20. Probability distributions for parameters LM 5.8

20.1 When can a parameter have a distribution?

(i) Sometimes a parameter is itself the outcome of a random process:
for example, the probability of heads varies across a population of coins,
or, the frequency of a genetic variant (allele) varies across the different genetic systems (loci) in our DNA.

(ii) In such cases, assigning a pmf/pdf to the parameter (P(heads) or allele frequency) makes sense. This
probability distribution, assigned from the process giving rise to the parameter, is known as the prior distribution.

(iii) Some believe that a prior distribution can always be assigned; and that this prior distribution expresses
beliefs about values of the parameter in the absence of data.

20.2 Example: from Stat340 final exam

(i) The setup: In a certain population, everyone is equally susceptible to colds. The number of colds suffered
by each person during each winter season can be modeled as the outcome of a Poisson random variable with
mean 5. A new cold prevention drug is introduced, which, for people for whom the new drug is effective
reduces the number of colds to the outcome of a Poisson random variable with mean 3. Unfortunately, the
drug is only effective in 20% of people.

(ii) For people taking the drug: \( \pi(5) = P(\theta = 5) = 0.8, \pi(3) = P(\theta = 3) = 0.2. \)

(iii) Fred decides to take the drug. Given that he gets 4 colds that winter, what is the conditional probability
that the drug is effective for Fred?

That is, we want \( \pi(\theta = 3 \mid X = 4) \) where \( X \) is the data, the outcome of the Poisson random variable.

(iv) Using Bayes’ Theorem:

\[
\pi(\theta = 3 \mid X = 4) = \frac{P(X = 4 \mid \theta = 3)\pi(\theta = 3)}{P(X = 4 \mid \theta = 3)\pi(\theta = 3) + P(X = 4 \mid \theta = 5)\pi(\theta = 5)} = \frac{0.2 \times 3^4 \times e^{-3}/(0.2 \times 3^4 \times e^{-3} + 0.8 \times 5^4 \times e^{-4})}{0.80655/(0.80655 + 3.36897)} = 0.19316.
\]

20.3 Using Bayes’ Theorem to get the posterior distribution

(i) Let the prior distribution for parameter \( \theta \) be \( \pi(\theta) \). Let the probability (pmf/pdf) of the data observations
\( x_1, \ldots, x_n \) be \( f(x_1, \ldots, x_n \mid \theta) \). Note this is just the likelihood, but with a slight change of notation, as \( \theta \) is now
a random variable. Then

\[
\pi(\theta \mid X_1 = x_1, \ldots, X_n = x_n) = f(x_1, \ldots, x_n \mid \theta) \pi(\theta) / \int_\theta f(x_1, \ldots, x_n \mid \theta) \pi(\theta) \, d\theta
\]

This is the posterior distribution for \( \theta \) given data \( x_1, \ldots, x_n \).

(ii) Suppose \( T \) is a sufficient statistic for \( \theta \). Then the likelihood factorizes as

\[
f(x_1, \ldots, x_n \mid \theta) = f(x_1, \ldots, x_n \mid T = t).f_T(t \mid \theta)
\]

where the first term does not depend on \( \theta \), so

\[
\pi(\theta \mid T = t) = \pi(\theta \mid X_1 = x_1, \ldots, X_n = x_n) = f_T(t \mid \theta) \pi(\theta) / \int_\theta f_T(t \mid \theta) \pi(\theta) \, d\theta
\]

This is the posterior distribution for \( \theta \) given the data \( x_1, \ldots, x_n \) (or given the value \( t \) of \( T \)).
21. Conjugate prior distributions LM Examples 5.8.3, 5.8.4

21.1 Normal data and Normal prior

Certain priors “match” a given data model, to give a posterior for \( \theta \) that is the same family as the prior. This can save a lot of work.

For example: \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \) with \( \sigma^2 \) known, so sufficient \( T = \overline{X_n} \sim N(\mu, \sigma^2/n) \).

Suppose the prior for \( \mu \) is \( N(0, \tau^2) \). Then

\[
\pi(\mu|\overline{X_n}) \propto \exp\left(-\frac{\mu^2}{2\tau^2} - \frac{n(\overline{X_n} - \mu)^2}{2\sigma^2}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\mu^2 - \frac{2\overline{X_n}\mu n}{\sigma^2}\right)
\]

so posterior for \( \mu \) is \( N(\overline{x_n}nK/\sigma^2, \varphi) \) where \( K = \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1} \).
Note if \( n \) is large, or \( \tau^2 \) is large, this is approx \( N(\overline{x_n}, \sigma^2/n) \).

21.2 Binomial data and Beta Prior LM Example 5.8.2

Suppose \( X_1, \ldots, X_n \) are Binomial(1, \( \theta \)), so sufficient \( T = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta) \).

Suppose prior for \( \theta \) is \( \Gamma(r + s)|1 - \theta|^{s-1}/\Gamma(r)\Gamma(s) \) on \( 0 \leq \theta \leq 1 \). (Beta(r,s) density).

Then \( \pi(\theta|t) \propto \theta^{r-1}(1 - \theta)^{s-1} \) so we know that the posterior must be \( \text{Beta}(t + r, n - t + s) \).

21.3 Poisson data and Gamma Prior LM Example 5.8.3

Suppose \( X_1, \ldots, X_n \) are Poisson(\( \theta \)), so sufficient \( T = \sum_{i=1}^n X_i \sim \text{Poisson}(n\theta) \).

Suppose prior for \( \theta \) is the Gamma pdf \( G(s, \beta) \). Then

\[
\pi(\theta|t) \propto \theta^{s-1}\exp(-\beta\theta), \exp(-n\theta)(n\theta)^t \propto \theta^{t+s-1}\exp(-(\beta + n)\theta)
\]

so the posterior for \( \theta \) is \( G(t + s, \beta + n) \).

21.4 The marginal distribution of the data random variables LM Example 5.8.4

Also of interest sometimes is the marginal probability of the data – integrating over the parameter:

For example: the overall probability Fred has \( k \) colds: \( P(k \text{ colds}) = (0.2 e^{-3}3^k + 0.8 e^{-5}5^k)/k! \).

In 21.1, for example

\[
f_{\overline{X_n}}(x) = \int_{\mu} f_{\overline{X_n}}(x|\mu)\pi(\mu) \, d\mu \propto \int_{\mu} \exp\left(-n(x - \mu)^2/(2\sigma^2) - \mu^2(2\tau^2)\right) \, d\mu
\]

\[
\propto \int_{\mu} \exp\left(-1/2K\mu - nxK/\sigma^2\right)^2 + (n^2x^2K/\sigma^4) - (nx^2)/(2\sigma^2) \, d\mu
\]

\[
\propto \exp\left(-Knx^2/(n\sigma^2) - (1/K)\right)/(2\sigma^2) \propto \exp\left(-Knx^2/(2\sigma^2\tau^2)\right)
\]

Thus, we find \( \overline{X_n} \) is Normal with mean 0 and variance \( \sigma^2\tau^2/(nK) = \left(\sigma^2/n\right) + \tau^2 \).

21.5 Continuing 21.3 LM Example 5.8.4

Suppose \( n = 1 \), i.e. we have a single Poisson observation, \( T \sim \text{Poisson}(\theta) \):

\[
P(T = t) = \int_0^\infty P(T = t | \theta) \pi(\theta) \, d\theta = \int_0^\infty \beta^t \theta^{t+s-1} \exp(-1 + \beta\theta) \, d\theta / t!\Gamma(s)
\]

\[
= \left(\frac{\beta}{t!\Gamma(s)}\right) \int_0^\infty \theta^{t+s-1} \exp(-1 + \beta\theta) \, d\theta = \left(\frac{\beta}{t!\Gamma(s)}\right)(\Gamma(t + s)/(\beta + 1)^{(t+s)})
\]

\[
= (\Gamma(t + s)/t!\Gamma(s))\left(\frac{\beta}{\beta + 1}\right)^s(\frac{1}{\beta + 1})^t = \left(\frac{t + s - 1}{t}\right)(\frac{\beta}{\beta + 1})^s(\frac{1}{\beta + 1})^t
\]

which is negative binomial!!
22: Bayesian estimation using loss functions LM Theorem 5.8.1


(i) Defn: Let \( w \) be an estimate for \( \theta \), based on the data \( x_1, \ldots, x_n \), or on the value of the sufficient statistics \( T = t \). Then the loss function, \( L(w, \theta) \) measures the cost of estimating by \( w \) when \( \theta \) is true value.

(ii) \( L(\theta, \theta) = 0 \), and \( L(w, \theta) \geq 0 \).

(iii) Defn; the Bayes risk is the posterior expected loss, where expectations is over the distribution of \( \theta \) given \( T = t \); \( R(w) = \int_\theta L(w, \theta) \pi(\theta | T = t) \ d\theta \).

(iv) Note: \( \theta \) is the random thing in this expression. The data are fixed.

22.2: Point estimation with squared error loss

(i) \( L(w, \theta) = (w-\theta)^2 \). Note the difference: m.s.e = \( E((W-\theta)^2) \) where \( W \) is random.

\[ \text{posterior expected loss} = R(w) = E(w-\theta)^2 \] where \( \theta \) is random.

(ii) Want to minimise posterior expected loss or \( R(w) = \int_\theta (w-\theta)^2 \pi(\theta | T = t) \ d\theta \).

Differentiating w.r.t. \( w \): \( R'(w) = \int_\theta 2(w-\theta)\pi(\theta | T = t) \ d\theta = 0 \)

or \( \int_\theta w \pi(\theta | T = t) \ d\theta = \int_\theta \pi(\theta | T = t) \ d\theta \)

or \( w = w \int_\theta \pi(\theta | T = t) \ d\theta = E(\theta | T = t) \)

The estimate is the mean of the posterior distribution.

(iii) Example: \( X_1, \ldots, X_n \) i.i.d. \( N(\mu, \sigma^2) \), where \( \sigma^2 \) is known. Then \( \overline{X}_n \) is sufficient for \( \mu \).

If prior distribution for \( \mu \) is \( N(0, \tau^2) \) then \( \pi(\theta | \overline{X}_n = \overline{x}_n) = N(\overline{x}_n K/\sigma^2, K) \) where \( K = (1/\tau^2 + \overline{X}_n^2)^{-1} \).

So Bayes estimate of \( \theta \) for squared error loss is \( \overline{x}_n K/\sigma^2 \).

22.3: Point estimation with absolute error loss

(i) \( L(w, \theta) = |w-\theta| \). Want to minimise posterior expected loss or \( R(w) = \int_\theta |w-\theta| \pi(\theta | T = t) \ d\theta \):

\[ R(w) = \int_{-\infty}^w (w-\theta) \pi(\theta | T = t) \ d\theta + \int_w^\infty (\theta-w) \pi(\theta | T = t) \ d\theta \]

\[ R'(w) = 0 + \int_w^\infty \pi(\theta | T = t) \ d\theta + 0 - \int_w^\infty \pi(\theta | T = t) \ d\theta = 0 \]

or \( P(\theta \leq w | T = t) = P(\theta \geq w | T = t) \)

The estimate is the median of the posterior distribution.

(ii) In the Normal example above: the posterior median is the same as the posterior mean, as the Normal distribution is symmetric.

22.4: Posterior interval estimation

(i) We can also make interval probability statements based on the posterior distribution: these are probabilities about the random \( \theta \). Contrast this with confidence intervals, where the probability statement is about the random \( T \).

(ii) Example: the Normal example again. \( \pi(\theta | \overline{X}_n = \overline{x}_n) = N(\overline{x}_n K/\sigma^2, K) \) where \( K = (1/\tau^2 + \overline{X}_n^2)^{-1} \).

\[ P(\overline{x}_n K/\sigma^2 - \sqrt{K} z_{\alpha/2} \leq \theta \leq \overline{x}_n K/\sigma^2 + \sqrt{K} z_{\alpha/2}) = 1 - \alpha \]

Or, \( (\overline{x}_n K/\sigma^2 - \sqrt{K} z_{\alpha/2}, \overline{x}_n K/\sigma^2 + \sqrt{K} z_{\alpha/2}) \) is a Bayesian posterior probability interval for \( \theta \).

(iii) If \( \tau^2 \) is very large: \( K \approx \sigma^2/n \), and the interval becomes \( (\overline{x}_n - \sigma z_{\alpha/2}/\sqrt{n}, \overline{x}_n + \sigma z_{\alpha/2}/\sqrt{n}) \).

This looks like our confidence interval for \( \mu \), but recall again the interpretation is quite different.