**Задание Д\_Ан - Решение задачи однофакторной бинарной классификации методом дискриминантного анализа**

Выполнил студент 2 курса

группы 09-715(1)

Санамян Артак Размикович

**Текст задания:**

1. Проверьте гипотезу о нормальности распределения каждой из двух выборок.
2. Проверьте гипотезу о равенстве дисперсий.
3. Используя метод дискриминантного анализа, сделайте прогноз значений переменной отклика для нескольких объектов из обучающей выборки.
4. Оцените качество прогноза.
5. Результаты оформите в виде Word-файла и прикрепите его здесь (Весь R-код поместите в конец файла как приложение).

Источник числовых данных, использованный для выполнения задания: <http://archive.ics.uci.edu/ml/datasets/Flags>

**Описание выполненной работы:**

Имеются данные о населении стран. Необходимо определить (преобладающую) религию в стране:

1 – Христиане;

2 – Мусульмане.

Проверим гипотезу о нормальности распределения выборки стран первого класса (Христиане) с помощью теста Шапиро–Уилка.

> # Проверим гипотезу о нормальности распределения выборки стран первого класса (христиане)

> shapiro.test(country1)

Shapiro-Wilk normality test

data: country1

W = 0.4205, p-value < 2.2e-16

Итак, при уровне значимости, например, α = 0.05 гипотеза должна быть отклонена, так как p-value < α.

Проверим гипотезу о нормальности распределения выборки стран второго класса (Мусульмане) с помощью теста Шапиро–Уилка.

> # Проверим гипотезу о нормальности распределения выборки стран второго класса (мусульмане)

> shapiro.test(country2)

Shapiro-Wilk normality test

data: country2

W = 0.62409, p-value = 2.198e-08

При уровне значимости α = 0.05 гипотеза должна быть отклонена, так как

p-value < α.

Проверим гипотезу о равенстве дисперсий с помощью критерия Фишера.

> # Проверим гипотезу о равенстве дисперсий

> var.test(country1, country2)

F test to compare two variances

data: country1 and country2

F = 0.79567, num df = 99, denom df = 35, p-value = 0.3804

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.4415956 1.3323486

sample estimates:

ratio of variances

0.795674

При уровне значимости α = 0.05 гипотеза должна быть отклонена, так как

p-value < α.

Найдем оценки параметров.

> # Найдём оценки параметров

> omuA=mean(country1)

> "Среднее выборочное значение населения стран для стран первого класса(христиане) = "

[1] "Среднее выборочное значение населения стран для стран первого класса(христиане) = "

> omuA

[1] 11.31

> omuF=mean(country2)

> "Среднее выборочное значение населения стран для стран второго класса(мусульмане) = "

[1] "Среднее выборочное значение населения стран для стран второго класса(мусульмане) = "

> omuF

[1] 19.47222

> osigA=sd(country1)

> "Среднее квадратическое отклонение населения стран для стран первого класса(христиане) = "

[1] "Среднее квадратическое отклонение населения стран для стран первого класса(христиане) = "

> osigA

[1] 29.15292

> osigF=sd(country2)

> "Среднее квадратическое отклонение населения стран для стран второго класса(мусульмане) = "

[1] "Среднее квадратическое отклонение населения стран для стран второго класса(мусульмане) = "

> osigF

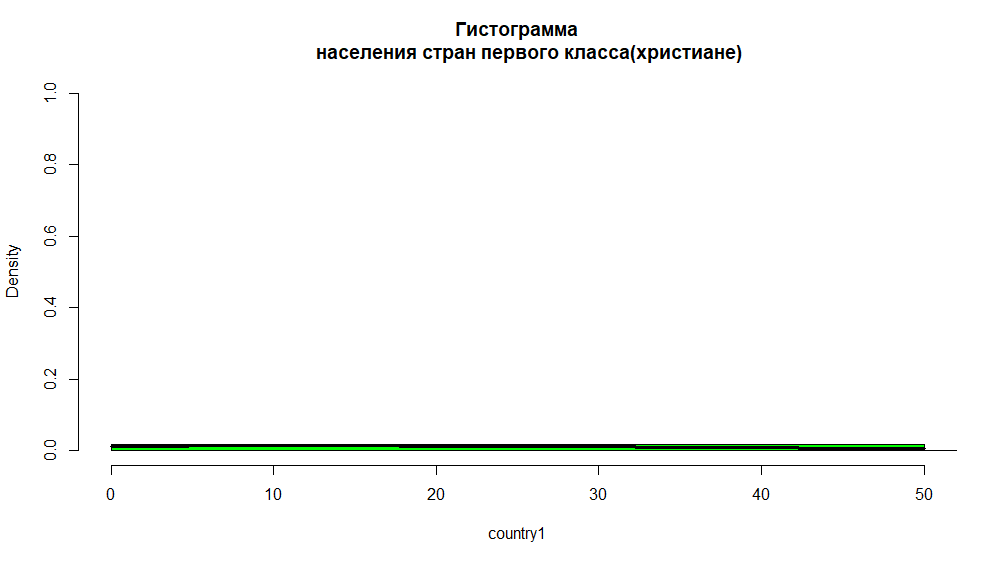
[1] 32.68244

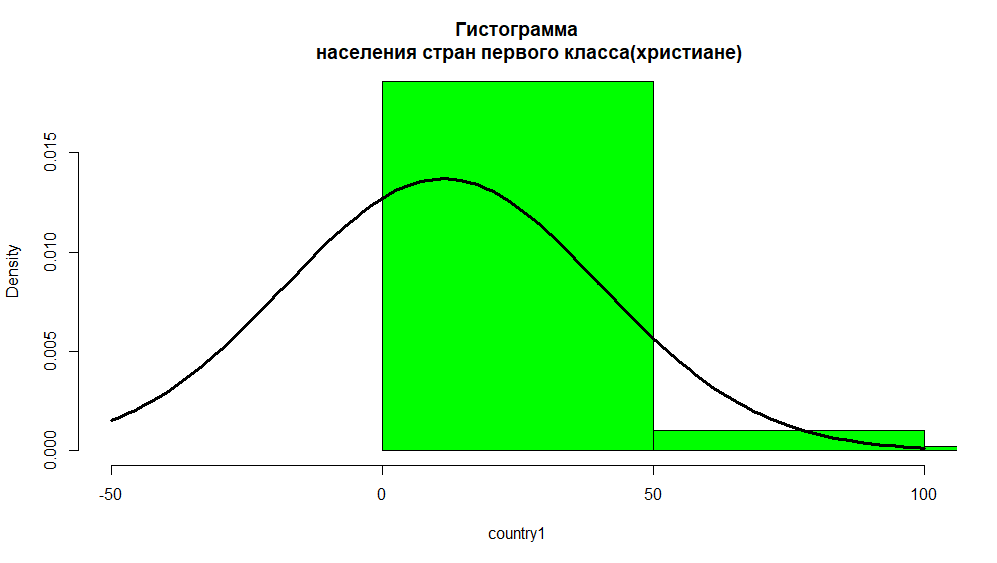
> osig = (osigA + osigF)/2

> osig

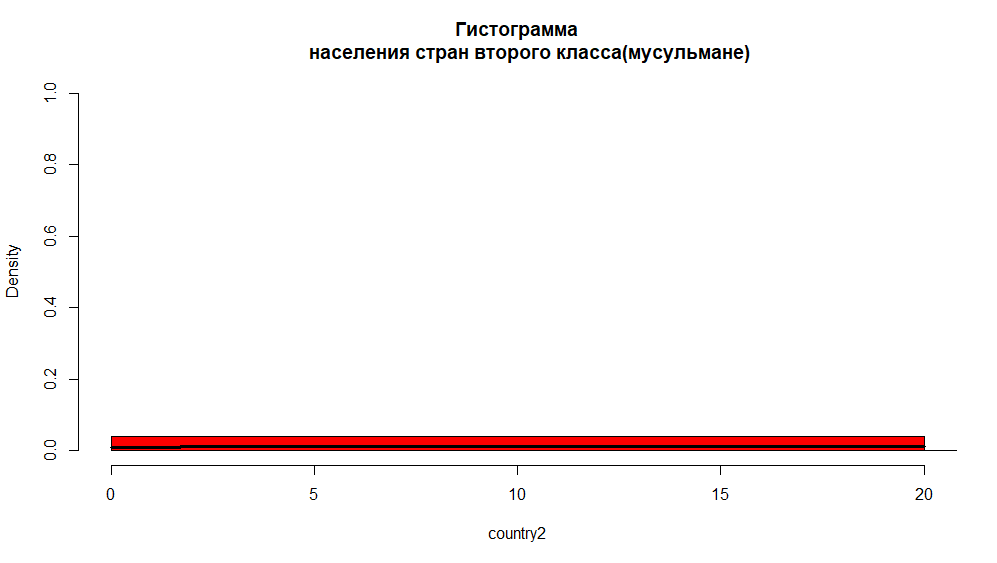
[1] 30.91768

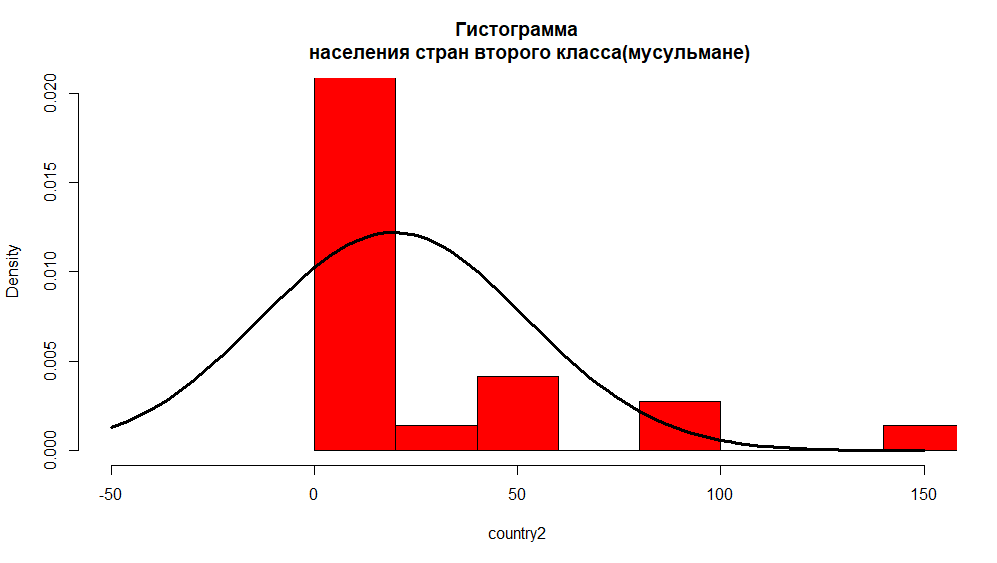
Построим гистограмму населения стран первого класса (Христиане), и совместим с графиком плотности.



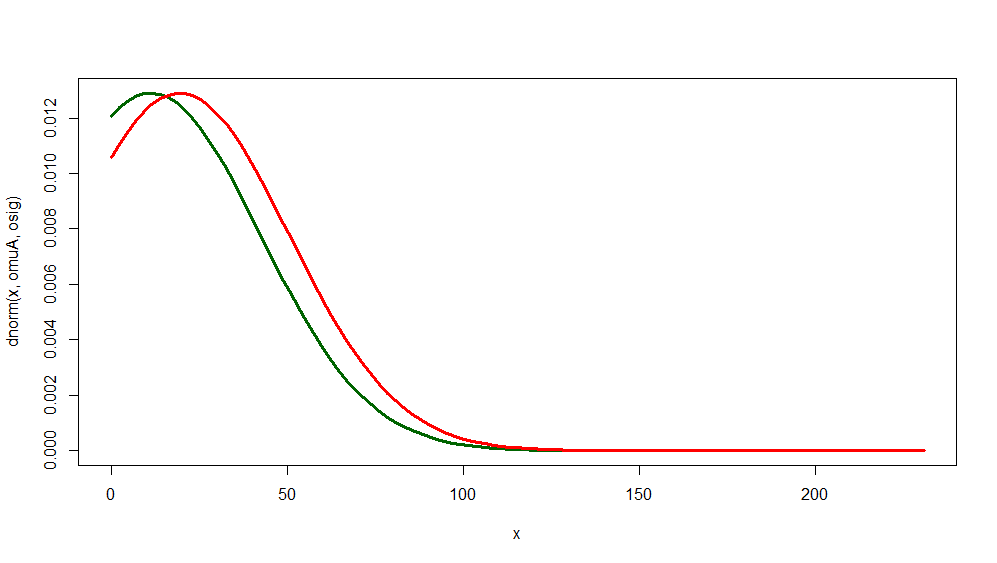


Построим гистограмму населения стран второго класса (Муссульмане), и совместим с графиком плотности.





Совместим графики плотности на одном рисунке:



Красная кривая соответствует распределению населения стран второго класса (Мусульмане), зелёная кривая - плотность распределения населения стран первого класса (Христиане). Как видно из графика, страны второго класса (Мусульмане) в среднем чуть более населены.

Разделом является абсцисса точки пересечения графиков плотностей распределения населения стран первого и второго класса (примерно 20).

Можно сделать вывод, что для стран, население которых находится в диапазоне примерно от 17 до 23, прогноз будет неточным.

Сделаем прогноз значений переменной отклика для нескольких стран из обучающей выборки.

> # Сделаем прогноз значений переменной отклика для нескольких

> # стран из обучающей выборки

> x = 119 # Brazil

> d=fdiscr(x)

> x

[1] 119

> d

[1] "Класс : 1"

> #############

> x = 157 # Indonesia

> d=fdiscr(x)

> x

[1] 157

> d

[1] "Класс : 2"

> ############

> x = 2 # Jamaica

> d=fdiscr(x)

> x

[1] 2

> d

[1] "Класс : 1"

Все предсказания оказались верными.

**Код программы:**

# Назначим рабочий директорий

setwd("C://Users/pc/Documents/Магистратура/2 семестр/R/task8 DA")

getwd()

#очищаем рабочее пространство

rm (list=ls())

# Прочтём данные о странах, принадлежащих 1 классу (христиане)

country1<- read.csv("country1.txt", sep=",", head=FALSE)

# Выделим фактор население

country1 = country1[,5]

# Прочтём данные о странах, принадлежащих 2 классу (мусульмане)

country2<- read.csv("country2.txt", sep=",", head=FALSE)

# Выделим фактор население

country2 = country2[,5]

# Проверим гипотезу о нормальности распределения выборки стран первого класса (христиане)

shapiro.test(country1)

# Проверим гипотезу о нормальности распределения выборки стран второго класса (мусульмане)

shapiro.test(country2)

# Проверим гипотезу о равенстве дисперсий

var.test(country1, country2)

# Найдём объемы выборок

nA = length(country1)

nA

nF = length(country2)

nF

n = nA + nF

n

# Априорные вероятности:

Pk = c(nA,nF)/n

Pk

# Найдём оценки параметров

omuA=mean(country1)

"Среднее выборочное значение населения стран для стран первого класса(христиане) = "

omuA

omuF=mean(country2)

"Среднее выборочное значение населения стран для стран второго класса(мусульмане) = "

omuF

osigA=sd(country1)

"Среднее квадратическое отклонение населения стран для стран первого класса(христиане) = "

osigA

osigF=sd(country2)

"Среднее квадратическое отклонение населения стран для стран второго класса(мусульмане) = "

osigF

osig = (osigA + osigF)/2

osig

hist(country1,col="green",freq=FALSE, main="Гистограмма

населения стран первого класса(христиане)", xlim=c(-50, 100), ylim=c(0, 0.018))

curve(dnorm(x,omuA,osigA),col="black",lwd=3,add=TRUE)

hist(country2,col="red",freq=FALSE, main="Гистограмма

населения стран второго класса(мусульмане)", xlim=c(-50, 150), ylim=c(0, 0.02))

curve(dnorm(x,omuF,osigF),col="black",lwd=3,add=TRUE)

# Совместим графики плотности на одном рисунке

# Для этого определим диапазон значений x, на котором будем делать построение

a=min(min(country1),min(country2))

b=max(max(country1),max(country2))

curve(dnorm(x,omuA,osig),a,b,col="darkgreen",lwd=3)

curve(dnorm(x,omuF,osig),a,b,col="red",lwd=3,add=TRUE)

# Напишем функцию для определения класса(религии) страны по населению страны

fdiscr = function (x)

{

vZ = log(Pk[1]) + x\*omuA/(osig^2) - (omuA^2)/(2\*osig^2);

vT = log(Pk[2]) + x\*omuF/(osig^2) - (omuF^2)/(2\*osig^2);

class <- ifelse(vZ > vT,"Класс : 1","Класс : 2");

return (class);

}

# Сделаем прогноз значений переменной отклика для нескольких

# стран из обучающей выборки

x = 119 # Brazil

d=fdiscr(x)

x

d

#############

x = 157 # Indonesia

d=fdiscr(x)

x

d

############

x = 2 # Jamaica

d=fdiscr(x)

x

d

###########

x = 125

d=fdiscr(x)

x

d

###########

x = 13.07

d=fdiscr(x)

x

d

###########

x = 14.21

d=fdiscr(x)

x

d