A Very Brief Summary of Statistical Inference, and Examples

Hilary Term 2007 Prof. Gesine Reinert

Data $\mathbf{x} = x_1, x_2, \ldots, x_n$, realisations of random variables $X_1, X_2, \ldots, X_n$

with distribution (model) $f(x_1, x_2, \ldots, x_n; \theta)$

Frequentist: $\theta \in \Theta$ unknown constant

1. Likelihood and Sufficiency

(Fisherian) Likelihood approach:

$$L(\theta) = L(\theta, \mathbf{x}) = f(x_1, x_2, \ldots, x_n; \theta)$$

Often: $X_1, X_2, \ldots, X_n$ independent, identically distributed (*i.i.d.*);

then $L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta)$

Summarize information about θ: find a *minimal sufficient statistic* $t(\mathbf{x})$; from the Factorization Theorem: $T = t(\mathbf{X})$ is sufficient for θ if and only if there exists functions $g(t, \theta)$ and $h(\mathbf{x})$ such that for all $\mathbf{x}$ and θ

$$f(\mathbf{x}, \theta) = g(t(\mathbf{x}), \theta) h(\mathbf{x}).$$

Moveover $T = t(\mathbf{X})$ is minimal sufficient when it holds that

$$\frac{f(\mathbf{x}, \theta)}{f(\mathbf{y}, \theta)} \text{ is constant in } \theta \iff t(\mathbf{x}) = t(\mathbf{y})$$

Example

Let $X_1, \ldots, X_n$ be a random sample from the truncated exponential distribution, where

$$f_{X_i}(x_i) = e^{\theta - x_i},\ x_i > \theta$$

or, using the indicator function notation,

$$f_{X_i}(x_i) = e^{\theta - x_i} I_{(\theta,\infty)}(x_i).$$

Show that $Y_1 = \min(X_i)$ is sufficient for θ.

Let $T = T(X_1, \ldots, X_n) = Y_1$. We need the pdf $f_T(t)$ of the smallest order statistic. Calculate the cdf for X_i,

$$F(x) = \int_\theta^x e^{\theta - z} dz = e^\theta [e^{-\theta} - e^{-x}] = 1 - e^{\theta - x}.$$

Now

$$P(T > t) = \prod_{i=1}^{n}(1 - F(t)) = (1 - F(t))^n.$$

Differentiating gives that $f_T(t)$ equals

$$n[1 - F(t)]^{n-1} f(t) = ne^{(\theta-t)(n-1)} \times e^{\theta-t} = ne^{n(\theta-t)}, t > \theta.$$

So the conditional density of $X_1, \ldots X_n$ given $T = t$ is

$$\frac{e^{\theta-x_1}e^{\theta-x_2}\ldots e^{\theta-x_n}}{ne^{n(\theta-t)}} = \frac{e^{-\sum x_i}}{ne^{-nt}},\ x_i \geq t,\ i = 1, \ldots n,$$

which does not depend on θ for each fixed $t = \min(x_i)$. Note that since $x_i \geq t, i = 1, \ldots, n$, neither the expression nor the range space depends on θ, so the first order statistic, $X_{(1)}$, is a sufficient statistic for θ.

Alternatively, use Factorization Theorem:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} e^{\theta - x_i} I_{(\theta,\infty)}(x_i) \\ &= I_{(\theta,\infty)}(x_{(1)}) \prod_{i=1}^{n} e^{\theta - x_i} \\ &= e^{n\theta} I_{(\theta,\infty)}(x_{(1)}) e^{-\sum_{i=1}^{n} x_i} \end{aligned}$$

with

$$g(t,\theta) = e^{n\theta} I_{(\theta,\infty)}(t)$$

and

$$h(\mathbf{x}) = e^{-\sum_{i=1}^{n} x_i}.$$

2. Point Estimation

Estimate θ by a function $t(x_1, \ldots, x_n)$ of the data; often by maximum-likelihood:

$$\hat{\theta} = arg \max_{\theta} L(\theta),$$

or by method of moments

Neither are unbiased in general, but the m.l.e. is asymptotically unbiased and asymptotically efficient; under some regularity assumptions,

$$\hat{\theta} \approx \mathcal{N}(\theta, I_n^{-1}(\theta)),$$

where

$$I_n(\theta) = \mathbf{E}\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right]$$

is the Fisher information (matrix)

Under more regularity,

$$I_n(\theta) = -\mathbf{E}\left(\frac{\partial^2 \ell(\theta, \mathbf{X})}{\partial \theta^2}\right)$$

If $\mathbf{x}$ is a random sample, then $I_n(\theta) = nI_1(\theta)$; often abbreviate $I_1(\theta) = I(\theta)$

The m.l.e. is a function of a sufficient statistic; recall that, for scalar θ, there is a nice theory using the Cramer-Rao lower bound and the Rao-Blackwell theorem on how to obtain minimum variance unbiased estimators based on a sufficient statistic and an unbiased estimator

Nice *invariance property:* The m.l.e. of a function $\phi(\theta)$ is $\phi(\hat{\theta})$

Example

Suppose $X_1, X_2, \ldots, X_n$ random sample from Gamma distribution

with density

$$f(x; c, \beta) = \frac{x^{c-1}}{\Gamma(c)\beta^c} e^{-\frac{x}{\beta}}, \quad x > 0,$$

where $c > 0, \beta > 0$; $\theta = (c, \beta)$;

$$\begin{aligned} \ell(\theta) &= -n \log \Gamma(c) - nc \log \beta \\ &\quad + (c-1) \log \prod x_i - \frac{1}{\beta} \sum x_i \end{aligned}$$

Put $D_1(c) = \frac{\partial}{\partial c} \log \Gamma(c)$, then

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= -\frac{nc}{\beta} + \frac{1}{\beta^2} \sum x_i \\ \frac{\partial \ell}{\partial c} &= -nD_1(c) - n \log \beta + \log(\prod x_i) \end{aligned}$$

Setting equal to zero yields

$$\hat{\beta} = \frac{\overline{x}}{\hat{c}},$$

where $\hat{c}$ solves

$$D_1(\hat{c}) - \log(\hat{c}) = \log([\prod x_i]^{1/n}/\overline{x})$$

(Could calculate that sufficient statistics is indeed $(\sum x_i, [\prod x_i)$, and is minimal sufficient)

Calculate Fisher information: Put $D_2(c) = \frac{\partial^2}{\partial c^2} \log \Gamma(c)$,

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \beta^2} &= \frac{nc}{\beta^2} - \frac{2}{\beta^3} \sum x_i \\
\frac{\partial^2 \ell}{\partial \beta \partial c} &= -\frac{n}{\beta} \\
\frac{\partial^2 \ell}{\partial c^2} &= -nD_2(c)
\end{aligned}$$

Use that $EX_i = c\beta$ to obtain

$$I(\theta) = n \begin{pmatrix} \frac{c}{\beta^2} & \frac{1}{\beta} \\ \\ \frac{1}{\beta} & D_2(c) \end{pmatrix}$$

Recall also that there is an iterative method to compute m.l.e.s, related to the Newton-Raphson method.

3. Hypothesis Testing

For simple null hypothesis and simple alternative, the Neyman-Pearson Lemma says that the most powerful tests are likelihood-ratio tests. These can sometimes be generalized to one-sided alternatives in such a way as to yield uniformly most powerful tests.

Example

Suppose as above that $X_1, \ldots, X_n$ is a random sample from the truncated exponential distribution, where

$$f_{X_i}(x_i; \theta) = e^{\theta - x_i},\ x_i > \theta.$$

Find a UMP test of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$:

Let $\theta_1 > \theta_0$, then the LR is

$$\frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_0)} \ = \ e^{n(\theta_1 - \theta_0)} I_{(\theta_1, \infty)}(x_{(1)}),$$

which increases with $x_{(1)}$, so we reject H_0 if the smallest observation, $T = X_{(1)}$, is large. Under H_0, we have calculated that the pdf of f_T is

$$ne^{n(\theta_0 - y_1)}, y_1 > \theta_0.$$

So for $t > \theta$

$$P(T > t) = \int_t^\infty n e^{n(\theta_0 - s)} ds = e^{n(\theta_0 - t)}.$$

Choose t such that

$$t = \theta_0 - \frac{\ln \alpha}{n}.$$

The LR test rejects H_0 if $X_{(1)} > \theta_0 - \frac{\ln \alpha}{n}$. The test is the same for all $\theta > \theta_0$, so is UMP.

For general null hypothesis and general alternative:

$H_0 : \theta \in \Theta_0$

$H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$

The (generalized) LR test uses the likelihood ratio statistic

$$T = \frac{\max_{\theta \in \Theta} L(\theta; \mathbf{X})}{\max_{\theta \in \Theta_0} L(\theta; \mathbf{X})}$$

rejects H_0 for large values of T; use chisquare asymptotics $2 \log T \approx \chi^2_p$, where $p = dim\Theta - dim\Theta_0$ (nested models only); or use score tests, which are based on the *score function* $\partial \ell / \partial \theta$, with asymptotics in terms of normal distribution $\mathcal{N}(0, I(\theta))$

An important example is Pearson's Chisquare test, which we derived as a score test, and we saw that it is asymptotically equivalent to the generalized likelihood ratio test

Example

Let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with parameter p;

$$P(X_i = k) = p(1-p)^k, \quad k = 0, 1, 2, \ldots.$$

Then

$$L(p) = p^n (1-p)^{\sum(x_i)}$$

and

$$\ell(p) = n \ln p + n\overline{x} \ln(1-p),$$

so that

$$\partial \ell / \partial p(p) = n \left(\frac{1}{p} - \frac{\overline{x}}{1-p} \right)$$

and

$$\partial^2 \ell / \partial p^2(p) = n \left(-\frac{1}{p^2} - \frac{\overline{x}}{(1-p)^2} \right).$$

We know that $\mathbf{E}X_1 = (1-p)/p$. Calculate the information

$$I(p) = \frac{n}{p^2(1-p)}.$$

Suppose $n = 20$, $\overline{x} = 3$, $H_0 : p_0 = 0.15$ and $H_1 : p_0 \neq 0.15$. Then

$\partial\ell/\partial p(0.15) = -62.7$ and $I(0.15) = 1045.8$. Test statistic

$$Z = -62.7/\sqrt{1045.8} = 1.9388.$$

Compare to 1.96 for a test at level $\alpha = 0.05$: do not reject H_0.

Example continued.

For a generalized LRT the test statistic is based on $2\log[L(\hat{\theta})/L(\theta_0)] \approx \chi^2_1$, and the test statistic is thus

$$2(\ell(\hat{\theta}) - \ell(\theta_0) = 2n(\ln(\hat{\theta}) - \ln(\theta_0) + \overline{x}(\ln(1-\hat{\theta}) - \ln(1-\theta_0)).$$

We calculate that the m.l.e. is

$$\hat{\theta} = \frac{1}{1+\overline{x}}.$$

Suppose again $n = 20$, $\overline{x} = 3$, $H_0 : \theta_0 = 0.15$ and $H_1 : \theta_0 \neq 0.15$.

The m.l.e. is $\hat{\theta} = 0.25$, and $\ell(0.25) = -44.98$; $\ell(0.15) = -47.69$.

Calculate

$$\chi^2 = 2(47.69 - 44.98) = 5.4$$

and compare to chisquare distribution with 1 degree of freedom: 3.84 at 5 percent level, so reject H_0.

4. Confidence Regions

If we can find a *pivot*, a function $t(\mathbf{X}, \theta)$ of a sufficient statistics whose distribution does not depend on θ, then we can find confidence regions in a straightforward manner.

Otherwise we may have to resort to approximate confidence regions, for example using the approximate normality of the m.l.e.

Recall that confidence intervals are equivalent to hypothesis tests with simple null hypothesis and one- or two-sided alternatives

Not to forget about:

Profile likelihood

Often $\theta = (\psi, \lambda)$ where ψ contains the parameters of interest; then base inference on profile likelihood for ψ,

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi).$$

Again we can use (generalized) LRT or score test; if ψ scalar, H_0 : $\psi = \psi_0$, $H_1^+ : \psi > \psi_0$, $H_1^- : \psi < \psi_0$, use test statistic

$$T = \frac{\partial \ell(\psi_0, \hat{\lambda}_0; \mathbf{X})}{\partial \psi},$$

where $\hat{\lambda}_0$ is the MLE for λ when H_0 true

Large positive values of T indicate H_1^+

Large negative values indicate H_1^-

$$T \approx \ell'_\psi - I^{-1}_{\lambda,\lambda} I_{\psi,\lambda} \ell'_\lambda \approx N(0, 1/I^{\psi,\psi}),$$

where $I^{\psi,\psi} = (I_{\psi,\psi} - I^2_{\psi,\lambda} I^{-1}_{\lambda,\lambda})^{-1}$ is the top left element of I^{-1}

Estimate by substituting the null hypothesis values; calculate the practical standardized form of T as

$$Z = \frac{T}{\sqrt{\mathrm{Var}(T)}} \approx \ell'_{\psi}(\psi, \hat{\lambda}_{\psi})[I^{\psi,\psi}(\psi, \hat{\lambda}_{\psi})]^{1/2}$$

$$\approx N(0,1)$$

Bias and variance approximations: the delta method

If we cannot calculate mean and variance directly: Suppose $T = g(S)$ where $ES = \beta$ and $\operatorname{Var} S = V$. Taylor

$$T = g(S) \approx g(\beta) + (S - \beta)g'(\beta).$$

Taking the mean and variance of the r.h.s.:

$$ET \approx g(\beta), \qquad \operatorname{Var} T \approx [g'(\beta)]^2 V.$$

Also works for vectors S, β, with T still a scalar

$\big(g'(\beta)\big)_i = \partial g / \partial \beta_i$ and $g''(\beta)$ matrix of second derivatives,

$$\operatorname{Var} T \approx [g'(\beta)]^T V g'(\beta)$$

and

$$ET \approx g(\beta) + \frac{1}{2} \operatorname{trace}[g''(\beta)V].$$

Exponential family

For distributions in the exponential family, many calculations have been standardized, see Notes.

A Very Brief Summary of Bayesian Inference, and Examples

Data $\mathbf{x} = x_1, x_2, \ldots, x_n$, realisations of random variables $X_1, X_2, \ldots, X_n$

with distribution (model) $f(x_1, x_2, \ldots, x_n|\theta)$

Bayesian: $\theta \in \Theta$ random; prior $\pi(\theta)$

1. Priors, Posteriors, Likelihood, and Sufficiency

Posterior distribution of θ given x is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Shorter: posterior $\propto$ prior $\times$ likelihood

The *(prior) predictive distribution* of the data x on the basis π is

$$p(x) = \int f(x|\theta)\pi(\theta)d\theta$$

Suppose data x_1 is available, want to predict additional data:

$$p(x_2|x_1) \;=\; \int f(x_2|\theta)\pi(\theta|x_1)d\theta$$

Note that x_2 and x_1 are assumed conditionally independent given θ

They are **not**, in general, unconditionally independent.

Example Suppose $y_1, y_2, \ldots, y_n$ are independent normally distributed random variables, each with variance 1 and with means $\beta x_1, \ldots, \beta x_n$, where β is an unknown real-valued parameter and $x_1, x_2, \ldots, x_n$ are known constants. Suppose prior $\pi(\beta) = \mathcal{N}(\mu, \alpha^2)$. Then the likelihood is

$$L(\beta) = (2\pi)^{-\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{(y_i - \beta x_i)^2}{2}}$$

and

$$\begin{aligned} \pi(\beta) &= \frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{(\beta-\mu)^2}{2\alpha^2}} \\ &\propto \exp\left\{ -\frac{\beta^2}{2\alpha^2} + \beta\frac{\mu}{\alpha^2} \right\} \end{aligned}$$

and the posterior of β given $y_1, y_2, \ldots, y_n$ can be calculated as

$$\begin{aligned}\pi(\beta|\mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\sum(y_i - \beta x_i)^2 - \frac{1}{2\alpha^2}(\beta - \mu)^2\right\} \\ &\propto \exp\left\{\beta\sum x_i y_i - \frac{1}{2}\beta^2\sum x_i^2 - \frac{\beta^2}{2\alpha^2} + \frac{\beta\mu}{\alpha^2}\right\}.\end{aligned}$$

Abbreviate $s_{xx} = \sum x_i^2$, $s_{xy} = \sum x_i y_i$, then

$$\begin{aligned}\pi(\beta|\mathbf{y}) &\propto \exp\left\{\beta s_{xy} - \frac{1}{2}\beta^2 s_{xx} - \frac{\beta^2}{2\alpha^2} + \frac{\beta\mu}{\alpha^2}\right\} \\ &= \exp\left\{-\frac{\beta^2}{2}\left(s_{xx} + \frac{1}{\alpha^2}\right) + \beta\left(s_{xy} + \frac{\mu}{\alpha^2}\right)\right\}\end{aligned}$$

which we recognize as

$$\mathcal{N}\left(\frac{s_{xy} + \frac{\mu}{\alpha^2}}{s_{xx} + \frac{1}{\alpha^2}}, \left(s_{xx} + \frac{1}{\alpha^2}\right)^{-1}\right).$$

Summarize information about θ: find a *minimal sufficient statistic* $t(\mathbf{x})$; $T = t(\mathbf{X})$ is sufficient for θ if and only if for all $\mathbf{x}$ and θ

$$\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x}))$$

As in frequentist approach; posterior (inference) based on sufficient statistic

Choice of prior

There are many ways of choosing a prior distribution; using a coherent belief system, e.g.. Often it is convenient to restrict the class of priors to a particular family of distributions. When the posterior is in the same family of models as the prior, i.e. when one has a conjugate prior, then updating the distribution under new data is particularly convenient.

For the regular k-parameter exponential family,

$$f(x|\theta) = f(x)g(\theta)exp\{\sum_{i=1}^{k} c_i\phi_i(\theta)h_i(x)\}, \quad x \in \mathcal{X},$$

conjugate priors can be derived in a straightforward manner, using the sufficient statistics.

Non-informative priors favour no particular values of the parameter over others. If Θ is finite, choose uniform prior; if Θ is infinite, there are several ways (some may be improper)

For a location density $f(x|\theta) = f(x-\theta)$, then the non-informative location-invariant prior is $\pi(\theta) = \pi(0)$ constant; usually choose $\pi(\theta) = 1$ for all θ (improper prior)

For a scale density $f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$ for $\sigma > 0$, the scale-invariant non-informative prior is $\pi(\sigma) \propto \frac{1}{\sigma}$; usually choose $\pi(\sigma) = \frac{1}{\sigma}$ (improper)

Jeffreys prior $\pi(\theta) \propto I(\theta)^{\frac{1}{2}}$ if the information $I(\theta)$ exists; they is invariant under reparametrization; may or may not be improper

Example continued. Suppose $y_1, y_2, \ldots, y_n$ are independent, $y_i \sim \mathcal{N}(\beta x_i, 1)$, where β is an unknown parameter and $x_1, x_2, \ldots, x_n$ are known. Then

$$\begin{aligned} I(\beta) &= -E\left(\frac{\partial^2}{\partial \beta^2} \log L(\beta, \mathbf{y})\right) \\ &= -E\left(\frac{\partial}{\partial \beta} \sum (y_i - \beta x_i) x_i\right) \\ &= s_{xx} \end{aligned}$$

and so the Jeffreys prior is $\propto \sqrt{s_{xx}}$; constant and improper. With this Jeffreys prior, the calculation of the posterior distribution is equivalent to putting $\alpha^2 = \infty$ in the previous calculation, yielding

$$\pi(\beta|\mathbf{y}) \text{ is } \mathcal{N}\left(\frac{s_{xy}}{s_{xx}}, (s_{xx})^{-1}\right).$$

If Θ is discrete, constraints

$$E_\pi g_k(\theta) = \mu_k, \quad k = 1, \ldots, m$$

choose the distribution with the maximum entropy under these constraints:

$$\tilde{\pi}(\theta_i) = \frac{exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}{\sum_i exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}$$

where the λ_i are determined by the constraints. There is a similar formula for the continuous case, maximizing the entropy relative to a particular reference distribution π_0 under constraints. For π_0 one would choose the "natural" invariant noninformative prior.

For inference, check influence of the choice of prior, for example by $\epsilon-$contamination class of priors.

2. Point Estimation

Under suitable regularity conditions, random sampling, when n is large, then the posterior is approximately $\mathcal{N}(\hat{\theta}, (nI_1(\hat{\theta}))^{-1})$, where $\hat{\theta}$ is the m.l.e.; provided that the prior is non-zero in a region surrounding $\hat{\theta}$. In particular, if θ_0 is the true parameter, then the posterior will become more and more concentrated around θ_0.

Often reporting the posterior distribution is preferred to point estimation.

Bayes estimators are constructed to minimize the posterior expected loss. Recall from *Decision Theory:* Θ is the set of all possible states of nature (values of parameter), $\mathcal{D}$ is the set of all possible decisions (actions); a loss function is any function

$$L : \Theta \times \mathcal{D} \to [0, \infty)$$

For point estimation: $\mathcal{D} = \Theta$, $L(\theta, d)$ loss in reporting d when θ is true.

Bayesian: For a prior π and data $x \in \mathcal{X}$, the posterior expected loss of a decision is

$$\rho(\pi, d|x) = \int_{\Theta} L(\theta, d)\pi(\theta|x)d\theta$$

For a prior π the integrated risk of a decision rule δ is

$$r(\pi, \delta) \ = \ \int_{\Theta}\int_{\mathcal{X}} L(\theta, \delta(x))f(x|\theta)dx\pi(\theta)d\theta$$

An estimator minimizing $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ that minimizes $\rho(\pi, \delta|x)$. A Bayes estimator associated with prior π, loss L, is any estimator δ^{π} which minimizes $r(\pi, \delta)$. Then $r(\pi) = r(\pi, \delta^{\pi})$ is the Bayes risk.

Valid for proper priors, and for improper priors if $r(\pi) < \infty$. If $r(\pi) = \infty$, define a generalized Bayes estimator as the minimizer, for every x, of $\rho(\pi, d|x)$

For strictly convex loss functions, Bayes estimators are unique. For squared error loss $L(\theta, d) = (\theta - d)^2$, the Bayes estimator δ^π associated with prior π is the posterior mean. For absolute error loss $L(\theta, d) = |\theta - d|$, the posterior median is a Bayes estimator.

Example. Suppose that $x_1, \ldots, x_n$ is a random sample from a Poisson distribution with unknown mean θ. Two models for the prior distribution of θ are contemplated;

$$\pi_1(\theta) = e^{-\theta}, \quad \theta > 0, \text{ and } \pi_2(\theta) = e^{-\theta}\theta, \quad \theta > 0.$$

Then calculate the Bayes estimator of θ under model 1, with quadratic loss, using

$$\pi_1(\theta|\mathbf{x}) \propto e^{-n\theta}\theta^{\sum x_i}e^{-\theta} = e^{-(n+1)\theta}\theta^{\sum x_i},$$

which we recognize as $Gamma(\sum x_i+1, n+1)$. The Bayes estimator is the expected value of the posterior distribution,

$$\frac{\sum x_i + 1}{n+1}.$$

For model 2,

$$\pi_2(\theta|\mathbf{x}) \propto e^{-n\theta}\theta^{\sum x_i}e^{-\theta}\theta = e^{-(n+1)\theta}\theta^{\sum x_i+1},$$

which we recognize as $Gamma(\sum x_i+2, n+1)$. The Bayes estimator is the expected value of the posterior distribution,

$$\frac{\sum x_i + 2}{n+1}.$$

Note that the first model has greater weight for smaller values of θ, so the posterior distribution is shifted to the left.

Recall that we have seen more in Decision Theory: least favourable distributions.

3. Hypothesis Testing

$H_0 : \theta \in \Theta_0$,

$$\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\} = \{1, 0\},$$

where 1 stands for acceptance; loss function

$$L(\theta, \phi) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \phi = 1 \\ a_0 & \text{if } \theta \in \Theta_0, \phi = 0 \\ 0 & \text{if } \theta \notin \Theta_0, \phi = 0 \\ a_1 & \text{if } \theta \notin \Theta_0, \phi = 1 \end{cases}$$

Under this loss function, the Bayes decision rule associated with a prior distribution π is

$$\phi^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0 | x) > \frac{a_1}{a_0 + a_1} \\ 0 & \text{otherwise} \end{cases}$$

The *Bayes factor* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is

$$\begin{aligned} B^\pi(x) &= \frac{P^\pi(\theta \in \Theta_0|x)/P^\pi(\theta \in \Theta_1|x)}{P^\pi(\theta \in \Theta_0)/P^\pi(\theta \in \Theta_1)} \\ &= \frac{p(x|\theta \in \Theta_0)}{p(x|\theta \in \Theta_1)} \end{aligned}$$

is the ratio of how likely the data is under H_0 and how likely the data is under H_1; measures the extent to which the data x will change the odds of Θ_0 relative to Θ_1

Note that we have to make sure that our prior distribution puts mass on H_0 (and on H_1). If H_0 is simple, this is usually achieved by choosing a prior that has some point mass on H_0 and otherwise lives on H_1.

Example revisited. Suppose that $x_1, \ldots, x_n$ is a random sample from a Poisson distribution with unknown mean θ; possible models for the prior distribution are

$$\pi_1(\theta) = e^{-\theta}, \quad \theta > 0, \text{ and } \pi_2(\theta) = e^{-\theta}\theta, \quad \theta > 0.$$

The prior probabilities of model 1 and model 2 are assessed at probability 1/2 each. Then the Bayes factor is

$$\begin{aligned} B(\mathbf{x}) &= \frac{\int_0^\infty e^{-n\theta}\theta^{\sum x_i}e^{-\theta}d\theta}{\int_0^\infty e^{-n\theta}\theta^{\sum x_i}e^{-\theta}\theta d\theta} \\ &= \frac{\Gamma(\sum x_i + 1)/((n+1)^{\sum x_i+1})}{\Gamma(\sum x_i + 2)/((n+1)^{\sum x_i+2})} \\ &= \frac{n+1}{\sum x_i + 1}. \end{aligned}$$

Note that in this setting

$$\begin{aligned} B(\mathbf{x}) &= \frac{P(\text{model 1}|\mathbf{x})}{P(\text{model 2}|\mathbf{x})} \\ &= \frac{P(\text{model 1}|\mathbf{x})}{1 - P(\text{model 1}|\mathbf{x})}, \end{aligned}$$

so that

$$P(\text{model 1}|\mathbf{x}) = (1 + B(\mathbf{x}))^{-1}.$$

Hence

$$\begin{aligned} P(\text{model 1}|\mathbf{x}) &= \left(1 + \frac{\sum x_i + 1}{n+1}\right)^{-1} \\ &= \left(1 + \frac{\overline{x} + \frac{1}{n}}{1 + \frac{1}{n}}\right)^{-1}, \end{aligned}$$

which is decreasing for $\overline{x}$ increasing.

For robustness: Least favourable Bayesian answers; suppose H_0 : $\theta = \theta_0$, $H_1 : \theta \neq \theta_0$, prior probability on H_0 is $\rho_0 = 1/2$. For G family of priors on H_1; put

$$\underline{B}(x, G) \;=\; \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta) d\theta}.$$

A Bayesian prior $g \in G$ on H_0 will then have posterior probability at least $\underline{P}(x, G) = \left(1 + \frac{1}{\underline{B}(x,G)}\right)^{-1}$ on H_0 (for $\rho_0 = 1/2$). If $\hat{\theta}$ is the m.l.e. of θ, and G_A the set of all prior distributions, then

$$\underline{B}(x, G_A) \;=\; \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}$$

and

$$\underline{P}(x, G_A) \;=\; \left(1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right)^{-1}.$$

4. Credible intervals

A $(1-\alpha)$ *(posterior) credible interval (region)* is an interval (region) of $\theta-$values within which $1-\alpha$ of the posterior probability lies

Often would like to find HPD (highest posterior density) region: a $(1-\alpha)$ credible region that has minimal volume

when the posterior density is unimodal, this is often straightforward

Example continued. Suppose $y_1, y_2, \ldots, y_n$ are independent normally distributed random variables, each with variance 1 and with means $\beta x_1, \ldots, \beta x_n$, where β is an unknown real-valued parameter and $x_1, x_2, \ldots, x_n$ are known constants; Jeffreys prior, posterior $\mathcal{N}\left(\frac{s_{xy}}{s_{xx}}, (s_{xx})^{-1}\right)$, then a 95%-credible interval for β is

$$\frac{s_{xy}}{s_{xx}} \pm 1.96\sqrt{\frac{1}{s_{xx}}}.$$

Interpretation in Bayesian: conditional on the observed $\mathbf{x}$; randomness relates to the distribution of θ

In contrast, frequentist confidence interval applies before $\mathbf{x}$ is observed; randomness relates to the distribution of $\mathbf{x}$

5. Not to forget about: *Nuisance parameters*

$\theta = (\psi, \lambda)$, where λ nuisance parameter

$\pi(\theta|x) = \pi((\psi, \lambda)|x)$

marginal posterior of ψ:

$$\pi(\psi|x) = \int \pi(\psi, \lambda|x) d\lambda$$

Just integrate out the nuisance parameter