R PRACTICAL 2

DAVID STEINSALTZ

8 December, 2015

(1) Make a plot of the graphs of the functions $y = x^p$ for $p = \frac{1}{4}, \frac{1}{2}, 1, 2, 3$, and $x \in [0, 2]$, all on the same set of axes, with different colours and line types. For an extra challenge use the function `legend` to add a legend that explains the plot.

(2) (a) Using R compute the following:
   (i) $\mathbb{P}\{X = 112\}$ where X is binomial with $n = 200$, $p = 0.6$.
   (ii) $\mathbb{P}\{X \geq 4\}$ where X is Poisson with parameter 8.
   (iii) $\mathbb{P}\{1 < X < 2\}$ where X is Exponential with parameter 2.
   (iv) $\mathbb{P}\{X < 2\}$ where X is normal with mean 3 and variance 7.

   For all examples but the last one, do the computation two ways, using your knowledge of the probability mass function or density, and using the appropriate distribution function in R. For the last one, find z such that the probability is the same as $\mathbb{P}\{Z < z\}$, where Z is standard normal (i.e., mean 0 and variance 1).

   (b) Make 1000 simulations of each distribution in the previous part. Plot a histogram of each set of simulations, and estimate the desired probability from the simulated outcomes.

   (c) *[optional]* How accurate might we expect these simulations to be? One way to estimate accuracy is to do multiple repetitions, and see how much the answers vary. Try this. (This is a version of the method called **bootstrap**.)

(3) Load the package `MASS` with the command `require(MASS)`. The data frame `hills` gives record times in 1984 for 35 Scottish hill races.
   (a) Look at the help file for this data set.
   (b) Try applying commands like `head`, `summary`, `dim`, `attributes`.
   (c) What happens when you plot `hills`?
   (d) Make a histogram of the `time` variable.
   (e) Compute the mean and SD of the times for those races where the climb was above the median, and those where it was below the median.

(4) ***E. coli* descriptive statistics**. In the following sequence of questions will be using data from the completely sequenced genome of *E. coli*, more precisely strain K-12, substrain MG 1655, version M52 short. *Adapted from a lab written by Terry Speed and Bin Yu.*

(a) **Background and raw data.**
The sequence data have come from the NCBI website: `www.ncbi.nlm.nih.gov`. You may download the data by typing
`load(url('http://steinsaltz.me.uk/DTC/ecoli.rda'))`.
What was originally a single string of length 4641652 has been processed by the command `strsplit(x,'')` to create a vector `ecp` of length 4641652, each entry consisting of one letter of A, C, G, T.

(b) Count the number of Adenines, Guanines, Cytosines and Thymines (As, Gs, Cs and Ts). Then there is the dinucleotide composition, that is, the number of AAs, AGs etc. (along one strand). There further is the trinucleotide composition, that is, the number of AAAs, AAGs etc.
   (i) If you had to assign a probability to observing an A at a stated position on the *E. coli* genome, what figure would you use and why? Under what assumptions does this seem appropriate?
   (ii) Type `L=length(ecp)` followed by `table(ecp[-L],ecp[-1])`. What is this telling you?
   (iii) Your are told there is an A at a position along the *E. coli* genome. What probability would you assign to it being followed by a G, and why?
   (iv) Does the *E. coli* composition data suggest that the event we observe a G at one site is independent (in some suitable sense) of the previous two bases? Explain fully, illustrating with appropriate data.

(c) **Purine counts.**
Divide up the data into about 46,400 blocks of 100 base pairs (bp). Count the number of purines (i.e. A or G). Do the same for the 4,640 blocks of 1,000 bp, and 464 blocks of 10,000 bp.
   (i) For each set of counts, calculate the mean and standard deviation of the number of purines per block, and draw histograms of these numbers.
   (ii) Compare the results of (a) across the different block sizes and comment.
   (iii) For each block size, calculate the fraction of the counts within 1, 2 and 3 standard deviations of the mean.
   (iv) Repeat (a), (b) and (c) for *proportions* (rather than counts) of purines in each block.

(d) Compute counts of TATAAT in blocks of 5,000 bp. Assuming that these counts follow a Poisson distribution, estimate the parameter of this distribution and obtain an estimate of the standard error of your parameter estimate. This can be done by either a formula or by simulation.