Lab09: Sampling

**Handout date:** Wednesday, day before *Halloween*, 2019

**Due date:** Wednesday, November 6, 2019 at the beginning of class as hardcopy

*This lab counts* 4 % *toward your total grade*

**Objectives:** In this lab you   
[a] empirically explore the central limit theorem from sampled data,

[b]evaluate the distribution of the sample mean,

[c] spatially sample provinces from an underlying population,

[b] estimate an area by using random points or random traverses.

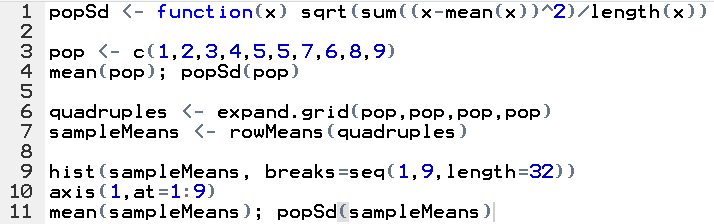
**Format of answer:** Your answers (graphs and verbal description) should be handed in as ***hard-copy*** in ***one*** document. Add a running title into the header of the document with the following information: ***your name***, ***Lab09*** and ***page numbers***. Label each answer properly starting with its task number. Maintain the sequence of questions. For your answers, use ***Word's Equation Editor***. Format any code and computer output properly before inserting it into the document with your answer. R-code and text output need to be in a ***monospaced*** font (i.e., fixed-pitch font) such as Courier New so proper spacing and alignments are preserved. Excessive, but irrelevant, output will lead to a deduction of your accumulated points.

# Part I: Central Limit Theorem

## Task 1: Evaluating the distribution of the sample mean (0.9 points)

Following the example on pages 273-275 in Burt, Barber and Rigby. Evaluate the distribution of the sample mean, which itself is a random variable.

You will fully enumerate all possible samples of length 4 with replacement from a population **pop** with 10 data values. For each possible sample its mean is calculated and the distribution of all sample means is evaluated. Use the code below:



[a] Give the *expectation* and *standard deviation* of the *population*. Why do you need to use the function **popSd( )** here rather than the standard  function **sd( )**? Hint: Check the denominator of the function. (0.2 points)

[b] What is the function **expand.grid( )** doing and how does its generated sample differ from the other enumeration function **combn( )**? (0.2 points)

[c] What is the plot in lines 9 and 10 exactly showing? Why do the tails of the distribution have so low frequencies? (0.2 points)

[d] How does the expectation of the sample means relate to the population expectation . (0.1 points)

[e] How does the standard error of the sample means relate to the standard deviation of the population . (0.2 points)

## Task 2: Explore Samples (1.5 points)

For task 2 show all relevant  code.

[a] Plot the *density function* of a beta distribution (see  function **dbeta( )**) with the shape parameters **shape1=1.5** and **shape2=6** in the default range of (you can use **x <- seq(0,1,by=0.01)**). Label the axes of your plot properly. (0.2 points)

Note: You may want to use the **plot( )** function with a properly set parameter for the **type**-option.

[b] Check what the theoretical ***expected value***, ***variance*** and ***skewness*** of this beta distribution are at **Wikipedia** and report their values. Note that the shape parameters in **Wikipedia** are called and . (0.2 points)

[c] Generate two sets of samples from the beta distribution. (0.2 points)

1. The first set consists of 10,000 samples of size .
2. The second set consists of 10,000 samples of size .

You can generate these sample with the statement **matrix(rbeta(*n*\*10000, …), nrow=10000, ncol = *n*)** .

Calculate the 10,000 means of the samples for both datasets and save them into a vector. You may use the function **rowMeans( )**. ***Do not show*** the 10,000 calculated means based on either sample size but show your code!

[d] Calculate the *mean* of the distribution of the sample means, the mean’s *standard error* and mean’s *skewness* for both sets (i.e., and ) and report the results in a properly formatted table. (0.2 points)

[e] Plot the histograms with the **hist( )** function of distributions of the sample means for both sets. Use the option **freq=F** for the histogram. Make sure that you use a proper bin-width and value range for both histograms. (0.2 points)

Superimpose a normal curve within the chosen value range of the histogram. Use the correct theoretical parameters for the mean and standard deviation of the normal curve. The following statements will perform the task:  
**xRange <- seq(…, by=0.001)  
lines(xRange, dnorm(xRange, mean=…, sd=…), col="red", lwd=2)**

[f] Discuss both distributions of the sample means with regards to the central limit theorem and the underlying beta distributed population with the given shape parameters: (0.3 points)

1. expected value of the distribution of the sample means;
2. standard error of the distribution of the sample means in dependence to the sample size ; and
3. shape of the distribution of the sample means compared to that of the normal distribution and their deviations from the normal distribution, measured here by the skewness, in relation to the sample size .

# Part II: Spatial Sampling

## Task 3: Randomly Sampling a Set of Regions (0.6 points)

For task 3 show all relevant  code.

[a] Open the file **Provinces.shp** from Lab03. Randomly sample [a] a set of 15, and [b] a set 35 provinces out of the 95 Italian provinces ***without replacement***. (0.3 points)  
Hints:

* Study the function **sample( )** in ’s online help.
* The syntax **idx <- sample(1:95,*…*)** and **Provinces.shp[idx, ]** performs the selection of the provinces.

[b] Generate a qualitative map for each set of sampled provinces, which shows your sampled and non-sampled provinces. (0.3 points)

## Task 4: Area Estimation by Sample Points and Traverses (1 point)

Below you find a map of a study region with a total area of 1 unit. This region has several lakes; one lake also has an island.

Note: the true total lake area happens to be 27 % of the study areas. In total 20 random points were placed into this study region. Furthermore, randomly 10 point pairs were connected by straight lines.   
Hint: This task can be done manually and using a ruler to evaluate the segment length of the traverses.

[a] Use the 20 sample points to estimate the total lake area. Explain the equation that you have used to estimate the area. (0.5 points)

[b] Use a ruler to estimate the total lake area based on the 10 random lines (labeled L:1 to L:10). Generate a table similar to that on page 286 of Burt, Barber and Rigby. (0.5 points)

