STAT2011 Statistical Models

Computer Exercise week 10

One of the fundamental statistical problems is to estimate the centre of symmetry of a symmetric, continuous probability distribution. Model data $x_1, x_2, \ldots, x_n$ as values taken by IID (continuous) random variables $X_1, X_2, \ldots, X_n$ whose PDF is symmetric about some unknown point θ.

Two estimators whose distributions are symmetric about θ (if indeed the "true" density is symmetric) are the sample mean and the sample median. We are going to examine the performance of each under two different scenarios, i.e. two different unknown true densities. The first is the standard normal distribution with PDF

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,, \quad \text{all real } x;$$

the second is the Laplace distribution with PDF

$$f(x) = \frac{1}{2} e^{-|x|}\,, \quad \text{all real } x.$$

Both are symmetric about zero; thus one way to assess the performance of each estimator is to simulate a large number of different samples, compute and save the estimates, and then compute the average squared error of each, which is just the mean square since the "true value" in each case is zero.

As with all simulation exercises, every time you `process()` your file you will generate different numbers and get different answers. If you wish to make specific comments about numbers obtained, you might want to put `set.seed(93458762839475)` or similar at the top of your file, so you get the same "random numbers" each time.

1. Generate a single pseudo-random sample of size 250 from the standard normal distribution using `x=rnorm(250)`. Draw a probability histogram (which has total area 1 unit) with (approximately) 25 bins using `hist(x,pr=T,n=25)`. Draw the standard normal density over the histogram using `curve(dnorm(x),add=T)` (**note**: you need to put all graphics commands for each plot all in the one `\graph...\end` block). Compute `mn=mean(x)` and `md=median(x)` and print them; they should both be "near zero".

2. Define `mn1=md1=0`. Write a `for`-loop which, for `i in 1:10000` performs the following steps:

   - generates a pseudo-random sample of size 250 from the normal distribution;
   - saves the sample mean in the `i`-th element of `mn1`;
   - saves the sample median in the `i`-th element of `md1`.

3. Compute the average squared error (ASE) of each vector of estimates, viewing them as estimates of the centre of symmetry. **Describe** their relative performances on the basis of the ASEs (**note:** we use ASE instead of MSE to emphasise the difference between empirical averages and theoretical means (i.e. expectations)).

4. Produce *comparative histograms* of the two vectors of estimates. These are both drawn to the same scale, to facilitate comparison:

   - determine the range of values included in both vectors combined: `r=range(c(mn1,md1))`;
   - create a (common) vector of breaks: `breaks=seq(from=r[1],to=r[2],len=50)`;
   - prepare the graph window for a 2-by-1 display (one-under-the-other): `par(mfrow=c(2,1))`;
   - create two histograms using e.g. `hist(mn1,pr=T,br=breaks,xlim=r,ylim=c(0,7))` and similar for `md1` (the last argument should be suitable but you can change the upper value if you need to – just make sure it is the same for both histograms).

   Comment on the two histograms.

5. The Laplace distribution can be interpreted as the distribution of SY, where Y has the standard exponential distribution on $(0, \infty)$, i.e. with pdf e^{-x} on $x > 0$ *independently* of a random sign S with distribution $P(S = +1) = P(S = -1) = \frac{1}{2}$. One can exploit this to generate a pseudo-random sample from the Laplace distribution as follows:

   - generate a standard-exponential pseudo-random sample of size 250 using `y=rexp(250)`;
   - generate pseudo-random signs: `s=sample(c(-1,1),repl=T,size=250)`;
   - `x=s*y`.

   Do this, draw a probability histogram as in question 1 and draw over the Laplace density using `curve(.5*exp(-abs(x)),add=T)`.

6. Write a loop as in question 2 but instead this time generate from the Laplace instead of the normal, and save your estimates in vectors `mn2` and `md2`. Compute ASEs again and **comment.**

7. Produce comparative histograms as for the normal case above, and **comment**.

8. **Comment** on the shapes of all 4 estimate histograms.