Answers to Exercises

2. (a) If $\mathsf{P}(A) = 0$ or $\mathsf{P}(A) = 1$.

(b) Take A as "odd red die", B as "odd blue die" and C as "odd sum".

(c) Take C as $\{HHH, THH, THT, TTH\}$.

3. (a) $\mathsf{P}(\text{homozygous}) = 1/3$; $\mathsf{P}(\text{heterozygous}) = 2/3$.

(b) By Bayes' Theorem

$$\mathsf{P}(BB \,|\, 7 \text{ black}) = \frac{(1/3)(1)}{(1/3)(1) + (2/3)(1/2^7)} = \frac{64}{65}.$$

(c) Similarly we see by induction (wih case $n = 0$ given by part (a)) that $\mathsf{P}(BB \,|\, \text{first } n \text{ black}) = 2^{n-1}/(2^{n-1}+1)$ since

$$\begin{aligned}\mathsf{P}(BB \,|\, \text{first } n+1 \text{ black}) &= \frac{\{2^{n-1}/(2^{n-1}+1)\}(1)}{\{2^{n-1}/(2^{n-1}+1)\}(1) + \{1/(2^{n-1}+1)\}(1/2)} \\ &= \frac{2^n}{2^n+1}.\end{aligned}$$

6. $\mathsf{P}(k=0) = (1-\pi)^n = (1-\lambda/n)^n \to \mathrm{e}^{-\lambda}$. More generally

$$\begin{aligned} p(k) &= \binom{n}{k} \pi^k (1-\pi)^{n-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\to \frac{\lambda^k}{k!} \exp(-\lambda). \end{aligned}$$

12. $\mathsf{E}X = n(1-\pi)/\pi$; $\mathcal{V}X = n(1-\pi)\pi^2$.

proof

When a joint probability density function is well defined and the expectations are integrable, we write for the general case

$$
\begin{aligned}
\mathrm{E}(X) &= \int x \Pr[X=x]\;dx \\
\mathrm{E}(X \mid Y=y) &= \int x \Pr[X=x \mid Y=y]\;dx \\
\mathrm{E}(\mathrm{E}(X \mid Y)) &= \int \left(\int x \Pr[X=x \mid Y=y]\;dx\right) \Pr[Y=y]\;dy \\
&= \int\int x \Pr[X=x, Y=y]\;dx\;dy \\
&= \int x \left(\int \Pr[X=x, Y=y]\;dy\right)\;dx \\
&= \int x \Pr[X=x]\;dx \\
&= \mathrm{E}(X)\,.
\end{aligned}
$$

2. $\overline{x} = 16.35525$, so assuming uniform prior, posterior is $\mathrm{N}(16.35525,\ 1/12)$. A 90% HDR is $16.35525 \pm 1.6449/\sqrt{12}$ or 16.35525 ± 0.47484, that is, the interval $(15.88041, 16.83009)$.

3. $x - \theta \sim \mathrm{N}(0, 1)$ and $\theta \sim \mathrm{N}(16.35525,\ 1/12)$, so $x \sim \mathrm{N}(16.35525,\ 13/12)$.

Continuous distribution, continuous parameter space

For the normal distribution $\mathcal{N}(\mu, \sigma^2)$ which has probability density function

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the corresponding probability density function for a sample of n independent identically distributed normal random variables (the likelihood) is

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

This family of distributions has two parameters: $\theta = (\mu, \sigma)$; so we maximize the likelihood, $\mathcal{L}(\mu, \sigma^2) = f(x_1, \ldots, x_n \mid \mu, \sigma^2)$, over both parameters simultaneously, or if possible, individually.

Since the logarithm function itself is a continuous strictly increasing function over the range of the likelihood, the values which maximize the likelihood will also maximize its logarithm (the log-likelihood itself is not necessarily strictly increasing). The log-likelihood can be written as follows:

$$\log\Big(\mathcal{L}(\mu, \sigma^2)\Big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\, x_i - \mu\,)^2$$

(Note: the log-likelihood is closely related to information entropy and Fisher information.)

We now compute the derivatives of this log-likelihood as follows.

$$0 = \frac{\partial}{\partial\mu}\log\Big(\mathcal{L}(\mu, \sigma^2)\Big) = 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2}.$$

where $\bar{x}$ is the sample mean. This is solved by

$$\widehat{\mu} = \bar{x} = \sum_{i=1}^{n}\frac{x_i}{n}.$$

This is indeed the maximum of the function, since it is the only turning point in μ and the second derivative is strictly less than zero. Its expected value is equal to the parameter μ of the given distribution,

$$\mathbb{E}\big[\,\widehat{\mu}\,\big] = \mu,$$

which means that the maximum likelihood estimator $\widehat{\mu}$ is unbiased.

Similarly we differentiate the log-likelihood with respect to σ and equate to zero:

$$0 = \frac{\partial}{\partial\sigma}\log\Big(\mathcal{L}(\mu, \sigma^2)\Big) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(\, x_i - \mu\,)^2.$$

which is solved by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Inserting the estimate $\mu = \widehat{\mu}$ we obtain

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i x_j.$$

To calculate its expected value, it is convenient to rewrite the expression in terms of zero-mean random variables (statistical error) $\delta_i \equiv \mu - x_i$. Expressing the estimate in these variables yields

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu - \delta_i)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mu - \delta_i)(\mu - \delta_j).$$

Simplifying the expression above, utilizing the facts that $\mathbb{E}\left[\ \delta_i\ \right] = 0$ and $\operatorname{E}\left[\ \delta_i^2\ \right] = \sigma^2$, allows us to obtain

$$\mathbb{E}\left[\ \widehat{\sigma}^2\ \right] = \frac{n-1}{n}\sigma^2.$$

This means that the estimator $\widehat{\sigma}^2$ is biased for σ^2. It can also be shown that $\widehat{\sigma}$ is biased for σ, but that both $\widehat{\sigma}^2$ and $\widehat{\sigma}$ are consistent.

Formally we say that the *maximum likelihood estimator* for $\theta = (\mu, \sigma^2)$ is

$$\widehat{\theta} = \left(\widehat{\mu}, \widehat{\sigma}^2\right).$$

Example

Suppose that we are given a sequence $(x_1, \dots, x_n)$ of IID $N(\mu, \sigma_v^2)$ random variables and a prior distribution of μ is given by $N(\mu_0, \sigma_m^2)$. We wish to find the MAP estimate of μ. Note that the normal distribution is its own conjugate prior, so we will be able to find a closed-form solution analytically.

The function to be maximized is then given by[3]

$$g(\mu)f(x \mid \mu) = \pi(\mu)L(\mu) = \frac{1}{\sqrt{2\pi}\sigma_m}\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_m}\right)^2\right)\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma_v}\right)^2\right),$$

which is equivalent to minimizing the following function of μ:

$$\sum_{j=1}^{n}\left(\frac{x_j-\mu}{\sigma_v}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_m}\right)^2.$$

Thus, we see that the **MAP estimator** for μ is given by[3]

$$\hat{\mu}_{\text{MAP}} = \frac{\sigma_m^2\, n}{\sigma_m^2\, n + \sigma_v^2}\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) + \frac{\sigma_v^2}{\sigma_m^2\, n + \sigma_v^2}\,\mu_0 = \frac{\sigma_m^2\left(\sum_{j=1}^{n} x_j\right) + \sigma_v^2\,\mu_0}{\sigma_m^2\, n + \sigma_v^2}.$$

which turns out to be a linear interpolation between the prior mean and the sample mean weighted by their respective covariances.

The case of $\sigma_m \to \infty$ is called a non-informative prior and leads to an improper probability distribution; in this case $\hat{\mu}_{\text{MAP}} \to \hat{\mu}_{\text{MLE}}$.

3. Take $\alpha/(\alpha+\beta) = 1/3$ so $\beta = 2\alpha$ and

$$\mathcal{V}\pi = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{2}{3^2(3\alpha+1)}$$

so $\alpha = 55/27 \cong 2$ and $\beta = 4$. Posterior is then $\mathrm{Be}(2+8, 4+12)$, that is $\mathrm{Be}(10, 16)$. The 95% values for $F_{32,20}$ are 0.45 and 2.30 by interpolation, so those for $F_{20,32}$ are 0.43 and 2.22. An appropriate interval for π is from $10 \times 0.43/(16 + 10 \times 0.43$ to $10 \times 2.22/(16 + 10 \times 2.22)$, that is $(0.21, 0.58)$.

Solution

Denoting $x' = x_{n+1}$ for short, the posterior predictive is

$$
\begin{aligned}
p(x'|x_{1:n}) &= \int p(x'|\theta)p(\theta|x_{1:n})d\theta \\
&= \int_0^\infty \theta e^{-\theta x'} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{(\alpha+1)-1} e^{-(\beta+x')\theta} d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(\beta+x')^{\alpha+1}} \int \mathrm{Gamma}(\theta \mid \alpha+1,\, \beta+x') d\theta \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(\beta+x')^{\alpha+1}}.
\end{aligned}
$$

The marginal likelihood is

$$
\begin{aligned}
p(x_{1:n}) &= \int p(x_{1:n}|\theta)p(\theta)d\theta \\
&= \int_0^\infty \theta^n e^{-\theta \sum x_i} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} d\theta \\
&= \frac{b^a}{\Gamma(a)} \int_0^\infty \theta^{a+n-1} \exp\big(-(b+\textstyle\sum x_i)\theta\big) d\theta \\
&= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n)}{(b+\sum x_i)^{a+n}} \int \mathrm{Gamma}\,\big(\theta \mid a+n,\, b+\textstyle\sum x_i\big) d\theta \\
&= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n)}{(b+\sum x_i)^{a+n}} = \frac{b^a}{\Gamma(a)} \frac{\Gamma(\alpha)}{\beta^\alpha}.
\end{aligned}
$$

The marginal likelihood can also be found by using Bayes' theorem: for any θ,

$$
p(x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{p(\theta|x_{1:n})} = \frac{\frac{b^a}{\Gamma(a)}\theta^{\alpha-1}e^{-\beta\theta}}{\mathrm{Gamma}(\theta|\alpha,\beta)} = \frac{\frac{b^a}{\Gamma(a)}}{\frac{\beta^\alpha}{\Gamma(\alpha)}}.
$$