theta_0; x)$. The Bayes factor is then just likelihood ratio

$$B = \frac{L(\theta_0; x)}{L(\theta_1; x)}.$$

If one hypothesis is simple and the other composite, for example, $H_0 : \theta = \theta_0; \quad H_1 : \pi(\theta|H_1),\ \theta \in \Theta$, with priors $P(H_0)$ and $P(H_1)$ for the two hypotheses, the Bayes factor is

$$B = \frac{L(x;\theta_0)}{\int_\Theta L(x;\theta)\pi(\theta|H_1)d\theta}$$

The denominator is just $\int_\Theta L(x;\theta)\pi(\theta)d\theta$ when $\pi(\theta|H_1)$ is a pdf.

Exercise Show that the posterior probability for H_0 is

$$P(H_0|x) = \frac{L(\theta_0;x)P(H_0)}{P(H_0)L(\theta_0;x) + P(H_1)\int_\Theta L(x;\theta)\pi(\theta)d\theta}$$

when $\pi(\theta|H_1)$ is a pdf, in this simple-composite comparison.

Example $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$, with σ^2 known.
$H_0 : \theta = 0$, $H_1 : \theta | H_1 \sim N(\mu, \tau^2)$. Bayes factor is P_0/P_1, where

$$
\begin{aligned}
P_0 &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right) \\
P_1 &= (2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\sum (x_i - \theta)^2\right) \\
&\qquad \times (2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\theta-\mu)^2}{2\tau^2}\right) d\theta.
\end{aligned}
$$

Completing the square in P_1 and integrating $d\theta$,

$$
\begin{aligned}
P_1 &= (2\pi\sigma^2)^{-n/2} \left(\frac{\sigma^2}{n\tau^2 + \sigma^2}\right)^{1/2} \\
&\qquad \times \exp\left[-\frac{1}{2}\left\{\frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2 + \frac{1}{\sigma^2}\sum (x_i - \bar{x})^2\right\}\right]
\end{aligned}
$$

so that

$$B = \left(1 + \frac{n\tau^2}{\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left\{\frac{n\bar{x}^2}{\sigma^2} - \frac{n}{n\tau^2+\sigma^2}(\bar{x}-\mu)^2\right\}\right]$$

Defining $t = \sqrt{n}\bar{x}/\sigma, \eta = -\mu/\tau, \rho = \sigma/(\tau\sqrt{n})$, this can be written as

$$B = \left(1 + \frac{1}{\rho^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left\{\frac{(t-\rho\eta)^2}{1+\rho^2} - \eta^2\right\}\right]$$

This example illustrates a problem choosing the prior. If we take a diffuse prior, for ρ so that $\rho \to 0$, then $B \to \infty$, giving overwelming support for H_0.

This is an instance of Lindley's paradox. The point here is that B compares the *models* $\theta = \theta_0$ and $\theta \sim \pi(\cdot|H_1)$, not the *sets*

θ_0 against $\theta \setminus \{\theta_0\}$. If the $\pi(\theta|H_1)$-prior becomes very diffuse then the *average* likelihood (ie $P(H_1|x)$, the marginal likelihood, which is the denominator of B) goes to zero, while $P(H_0|x) = L(\theta_0; x)$ is fixed.

Decision Theory (see Young and Smith 2005, Ch 2, GJJ Ch 6)

Decision Theory sits 'above' Bayesian and classical statistical inference and gives us a basis to compare different approaches to statistical inference.

We make decisions by applying rules to data. Decisions are subject to risk. A risk function specifies the expected loss which follows from the application of a given rule, and this is a basis for comparing rules. We may choose a rule to minimize the maximum risk, or we may choose a rule to minimize the average risk.

Decision Theory Terminology (examples from Point Estimation)

θ is the 'true state of nature', $X \sim f(x;\theta)$ is the data.

The Decision rule is δ. If $X = x$, adopt the action $\delta(x)$ given by the rule.

Example: A single parameter θ is estimated from data $X = x$ by $\widehat{\theta}(x)$. The rule $\widehat{\theta}$ is the functional form of the estimator, it's value, the action.

The Loss function $L_S(\theta, \delta(x))$ measures the loss from action $\delta(x)$ when θ holds.

Example: $L_S(\theta, \widehat{\theta}(x))$ is the loss function which increases for $\widehat{\theta}(x)$ being away from θ. Here are three common loss functions.

1. Zero-One loss

$$L_S(\theta, \widehat{\theta}(x)) = \begin{cases} 0 & |\widehat{\theta}(x) - \theta| < b \\ a & |\widehat{\theta}(x) - \theta| \geq b \end{cases}$$

where a, b are constants.

2. Absolute error loss

$$L_S(\theta, \widehat{\theta}(x)) = a|\widehat{\theta}(x) - \theta|$$

where $a > 0$.

3. Quadratic loss

$$L_S(\theta, \widehat{\theta}(x)) = (\widehat{\theta}(x) - \theta)^2.$$

Definition The risk function $R(\theta, \delta)$ is defined as

$$R(\theta, \delta) = \int L_S(\theta, \delta(x)) f(x; \theta) dx,$$

ie, the expected loss.

Example: in the context of point estimation, with Quadratic Loss, the risk function is the mean square error,

$$R(\theta, \widehat{\theta}) = \mathbb{E}[(\widehat{\theta}(X) - \theta)^2].$$

Definition A procedure δ_1 is inadmissible if there exists another procedure δ_2 such that

$$R(\theta, \delta_1) \geq R(\theta, \delta_2), \text{ for all } \theta \in \Theta$$

with $R(\theta, \delta_1) > R(\theta, \delta_2)$ for at least some θ. A procedure which is not inadmissible is admissible.

Example: suppose $X \sim U(0, \theta)$. Consider estimators of the form $\widehat{\theta}(x) = ax$ (this is a family of decisions rules indexed by a). Show that $a = 3/2$ is a necessary condition for the rule $\widehat{\theta}$ to be admissible for quadratic loss.

$$\begin{aligned} R(\theta, \widehat{\theta}) &= \int_0^\theta (ax - \theta)^2 \frac{1}{\theta} dx \\ &= (a^2/3 - a + 1)\theta^2 \end{aligned}$$

and R is minimized at $a = 3/2$. This does not show $\widehat{\theta}(x) = 3x/2$ is admissible here. It does show that if a takes any other value then $\widehat{\theta}(x) = ax$ is not admissible.