Computer Lab 1
Computational Statistics

Linköpings Universitet, IDA, Statistik

2020/11/04

Kurskod och namn:	732A90 Computational Statistics
Datum:	2020/11/03—2020/11/10 (lab session 4 November 2020)
Delmomentsansvarig:	Krzysztof Bartoszek, Martin Beneš, Filip Ekström, Oriol Garrobé Guilera
Instruktioner:	This computer laboratory is part of the examination for the Computational Statistics course Create a group report, (that is directly presentable, if you are a presenting group), on the solutions to the lab as a **.PDF** file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments. **All R code should be included as an appendix into your report.** A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes. In the report reference **ALL** consulted sources and disclose **ALL** collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to marbe619@student.liu.se, filip.ekstrom@liu.se, origa255@student.liu.se, or krzysztof.bartoszek@liu.se), by **23:59 10 November 2020** at latest. Notice there is a final deadline of **23:59 31 January 2021** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year. The seminar for this lab will take place **25 November 2020**. The report has to be written in English.

Question 1: Be careful when comparing

Consider the following two R code snippets

```
x1<-1/3;x2<-1/4
if (x1-x2==1/12){
print("Subtraction is correct")
}else{
print("Subtraction is wrong")
}
```

and

```
x1<-1;x2<-1/2
if (x1-x2==1/2){
print("Subtraction is correct")
}else{
print("Subtraction is wrong")
}
```

1. Check the results of the snippets. Comment what is going on.
2. If there are any problems, suggest improvements.

Question 2: Derivative

From the defintion of a derivative a popular way of computing it at a point x is to use a small ϵ and the formula

$$f'(x) = \frac{f(x+\epsilon) - f(x)}{\epsilon}.$$

1. Write your own R function to calculate the derivative of $f(x) = x$ in this way with $\epsilon = 10^{-15}$.
2. Evaluate your derivative function at $x = 1$ and $x = 100000$.
3. What values did you obtain? What are the true values? Explain the reasons behind the discovered differences.

Question 3: Variance

A known formula for estimating the variance based on a vector of n observations is

$$\mathrm{Var}(\vec{x}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

1. Write your own R function, `myvar`, to estimate the variance in this way.
2. Generate a vector $x = (x_1, \ldots, x_{10000})$ with 10000 random numbers with mean 10^8 and variance 1.
3. For each subset $X_i = \{x_1, \ldots, x_i\}$, $i = 1, \ldots, 10000$ compute the difference $Y_i = \texttt{myvar}(X_i) - \texttt{var}(X_i)$, where $\texttt{var}(X_i)$ is the standard variance estimation function in R. Plot the dependence Y_i on i. Draw conclusions from this plot. How well does your function work? Can you explain the behaviour?
4. How can you better implement a variance estimator? Find and implement a formula that will give the same results as `var()`?

Question 4: Binomial coefficient

The binomial coefficient "n choose k" is defined as

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{(k+1)(k+2)\cdots(n-1)n}{(n-k)!},$$

where n and k are an arbitrary pair of integers satisfying $0 \le k \le n$. Consider the three below `R` expressions for computing the binomial coefficient. They all use the `prod()` function, which computes the product of all the elements of the vector passed to it.

A) `prod(1:n) / (prod(1:k) * prod(1:(n-k)))`
B) `prod((k+1):n) / prod(1:(n-k))`
C) `prod(((k+1):n) / (1:(n-k)))`

1. Even if overflow and underflow would not occur these expressions will not work correctly for all values of n and k. Explain what is the problem in A, B and C respectively.
2. In mathematical formulae one should suspect overflow to occur when parameters, here n and k, are large. Experiment numerically with the code of A, B and C, for different values of n and k to see whether overflow occurs. Graphically present the results of your experiments.
3. Which of the three expressions have the overflow problem? Explain why.