Lab 4

Kuzyshyn, Mishchenko

Tasks 1 - 4

Generating sample data

All of this is done according to the formula in the statement of the problem.

```
n = 100
m = 50
k <- 1:(m + n)
id <- 56  # Team ID
a.data = sapply(k, function(x) (x * log(x^2 * id + pi))%%1)
x = qnorm(a.data[1:n])
y = qnorm(a.data[(n + 1):(n + m)])
```

Task 1

As variance is unknown, it's convenient to use the t-test here.

As we have one-sided alternative, the rejection region would be:

$C_\alpha = \{x \in R^{100} | t(x) \leq t^{99}_{0.05}\}$

```
# As std is unknown, let's estimate it firstly
S <- sd(x)
sample_mean <- mean(x)
student <- (sample_mean - 0)* sqrt(100) / S
t <- qt(0.05, df=99)

if (student <= t){
  print("The test concluded that the that the null hypothesis should be rejected at alpha = 0.05")
} else{
    print("The test concluded that the that the null hypothesis should not be rejected at alpha = 0.05"
}
```

```
## [1] "The test concluded that the that the null hypothesis should not be rejected at alpha = 0.05"
```

```
t.test(x, mu = 0, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.5859, df = 99, p-value = 0.05798
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##         -Inf 0.006334158
## sample estimates:
```

```
##  mean of x
## -0.1348022
```

As we can see, the p-value is slightly bigger than α so we do not reject our null hypothesis in this case.

Task 2

Here, we test for the strict equality of μ_1 and μ_2 with variances assumed known: $\sigma_1^2 = \sigma_2^2 = 2$.

For this, we use the GLRT with test statistics $2log\mathbf{L}_{x,y}(H_0, H_1)$ in comparison to $\chi^{(1)}_{1-\alpha}$, with test size equal to 0.05.

```
alpha <- 0.05
var <- 2

chisq_statistic <- (m * n / (m + n)) * ((mean(x) - mean(y))^2 / var)
```

```
## [1] "GLRT value: 2.474659"
```

```
## [1] "Chi-squared distribution quantile of size 0.95: 3.841459"
```

```
## [1] "The test concluded that the that the null hypothesis should be accepted at alpha = 0.05."
```

You can also transform the coefficient into a z-score by using its square root, at which point you can use the z-test of size 1 - alpha / 2.

```
norm_statistic <- sqrt(chisq_statistic)
```

```
## [1] "z-score: 1.573105"
```

```
## [1] "Normal distribution quantile of size 0.975: 1.959964"
```

```
## [1] "The test concluded that the that the null hypothesis should be accepted at alpha = 0.05"
```

The p-value of these tests is identical, since they themselves are merely transformations of each other, and is exactly:

```
pvalue <- 2 * pnorm(-norm_statistic)
```

```
## [1] "P-value: 0.115695"
```

Since usually null hypotheses are rejected for a p-value less than 0.05, we cannot reject it here.

Thus, we conclude that μ_1 and μ_2 must be equal with $\sigma^2 = 2$.

Task 3

Here, we test for $\sigma_1^2 = 1$ with $\mu_1 = 0$.

This is a two-sided chi-squared test with the statistics of $V = \sum_{k=1}^{n}(X_k - \mu)^2/\sigma_0^2$

```
alpha <- 0.05
mu <- 0
var <- 1
chisq_statistic <- sum(((x - mu) ^ 2) / var)
```

```
## [1] "V: 73.347771"
```

```
## [1] "Chi-squared distribution quantile of size 0.025: 0.000982"
```

```
## [1] "Chi-squared distribution quantile of size 0.975: 5.023886"
```

```
## [1] "The test concluded that the that the null hypothesis should be rejected at alpha = 0.05"
```

The p-value is twice the smaller of the chi-squared c.d.f. at V(X) or 1 minus the same value.

```
pvalue <- 2 * min(pchisq(chisq_statistic, df=1),  1 - pchisq(chisq_statistic, df=1))
```

```
## [1] "P-value: 0.000000"
```

In this case, however, the statistic V(X) is so large that the p-value is too small for R's handler of float values.

Task 4

Problem 4

In this problem, both means and both variances are unknown, so, according to hint, we should use the f-test here.

Hence, our test statistics is $F = Var(x)/Var(y)$ and the rejection region would be:

$C_\alpha = \{x \in R^n | F > F_{0.05,99,49}\}$

```
S_x = var(x)
S_y = var(y)
f_statistic <- S_x / S_y
f <- qf(alpha, 99, 49)

if (f_statistic > f){
  print("The test concluded that the that the null hypothesis should be rejected at alpha = 0.05")
} else{
    print("The test concluded that the that the null hypothesis should not be rejected at alpha = 0.05"
}
```

```
## [1] "The test concluded that the that the null hypothesis should not be rejected at alpha = 0.05"
```

```
pf(f_statistic, 99, 49)
```

```
## [1] 0.003558429
```

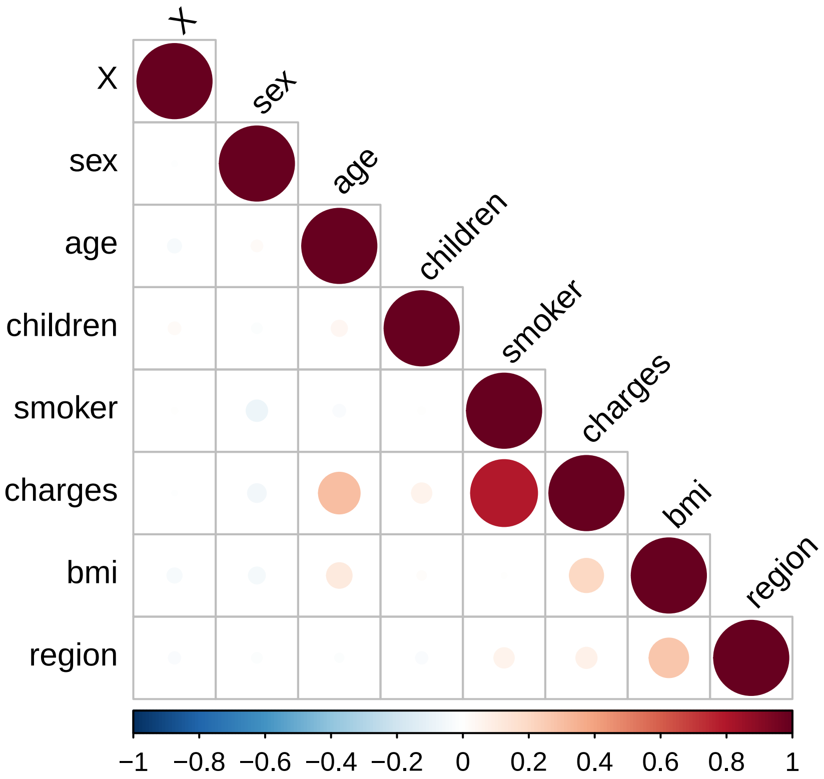
P-value turned out to be 0.003558429, so we can safely reject the null hypothesis, since it is much smaller than the 0.05 threshold.

Problem 5

```
dataframe <- read.csv("insurance.csv")

rquery.cormat(dataframe)
```

```
## corrplot 0.84 loaded
```



```
## $r
##                X    sex    age children smoker charges  bmi region
## X              1
## sex      -0.0037      1
## age       -0.031  0.021      1
## children   0.025 -0.017  0.042        1
## smoker    0.0052 -0.076 -0.025   0.0077      1
## charges  -0.0034 -0.057    0.3    0.068   0.79       1
## bmi       -0.036 -0.046   0.11    0.013 0.0038     0.2    1
## region    -0.023 -0.017 -0.012   -0.023  0.068   0.074 0.27      1
##
## $p
##             X    sex     age children   smoker charges     bmi region
## X           0
## sex      0.89      0
## age      0.25   0.45       0
## children 0.36   0.53    0.12        0
## smoker   0.85 0.0053    0.36     0.78        0
## charges   0.9  0.036 4.9e-29    0.013 8.3e-283       0
## bmi      0.19   0.09 6.2e-05     0.64     0.89 2.5e-13       0
## region    0.4   0.53    0.67      0.4    0.012  0.0068 8.7e-24      0
##
## $sym
##          X sex age children smoker charges bmi region
## X        1
## sex        1
## age            1
## children           1
## smoker                      1
## charges                     ,      1
## bmi                                        1
## region                                         1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

As we can see from the matrix above, the biggest correlation with charges are the following parameters: 1) Whether a person is smoker or not(correlation +0.79) 2) The age of a person(+0.3) The rest are almost uncorrelated, so we will not count them, in order to save computational power.

```r
multi.fit <- lm(charges ~ smoker + age, data=dataframe)
summary(multi.fit)
```

```
##
## Call:
## lm(formula = charges ~ smoker + age, data = dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16088.1  -2046.8  -1336.4   -212.7  28760.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2391.63     528.30  -4.527 6.52e-06 ***
## smoker      23855.30     433.49  55.031  < 2e-16 ***
## age           274.87      12.46  22.069  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6397 on 1335 degrees of freedom
## Multiple R-squared:  0.7214, Adjusted R-squared:  0.721
## F-statistic:  1728 on 2 and 1335 DF,  p-value: < 2.2e-16
```

a) $\hat{a} = -2391.63$(sum of all $\hat{a}_1, \hat{a}_2...$) $\hat{b}_1 = 23855.30$, $\hat{b_2} = 274.87$. Standard error is equal to 6397

b) As the p-value is much less than 0.01, we can reject the null hypothesis. Hence there is significant relationship between the charges, and age with smoker variables.

c) The determination coefficient r^2 is equal to 0.7214008. According to multiple articles that we have read, the R square of .70 is generally considered very good, so we can say that our linear model is adequate.

d)

```r
data_1 <- data.frame(age=0, smoker=0)
data_2 <- data.frame(age=20, smoker=20)

predict(multi.fit, newdata=data_1, interval="prediction")
```

```
##         fit       lwr      upr
## 1 -2391.626 -14983.13 10199.88
```

```r
predict(multi.fit, newdata=data_2, interval="prediction")
```

```
##        fit      lwr      upr
## 1 480211.9 459216.9 501206.8
```

The confidence interval for x=20 is gigantic because smoker is a boolean variable that can not be 20)