Solutions to APTS Assessment on Statistical Inference

Simon Shaw, s.shaw@bath.ac.uk
University of Bath

Warwick, 13-16 December 2022

Principles for Statistical Inference

1. **Consider Birnbaum's Theorem, (WIP $\wedge$ WCP) $\leftrightarrow$ SLP. In lectures, we showed that (WIP $\wedge$ WCP) $\rightarrow$ SLP but not the converse. Hence, show that SLP $\rightarrow$ WIP and SLP $\rightarrow$ WCP.**

   The Strong Likelihood Principle (SLP) states that if $\mathcal{E}_1 = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \,|\, \theta)\}$ and $\mathcal{E}_2 = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \,|\, \theta)\}$ are two experiments with the same parameter θ and if $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$ for some $c > 0$ for all $\theta \in \Theta$ then $Ev(\mathcal{E}_1, x_1) = Ev(\mathcal{E}_2, x_2)$.

   (a) SLP $\rightarrow$ WIP.
   The Weak Indifference Principle (WIP) states that for the experiment $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$ if $f_X(x \,|\, \theta) = f_X(x' \,|\, \theta)$ for all $\theta \in \Theta$ then $Ev(\mathcal{E}, x) = Ev(\mathcal{E}, x')$.

   In the SLP, let $\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}$ and suppose that $f_X(x \,|\, \theta) = f_X(x' \,|\, \theta)$ for all $\theta \in \Theta$. Hence, taking $c(x, x') = 1$, the SLP implies that $Ev(\mathcal{E}, x) = Ev(\mathcal{E}, x')$ which is the WIP.

   (b) SLP $\rightarrow$ WCP.
   The Weak Conditionality Principle (WCP) states that if $\mathcal{E}^*$ is the mixture of the experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ according to mixture probabilities $p_1, p_2 = 1 - p_1$ then $Ev(\mathcal{E}^*, (i, x_i)) = Ev(\mathcal{E}_i, x_i)$.

   For the mixture experiment we have $f^*((i, x_i) \,|\, \theta) = p_i f_{X_i}(x_i \,|\, \theta)$ for all $\theta \in \Theta$. Applying the SLP with $c((i, x_i), x_i) = p_i$ gives $Ev(\mathcal{E}^*, (i, x_i)) = Ev(\mathcal{E}_i, x_i)$ which is the WCP.

2. [1]**Suppose that we have two discrete experiments $\mathcal{E}_1 = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \,|\, \theta)\}$ and $\mathcal{E}_2 = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \,|\, \theta)\}$ and that, for $x_1' \in \mathcal{X}_1$ and $x_2' \in \mathcal{X}_2$,**

$$f_{X_1}(x_1' \,|\, \theta) \quad = \quad c f_{X_2}(x_2' \,|\, \theta) \qquad (1)$$

   for all θ where c is a positive constant not depending upon θ (but which may depend on x_1', x_2') and $f_{X_1}(x_1' \,|\, \theta) > 0$. We wish to consider estimation of

[1]See Section 5 of Berger, J. (1985). In defense of the likelihood principle: Axiomatics and coherency. *Bayesian Statistics 2* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, Eds.), 33-66. North-Holland.

θ under a loss function $L(\theta, d)$ which is strictly convex in d for each θ. Thus, for all $d_1 \neq d_2 \in \mathcal{D}$, the decision space, and $\alpha \in (0, 1)$,

$$L(\theta, \alpha d_1 + (1-\alpha)d_2) \quad < \quad \alpha L(\theta, d_1) + (1-\alpha)L(\theta, d_2).$$

For the experiment $\mathcal{E}_j$, $j = 1, 2$, for the observation x_j we will use the decision rule $\delta_j(x_j)$ as our estimate of θ so that

$$\text{Ev}(\mathcal{E}_j, x_j) \quad = \quad \delta_j(x_j).$$

Suppose that the inference violates the strong likelihood principle so that, whilst equation (1) holds, $\delta_1(x_1') \neq \delta_2(x_2')$.

(a) **Let $\mathcal{E}^*$ be the mixture of the experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ according to mixture probabilities $1/2$ and $1/2$. For the outcome (j, x_j) the decision rule is $\delta(j, x_j)$. If the Weak Conditionality Principle (WCP) applies to $\mathcal{E}^*$ show that**

$$\delta(1, x_1') \quad \neq \quad \delta(2, x_2').$$

Under the WCP, $\text{Ev}(\mathcal{E}^*, (j, x_j)) = \text{Ev}(\mathcal{E}_j, x_j)$ so that $\delta(j, x_j) = \delta_j(x_j)$. Thus, if $\delta_1(x_1') \neq \delta_2(x_2')$ it immediately follows that $\delta(1, x_1') \neq \delta(2, x_2')$.

(b) **An alternative decision rule for $\mathcal{E}^*$ is**

$$\delta^*(j, x_j) \quad = \quad \begin{cases} \frac{c}{c+1}\delta(1, x_1') + \frac{1}{c+1}\delta(2, x_2') & \text{if } x_j = x_j' \text{ for } j = 1, 2, \\ \delta(j, x_j) & \text{otherwise.} \end{cases} \tag{2}$$

Show that if the WCP applies to $\mathcal{E}^*$ then δ^* dominates δ so that δ is inadmissible.
[Hint: First show that $R(\theta, \delta^*) = \frac{1}{2}\mathbb{E}[L(\theta, \delta^*(1, X_1)) \,|\, \theta] + \frac{1}{2}\mathbb{E}[L(\theta, \delta^*(2, X_2)) \,|\, \theta]$.]

In the mixture experiment the pair (j, x_j) are random and the classical risk for δ^* is

$$\begin{aligned} R(\theta, \delta^*) &= \mathbb{E}[L(\theta, \delta^*(J, X_J)) \,|\, \theta] \\ &= \sum_j \sum_{x_j} L(\theta, \delta^*(j, x_j)) f^*((j, x_j) \,|\, \theta) \\ &= \sum_j \sum_{x_j} L(\theta, \delta^*(j, x_j)) \frac{1}{2} f_{X_j}(x_j \,|\, \theta) \\ &= \frac{1}{2}\mathbb{E}[L(\theta, \delta^*(1, X_1)) \,|\, \theta] + \frac{1}{2}\mathbb{E}[L(\theta, \delta^*(2, X_2)) \,|\, \theta]. \end{aligned} \tag{3}$$

In an identical fashion it follows that

$$R(\theta, \delta) \quad = \quad \frac{1}{2}\mathbb{E}[L(\theta, \delta(1, X_1)) \,|\, \theta] + \frac{1}{2}\mathbb{E}[L(\theta, \delta(2, X_2)) \,|\, \theta]. \tag{4}$$

Now, for each $j = 1, 2$, as $\delta^*(j, x_j) = \delta(j, x_j)$ for all $x_j \neq x_j'$,

$$\begin{aligned} \mathbb{E}[L(\theta, \delta^*(j, X_j)) \,|\, \theta] &= \sum_{x_j} L(\theta, \delta^*(j, x_j)) f_{X_j}(x_j \,|\, \theta) \\ &= \sum_{x_j} L(\theta, \delta(j, x_j)) f_{X_j}(x_j \,|\, \theta) + \{L(\theta, \delta^*(j, x_j')) - L(\theta, \delta(j, x_j'))\} f_{X_j}(x_j' \,|\, \theta) \\ &= \mathbb{E}[L(\theta, \delta(j, X_j)) \,|\, \theta] + \{L(\theta, \delta^*(j, x_j')) - L(\theta, \delta(j, x_j'))\} f_{X_j}(x_j' \,|\, \theta). \end{aligned} \tag{5}$$

Substituting, for each j, equation (5) into (3) and using (4) gives

$$\begin{aligned} R(\theta, \delta^*) &= R(\theta, \delta) + \frac{1}{2}\{L(\theta, \delta^*(1, x_1')) - L(\theta, \delta(1, x_1'))\} f_{X_1}(x_1' \,|\, \theta) + \\ &\qquad \frac{1}{2}\{L(\theta, \delta^*(2, x_2')) - L(\theta, \delta(2, x_2'))\} f_{X_2}(x_2' \,|\, \theta) \\ &= R(\theta, \delta) + \frac{1}{2}\{L(\theta, \delta^*(1, x_1')) - L(\theta, \delta(1, x_1'))\} f_{X_1}(x_1' \,|\, \theta) + \\ &\qquad \frac{1}{2c}\{L(\theta, \delta^*(2, x_2')) - L(\theta, \delta(2, x_2'))\} f_{X_1}(x_1' \,|\, \theta) \quad (6) \end{aligned}$$

using equation (1). Now, from equation (2), $\delta^*(1, x_1') = \delta^*(2, x_2')$ and so, for all θ, $L(\theta, \delta^*(1, x_1')) = L(\theta, \delta^*(2, x_2'))$. Hence, (6) becomes

$$\begin{aligned} R(\theta, \delta^*) &= R(\theta, \delta) \\ &\quad + \frac{f_{X_1}(x_1' \,|\, \theta)}{2c}\{(c+1)L(\theta, \delta^*(1, x_1')) - cL(\theta, \delta(1, x_1')) - L(\theta, \delta(2, x_2'))\} \\ &= R(\theta, \delta) + \frac{(c+1) f_{X_1}(x_1' \,|\, \theta)}{2c} A(\theta) \quad (7) \end{aligned}$$

where

$$\begin{aligned} A(\theta) &= L(\theta, \delta^*(1, x_1')) - \frac{c}{c+1} L(\theta, \delta(1, x_1')) - \frac{1}{c+1} L(\theta, \delta(2, x_2')) \\ &= L\left(\theta, \frac{c}{c+1}\delta(1, x_1') + \frac{1}{c+1}\delta(2, x_2')\right) - \\ &\quad \left(\frac{c}{c+1} L(\theta, \delta(1, x_1')) + \frac{1}{c+1} L(\theta, \delta(2, x_2'))\right) \\ &< 0 \end{aligned}$$

by the strict convexity of $L(\theta, d)$ in d for each θ as $\delta(1, x_1') \neq \delta(2, x_2')$. Hence, using equation (7), we have for each θ that

$$R(\theta, \delta^*) \quad < \quad R(\theta, \delta)$$

so that δ^* dominates δ and thus δ is inadmissible.

(c) **Comment on the result of part (b).**

Part (b) shows that if we use a decision rule which violates the SLP but retains the WCP then the corresponding decision rule of the mixture experiment, δ, also violates the SLP as $\delta(1, x_1') \neq \delta(2, x_2')$. Moreover, this rule is inadmissible and is dominated by a rule, δ^*, which does satisfy $\delta^*(1, x_1') = \delta^*(2, x_2')$ and so respects the SLP for the outcomes x_1', x_2'. As δ is inadmissible then we would not want to use it which suggests that violating the SLP is not advisable (if we accept the WCP) or a justification for not applying the WCP is required.

Statistical Decision Theory

3. **Suppose we have a hypothesis test of two simple hypotheses**

$$H_0 : X \sim f_0 \quad \text{versus} \quad H_1 : X \sim f_1$$

so that if H_i is true then X has distribution $f_i(x)$. It is proposed to choose between H_0 and H_1 using the following loss function.

		Decision H_0	Decision H_1
Outcome	H_0	c_{00}	c_{01}
	H_1	c_{10}	c_{11}

where $c_{00} < c_{01}$ and $c_{11} < c_{10}$. Thus, $c_{ij} = L(H_i, H_j)$ is the loss when the 'true' hypothesis is H_i and the decision H_j is taken. Show that a decision rule $\delta(x)$ for choosing between H_0 and H_1 is admissible if and only if

$$\delta(x) = \begin{cases} H_0 & \text{if } \frac{f_0(x)}{f_1(x)} > c, \\ H_1 & \text{if } \frac{f_0(x)}{f_1(x)} < c, \\ \text{either } H_0 \text{ or } H_1 & \text{if } \frac{f_0(x)}{f_1(x)} = c, \end{cases}$$

for some critical value $c > 0$.

For the prior distribution $\pi = (\pi_0, \pi_1)$ where $\pi_i > 0$, let $\pi^* = (\pi_0^*, \pi_1^*)$ denote the posterior distribution so that

$$\begin{aligned} \pi_0^* &= \mathbb{P}(H_0 \,|\, X = x) \\ &= \frac{f_0(x)\pi_0}{f_0(x)\pi_0 + f_1(x)\pi_1}, \\ \pi_1^* &= \mathbb{P}(H_1 \,|\, X = x) \\ &= \frac{f_1(x)\pi_1}{f_0(x)\pi_0 + f_1(x)\pi_1}. \end{aligned}$$

As we also have $f_i(x) > 0$ for all $x \in \mathcal{X}$ then $\pi_i^* > 0$. We calculate the posterior risk under the two decisions H_0 and H_1.

$$\begin{aligned} \rho(\pi^*, H_0) &= L(H_0, H_0)\pi_0^* + L(H_1, H_0)\pi_1^* \\ &= c_{00}\pi_0^* + c_{10}\pi_1^*, \qquad (8) \\ \rho(\pi^*, H_1) &= L(H_0, H_1)\pi_0^* + L(H_1, H_1)\pi_1^* \\ &= c_{01}\pi_0^* + c_{11}\pi_1^*. \qquad (9) \end{aligned}$$

Thus,

$$\begin{aligned} \rho(\pi^*, H_0) < \rho(\pi^*, H_1) &\iff c_{00}\pi_0^* + c_{10}\pi_1^* < c_{01}\pi_0^* + c_{11}\pi_1^* \\ &\iff (c_{00} - c_{01})\pi_0^* < (c_{11} - c_{10})\pi_1^* \\ &\iff \frac{\pi_0^*}{\pi_1^*} > \frac{c_{11} - c_{10}}{c_{00} - c_{01}} = \frac{c_{10} - c_{11}}{c_{01} - c_{00}} \end{aligned}$$

since $c_{00} - c_{01} < 0$ and $\pi_1^* > 0$. Using equations (8) and (9) we thus have

$$\begin{aligned} \rho(\pi^*, H_0) < \rho(\pi^*, H_1) &\iff \frac{f_0(x)\pi_0}{f_1(x)\pi_1} > \frac{c_{10} - c_{11}}{c_{01} - c_{00}} \\ &\iff \frac{f_0(x)}{f_1(x)} > \frac{(c_{10} - c_{11})\pi_1}{(c_{01} - c_{00})\pi_0} = c \qquad (10) \end{aligned}$$

since $\pi_0/\pi_1 > 0$ and thus $c > 0$. The analogous arguments show that

$$\rho(\pi^*, H_0) > \rho(\pi^*, H_1) \iff \frac{f_0(x)}{f_1(x)} < c \tag{11}$$

$$\rho(\pi^*, H_0) = \rho(\pi^*, H_1) \iff \frac{f_0(x)}{f_1(x)} = c \tag{12}$$

The decision rule $\delta(x)$ is chosen to minimise the posterior risk and so is H_0 when (10) holds, H_1 when (11) holds and is indifferent between H_0 and H_1 when (12) holds.

Wald's Complete Class Theorem states that a decision rule is admissible if and only if it is a Bayes rule for some prior distribution π with strictly positive values. Thus, all admissible decision rules have the form of $\delta(x)$.

4. **Let $X_1, \ldots, X_n$ be exchangeable random variables so that, conditional upon a parameter θ, the X_i are independent. Suppose that $X_i \,|\, \theta \sim N(\theta, \sigma^2)$ where the variance σ^2 is known, and that $\theta \sim N(\mu_0, \sigma_0^2)$ where the mean μ_0 and variance σ_0^2 are known. We wish to produce a point estimate d for θ, with loss function**

$$L(\theta, d) = 1 - \exp\left\{-\frac{1}{2}(\theta - d)^2\right\}. \tag{13}$$

   (a) **Let $f(\theta)$ denote the probability density function of $\theta \sim N(\mu_0, \sigma_0^2)$. Show that $\rho(f, d)$, the risk of d under $f(\theta)$, can be expressed as**

$$\rho(f, d) = 1 - \frac{1}{\sqrt{1 + \sigma_0^2}} \exp\left\{-\frac{1}{2(1 + \sigma_0^2)}(d - \mu_0)^2\right\}.$$

   We calculate the risk of decision d under $f(\theta)$,

$$\begin{aligned}
\rho(f, d) &= \mathbb{E}\left[1 - \exp\left\{-\frac{1}{2}(\theta - d)^2\right\} \,\middle|\, \theta \sim f(\theta)\right] \\
&= 1 - \mathbb{E}\left[\exp\left\{-\frac{1}{2}(\theta - d)^2\right\} \,\middle|\, \theta \sim f(\theta)\right] \\
&= 1 - \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(\theta - d)^2\right\} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right\} d\theta \\
&= 1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2}\left((\theta - d)^2 + \frac{1}{\sigma_0^2}(\theta - \mu_0)^2\right)\right\} d\theta. \qquad (14)
\end{aligned}$$

   Now, using the result that

$$(\theta - a)^2 + b(\theta - c)^2 = (1 + b)\left(\theta - \frac{a + bc}{1 + b}\right)^2 + \left(\frac{b}{1 + b}\right)(a - c)^2$$

   for any $a, b, c \in \mathbb{R}$ with $b \neq -1$ we have that

$$\begin{aligned}
(\theta - d)^2 + \frac{1}{\sigma_0^2}(\theta - \mu_0)^2 &= \left(\frac{1 + \sigma_0^2}{\sigma_0^2}\right)\left(\theta - \frac{\sigma_0^2 d + \mu_0}{1 + \sigma_0^2}\right)^2 + \frac{1}{1 + \sigma_0^2}(d - \mu_0)^2 \\
&= \left(\frac{1 + \sigma_0^2}{\sigma_0^2}\right)(\theta - \tilde{\mu})^2 + \frac{1}{1 + \sigma_0^2}(d - \mu_0)^2 \qquad (15)
\end{aligned}$$

where $\tilde{\mu} = \dfrac{\sigma_0^2 d + \mu_0}{1+\sigma_0^2}$. Substituting equation (15) into (14) gives

$$\rho(f,d) \;=\; 1 - \exp\left\{\frac{-1}{2(1+\sigma_0^2)}(d-\mu_0)^2\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1+\sigma_0^2}{2\sigma_0^2}(\theta-\tilde{\mu})^2\right\} d\theta \quad (16)$$

We recognise the integrand as a kernel of a $N(\tilde{\mu}, \sigma_0^2/(1+\sigma_0^2))$ distribution. Thus, as

$$\int_{-\infty}^{\infty} \frac{\sqrt{1+\sigma_0^2}}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1+\sigma_0^2}{2\sigma_0^2}(\theta-\tilde{\mu})^2\right\} d\theta \;=\; 1,$$

equation (16) becomes

$$\rho(f,d) \;=\; 1 - \frac{1}{\sqrt{1+\sigma_0^2}} \exp\left\{-\frac{1}{2(1+\sigma_0^2)}(d-\mu_0)^2\right\}$$

as required.

(b) **Using part (a), show that the Bayes rule of an immediate decision is $d^* = \mu_0$ and find the corresponding Bayes risk.**

$\rho(f,d)$ is minimised when $\dfrac{1}{\sqrt{1+\sigma_0^2}} \exp\left\{-\dfrac{1}{2(1+\sigma_0^2)}(d-\mu_0)^2\right\}$ is maximised. This is when $d^* = \mu_0$. The corresponding Bayes risk is

$$\rho^*(f) \;=\; \rho(f,d^*) \;=\; 1 - \frac{1}{\sqrt{1+\sigma_0^2}}.$$

(c) **Find the Bayes rule and Bayes risk after observing $x = (x_1, \ldots, x_n)$. Express the Bayes rule as a weighted average of d^* and the maximum likelihood estimate of θ, $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, and interpret the weights.**

As $X_i \,|\, \theta \sim N(\theta, \sigma^2)$ then

$$\begin{aligned} f(x\,|\,\theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(\theta^2 - 2x_i\theta)\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(n\theta^2 - 2n\overline{x}\theta)\right\} \end{aligned}$$

where the proportionality is with respect to θ. Hence, as $\theta \sim N(\mu_0, \sigma_0^2)$,

$$\begin{aligned} f(\theta\,|\,x) &\propto f(x\,|\,\theta)f(\theta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(n\theta^2 - 2n\overline{x}\theta)\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta^2 - 2\mu_0\theta)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\left[\theta^2 - 2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right)\theta\right]\right\}, \end{aligned}$$

which we recognise as the kernel of a $N(\mu_n, \sigma_n^2)$ where

$$\mu_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right), \qquad \sigma_n^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$$

so that $\theta \,|\, x \sim N(\mu_n, \sigma_n^2)$. Thus, we have conjugacy. The solution of $[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$ will be identical to that of $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$ but with revised hyperparameters $\mu_0 \mapsto \mu_n$ and $\sigma_0^2 \mapsto \sigma_n^2$.

The Bayes rule after observing x is thus

$$d^*(x) = \mu_n = \lambda\mu_0 + (1-\lambda)\overline{x}$$

where $\lambda = \dfrac{(1/\sigma_0^2)}{(1/\sigma_0^2) + (n/\sigma^2)}$. Thus, $d^*(x)$ is a weighted average of $d^* = \mu_0$ and $\overline{x}$ weighted according to their respective precisions. The corresponding Bayes risk is

$$\rho^*(f(\theta \,|\, x)) = 1 - \frac{1}{\sqrt{1+\sigma_n^2}}.$$

(d) **Suppose now, given data y, the parameter θ has the general posterior distribution $f(\theta \,|\, y)$. We wish to use the loss function $L(\theta, d)$, as given in equation (13), to find a point estimate d for θ. By considering an approximation of $L(\theta, d)$, or otherwise, what can you say about the corresponding Bayes rule?**

To first-order, $e^z = 1 + z$ so that

$$\begin{aligned} L(\theta, d) &\approx 1 - \left[1 - \frac{1}{2}(\theta - d)^2\right] \\ &= \frac{1}{2}(\theta - d)^2 \\ &\propto (\theta - d)^2. \end{aligned}$$

Thus, $L(\theta, d)$ is approximately proportional to quadratic loss and so the Bayes rule may be equivalently found by considering the loss function to be quadratic loss. For the decision problem $[\Theta, \mathcal{D}, \pi(\theta), (\theta-d)^2]$ the Bayes rule is $\mathbb{E}(\theta \,|\, \theta \sim \pi(\theta))$ so for $\pi(\theta) = f(\theta \,|\, y)$ the corresponding Bayes rule is $\mathbb{E}(\theta \,|\, Y)$ which is thus the approximate Bayes rule for the loss function given in equation (13).

Confidence sets and p-values

5. **Show that if p is a family of significance procedures then**

$$p(x; \Theta_0) = \sup_{\theta \in \Theta_0} p(x; \theta)$$

is a significance procedure for the null hypothesis $\Theta_0 \subset \Theta$, that is that $p(X; \Theta_0)$ is super-uniform for every $\theta \in \Theta_0$.

Notice that, for all $\theta \in \Theta_0$,

$$p(x; \Theta_0) \leq u \implies p(x; \theta) \leq u.$$

Thus, by the containment rule, for all $\theta \in \Theta_0$,

$$\begin{aligned} \mathbb{P}(p(X; \Theta_0) \leq u \,|\, \theta) &\leq \mathbb{P}(p(X; \theta) \leq u \,|\, \theta) & (17) \\ &\leq u & (18) \end{aligned}$$

where equation (18) follows from (17) as p is a family of significance procedures. Hence, $p(X; \Theta_0)$ is super-uniform for every $\theta \in \Theta_0$.

6. **Suppose that, given θ, $X_1, \ldots, X_n$ are independent and identically distributed $N(\theta, 1)$ random variables so that, given θ, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\theta, 1/n)$.**

   (a) **Consider the test of the hypotheses**

   $$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta = 1$$

   using the statistic $\overline{X}$ so that large observed values $\overline{x}$ support H_1. For a given n, the corresponding p-value is

   $$p_n(\overline{x}; 0) \quad = \quad \mathbb{P}(\overline{X} \geq \overline{x} \,|\, \theta = 0).$$

   We wish to investigate how, for a fixed p-value, the likelihood ratio for H_0 versus H_1,

   $$LR(H_0, H_1) \quad := \quad \frac{f(\overline{x} \,|\, \theta = 0)}{f(\overline{x} \,|\, \theta = 1)}$$

   changes as n increases.

   (i) **Use R to create a plot of $LR(H_0, H_1)$ for each $n \in \{1, \ldots, 20\}$ where, for each n, $\overline{x}$ is the value which corresponds to a p-value of 0.05.**

   For $p = 0.05$, for each n, we want to find $\overline{x}$ such that $\mathbb{P}(\overline{X} \geq \overline{x} \,|\, \theta = 0) = 0.05$, that is $\overline{x}$ is the 95th quantile of $N(0, 1/n)$. The following R code can be used to create Figure 1; a log-scale has been used to present the plot slightly more attractively though this is not necessary.

```
alpha <- 0.05
nseq <- 1:20
logBF <- sapply(nseq, function(n){
sd <- 1 / sqrt(n)
z <- qnorm(1 - alpha, mean = 0, sd = sd)
dnorm(z, mean = 0, sd = sd, log = TRUE) -
dnorm(z, mean = 1, sd = sd, log = TRUE)
})
plot(nseq, exp(logBF), type = "b", pch = 16, log = "xy",
ylim = c(0.2, 15),
xlab = "Number of observations, n",
ylab = expression(paste("Likelihood ratio for ", H[0],
" versus ", H[1])), xpd = NA)
abline(h = 1, lty = 2)
```

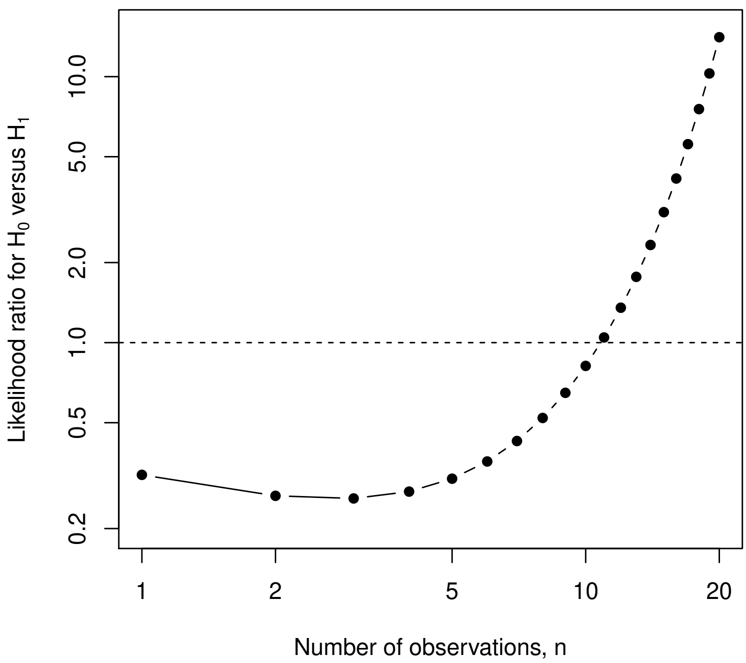

Figure 1: The likelihood ratio for the hypothesis test $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ where $\overline{X} \sim N(\theta, 1/n)$ and the p-value is fixed at 0.05.

(ii) **Comment on your plot, in particular on what happens to the likelihood ratio as n increases. What is the implication for hypothesis testing and the corresponding (fixed) p-value?**

Figure 1 shows that for small n a small p-value for H_0 such as 0.05 corresponds to a likelihood ratio for H_0 versus H_1 of less than one, and so 'rejecting H_0 in favour of H_1 ' is supported by the evidence from the observations. But as n increases a p-value of 0.05 for H_0 comes to correspond to a likelihood ratio that strongly favours H_0 over H_1. By the time $n = 20$ the likelihood ratio already exceeds 10.

We conclude that a fixed threshold for a p-value is a very poor way of distinguishing between hypotheses. The moral of this story is that where there is an explicit H_1 it should be used in a Neyman-Pearson test based on the likelihood ratio and with careful consideration of both size and power. In medical science the 'minimal clinically important difference' is the smallest gap between H_0 and H_1 that is interesting. It is used to do design calculations for sample size, but it can also be used to do hypothesis testing, rather than just p-valuing H_0.

(b) **Consider the test of the hypotheses**

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0$$

using once again $\overline{X}$ as the test statistic.

(i) **Suppose that $\overline{x} > 0$. Show that**

$$lr(H_0, H_1) \;:=\; \min_{\theta>0} \frac{f(\overline{x}\,|\,\theta = 0)}{f(\overline{x}\,|\,\theta)} \;=\; \exp\left\{-\frac{n}{2}\overline{x}^2\right\}.$$

Since $\overline{X} \sim N(\theta, 1/n)$ then

$$\begin{aligned} lr(H_0, H_1) &= \min_{\theta>0} \exp\left\{-\frac{n}{2}\left[(\overline{x}-0)^2 - (\overline{x}-\theta)^2\right]\right\} \\ &= \exp\left\{-\frac{n}{2}\overline{x}^2\right\} \end{aligned}$$

if $\overline{x} > 0$ (and is equal to one otherwise).

(ii) **Use R to create a plot of $lr(H_0, H_0)$ for a range of p-values for H_0 from 0.001 to 0.1.**[2] **Comment on whether the conventional choice of 0.05 is a suitable threshold for choosing between hypotheses, or whether some other choice might be better.**[3]

The aim of this question is for fixed n to investigate how the likelihood ratio changes with the p-value. For each p-value α, $\overline{x}$ is the $100(1-\alpha)$th quantile of $N(0, 1/n)$. The following R code, taking $n = 1$, can be used to create Figure 2.

```
pseq <- c(0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1)
z <- qnorm(1 - pseq, mean = 0, sd = 1)
ell <- pmin(1, exp(-(1/2)*z^2))
plot(pseq, ell, type = "b", pch = 16, log = "xy",
xlab = "P-value", ylab = "Lower bound on likelihood ratio")
```

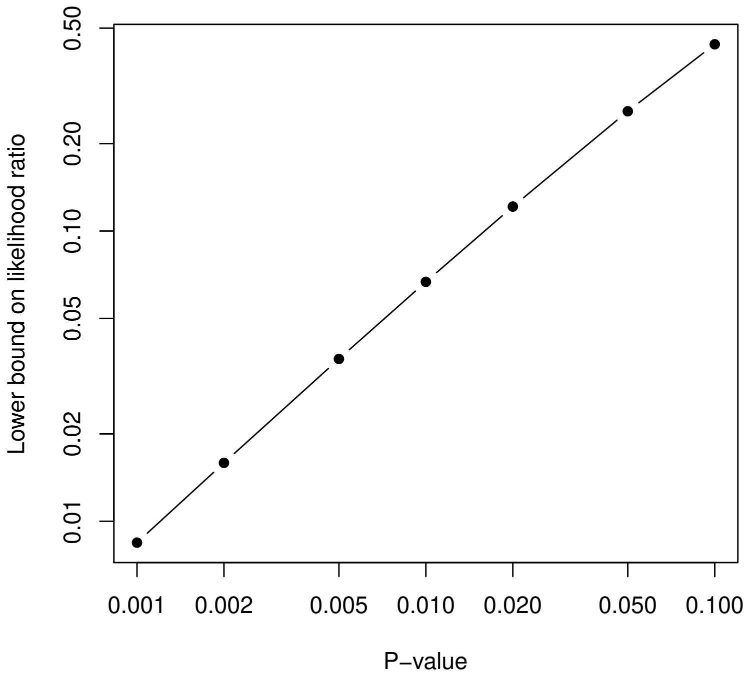

Figure 2: Lower bound on the likelihood ratio as a function of the p-value for H_0 for the hypothesis test $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ where $\overline{X} \sim N(\theta, 1/n)$.

In this case, a p-value of 0.05 corresponds to a lower bound on the likelihood ratio of 0.26. If we agree that a likelihood ratio of 1/20 is starting to get interesting, then a p-value of a bit less than 0.01 is suggested for this model and these hypotheses.

[2] The plot doesn't depend upon the actual choice of n and so you may choose $n = 1$.

[3] For the origins of the use of 0.05 see Cowles, M. and C. Davis (1982). On the origins of the .05 level of statistical significance. *American Psychologist 37(5)*, 553-558.