Problem set 4: Empirical distributions and the bootstrap

S681

Upload your typed answers to Canvas as a PDF by 11:59 pm, Sunday 17th March. Include R code where applicable.

1. Recall that a Poisson distribution with parameter λ has probability mass function

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Does the number of children a woman has follow a Poisson distribution? We collect data from 1761 adult German women, and count their children:

Number of children	0	1	2	3	4	5	6	7	8	9	10
Women with this number of children	398	455	575	227	65	28	9	3	0	1	0

Perform a goodness-of-fit test to see whether this data can be well-modeled using a Poisson distribution, stating the value of your test statistic, your P-value, and your conclusion.

2. The data set `illinois-rainstorms.txt` gives the rainfall (in inches) in a sample of 227 rainstorms in Illinois.

On the same graph, plot:

(a) A 95% pointwise confidence band for the CDF of rainfall;

(b) A 95% simultaneous confidence band for the CDF of rainfall.

and clearly indicate on the graph which band is which.

3. Here are survival times (in days) for a sample of HIV patients (a "+" indicates the patient was still alive at the last time of observation):

$$22, 90, 256, 320+, 428, 670+, 910, 997, 1070, 1081, 1197, 1355+, 1560, 1933, 2202$$

Use R to obtain the Kaplan-Meier estimate of the survival function, and plot it along with pointwise confidence limits calculated using a method of your choice.

4. For the Illinois rainstorms data:

(a) Find the sample mean $\bar{x}$, and use the bootstrap to estimate the mean squared error of $\bar{x}$ as an estimate of the population mean.

(b) Find the sample standard deviation s, and use the bootstrap to estimate the mean squared error of s as an estimate of the population standard deviation.

(c) Find the sample coefficient of variation $s/\bar{x}$, and use the bootstrap to estimate the mean squared error of $s/\bar{x}$ as an estimate of the population coefficient of variation.

5. The file `rabbits.txt` contains the eosinophil counts of a bunch of rabbits.

   (a) Find a 95% bootstrap Studentized t-pivot confidence interval for the mean eosinophil count of rabbits, *without* using the `boot` package.

   (b) Find a 95% bootstrap BCA confidence interval for the mean eosinophil count of rabbits, using whatever packages you wish.

6. The file `geyser.txt` contains two variables: `waiting`, which gives the waiting times (in minutes) between eruptions of Old Faithful, and `duration`, which gives the duration (in minutes) of the eruption following that waiting time.

   (a) Plot the data and add a regression line to predict duration from waiting time. Does it look like the key assumptions required for classical regression inference (linearity and homoskedasticity) are met?

   (b) Even if classical assumptions are not met, we can still use the bootstrap to do inference. Find a 95% bootstrap confidence interval for the slope parameter of the linear regression, and carefully explain what this interval means.

7. (Computationally expensive; set aside lots of computing time.) Suppose we wish to find a 95% confidence interval for the mean from an IID sample of size 20 from a chi-square distribution with 1 degree of freedom. (You can generate such a sample using `rchisq()`.) Perform simulations to estimate the level of coverage of

   (a) The percentile bootstrap

   (b) The residual (basic) bootstrap

   (c) The BCa bootstrap

   (d) The Studentized (t-pivot) bootstrap

   Note: The number of bootstrap replications has little effect on relative accuracy, so keep B to a moderate value like 1000.

8. For the Illinois rainstorms data:

   (a) Find the maximum likelihood estimates of the shape and rate parameters of a gamma distribution fitted to the data.

   (b) Plot (on the same graph):

   - The empirical CDF of the data;
   - The CDF of the gamma distribution you estimated.

9. For the `geyser.txt` data:

   (a) Choose a bandwidth for a Gaussian kernel to estimate the PDF of waiting time, stating how how you (or R) calculated the bandwidth. Plot a Gaussian kernel density estimate of the PDF of waiting time using this bandwidth.

   (b) Your friend is allergic to normal distributions, and asks you to instead create a kernel density estimate using a uniform kernel with bandwidth 4 (meaning the kernel stretches two minutes up and down from each observation.) Plot such an estimate. Hint: Make sure the area under the curve is 1.

10. For the `geyser.txt` data:

    (a) Plot a conditional density estimate of eruption duration given the previous waiting time. Describe what your plot tells you.

    (b) Plot a conditional density estimate of waiting time given the *previous* eruption duration. (You'll have to manipulate the data to get the previous duration to line up with the following waiting time.) Describe what your plot tells you.