Classical and Bayesian inference

AMS 132

The Prior Distribution

Definition

Suppose that one has a statistical model with parameter θ. If one treats θ as random, then the distribution that one assigns to θ before observing the other random variables of interest is called its *prior distribution*.
If the parameter space is at most countable, then the prior distribution is discrete and its probability function (p.f) is called *prior p.f.* of θ. If the prior distribution is a continuous distribution, then its probability density function (p.d.f.) is called *prior p.d.f.* of θ. The prior p.f. or p.d.f. will be denoted by $\xi(\theta)$.

- When θ is random, then "prior distribution" is another name of marginal distribution of the parameter.
- The prior distribution of a parameter θ must be a probability distribution over the parameter space Ω.
- We assume that, before the experimental data have been collected or observed, the experimenters past experience and knowledge will lead him to believe that θ is more likely to lie in certain regions of Ω than in others.

The Prior Distribution

Example (Fair or Two-Headed Coin)

Consider the experiment of tossing a coin, and let θ denote the probability of obtaining a head in each trial.

a) Describe the statistical model.

b) Define a prior distribution for θ (See `R`).

c) Suppose that it is known that the coin either is fair or has a head on each side. Write the prior distribution for θ when the prior probability that the coin is fair is 0.8.

The Prior Distribution

Some terminology

- In what follows we will consider statistical problems in which parameters, denoted θ, are random variables of interest. θ can denote one or more parameters.
- When θ is random, $f(x \mid \theta)$ is the conditional p.f. or p.d.f. of observation X given θ.
- When θ is random, we say that $X_1, \ldots, X_n$ are conditionally i.i.d. r.v.s given θ with conditional p.f. or p.d.f. $f(x \mid \theta)$. This is denoted by $X_i \mid \theta \overset{i.i.d.}{\sim} F(\cdot \mid \theta)$, where F is the distribution of X_i.
- $X_i \mid \theta \overset{i.i.d.}{\sim} F(\cdot \mid \theta)$ will be understood equivalently as $X_1, \ldots, X_n$ form a random sample with p.f. or p.d.f. $f(x \mid \theta)$.

Sensitivity analysis

- A *sensitivity analysis* is to compare the posterior distribution that arise from considering several different prior distributions in order to see how much effect the prior distribution has on the answers to important questions.
- If there are a lot of data, then different experimenters that do not agree on a prior distribution becomes less important.
- If the prior distributions being compared are very spread out, then it is not going to matter much which one is specified.

Improper priors

- There are some calculations available that attempt to capture the idea that the data contain much more information than is available a priori.
- These calculations result in defining a prior distribution for θ such that

$$\int_{\Omega} \xi(\theta) d\theta = \infty,$$

which violates the definition of a prior. Such priors are called *improper priors*.

The Posterior Distribution

Definition (Posterior Distribution)

Consider a statistical inference problem with parameter θ and random variables $X_1, \ldots, X_n$ to be observed. The conditional distribution of θ given $X_1, \ldots, X_n$ is called the *posterior distribution* of θ. The conditional p.f. or p.d.f. of θ given $X_1 = x_1, \ldots, X_n = x_n$ is called the *posterior p.f. or p.d.f.* of θ and is denoted $\xi(\theta \mid x_1, \ldots, x_n)$.

- When parameters are random, the name “posterior distribution” is another name for conditional distribution of the parameter given the data.
- Bayes theorem tells us how to compute the posterior p.f. or p.d.f of θ after observing data.

The Posterior Distribution

Theorem (Posterior Distribution)

Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x \mid \theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f. or p.f. of θ is $\xi(\theta)$. Then the posterior p.d.f. or p.f. of θ is

$$\xi(\theta \mid x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f(x_i \mid \theta)\xi(\theta)}{g_n(x_1, \ldots, x_n)},$$

for $\theta \in \Omega$ where g_n is the marginal joint p.d.f. or p.f. of $X_1, \ldots, X_n$.

Proof: ...

Using vector notation, where $\boldsymbol{x} = (x_1, \ldots, x_n)$, then

$$\xi(\theta \mid \boldsymbol{x}) = \frac{\prod_{i=1}^{n} f(x_i \mid \theta)\xi(\theta)}{g_n(\boldsymbol{x})}.$$

The Posterior Distribution

Example (Lifetime of components)

Suppose that the distribution of the lifetimes of certain components is the exponential distribution with parameter θ, and the prior distribution for θ is the gamma distribution with parameters a and b. Suppose also that the lifetimes $x_1 = 0.92, x_2 = 0.72, x_3 = 1.27, x_4 = 0.69, x_5 = 1.97$ of a random sample of $n = 5$ components are observed.

a) Find the posterior distribution of θ.

b) Plot the prior and posterior distributions of θ when $a = 5$, $b = 1$, and when $a = 1$ and $b = 0.2$ (See R)

The Likelihood Function

- The posterior p.f. or p.d.f. of θ can be written as $\xi(\theta \mid \boldsymbol{x}) \propto f_n(\boldsymbol{x} \mid \theta)\xi(\theta)$, where $\propto$ is the proportionality symbol.
- The normalizing constant can be determined at any time.

Definition

When the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x} \mid \theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \ldots, x_n$, it is called the *likelihood function*.

Example (Lifetime of components (cont.))

- Plot the likelihood function (See R).
- Find the posterior distribution of θ (See R).

Sequential Observations and Prediction

- In many experiments, the observations $X_1, \ldots, X_n$, which form the random sample, must be obtained sequentially, that is, one at a time.
- If that is the case, it can be shown that $\xi(\theta \mid \boldsymbol{x}) \propto f(x_n \mid \theta)\xi(\theta \mid x_1, \ldots, x_{n-1})$.
- Alternatively, after all n values $x_1, \ldots, x_n$ have been observed, we could calculate the posterior p.d.f. in the usual way, and the posterior distribution would be the same (Check it as an exercise!).
- When the observations are sequentially sampled the proportionality constant has a useful interpretation as 1 over the conditional p.d.f. or p.f. of X_n given $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$.
- Now we can find the p.d.f. or p.f. of a new random variable, X_{n+1} given that we have observed $X_1 = x_1, \ldots, X_n = x_n$.

Example (Lifetime of components (cont.))

Let X_6 be the lifetime of a new component. Compute $P(X_6 > 1.5 \mid x_1, \ldots, x_5)$ (See R).