Problem 4 (2 credits)

0
1
2

Question Nr. 0EA89DS9L0PY63ZE3

The `brexit_polls` dataset from the dslabs package contains poll outcomes for 127 polls performed by different pollsters either online or by telephone (poll_type).

```
data(brexit_polls)
head(brexit_polls)
```

```
##    startdate    enddate   pollster poll_type samplesize remain leave undecided
## 1 2016-06-23 2016-06-23     YouGov    Online       4772   0.52  0.48      0.00
## 2 2016-06-22 2016-06-22    Populus    Online       4700   0.55  0.45      0.00
## 3 2016-06-20 2016-06-22     YouGov    Online       3766   0.51  0.49      0.00
## 4 2016-06-20 2016-06-22 Ipsos MORI Telephone       1592   0.49  0.46      0.01
## 5 2016-06-20 2016-06-22    Opinium    Online       3011   0.44  0.45      0.09
## 6 2016-06-17 2016-06-22     ComRes Telephone       1032   0.54  0.46      0.00
##   spread
## 1   0.04
## 2   0.10
## 3   0.02
## 4   0.03
## 5  -0.01
## 6   0.08
```

You are interested in generating a table that shows the number of polls done online and by telephone for the pollsters YouGov, Ipsos MORI and Opinium. Write R code using the library data.table to create such table with the same column names (header displayed below).

```
##      pollster N_polls_online N_polls_telephone
## 1: Ipsos MORI              0                 7
## 2:    Opinium              9                 0
## 3:     YouGov             25                 1
```



```
brexit_polls <- as.data.table(brexit_polls)
dt_1 <- brexit_polls[pollster %in% c("Ipsos MORI","Opinium","YouGov")][poll_type == "Online"][, .(N_polls_online = .N), by = pollster]
dt_2 <- brexit_polls[pollster %in% c("Ipsos MORI","Opinium","YouGov")][poll_type == "Telephone"][, .(N_polls_telephone = .N), by =
pollster]

dt_3 <- merge(dt_1,dt_2, by = "pollster", all = T)
dt_3[is.na(dt_3)] <- 0
```

Problem 5 (2 credits)

0
1
2

Question Nr. 9C2892V3965B2346Y4957F1423J198

Load the column 'global_history' from the dataset 'nyc_regents_scores'. Compute its median. Additionally obtain a 80% equi-tailed bootstrap confidence interval for this quantity. Run 999 bootstrap iterations. Provide R code and the lower and upper bound of the interval rounded to two significant digits using signif(...,digits=2).

Load the data using the following code:

```
dt <- as.data.table(dslabs::nyc_regents_scores)
dt <- na.omit(dt)
```

```
dt <- as.data.table(dslabs::nyc_regents_scores)
dt <- na.omit(dt)

median_global <- median(dt$global_history)
# number of random simulations
R <- 999

# initialize T_boot with missing values
# (safer than with 0's)
T_bootstrap <- rep(NA, 1000)

# iterate for i=1 to R
for(i in 1:R){
  # sample the original data with same size with replacement
  dt_boot <- dt$global_history[sample(nrow(dt$global_history), replace=TRUE)]
  # store the difference of medians in the i-th entry of T_permuted
  T_bootstrap[i] <- diff_median(dt_boot)
}
```

Problem 7 (2 credits)

0
1
2

Question Nr. 8NC1OH0M16TJ23YV57WP3

Consider only the features 'PLEKHJ1' and 'OTC' from the 'tissue_gene_expression' dataset. Provide R code that allowes determining, e.g. with an appropriate plot, the feature with the highest recall at a false positive rate of 0.43 as a predictor of gene expression in liver (variable y == "liver"). Write down the name of this feature and justify your choice.

Load the data using the following code:

```
library(dslabs)
dt <- as.data.table(tissue_gene_expression$x)
dt[, y := tissue_gene_expression$y]
```

```
dt <- dt[,c("PLEKHJ1","OTC","y")]
log_model_2 <- glm(y ~ PLEKHJ1, data=dt, family = "binomial")
log_model_3 <- glm(y ~ OTC, data=dt, family = "binomial")

pred_1 <- dt[, pred_PLEK := round(predict(log_model_2, dt, type ="response"))]
pred_2 <- dt[, pred_OLC := round(predict(log_model_3, dt, type ="response"))]
```

Problem 8 (1 credit)

0
1

Question Nr. 0RCKP4Y75EPRJB0YNDT76LF3VA

Consider a multivariate dataset with 6 observations denoted as W,F,R,P,S and T.

A first clustering method gives the two clusters {P} and {W,F,R,S,T}. Applying k-means clustering yields two clusters {S} and {W,F,R,P,T}. Applying hierarchical clustering yields two clusters {R,T} and {W,F,P,S}. Which of the k-mean clustering and the hierarchical clustering yielded the grouping of the elements most similar to the first clustering? Base your answer on a metric learned in the lecture.

```
dt <- data.table(element = c("w","f","r","p","s","t"),
                 first_cluster = c(2,2,2,1,2,2),
                 k_means = c(2,2,2,2,1,2),
                 hc = c(2,2,1,2,2,1))
dt

my_rand_index <- function(cl1, cl2) {
  ## enumerate all pairs
  pairs = lapply(1:(length(cl1) - 1), function(i)
    lapply((i + 1):length(cl1), function(j) return(c(i, j))))
  pairs = t(matrix(unlist(pairs), nrow=2))
  ## label pairs as same or different in cl1
  same.cl1 = cl1[pairs[,1]] == cl1[pairs[,2]]
  ## label pairs as same or different in cl2
  same.cl2 = cl2[pairs[,1]] == cl2[pairs[,2]]
  ## compare the labels
  same.in.both = sum(same.cl1 & same.cl2)
  diff.in.both = sum(!same.cl1 & !same.cl2)
  ## compute the Rand index
  return((same.in.both + diff.in.both) / nrow(pairs))
}

my_rand_index(dt$first_cluster,dt$k_means)
my_rand_index(dt$first_cluster,dt$hc)

#Based on the rand index results k means resembles the most with the first clustering method. (rand index of k means = 0.46)
```

Problem 9 (2 credits)

Question Nr. 4MCTSENSCXFE7Z8DW

Consider the R code below that defines the variables A,B,C and D:

```
n <- 1000
sigma <- function(z){ 1 / (1 + exp(-z)) }
B <- sigma(rnorm(n))
C <- rbinom(n, size=5, prob=0.5)
D <- rbinom(n, size=10, prob=sigma(C))
A <- rnorm(n, mean=B*D)
```

0 1

a)

Is B statistically independent of C? Justify. No statistical test nor plot is required nor shall be the basis for your justification.

It depends but only considering the relationship between B and C, we can conlude that they are independent since they are constituted from different distributions.

0 1

b)

Given A, is B statistically independent of C? Justify. No statistical test nor plot is required nor shall be the basis for your justification.

Yes they are still independent. A is dependent on both B and C however this does not neccesarily makes B and C dependent. This is a typical case of common consequence.

Problem 10 (1 credit)

0
1

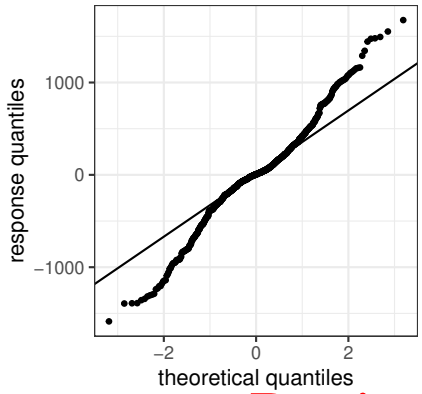
Question Nr. 5LU05CH27RX6WP76FN6

We consider a linear regression model parameterized as

$$y_i = \alpha + \beta \cdot x_i + \epsilon_i$$

where $i = 1...N$ denotes the data point indices, y_i is the response variable, α and β the coefficients, x_i the explanatory variable and ϵ_i the error term. Let $\hat{y}_i$ be the i-th fitted value and $\hat{\epsilon}_i$ be the i-th residual.

Does the following plot provide evidence against the assumptions of the linear regression? Justify.



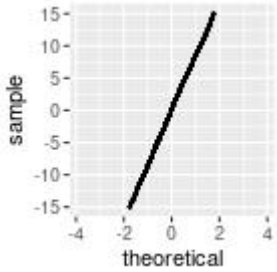
One of the assumptions of linear regression is normality. By checking this QQ plot we can see that data is not normally distributed around the tails, so therefore this plot provides an evidence to the case of non normality.

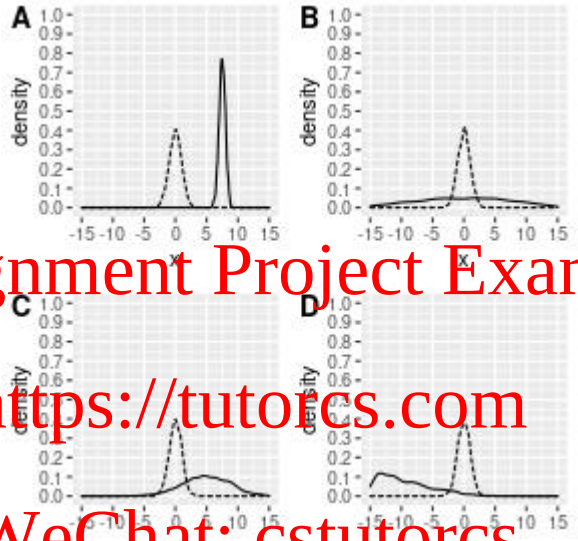
Problem 11 (1 credit)

0
1

Question Nr. 8XW79VF9BD5ZN54PS1

Which density plot A, B, C, or D corresponds to the Q-Q plot (i.e. Q-Q plot against the standard Normal distribution) depicted below? The standard Normal distribution, i.e. the Gaussian distribution with mean 0 and variance 1, is shown in the density plots below with a dashed line. Justify.





B. First, we are looking for a zero mean distribution. Therefore, A, D, C is eliminated. Also we have values ranging from -15 to 15 but in A,C,D we can clearly see that this is not the case.

Problem 12 (2 credits)

0
1
2

Question Nr. 7IS16XJ89VA32GH51VZ2

Consider the variable "fractal_dim_mean" of the "brca" dataset. A researcher wants to find a set of the other variables from the matrix 'brca$x' associating with the variable 'fractal_dim_mean' according to Spearman's correlation such that the set is as large as possible but that, on average, at most 10% of the reported associations are false positives. Identify this set. Provide code, report the size of this set, and justify your answer. Do not mind warnings, if any, about exact p-values with ties.

Load the data using the following code:

```
library(dslabs)
dt <- as.data.table(brca$x)
```

```
res <- cor(dt)
res <- round(res, 2)
```

According to correlation table, fractal_dim_worst have the highest correlation among all variables.

Problem 13 (2 credits)

Question Nr. 5I6747B4958E6546U0942G0402V730

Alice and Bianca would like to predict the variable 'fertility' from the dataset 'gapminder' using linear regression. In this question, assume that the assumptions of linear regression hold in every regression performed. Load the dataset using the command:

```
gap <- data.table(na.omit(dslabs::gapminder))
gap[, gdp_log10 := log10(gdp)]
gap[, infant_mortality_log10 := log10(infant_mortality)]
gap[, population_log10 := log10(population)]
```

a) Alice has proposed a model using only 'gdp_log10' as a predictor. Bianca, argues that we also need to include 'population_log10'. How much more of the variance is explained by Bianca's model? Provide R code and the added fraction of the variance rounded to 2 significant digits.

b) Alice disagrees, arguing that this additional variable does not significantly improve the model. Settle the debate with an appropriate test. Provide R code as well as the p-value rounded to 2 significant digits.

0
1

a)

```
pca <- princomp(gap[, .(gdp_log10,population_log10)])
pca
summary(pca)

# Proportion of Variance Exp.

#gdp_log10 = 0.88
#population_log10 = 0.12
```


0
1

b)

```
l <- lm(fertility ~ gdp_log10 + population_log10, data=gap)
summary(l)

# With a regression model fit, we can see that actually both predictors do a statistically significant job to predict fertility. Both have p values of  0.00000000000002
```

Problem 14 (2 credits)

0
1
2

Question Nr. 7OY93DP39F99BP58DY6

Consider the "brca" dataset from dslabs package. Fit a logistic regression model which predicts the response variable "outcome" given the feature ' smoothness_se '. Assume that all assumptions of the logistic regression model are met. Starting from an original probability of 10%, of malignant (cancer) how much does the probability of developping a malignant (cancer) increase when the feature' smoothness_se ' increases by 0.1 . Provide R code that determines this probability and explicitly state this probability.

Load the data using the following code:

```
library(dslabs)
dt <- as.data.table(brca$x)
dt[, outcome := brca$y]
```

```
log_model <- glm(outcome ~ smoothness_se, data=dt, family = "binomial")
log_model
summary(log_model)

a <- predict(log_model, data.table(smoothness_se = -0.1789 + 0.1, type ="response"))
probability_increase <- exp(a)/ 100
```