Lab 1

- Generate in R the following sequences (do not type them by hand):
  i) 1, 3, 5, . . . , 999.
  ii) 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3
  iii) 1, 1, 1, 2, 2, 3
  iv) 1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1
  v) 1, 1/2, 1/3, . . . , 1/10 (use the function *fractions()* in the MASS package)
  vi) 1, 8, 27, 64, 125, 216

- Compute in R and discuss the results:
  i) seq(0,10.5,by=1)
  ii) -0.5:10
  iii) 0:10-0.5
  iv) seq(0.5, 9.5)
  v) 10:22/10
  vi) 10/2:22
  vii) (10/2):22
  viii) (1:2)*(0:3)
  ix) r=1:5; s=-2:2; s/r; r/s; s/s

Lab 2

Exercises.

1. We consider the data set *Cars93* from the *MASS* package.
   i) Build a dataframe that contains only the variables *Manufacturer*, *Make*, *Price*, *Passengers* and *Origin*.
   ii) Select and print only the data for the cars built in the US. How many such cars are in the data set?
   iii) Select the car models that are produced by the manufacturer Ford.
   iv) Select and print the data available for the cars that can carry at least 5 passangers, sorted in an ascending order with respect to the price.
2. Consider the vector $x = c(1, 4, 2, 6, 8, 9, 11, 10, 21)$. Compute the sum of the elements which have indices that are a multiply of 3.
3. Use R to identify the elements of the sequence $\{2^1, 2^2, \ldots, 2^{15}\}$ that exceed the corresponding elements of the sequence $\{1^3, 2^3, \ldots, 15^3\}$.
4. An experiment had 10 different trials. Create a character vector with 10 different names for the trials, e.g., "*Trial* 1", ... Use the function *paste()*.

5. The *mtcars* data set records information about cars from 1972. The values are coded using numbers. Recoding as factors can be more informative for the user. Recode the *am* variable, with 0 being "automatic" and 1 being "manual". How many cars have automatic (manual, respectively) transmission.

6. Consider the built-in *USArrests* dataframe.
   i) Determine the number of rows and columns for this dataframe. What are the types of the variables? Use the *str()* function.
   ii) Calculate the mean of each variable of this dataframe.
   iii) Find the average per capita murder rate (*Murder*) in regions where the percentage of the population living in urban areas (*UrbanPop*) exceeds 77%. Compare this with the average per capita murder rate where urban area population is less than 50%.

7. Generate a sequence t of equidistant values from 0 to 20, with a step of 0.01. Plot the points with coordonates $(\sqrt{t}\sin(2\pi t), \sqrt{t}\cos(2\pi t))$ using the *plot()* function. Connect the points with segments. Use the Help menu or window to understand the syntax of this function.

Lab 3

Solve the following exercises.

1. Write a function in R that determines if a number given as an argument is prime or not. Verify beforehand whether the number is a positive integer.

2. Let F_n denote the nth Fibonacci number.

   (a) Define a function that returns the first n terms of the Fibonacci sequence.

   (b) Define a function that returns the n^{th} term of the sequence $G_n = \frac{F_{n+1}}{F_n}$, $n \geq 1$.

   (b) Plot the first 30 terms of this sequence using the **plot()** function. To what real number does the sequence appear to be converging? Hint: golden ratio.

Bisection is an algorithm for finding a zero of a function. The algorithm starts with two values x_1 and x_2 for which $f(x_1)$ and $f(x_2)$ have opposite signs. If $f(x)$ is a continuous function, then we know a root must lie somewhere between these two values. We find it by evaluating $f(x)$ at the midpoint, and selecting whichever half of the interval still brackets the root. We then start over, and repeat the calculation until the interval is short enough to give us our answer to the required precision.

3. (a) Write a function to implement the bisection algorithm.
   (b) Use your function to find a root to 6 decimal place accuracy for $f(x) = x^3 + 2x^2 - 7$, given that the root is known to lie between 0 and 2.
   (c) Calculate how many steps you need to find the root at (b).

4. Let $f(x) = x$ for $0 \leq x \leq \sqrt{2}$ be the PDF of a triangular random variable X.
   (a) Define a function that computes $P(X \leq b)$, for $0 \leq b \leq \sqrt{2}$. What is the name of this function.
   (b) Find the mean and variance of X by using the **integrate()**.
   (c) Define the quantile function for this distribution.
   (d) plot the PDF and CDF of the random variable X.

5. A sample of 100 people is drawn from a population of $600,000$. If it is known that 40% of the population has a specific attribute, what is the probability that 35 or fewer in the sample have that attribute? Use built-in functions in R for the appropriate distribution.

6. Suppose that the population of adult, male black bears has weights that are approximately distributed as $Normal(350, 75)$. What is the probability that a randomly observed male bear weighs more than 450 pounds? What is the $95th$ percentile?

7. For the Student t distribution, we can see that as the degrees of freedom get large, the density approaches the normal. To investigate, plot the standard normal density with the command **curve(dnorm(x), -4, 4)** and add densities for the t-distribution with $k = 5, 10, 25, 50,$ and 100 degrees of freedom. These can be added as follows:

   k = 8; curve(dt(x,df=k), add=TRUE)

Lab 4

1. Historically, a certain baseball player has averaged three hits every ten official at bats (he's a 300 hitter). Assume a binomial model for the number of hits in a 600-at-bat season. What is the probability the player has a batting average higher than 0.350? Use the normal approximation to answer.

2. An elevator can safely hold 3,500 pounds. A sign in the elevator limits the passenger count to 15. If the adult population has a mean weight of 180 pounds with a 25-pound standard deviation, how unusual would it be, if the central limit theorem applied, that an elevator holding 15 people would be carrying more than 3,500 pounds?

   (See Chapter 6 in the book "Using R for introductory statistics", John Verzani.)

3. If Z is $Normal(0, 1)$, find the following:

   1. $P(Z \leq 2.2)$.
   2. $P(-1 < Z \leq 2)$
   3. $P(Z > 2.5)$
   4. b such that $P(-b < Z \leq b) = 0.90$

4. For the Student t distribution, we can see that as the degrees of freedom get large, the density approaches the normal. To investigate, plot the standard normal density and add densities for the t-distribution with $k = 5, 10, 25, 50,$ and 100 degrees of freedom.

5. Consider the buil-in data set *infert* and determine the following contigency tables:

    1. for the variables *education* and *spontaneous*;
    2. for the variables *education*, *induced* and *spontaneous*;
    3. add the marginal frequencies to both crosstabs and extract the ones computed on rows.

6. Use the data set "iris" to compute:

    1. mean, median, mode for *Petal.Length*;
    2. range, variance, standard deviation, IQR for *Sepal.Length* for the species "versicolor"; are there any outliers?
    3. histogram and qqplot for the variable *Petal.Width*; does the data seem to have a normal distribution?
    4. do a boxplot of *Petal.Length* with respect to the three different species.

Lab 6

1. Consider the data given in the file poverty.txt. This data set of size n = 51 are for the 50 states and the District of Columbia in the United States. The variables are y = year 2002 birth rate per 1000 females, aged 15 to 17 years old, and x = poverty rate, which is the percent of the state's population living in households with incomes below the federally defined poverty level. (Data source: Mind On Statistics, 3rd edition, Utts and Heckard).

   Read the data in R using the function read.table(). Construct a scatter plot of the data and compute the correlation coefficient of the two variables. Determine a linear regression model for the two variables. Plot the regression line over the scatter plot. Analyze the model and say whether or not the assumptions are met. Predict the birth rate for a value of 12.7 for the poverty rate.

2. Using the data in the "mtcars" data frame, perform a multiple linear regression on the "mpg" variable with respect to the variables "disp", "hp", "wt" (extract the data first into a new data frame). Compute the correlation coefficient for every pair of variables using the function cor(). Make sure that the assumptions are met. Redo the model by excluding the variable that does failes the t-test. Predict the mileage for a car with hp $= 102$ and wt $= 2.91$.

Choose one of the following problems.

3. Consider the data set logregr1.csv which has 200 observations and the outcome variable used will be "hon", indicating if a student is in an honors class or not. Import the data in R using the function "read.csv()" and perform logistic regression including the variables "female" and "math" (a math score). Discuss your findings and say with what percentage do the odds of being in honors class increase with an increase of one unit in each variable separately.

4. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable. Read the data into R from the source https://stats.idre.ucla.edu/stat/data/binary.csv and perform logistic regression. Discuss your findings.

Lab 7

1. The United States Department of Energy conducts weekly phone surveys on the price of gasoline sold in the United States. Suppose one week the sample average was \$4.03, the sample standard deviation was \$0.42, and the sample size was 800. Perform a one-sided significance test of $H_0 : \mu = 4.00$ against the alternative $H_a : \mu > 4.00$. Find the p-value of the test.

2. The **babies** (**UsingR**) data set contains data covering many births. Information included is the gestation period and a factor indicating whether the mother was a smoker.

   a) Test, at the significance level of 0.05, whether the gestation period for smoking and non-smoking mothers is the same.
   b) Is there a difference between the ages of mothers and fathers?

3. In the built-in data set named **quine**, children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school. Assuming that the data in quine follows the normal distribution, find out if there is a difference in female proportions with respect to ethnicity. Perform a χ^2 test for proportions, $\alpha = 0.05$.

4. Students wishing to graduate must achieve a specific score on a standardized test. Those failing must take a course and then attempt the test again. Suppose 12 students are enrolled in the extra course and their two test scores are given by

Student	scores											
Pre-test	17	12	20	12	20	21	23	10	15	17	18	18
Post-test	19	25	18	18	26	19	27	14	20	22	16	18

Assuming these students represent a random sample from some larger population, perform a t-test to see if there would be any improvement in the population mean scores following such a class. If you assume equal variances or a paired test, explain why.

5. The data set **fat** (**UsingR**) contains several body measurements that can be done using a scale and a tape measure. These can be used to predict the body-fat percentage (**body.fat**). Measuring body fat requires a special apparatus; if our resulting model fits well, we have a low-cost alternative. Fit the variable **body.fat** using each of the variables age, weight, height, BMI, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist. Which variables are marked as statistically significant by the marginal

t-tests? Use the **stepAIC** function to select a submodel. For this submodel, what is the adjusted R^2?

6. The data set **hall.fame** (**UsingR**) contains statistics for several Major League Baseball players over the years. We wish to see which factors contribute to acceptance into the Hall of Fame. To do so, we will look at a logistic regression model for acceptance modeled by the batting average (BA), the number of lifetime home runs (HR), the number of hits (hits), and the number of games played (games). First, make a binary variable for Hall of Fame membership. Now, fit a logistic regression model of hfm modeled by the variables above. Which variables are flagged as significant? Run **stepAIC**. Which model is selected?