TAE 2014
Statistics exercises

Alejandro Vaquero Avilés-Casco

September 18, 2014

1 Exercise 1: Set of sufficient statistics

Show that

$$S = \left\{ n; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}$$

is a set of sufficient statistics with $X = N(x|\mu, \sigma)$ and $e(n) \rightarrow \mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.

1.1 Solution

The normal distribution is given by

$$N(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

where μ is the mean of the distribution and σ^2 is its variance. For our data

$$p(data|\mu, \sigma) \propto \sigma^{-n} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}, \tag{2}$$

because it is an independent set of measurements, and the probability of obtaining the whole set $\mathbf{x}$ is equal to the product of probabilities of obtaining each one of the data x_i. We expand the exponent of the distribution and try to change the dependence on the data to a dependence on the elements of the set S:

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x} + \bar{x} - \mu}{\sigma}\right)^2 =$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu)$$

The last term vanishes because $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ by definition, and the sum is weighted by a constant $(\bar{x} - \mu)$. Now we substitute by the elements of S

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{ns^2 + n(\bar{x} - \mu)^2}{\sigma^2},$$

and if we translate this result to the exponential,

$$p(data|\mu,\sigma) \propto \sigma^{-n} e^{-\frac{1}{2\sigma^2}\left(ns^2+n(\bar{x}-\mu)^2\right)}, \tag{3}$$

hence the set S is a set of sufficient statistics.

2 Exercise 2: Inferences on parameters

Being $\{\mu,\sigma\}$ location and scale parameters, take $\pi(\mu,\sigma) \propto \frac{1}{\sigma}$ as (improper) prior and show that

$$T = \sqrt{n-1}\left(\frac{\mu-\bar{x}}{s}\right) \sim St\,(t|n-1) \qquad Z = n\left(\frac{s^2}{\sigma^2}\right) \sim \chi^2\,(z|n-1)$$

$\rightarrow$ Inferences on μ	$\rightarrow$ Inferences on σ
$E[T] = 0 \quad (n>2)$	$E[Z] = n-1 \quad (n>1)$
$V[T] = (n-1)(n-3)^{-1} \quad (n>3)$	$V[Z] = 2(n-1) \quad (n>1)$

2.1 Solution T

We want to make inferences on μ, therefore, applying the bayes rule we get

$$p(\mu,\sigma|data) \propto p(data|\sigma,\mu)\pi(\mu,\sigma). \tag{4}$$

The idea is to integrate out σ, so our probabilities refer only to μ, without caring about the value of σ

$$p(\mu|data) = \int p(\mu,\sigma|data)d\sigma \propto \int p(data|\sigma,\mu)\pi(\mu,\sigma)d\sigma.$$

Our input is (3) from the former exercise, and the prior in this case is $\pi(\mu,\sigma) = 1/\sigma$, thence, we integrate the rhs of (4) in σ to get the distribution of μ

$$\int_0^\infty p(\mu,\sigma|x)d\sigma \sim \int_0^\infty d\sigma\, \sigma^{-(n+1)} e^{-\frac{1}{2\sigma^2}\left(ns^2+n(\bar{x}-\mu)^2\right)}$$

The range of integration in σ ranges from 0 to inf, for negative values doesn't make any sense. We change variables to evaluate the integral,

$$u = \frac{1}{\sigma^2}, \qquad du = -\frac{2}{\sigma^3}d\sigma,$$

$$\sigma = u^{-\frac{1}{2}}, \qquad d\sigma = -\frac{1}{2}u^{-\frac{3}{2}}du,$$

and the integral becomes

$$\int_0^\infty p(\mu,\sigma|x)d\sigma \sim -\frac{1}{2}\int_\infty^0 du\, u^{\frac{n}{2}-1} e^{-\frac{u}{2}\left(ns^2+n(\bar{x}-\mu)^2\right)}$$

The integral is of the kind

$$\int_\infty^0 x^n e^{-ax} \propto \frac{1}{a^{n+1}},$$

so the final result becomes

$$\int_0^\infty p(\mu,\sigma|x)d\sigma \propto \frac{1}{\left(ns^2+n\left(\bar{x}-\mu\right)^2\right)^{\frac{n}{2}}} \propto \frac{1}{\left[1+\left(\frac{\mu-\bar{x}}{s}\right)^2\right]^{\frac{n}{2}}} = \left[1+\frac{t^2}{n-1}\right]^{-\frac{n}{2}}$$

which is the t-Student distribution for $n-1$ degrees of freedom, exactly what we wanted to prove.

2.2 Solution Z

In the second case, we want to make an inference on σ, therefore we integrate out μ

$$p(\mu,\sigma|data) \propto p(data|\sigma,\mu)\pi(\mu,\sigma). \tag{5}$$

From (3) we already know the rhs, now we need to integrate it

$$\int_{-\infty}^\infty p(\mu,\sigma|x)d\mu \sim \frac{1}{\sigma^{n+1}} e^{-\frac{ns^2}{2\sigma^2}} \int_{-\infty}^\infty e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} \tag{6}$$

The last integral is gaussian, and we can easily evaluate it

$$\int_{-\infty}^\infty e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2} = \sqrt{\frac{2\pi}{n}}\sigma, \tag{7}$$

so the final expression becomes

$$\int_{-\infty}^\infty p(\mu,\sigma|x)d\mu \sim \frac{1}{\sigma^n} e^{-\frac{ns^2}{2\sigma^2}}. \tag{8}$$

Again, we change variables

$$z = \frac{ns^2}{\sigma^2}, \qquad \frac{1}{\sigma^n} = \left(\frac{z}{ns^2}\right)^{\frac{n}{2}}$$

$$d\sigma = \frac{z^{-\frac{3}{2}}}{ns^2}dz$$

to obtain

$$\int p(\mu,\sigma|x)d\mu \propto z^{\frac{n-1}{2}-1} e^{-\frac{z}{2}} \tag{9}$$

which is the χ^2 distribution for $n-1$ degrees of freedom. It is necessary to take into account the Jacobian of the transformation $z\left(\sigma\right)$, for at the end, what we get is the pdf $p(z|x)dz$, and in order to obtain a probability of having a measurement of z between the values z_1 and z_2, we should integrate the expression; but of course for this we need to take into account that our variable of integration is z, and not σ, hence the Jacobian.

3 Exercise 3: Binomial distribution

There is a detector which detected n events, we would like to guess the efficiency θ of the detector. The total number of events is N and obviously the rule $n \leq N$ holds. The probability distribution in this case is binomial

$$p(n|N,\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n} \tag{10}$$

and one can relate the number of measured events n with the number of total events N though the efficiency θ

$$E[n] = N\theta \tag{11}$$

Find a suitable prior $\pi(\theta)$ that would allow us to estimate $p(\theta|n,N)$.

3.1 Solution

In order to calculate the distribution of the efficiency as a function of the number of events n, we would like to guess a good prior $\pi(\theta)$ and apply the Bayes rule

$$p(\theta|n,N) \propto p(n|\theta,N)\,\pi(\theta) \tag{12}$$

so one can make inferences on θ with the data collected n. One way to guess (although not the only one) a prior is to use the Fisher matrix

$$w = -\ln p(n|\theta,N) = w_0 - n\ln\theta - (N-n)\ln(1-\theta). \tag{13}$$

Other choices are possible, but being the Fisher matrix invariant under reparametrizations, it is a good candidate for a guess of our prior

$$I(\theta) = E_X\left[\frac{\partial^2 \ln p(n|\theta,N)}{\partial\theta^2}\right] = E_X\left[\frac{n}{\theta^2} + \frac{N-n}{(1-\theta)^2}\right]$$

To evaluate $E_X[]$ we make use of (11)

$$I(\theta) = E_X\left[\frac{n}{\theta^2} + \frac{N-n}{(1-\theta)^2}\right]\Bigg|_{n=N\theta} = \frac{N}{\theta} + \frac{N}{1-\theta} \sim \frac{1}{\theta(1-\theta)}$$

The prior is related to the Fisher matrix as

$$\pi(\theta) \sim [I(\theta)]^{\frac{1}{2}} = \frac{1}{(\theta(1-\theta))^{\frac{1}{2}}} \tag{14}$$

And we can calculate the distribution of the efficiency according to out measured data as (12).

4 Exercise 4

We have a sample

$$e(n) \to \mathbf{x} = \{x_1, x_2, \ldots, x_n\},$$

that follows an exponential distribution, with $\bar{x} = \frac{1}{n}\sum x_i$

$$p\left(x_i|\tau\right) = \frac{1}{\tau}e^{-\frac{x_i}{\tau}} \tag{15}$$

and the average of each one of the variables is given by

$$E\left[x_i\right] = \tau \tag{16}$$

Again, we must find a suitable prior $\pi\left(\tau\right)$ that would allow us to estimate $p\left(\tau|\bar{x}\right)$.

4.1 Solution

First we calculate the probability distribution function of the whole set of data. Since the events are completely uncorrelated, the pdf of the total set will be equal to the products of the pdfs of each one of the data

$$p(\bar{x}|\tau) = \frac{1}{\tau^n}e^{-\sum\frac{x_i}{\tau}} = \frac{1}{\tau^n}e^{-n\frac{\bar{x}}{\tau}} \tag{17}$$

We calculate the Fisher matrix of the pdf

$$w = -\ln p(\bar{x}|\tau) = n\frac{\bar{x}}{\tau} + n\ln\tau,$$

$$I\left(\tau\right) = E_X\left[\frac{\partial^2 \ln p(\bar{x}|\tau)}{\partial\tau^2}\right] = E_X\left[n\frac{2\bar{x}}{\tau^3} - \frac{n}{\tau^2}\right].$$

Using (11) and the definition of $\bar{x}$ we can remove the $\bar{x}$ from the Fisher matrix

$$I\left(\tau\right) = E_X\left[n\frac{2\bar{x}}{\tau^3} - \frac{n}{\tau^2}\right]\Bigg|_{\bar{x}=\tau} \propto \frac{2}{\tau^2} - \frac{1}{\tau^2} = \frac{1}{\tau^2}.$$

From here, the calculation of the prior is straightforward

$$\pi\left(\tau\right) \sim \left[I\left(\tau\right)\right]^{\frac{1}{2}} = \frac{1}{\tau},$$

and the pdf of τ is given by

$$p(\tau|\bar{x}) = p(\bar{x}|\tau)\pi\left(\tau\right) = \frac{1}{\tau^{n+1}}e^{-n\frac{\bar{x}}{\tau}} \tag{18}$$