Inference Tutorial 7

1. The following are observations of, x, the length of time in excess of 14 days that it took for a random sample of 10 grasshoppers infected with a fungal disease to die, where the times are measured from infection.

   `1.6 12.1 5.2 6.2 6.8 9.6 17.8 24.4 4.1 7.8`

   ($\sum x_i = 95.6$ and $\sum \log(x_i) = 20.154$). It is believed that a gamma distribution may be a good model here, so that the p.d.f. of x is

   $$f(x) = \frac{1}{\beta^\alpha (\alpha-1)!} x^{\alpha-1} e^{-x/\beta} \quad x > 0$$

   In this case the parameter β could be any positive real number, but α can only take positive integer values (because of the requirements of the simulation model in which the estimated parameters will eventually be used).

   Find the maximum likelihood estimates of α and β (a pencil paper and calculator exercise).

   Solution

   $$l(\alpha, \beta) = -n\alpha \log(\beta) - n \log[(\alpha-1)!] + (\alpha-1)\sum \log(x_i) - \sum x_i/\beta$$

   $$\frac{\partial l}{\partial \beta} = \frac{-n\alpha}{\beta} + \frac{\sum x_i}{\beta^2} = 0 \Rightarrow \hat{\beta} = \frac{\sum x_i}{n\alpha}$$

   Now $\sum x_i = 95.6$ and $\sum \log(x_i) = 20.154$, so ...

   $$\alpha = 1 \Rightarrow \hat{\beta} = 9.56 \Rightarrow l = -10\log(9.56) - 10 = -32.58$$

   $$\alpha = 2 \Rightarrow \hat{\beta} = 4.78 \Rightarrow l = -20\log(4.78) + \sum \log(x_i) - 20 = -31.134$$

   $$\alpha = 3 \Rightarrow \hat{\beta} = 3.1867 \Rightarrow l = -30\log(3.1867) - 10\log(2) + 2\sum \log(x_i) - 30 = -31.392$$

   etc...

   $$\Rightarrow \hat{\alpha} = 2, \quad \hat{\beta} = 4.78.$$

2. Suppose that you have n independent measurements, t_i, and believe that the probability density function for the t_is is of the form:

   $$f(t) = ke^{-t^2/\theta} \quad t \geq 0$$

   where θ and k are the same for all i.

   (a) By considering the p.d.f of the normal distribution, show that:

   $$k = \sqrt{\frac{4}{\pi\theta}}$$

   (b) Show that $E(t) = \sqrt{\theta/\pi}$ and that var$(t) = \theta/2 - \theta/\pi$ (for the latter you should not need to perform any integration). Finally, use the fact that if $X \sim N(0, \sigma^2)$ then $E(X^4) = 3\sigma^4$, to show that var$(t^2) = \theta^2/2$.

   (c) Obtain an expression for the maximum likelihood estimator for θ.

   (d) Show that your estimator is unbiased.

   (e) Does an unbiased estimator exist that has smaller variance than your estimator, at finite sample sizes? Explain your answer.

   Solution

   (a) Since the normal pdf is symmetric about μ then if $\mu = 0$:

   $$\int_0^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/(2\sigma^2)} dt = \frac{1}{2}$$

   Setting $2\sigma^2 = \theta$ it's obvious that the honesty condition:

   $$\int_0^\infty ke^{-t^2/\theta} dt = 1$$

   requires that $k = \sqrt{4/(\pi\theta)}$.

(b)

$$E(t) = \int_0^\infty \sqrt{\frac{4}{\pi\theta}} t e^{-t^2/\theta} dt = \left[-\sqrt{\frac{4}{\pi\theta}} \frac{\theta}{2} e^{-t^2/\theta} \right]_0^\infty = \sqrt{\frac{\theta}{\pi}}$$

Let $\pi(t)$ denote the pdf of $N(0, \sigma^2)$. We know that $\int_{-\infty}^{\infty} t^2 \pi(t) dt = \sigma^2$. Hence by symmetry $\int_0^\infty t^2 \pi(t) = \sigma^2/2$, and $\int_0^\infty t^2 2\pi(t) = \sigma^2$. But $2\pi(t)$ is the pdf of interest in this question, if we set $\sigma^2 = \theta/2$. So $E(t^2) = \theta/2$. Hence var$(t) = E(t^2) - E(t)^2 = \theta/2 - \theta/\pi$. By similar reasoning we have $E(t^4) = 3\theta^2/4$, so var$(t^2) = E(t^4) - E(t^2)^2 = 3\theta^2/4 - \theta^2/4 = \theta^2/2$.

(c) By independence, the joint p.d.f. is:

$$\prod_{i=1}^{n} \sqrt{\frac{4}{\pi\theta}} e^{-t_i^2/\theta}$$

so the log- likelihood is:

$$l(\lambda) = -\frac{1}{2} \sum_i \log \theta + \frac{1}{2} \sum_i \log (4/\pi) - \sum_i t_i^2/\theta$$

differentiating and setting to zero gives:

$$\frac{\partial l}{\partial \theta} = -\frac{1}{2} \sum_i \frac{1}{\theta} + \sum_i \frac{t_i^2}{\theta^2} = 0 \Rightarrow \hat{\theta} = \frac{2 \sum t_i^2}{n}$$

(and the second derivative is clearly negative.)

(d) It's unbiased:

$$E(\hat{\theta}) = \frac{2 \sum_i E(t_i^2)}{n} = \theta.$$

(e) The (exact) variance is:

$$\text{var}(\hat{\theta}) = \frac{4}{n^2} \sum_i \text{var}(t_i^2) = \frac{2\theta^2}{n}$$

To find out whether this is the minimum possible for an unbiased estimator, we need to evaluate the CR lower bound

$$-E\left(\frac{\partial^2 l}{\partial \theta^2}\right)^{-1} = \left(-\frac{1}{2} \frac{n}{\sigma^2} + \frac{2 \sum_i E(t_i^2)}{\theta^3} \right)^{-1} = \frac{2\theta^2}{n}.$$

So the variance of our estimator achieves the CR lower bound, implying that a lower variance unbiased estimator does not exist.

3. `bc <- read.table("https://people.maths.bris.ac.uk/~sw15190/TOI/bccd.dat",header=TRUE)` reads into R some data from a study on the side effects of radio-therapy on women with breast cancer. The data relate to the onset of breast retardation after the commencement of radiotherapy for early stage breast cancer. Due to the nature of the study the point of onset of retardation is not known — all that is known is that it lies within some interval, and for some patients only the lower limit of the interval is known (these patients may never have experienced retardation, or may have left the study or died). In assessing treatment regimes it is helpful to have a model for the time of onset of retardation, and a very simple first model might be that the onset time, t_i, is an observation of a random variable with p.d.f

$$f(t_i) = \lambda e^{-\lambda t_i}.$$

where λ is a rate parameter to be estimated. Notice that t_i is not observed directly, we only know the interval it occurs in, and our likelihood will need to reflect this fact. The columns of `bccd.dat` are headed `ok` for the last time at which there was known to be no deterioration, and `not` for the first time at which deterioration was observed — the actual time of onset will be between these two values.

(a) Derive an expression for the probability that t_i lies between any two values a and b where $b > a$.

(b) Hence write down an expression for the likelihood of the parameter λ.

(c) Write an R function to evaluate the log likelihood of λ (or better still $p = \log \lambda$, since λ is inherently positive). Note that R can handle numbers with value `inf`.

(d) By appropriate use of `optim` obtain a 95% CI for λ (a starting value of 0.01 is ok).

(e) From the above CI obtain a 95% CI for the mean time of onset of retardation.

Solution

(a)

$$\Pr[a < T_i < b] = \int_a^b \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_a^b = e^{-\lambda a} - e^{-\lambda b}$$

(b) Let a_i and b_i be respectively the lower and upper interval limits for subject i.

$$L(\lambda) = \prod_{i=1}^{n} \left(e^{-\lambda a_i} - e^{-\lambda b_i}\right)$$

(c)
```
lli <- function(p,a,b) {
# negative log likelihood for breast retardation interval data
# p is parameter a is array of lower limits, b array of upper limits
  lambda <- exp(p)
  -sum(log(exp(-lambda*a)-exp(-lambda*b)))
}
```

(d)
```
bc <- read.table("https://people.maths.bris.ac.uk/~sw15190/TOI/bccd.dat",header=TRUE)
p <- log(0.01)
bc.fit <- optim(p,lli,method="BFGS",hessian=TRUE,a=bc$ok,b=bc$not)
p.hat <- bc.fit$par
sig.p <- (1/bc.fit$hessian)^0.5
ci <- exp(c(p.hat - 1.96*sig.p , p.hat + 1.96*sig.p))
```
which gives a 95% CI for λ of

```
> ci
[1] 0.0106158 0.0249902
```

(e) The m.l.e. of the expected time to retardation is $1/\hat{\lambda}$ (by invariance). So a 95% CI for the mean time can be obtained by taking the reciprocals of the end points of the interval for λ.

```
> sort(1/ci)
[1] 40.01569 94.19920
```

(d-e) Alternatively, a Wilks interval could be computed (what range of parameter values would be accepted for the null hypothesis using a GLRT).

```
crit <- qchisq(.95,df=1)/2

ll <- p <- seq(-5,-2,length=1000)
for (i in 1:length(p)) { ## find log lik for various parameter values
  ll[i] <- -lli(p=p[i],a=bc$ok,b=bc$not)
}
## find the range of parameter values giving a log likelihood high
## enough to accept....
wilks.ci <- exp(range(p[ll > -bc.fit$value - crit]))
wilks.ci
sort(1/wilks.ci) ## interval for expected onset time.
```