Examination Computational Statistics

Linköpings Universitet, IDA, Statistik

Course code and name:	732A90 Computational Statistics
Date:	2017/11/09, 8–12
Assisting teacher:	Krzysztof Bartoszek
Allowed aids:	Printed books and 100 page computer document
Grades:	A= $[18 - 20]$ points B= $[15.5 - 18)$ points C= $[10.5 - 15.5)$ points D= $[8.5 - 10.5)$ points E= $[7 - 8.5)$ points F= $[0 - 7)$ points **(FAIL)**
Instructions:	Provide a detailed report that includes plots, conclusions and interpretations. Give motivated answers to the questions. If an answer is not motivated, the points are reduced. Provide all necessary codes in an appendix. In a number of questions you are asked to do plots. Make sure that they are informative, have correctly labelled axes, informative axes limits and are correctly described. Points may be deducted for poorly done graphs. Name your solution files as: **[your exam account]_[own file description].[format]** There are **TWO** assignments (with sub–questions) to solve.

NOTE: If you fail to do a part on which subsequent question(s) depend on describe (maybe using dummy data, partial code e.t.c.) how you would do them given you had done that part. You *might* be eligible for partial points.

Assignment 1 (10p)

In the course we considered the sample average

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

but if you think about this more carefully you can see that the observations have to be sampled from a space, let us call it $\mathcal{M}$ where addition is defined and dividing by n makes sense (e.g. if one asks about an "average" human face this formula for the mean could make little sense). If instead the sample comes from a general metric space $(\mathcal{M}, d)$, where d is a function that measures the distance between any two elements from $\mathcal{M}$, then one can consider the *Fréchet mean* defined as

$$m(\mathbf{X}) := \arg\min_{y\in\mathcal{M}} \sum_{i=1}^{n} w_i d(y, X_i)$$

where $\mathbf{X} = \{X_i, \ldots, X_n\}$ is the observed sample. The numbers $\{w_i\}$ are a vector of weights for each observation (e.g. you want to decrease an outlier's influence). Briefly the Fréchet mean is the point that is in the "centre" of your observed sample, where centre has to be understood with respect to the way you calculate distance using the function $d(\cdot,\cdot)$. The notation $\arg\min_{y\in\mathcal{M}} f(y)$ indicates the value of $y \in \mathcal{M}$ that minimizes the function $f(y)$.

Question 1.1 (3p)

Write two separate simulators. One that simulates from the standard normal, $\mathcal{N}(0,1)$, distribution (using only `R`'s `runif()`) and another that simulates from the geometric distribution with a user provided parameter p (using only `R`'s `sample(c(0,1),1,replace=FALSE,prob=c(1-p,p))`). Point(s) will be deducted for using other generators, in particular (but not limited to) `rnorm()` or `rgeom()`.
Tip: Recall that if $\theta \sim \text{Unif}[0, 2\pi]$ and $D \sim \text{Unif}[0, 1]$, then the pair

$$X_1 = \sqrt{-2\ln D}\cos\theta, \quad X_2 = \sqrt{-2\ln D}\sin\theta$$

is a pair of independent and $\mathcal{N}(0,1)$ distributed random variables.
Tip: The geometric distribution can be interpreted as the number of tries to observing the first success.

Question 1.2 (5p)

Using your implemented simulators generate a bivariate (U, V) sample of 10 and 50 observations, where U and V are independent $U \sim \mathcal{N}(0,1)$ and $V \sim$ geometric with parameter $p = 1/3$ distributions. Using `optim()` find the Fréchet means of all 3 samples separately for $d(\vec{x}, \vec{y}) := |x_1 - y_1| + |x_2 - y_2|$ and $d(x, y) := (x_1 - y_1)^2 + (x_2 - y_2)^2$. You may take $w_i = 1$ everywhere. You are free to choose the optimization method.
Tip: Take the sample mean or median as the starting point of the optimization.

Question 1.3 (2p)

Make a scatterplot of your samples. On the plot indicate the estimated Fréchet mean, the sample average and sample median (calculate the median separately in both dimensions). Can you draw any conclusions? If you find it difficult to make conclusions try generating samples of size 30 and 100, calculate the Fréchet means (using `optim()`) and do the same plots.

Assignment 2 (10p)

Question 2.1 (2p)

Assume you have a routine for generating uniform numbers Unif$[0, 1]$. Implement an algorithm for sampling a uniform integer between 1 and n, i.e. $X \sim \text{Unif}\{1, \ldots, n\}$. Point(s) will be deducted for using other generators, in particular (but not limited to) for using `sample()`.

Question 2.2 (5p)

Consider a regular lattice of size $m \times m$. A robot should walk from vertex $(1, 1)$ to (m, m). It is only allowed to walk horizontal or vertical edges. When the robot reaches a vertex, it takes a random direction, including the edge where it entered. For an internal vertex this means 4 equally likely directions. At the outermost vertices and corners there are less options (2 or 3), but these (2 or 3 directions) are still equally likely. The horizontal and vertical edges have distance 1. Figure 1 illustrates this for $m = 4$.

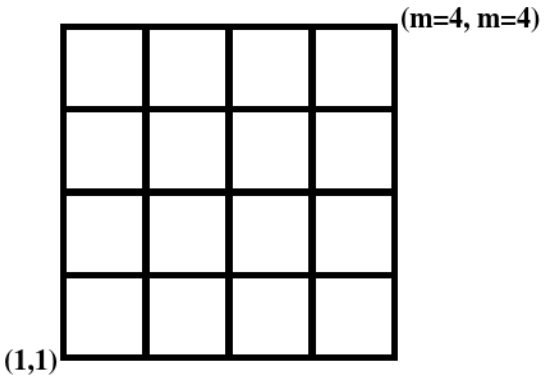

Figure 1: Illustration of vertices and edges. The robot walks from $(1, 1)$ to $(m = 4, m = 4)$.

Let T be the distance walked by the robot before reaching vertex (m, m). What is the probability of the event $T = 4m$, for $m = 4$ and $m = 8$. The distribution of T may be explored by Monte Carlo sampling. It can also be found analytically but this is outside the scope of the course. **Save your simulations for Question 2.3!**

TIP: Your generated sample should be of size about 1000 to get a decent estimate of the probability. If your implementation is OK, a sample of this size is generated quickly. For testing purposes take about 20 repeats while for your report take about 1000 repeats (try 100 if 1000 takes too long).

Question 2.3 (3p)

What is the sample mean of the length your simulated walk from $(1, 1)$ to (m, m)? Use `R`'s `boot()`, `boot.ci()`, in `library(boot)`, function to obtain 95% bootstrap confidence intervals for the mean. Report all the calculated bootstrap confidence intervals. Provide a short (maximum four sentences) intuitive description how the bootstrap and bootstrap confidence interval work (using formulae is strongly discouraged).
TIP: Take about 1000 bootstrap replicates.