# **Исследовательский проект**

**Цель работы –** проведение исследования с применением изученных в рамках курса методов графического анализа, проверки нормальности распределения, расчёта и интерпретации описательных статистик, кластерного и корреляционного анализов на основе данных о водородном показателе ткани мозга различных людей.

# **Постановка исследовательской задачи**

В июне 2013 года группа учёных Национальной академии наук США в составе Джун Ли, Блюн Банни, Фан Мег и других опубликовала данные исследования о циркадных паттернах экспрессии генов в мозге человека и нарушениях при большом депрессивном расстройстве. Говоря доступным языком, исследователи стремились найти взаимосвязь между нарушением циклических колебаний активности мозга человека и расстройствами психи, такими как депрессия. Существует теория, что биологические ритмы людей с депрессией нарушены на клеточном уровне. Гены пациентов, страдающих от подавленного состояния, не синхронизированы с суточным циклом. Учёным удалось успешно доказать наличие взаимосвязи, идентифицировать потенциально «проблемные» гены и многое другое, что в дальнейшем позволит разработать новые биомаркеры болезни и более эффективно лечить пациентов с депрессией.

Исследователи академии собрали данные о характеристиках мозга, в том числе активности генов, людей посмертно. Информация, которую они опубликовали в открытом источнике, также может использоваться для смежных исследований.

Данная работа сконцентрирована на ответе на ключевой вопрос: влияет ли возраст человека, его пол, этнос или причина смерти, то есть ранее сопутствующие заболевания, на уровень кислотности тканей мозга? Повышенная кислотность крови и ткани мозга повышает интенсивность переживания страха, вызывает панику. Пониженная кислотность оказывает успокаивающее действие.

Были сформированы следующие гипотезы:

1. Возраст человека обратно пропорционально влияет уровень кислотности ткани мозга, то есть пожилые люди менее тревожны;
2. Женщины обладают более высоким уровнем кислотности ткани мозга, то есть более тревожны чем мужчины;
3. Этнос не влияет на уровень кислотности ткани мозга.

В заключении работы будет рассмотрены результаты исследования. Более подробно об исследовании НАН США: https://www.pnas.org/content/110/24/9950

# **Описание данных**

Набор данных содержит 54 наблюдения по следующим 5-ти переменным:

1. **pH** – водородный показатель или уровень кислотности ткани мозга, тип данных – numeric.
2. **Sex** – пол, принимает значения F или М, тип данных – factor.
3. **Ethnicity** – этнос, принимает значения AfricanAmerican, Asian, Caucasian, или PacificIslander, то есть афроамериканец, азиат, кавказец или житель тихоокеанских островов соотвественно, тип данных – factor.
4. **Age** – возраст человека на момент смерти, тип данных – numeric.
5. **DeathType** – причина смерти, принимает значения Cardiac, Other, или Suicide, то есть сердечно-сосудистый приступ, другие или суицид.

Источник информации: https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/BrainpH.html

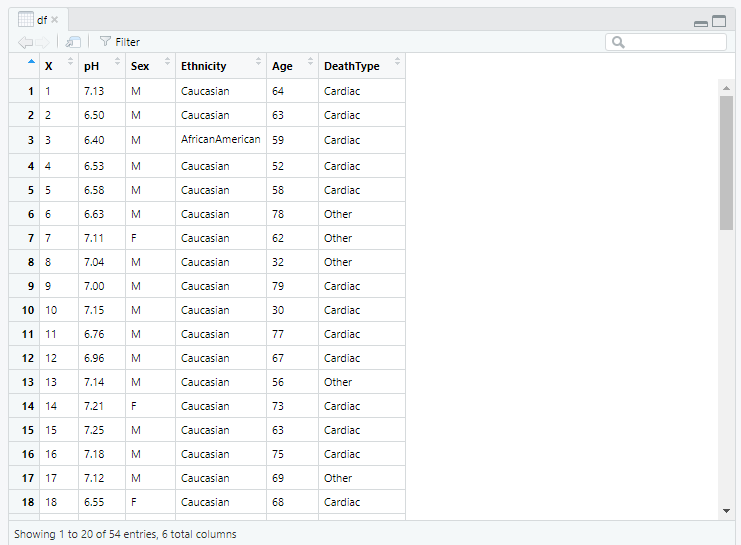
# **Предварительный анализ данных**

Набор данных содержится в файле «BrainpH.csv» в рабочей директории программы. Перед началом работы загрузим выборку в переменную df при помощи функции read.csv:

df <- read.csv("BrainpH.csv", sep = ",")

View(df)

Получим следующий результат:



Переименуем название первой колонки таблицы:

names(df)[1] <- "N"

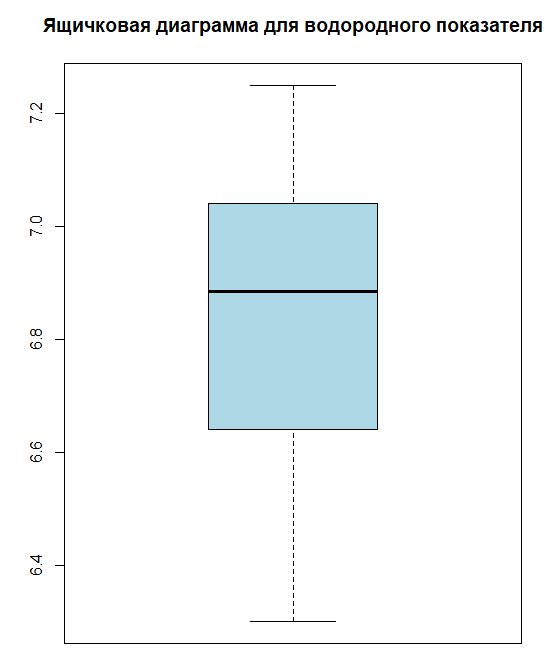
Проведём предварительный анализ данных с выявлением и исключением имеющихся выбросов, анализом описательных статистик и графическим анализом переменных.

## **Выявление аномальных значений**

На первый взгляд выборка не содержит выбросов. Проверим наличие в наборе данных аномальных значений количественных переменных при помощи построения ящичковых диаграмм, используя функцию boxplot, а также функцию vioplot для построения скрипичных диаграмм.

Переменная pH

boxplot(df$pH, col = "light blue", main = "Ящичковая диаграмма для водородного показателя")



Вывод: на основе ящичковой диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной pH.

vioplot(df$pH, names = "pH", col = "Gold")

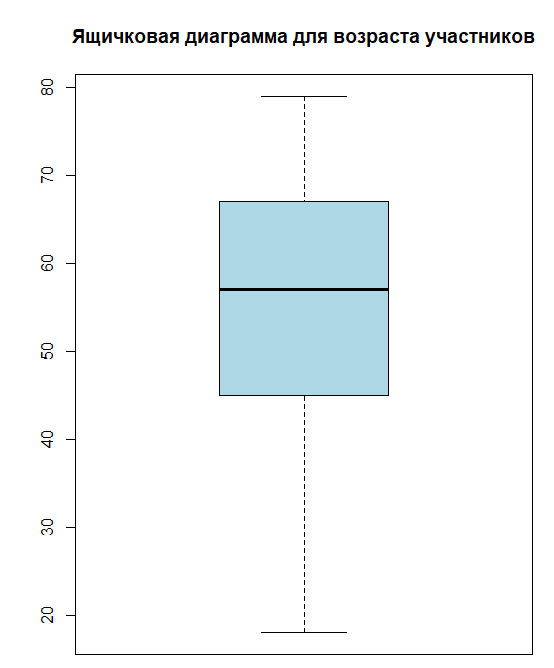
title("Скрипичная диаграмма для водородного показателя")



Вывод: на основе скрипичной диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной pH.

Переменная Age

boxplot(df$Age, col = "light blue", main = "Ящичковая диаграмма для возраста участников")



Вывод: на основе ящичковой диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной Age.

vioplot(df$Age, names = "Возраст, лет", col = "Gold")

title("Скрипичная диаграмма для возраста")

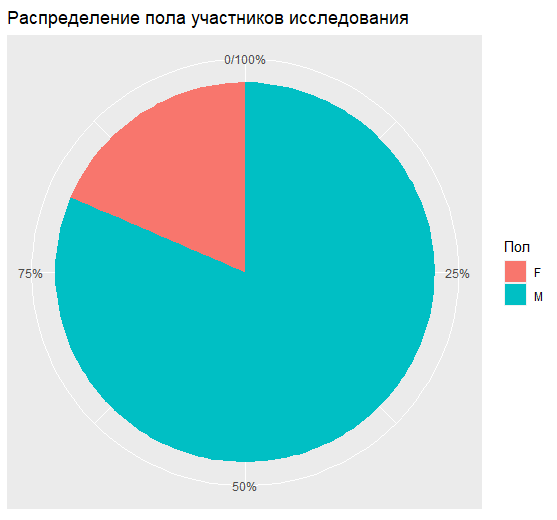


Вывод: на основе скрипичной диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной Age.

Для выявления выбросов и графического анализа качественных переменных построим соответствующие круговые диаграммы при помощи функции ggplot.

Переменная Sex

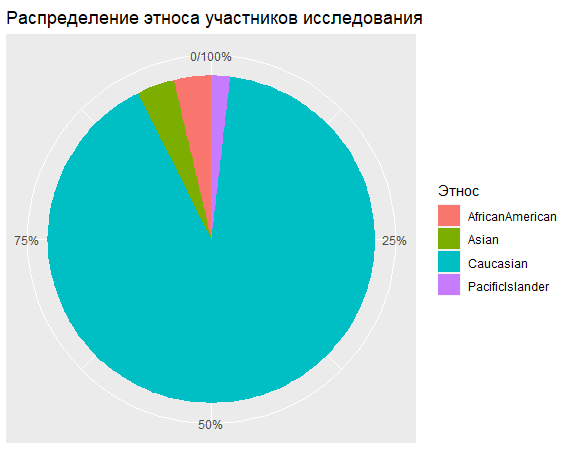
ggplot(df, aes(x=factor(1), fill=Sex)) + geom\_bar(width=1) + coord\_polar(theta="y") + labs(fill = "Пол", title = "Распределение пола участников исследования", x = "", y = "") + scale\_y\_continuous(breaks = seq(0, length(df$Sex), length(df$Sex)/4), labels = c("0", "25%", "50%", "75%", "100%")) + theme(axis.text.y = element\_blank(), axis.title.y = element\_blank(), axis.ticks.y = element\_blank(), axis.title.x = element\_blank())



Вывод: на основе круговой диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной Sex. Участников мужского пола в несколько раз больше, чем участников женского пола.

Переменная Ethnicity

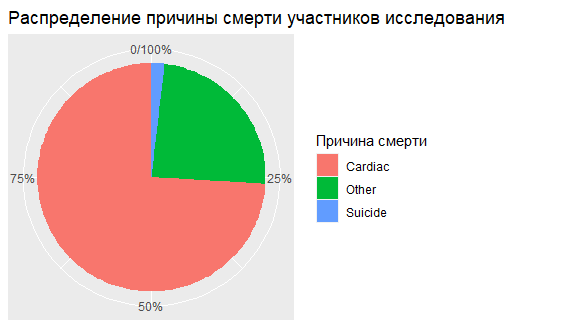
ggplot(df, aes(x=factor(1), fill=df$Ethnicity)) + geom\_bar(width=1) + coord\_polar(theta="y") + labs(fill = "Этнос", title = "Распределение этноса участников исследования") + scale\_y\_continuous(breaks = seq(0, length(df$Ethnicity), length(df$Ethnicity)/4), labels = c("0", "25%", "50%", "75%", "100%")) + theme(axis.text.y = element\_blank(), axis.title.y = element\_blank(), axis.ticks.y = element\_blank(), axis.title.x = element\_blank())



Вывод: на основе круговой диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной Ethnicity. Участников кавказского этноса намного больше, чем участников других этносов, что следует учитывать в дальнейшем исследовании.

Переменная DeathType

ggplot(df, aes(x=factor(1), fill=df$DeathType)) + geom\_bar(width=1) + coord\_polar(theta="y") + labs(fill = "Причина смерти", title = "Распределение причины смерти участников исследования") + scale\_y\_continuous(breaks = seq(0, length(df$DeathType), length(df$DeathType)/4), labels = c("0", "25%", "50%", "75%", "100%")) + theme(axis.text.y = element\_blank(), axis.title.y = element\_blank(), axis.ticks.y = element\_blank(), axis.title.x = element\_blank())



Вывод: на основе круговой диаграммы можно утверждать, что выборка не содержит нетипичных значений переменной DeathType. Участников, причиной смерти которых стал сердечно-сосудистый приступ, намного больше, чем других, что следует учитывать в дальнейшем исследовании.

**Анализ описательных статистик переменных**

Для получения описательных статистик переменных воспользуемся функцией summary для определения максимального, минимального и среднего значения, медианы и квантилей, kurtosis и skewness для расчёта коэффициента эксцесса и асимметрии соответственно, var и sd – дисперсии и стандартного отклонения.

Для нахождения моды распределения использована следующая пользовательская функция:

getmode <- function(v) {

+ uniqv <- unique(v)

+ uniqv[which.max(tabulate(match(v, uniqv)))]

+ }

Переменная pH

summary(df$pH)

Min. 1st Qu. Median Mean

6.300 6.650 6.885 6.866

3rd Qu. Max.

7.037 7.250

cat(kurtosis(df$pH, na.rm = TRUE))

-0.8619844

cat(skewness(df$pH, na.rm = TRUE))

-0.3367862

cat(sd(df$pH) / mean(df$pH) \* 100)

3.390993

cat(getmode(df$pH))

6.76

cat(var(df$pH))

0.05421244

cat(sd(df$pH))

0.2328356

Вывод: минимальное значение водородного показателя ткани мозга – 6.3 pH, максимальное – 7.25 pH, среднее значение показателя – 6.866 pH, медиана распределения –6.885 pH, то есть вероятность получить значение показателя больше 6.885 равна вероятности получить значение возраста меньше 6.885, нет отсутствующих записей о значении уровня кислотности, с вероятностью 25% значение показателя меньше либо равно 6.65 и с вероятностью 75% значение показателя меньше либо равно 7.037. Дисперсия распределения, мера разброса значений случайной величины относительно её математического ожидания – 0.05421244, стандартное отклонение, то есть показатель рассеивания значений случайной величины относительно её математического ожидания – 0.2328356. График распределения переменной имеет пологую вершину, так коэффициент эксцесса меньше нуля, и левостороннюю асимметрию, так как коэффициент асимметрии меньше нуля. Коэффициент вариации меньше 33%, следовательно переменная pH является однородной. Наиболее часто встречающееся значение – 6.76 pH.

Переменная Age

summary(df$Age)

Min. 1st Qu. Median Mean

18.00 45.00 57.00 55.87

3rd Qu. Max.

66.75 79.00

cat(kurtosis(df$Age, na.rm = TRUE))

-0.5112347

cat(skewness(df$Age, na.rm = TRUE))

-0.3930908

cat(sd(df$Age) / mean(df$Age) \* 100)

25.18084

cat(getmode(df$Age))

63

cat(var(df$Age))

197.9263

cat(sd(df$Age))

14.06863

Вывод: самому молодому участнику исследования было 18 лет, самому пожилому – 79 лет, среднее значение возраста – 55.87 лет, медиана распределения – 57 лет, то есть вероятность получить значение возраста больше 57 равна вероятности получить значение возраста меньше 57, нет отсутствующих записей о возрасте участников, с вероятностью 25% возраст участников меньше либо равен 45 и с вероятностью 75% возраст меньше либо равен 66.75. Дисперсия распределения – 197.9263, стандартное отклонение – 14.06863. График распределения переменной имеет пологую вершину, так коэффициент эксцесса меньше нуля, и левостороннюю асимметрию, так как коэффициент асимметрии меньше нуля. Коэффициент вариации меньше 33%, следовательно переменная Age является однородной. Наиболее часто встречающееся значение – 63 года.

Переменная Sex

summary(df$Sex)

F M

10 44

getmode(df$Sex)

[1] M

Levels: F M

Вывод: в исследовании принимало участие 10 женщин и 44 мужчины. Мода распределения – M.

Переменная Entnicity

summary(df$Ethnicity)

AfricanAmerican Asian

2 2

Caucasian PacificIslander

49 1

getmode(df$Ethnicity)

[1] Caucasian

4 Levels: AfricanAmerican ... PacificIslander

Вывод: в исследовании принимало участие 2 афро-американца, 2 азиата, 49 кавказцев и 1 островитянин, из чего следует, что нет достаточных данных для сравнения показателей участников разных этносов, включая жителей Тихоокеанских островов, и построения соответствующих гипотез. Мода распределения – Caucasian.

Переменная DeathType

summary(df$DeathType)

Cardiac Other Suicide

40 13 1

getmode(df$DeathType)

[1] Cardiac

Levels: Cardiac Other Suicide

Вывод: причиной смерти 40 участников послужил сердечный приступ, 1 участника – суицид, 13 участников – другие причины. Стоит отметить, что в выборке недостаточно данных об участниках, причина смерти которых – суицид, следовательно, есть возможность проверить гипотезы на основе причины смерти, только исключая суицид. Мода распределения – Cardiac.

**Проверка нормальности распределения переменных**

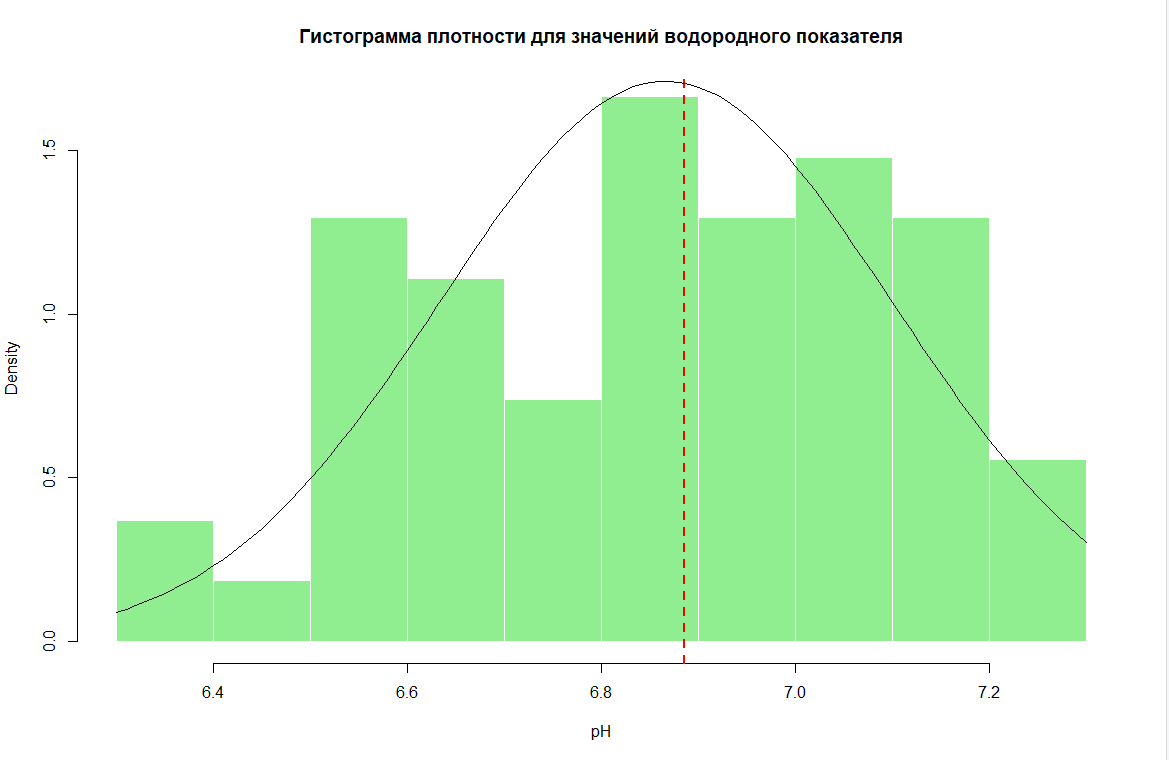
Для оценки нормальности распределения применяются функции hist, abline и curve для построения гистограммы, функции qqnorm и qqline – для построения QQ-Plot. Также для проверки используются критерии Шапиро-Уилка, так как он наиболее подходящий для небольших выборок, Адерсона-Дарлинга и ХИ-квадрат Пирсона, так они считаются достаточно «мощными» критериями.

Переменная pH

hist(df$pH, col = 'light green', border = 'white', main = "Гистограмма плотности для значений водородного показателя", xlab = "pH", freq = FALSE)

abline(v = median(df$pH), col = "red", lwd = 2, lty = 2)

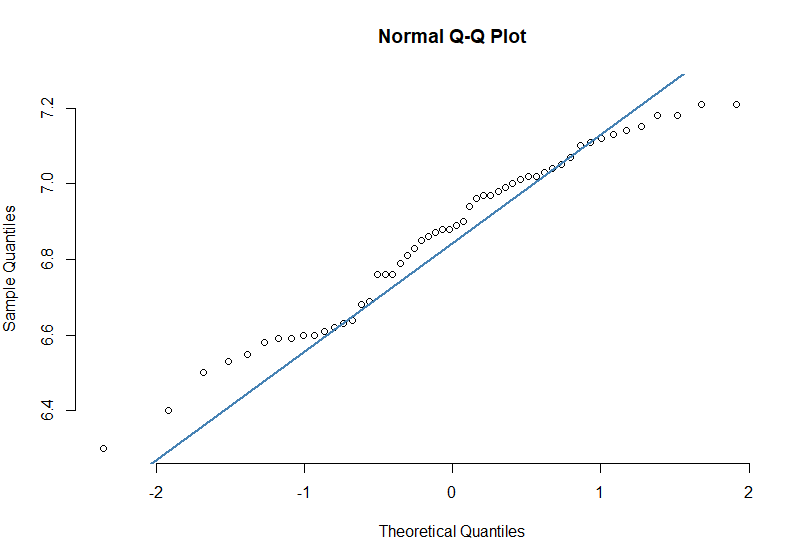
curve(dnorm(x, mean(df$pH), sd = sd(df$pH)), add = TRUE)



Вывод: на основании графика нельзя утверждать, что распределения является нормальным.

qqnorm(df$pH, pch = 1, frame = FALSE)

qqline(df$pH, col = "steelblue", lwd = 2)



Вывод: на графе показано. что большинство значений эмпирических квантилей схожи со значениями теоретических квантилей, следовательно, можно утверждать, что распределение переменной pH является нормальным.

shapiro.test(df$pH)

Shapiro-Wilk normality test

data: df$pH

W = 0.96331, p-value = 0.09696

ad.test(df$pH)

Anderson-Darling normality test

data: df$pH

A = 0.65623, p-value = 0.08213

pearson.test(df$pH)

Pearson chi-square normality test

data: df$pH

P = 7.1111, p-value = 0.4174

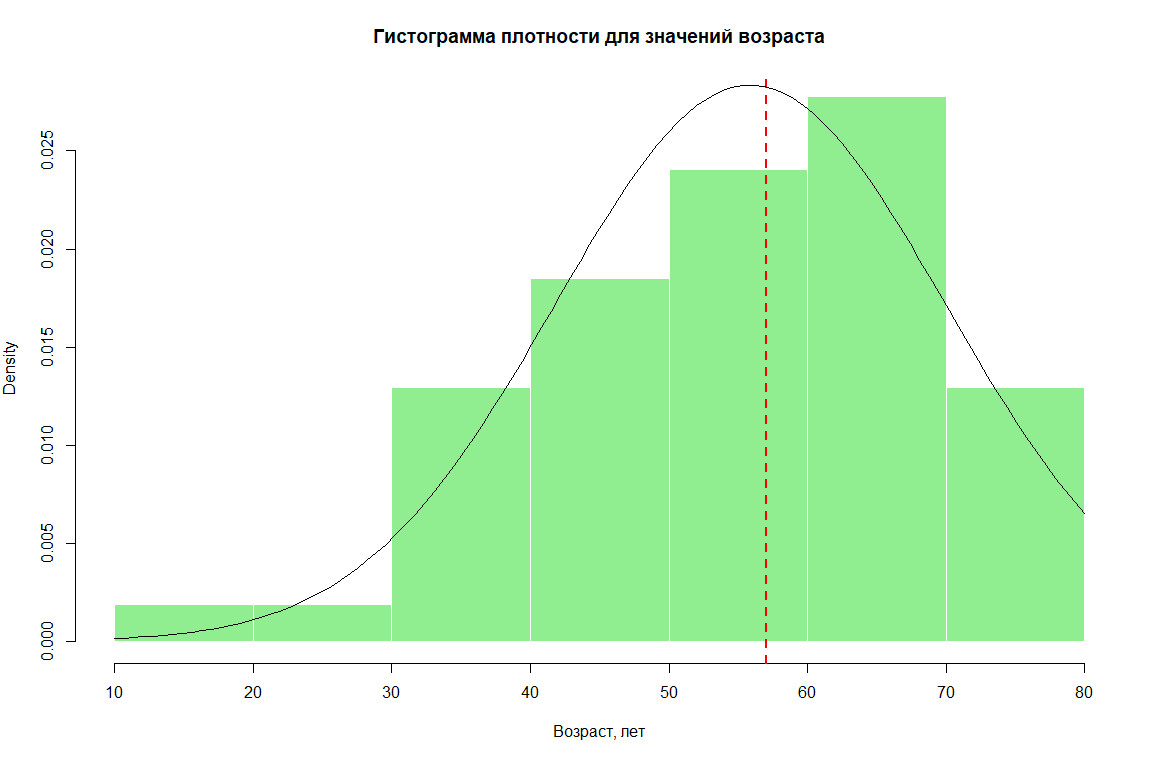
Вывод: p-value > 0.05 в трёх проведённых тестах – нулевая гипотеза не отвергается, значит, анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение.

Переменная Age

hist(df$Age, col = 'light green', border = 'white', main = "Гистограмма плотности для значений возраста", xlab = "Возраст, лет", freq = FALSE)

> abline(v = median(df$Age), col = "red", lwd = 2, lty = 2)

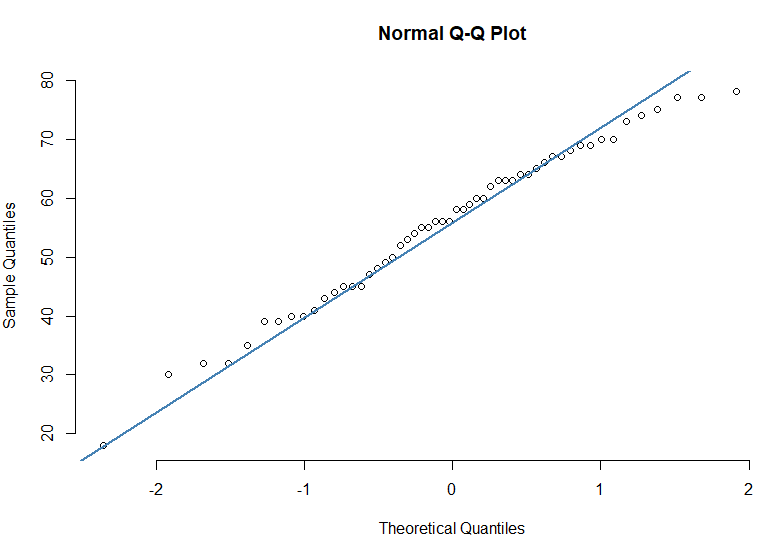
> curve(dnorm(x, mean(df$Age), sd = sd(df$Age)), add = TRUE)



Вывод: на основании графика можно утверждать, что распределения является нормальным.

qqnorm(df$Age, pch = 1, frame = FALSE)

qqline(df$Age, col = "steelblue", lwd = 2)



Вывод: на графе показано. что большинство значений эмпирических квантилей схожи со значениями теоретических квантилей, следовательно, можно утверждать, что распределение переменной Age является нормальным.

shapiro.test(df$Age)

Shapiro-Wilk normality test

data: df$Age

W = 0.97421, p-value = 0.2934

ad.test(df$Age)

Anderson-Darling normality test

data: df$Age

A = 0.35514, p-value = 0.4478

pearson.test(df$Age)

Pearson chi-square normality test

data: df$Age

P = 2.2963, p-value = 0.9416

Вывод: p-value > 0.05 в трёх проведённых тестах – нулевая гипотеза не отвергается, значит, анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение.

На основании проведённого анализа можно утверждать, что данные имеют нормальное распределение, следовательно, допустимо применение статистических методов, которые исходят из предположения нормальности распределения, например, критерия Стьюдента для проверки гипотезы о равенстве средних значений в двух независимых выборках.

# **Кластерный анализ**

Перед кластеризацией отберём из выборки те столбцы, который необходимы для анализа: pH, Sex и Age, так как они наиболее информативны.

data <- df[2:5]

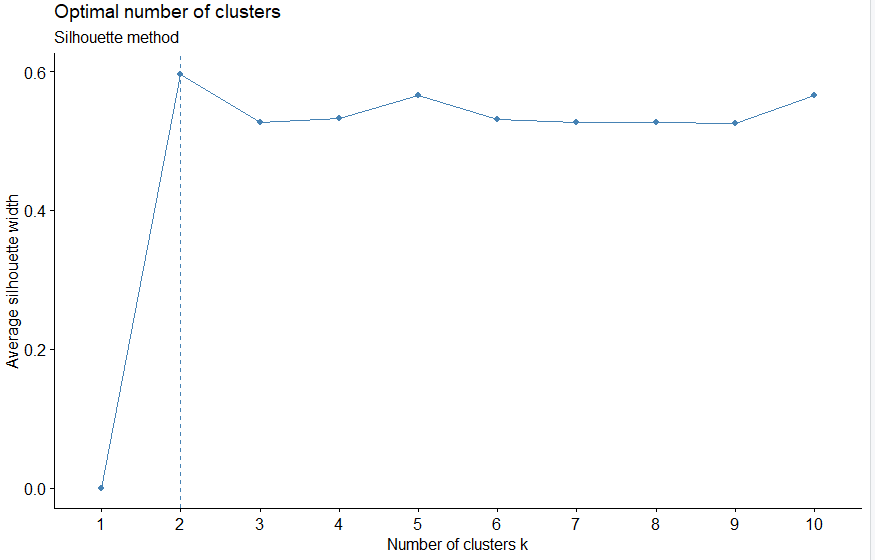
data <- data[, -3]

Преобразуем столбец Sex к количественному типу, 1 – участник женщина, 0 – мужчина.

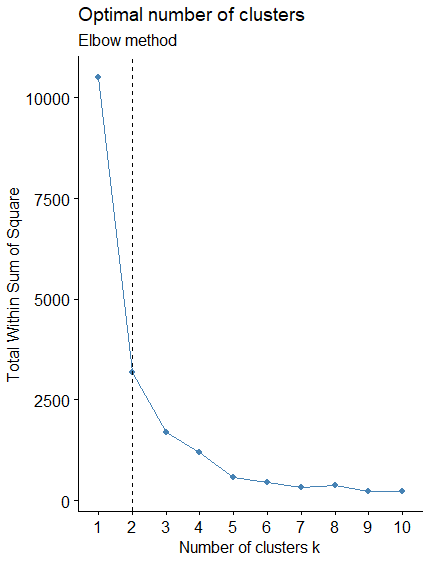
data$Sex <- as.integer(ifelse(data$Sex == 'F', 1, 0))

Используем «метод согнутого колена» и «силуэтным метод» для определения оптимального количества кластеров.

fviz\_nbclust(data, kmeans, method = "silhouette") + labs(subtitle = "Silhouette method")



fviz\_nbclust(data, kmeans, method = "wss") + labs(subtitle = "Elbow method") + geom\_vline(xintercept = 2, linetype = 2)



Вывод: на основе построенных графиков можно утверждать, что оптимальным будет распределение данных на два кластера.

Проведём кластеризацию данных методом k-mean при помощи функции kmeans:

cl <- kmeans(data,2)

cl

K-means clustering with 2 clusters of sizes 20, 34

Cluster means:

pH Sex Age

1 6.847000 0.1500000 40.70000

2 6.877647 0.2058824 64.79412

Clustering vector:

[1] 2 2 2 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1

[27] 2 1 1 2 2 1 1 2 2 1 1 2 1 1 1 2 2 2 2 1 1 1 1 2 2 2

[53] 1 1

Within cluster sum of squares by cluster:

[1] 1271.691 1919.038

(between\_SS / total\_SS = 69.6 %)

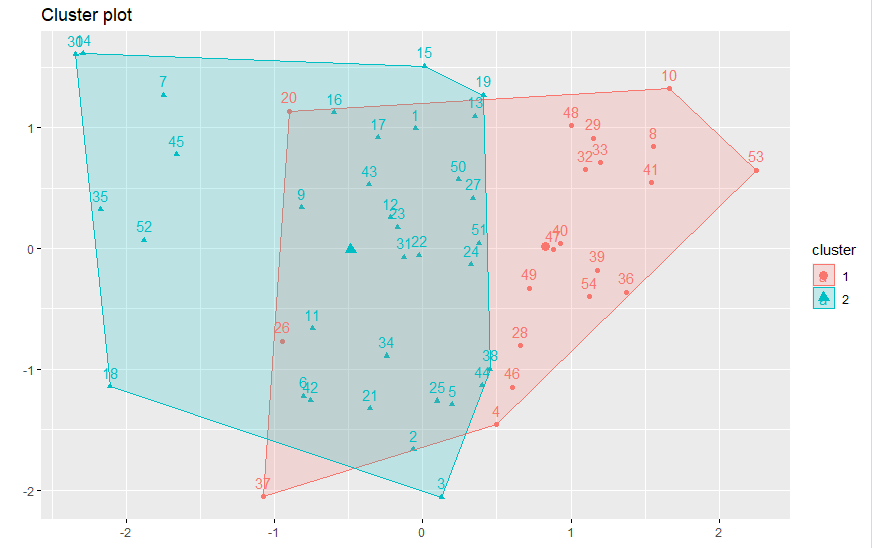
Available components:

[1] "cluster" "centers" "totss"

[4] "withinss" "tot.withinss" "betweenss"

[7] "size" "iter" "ifault"

Так как метод k-средних является не иерархическим методом кластеризации, а итерационным, то для визуализации результатов воспользуемся не дендрограммой, а другим графиком:



Проверим гипотезы о нормальном распределении, равенстве дисперсий и среднего значения показателя pH (уровень кислотности) в разных кластерах. Для проверки нормальности используется критерий Шапиро-Уилка, так как он наиболее подходящий для небольших выборок. Для проверки равенства дисперсий и среднего используется критерий Фишера и Стьюдента соответственно.

Добавим столбец с номером кластера в таблицу:

test <- matrix(unlist(cl["cluster"], F))

data <- data %>% mutate(groups2 = factor(test))

Критерий Шапиро-Уилка:

shapiro.test(subset(data$pH, groups2 == 1))

Shapiro-Wilk normality test

data: subset(data$pH, groups2 == 1)

W = 0.9396, p-value = 0.2356

shapiro.test(subset(data$pH, groups2 == 2))

Shapiro-Wilk normality test

data: subset(data$pH, groups2 == 2)

W = 0.94549, p-value = 0.08994

Вывод: p-value > 0.05 в обоих кластерах – нулевая гипотеза не отвергается, значит, параметр pH происходит из генеральной совокупности, имеющей нормальное распределение.

Сравним дисперсии при помощи функции var.test:

var.test(data$pH[groups2 == 1], data$pH[groups2 == 2])

F test to compare two variances

data: data$pH[groups2 == 1] and data$pH[groups2 == 2]

F = 0.8508, num df = 19, denom df = 33, p-value =

0.7241

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

95 percent confidence interval:

0.3926754 2.0180085

sample estimates:

ratio of variances

0.8507996

Вывод: p-value > 0.05 – нулевая гипотеза не отвергается, значит, параметр pH имеет равные дисперсии в обоих выборках.

Сравним среднее в двух выборках, учитывая, что дисперсии равны:

with(data, t.test (data$pH[groups2 == 1], data$pH[groups2 == 2], var.equal = TRUE))

Two Sample t-test

data: data$pH[groups2 == 1] and data$pH[groups2 == 2]

t = -0.46361, df = 52, p-value = 0.6449

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1632959 0.1020018

sample estimates:

mean of x mean of y

6.847000 6.877647

Вывод: p-value > 0.05 – нулевая гипотеза не отвергается, значит, параметр pH имеет равное среднее в обоих выборках.

Можно сделать вывод о том, что распределение параметра pH в двух выборках однородно.

# **Корреляционный анализ**

Построим в каждом кластере корреляционные поля и рассчитаем парные и частные коэффициенты корреляции, проверим гипотезу о значимости коэффициентов корреляции на 5% уровне.

1-ый кластер:

Построим корреляционную матрицу для переменных выборки:

cor(subset(data, groups2 == 1)[1:3])

pH Sex Age

pH 1.0000000 -0.3815611 -0.5549191

Sex -0.3815611 1.0000000 0.2620127

Age -0.5549191 0.2620127 1.0000000

Получим вектор коэффициентов корреляции по модулю:

coeffs <- abs(cor(data[1:3]))

Зададим цвета (автоматическое разбиение по вектору коэффициентов):

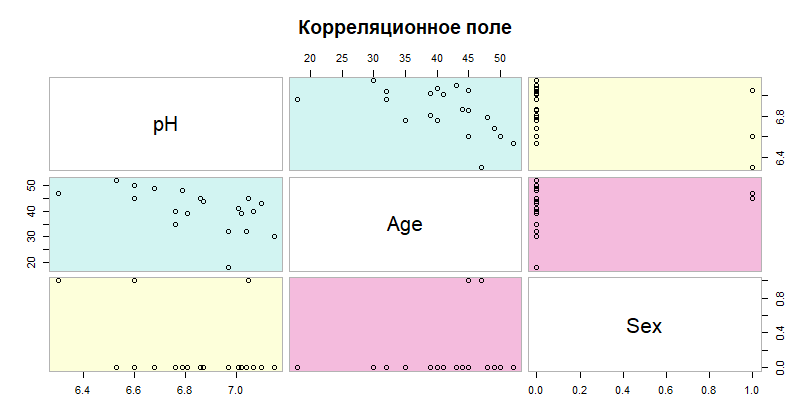
colors <- dmat.color(coeffs)

Отсортируем так, чтобы графики, где связь переменных наибольшая, были ближе к диагонали:

order <- order.single(coeffs)

Построим график:

cpairs(subset(data, groups2 == 1)[1:3], order, panel.colors = colors, gap = .5, main = "Корреляционное поле (1-ый кластер)" )



Вывод: на основе корреляционной матрицы можно сделать вывод о том, что между переменными pH и Age существует слабая обратная взаимосвязь, то есть чем старше участник исследования, тем меньше водородный показатель тканей его мозга. Между другими переменными не существует сильных прямых или обратных взаимосвязей, так коэффициенты матрицы меньше 0.5 и больше – 0.5 соответственно. Пол участника не влияет на уровень кислотности тканей мозга.

Проверим гипотезу о значимости парных коэффициентов корреляции при помощи функции rcorr, которая возвращает массив коэффициентов корреляции и p-value для коэффициентов.

rcorr(as.matrix(subset(data, groups2 == 1)[1:3]), type = 'pearson')

pH Sex Age

pH 1.00 -0.38 -0.55

Sex -0.38 1.00 0.26

Age -0.55 0.26 1.00

n= 20

P

pH Sex Age

pH 0.0969 0.0111

Sex 0.0969 0.2644

Age 0.0111 0.2644

Вывод: P-value > 0.05 для коэффициентов r(pH, Sex) и r(Sex, Age)– нулевая гипотеза не отвергается, значит, коэффициенты корреляции незначимы. P-value < 0.05 для r(pH, Age) – нулевая гипотеза не принимается, значит, коэффициент корреляции значим.

Проверим гипотезу о значимости частных коэффициентов корреляции при помощи функции pcorr, которая возвращает массив коэффициентов корреляции и p-value для коэффициентов.

pcor(subset(data, groups2 == 1)[1:3])

$estimate

pH Sex Age

pH 1.0000000 -0.29416187 -0.50999904

Sex -0.2941619 1.00000000 0.06538282

Age -0.5099990 0.06538282 1.00000000

$p.value

pH Sex Age

pH 0.0000000 0.2215389 0.0256941

Sex 0.2215389 0.0000000 0.7902907

Age 0.0256941 0.7902907 0.0000000

$statistic

pH Sex Age

pH 0.000000 -1.2690069 -2.4445956

Sex -1.269007 0.0000000 0.2701583

Age -2.444596 0.2701583 0.0000000

Вывод: P-value > 0.05 для коэффициентов r(pH, Sex) и r(Sex, Age)– нулевая гипотеза не отвергается, значит, коэффициенты корреляции незначимы. P-value < 0.05 для r(pH, Age) – нулевая гипотеза не принимается, значит, коэффициент корреляции значим.

2-ый кластер:

Построим корреляционную матрицу для переменных выборки:

cor(subset(data, groups2 == 1)[1:3])

pH Sex Age

pH 1.00000000 0.2009111 0.09528571

Sex 0.20091115 1.0000000 0.16919616

Age 0.09528571 0.1691962 1.00000000

Получим вектор коэффициентов корреляции по модулю:

coeffs <- abs(cor(data[1:3]))

Зададим цвета (автоматическое разбиение по вектору коэффициентов):

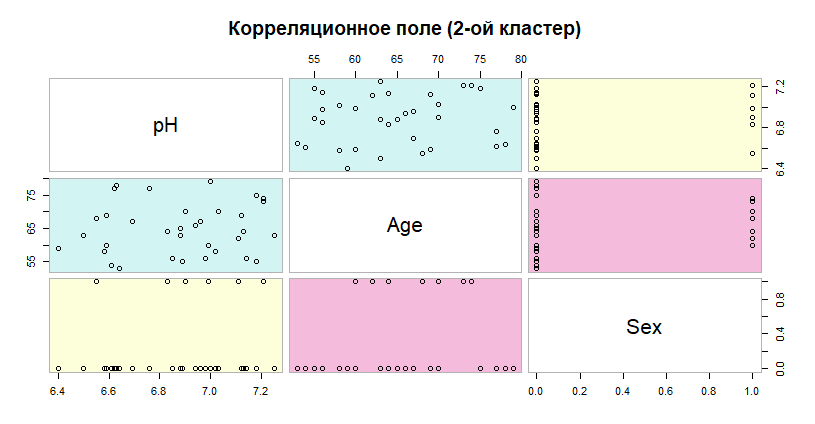
colors <- dmat.color(coeffs)

Отсортируем так, чтобы графики, где связь переменных наибольшая, были ближе к диагонали:

order <- order.single(coeffs)

Построим график:

cpairs(subset(data, groups2 == 2)[1:3], order, panel.colors = colors, gap = .5, main = "Корреляционное поле (1-ой кластер)" )



На основе корреляционной матрицы можно сделать вывод о том, что между другими переменными не существует сильных прямых или обратных взаимосвязей, так коэффициенты матрицы меньше 0.5 и больше – 0.5 соответственно.

Проверим гипотезу о значимости парных коэффициентов корреляции при помощи функции rcorr, которая возвращает массив коэффициентов корреляции и p-value для коэффициентов.

rcorr(as.matrix(subset(data, groups2 == 2)[1:3]), type = 'pearson')

pH Sex Age

pH 1.0 0.20 0.10

Sex 0.2 1.00 0.17

Age 0.1 0.17 1.00

n= 34

P

pH Sex Age

pH 0.2546 0.5919

Sex 0.2546 0.3388

Age 0.5919 0.3388

Вывод: P-value > 0.05 для всех коэффициентов – нулевая гипотеза не отвергается, значит, коэффициенты корреляции незначимы.

Проверим гипотезу о значимости частных коэффициентов корреляции при помощи функции pcorr, которая возвращает массив коэффициентов корреляции и p-value для коэффициентов.

pcor(subset(data, groups2 == 2)[1:3])

$estimate

pH Sex Age

pH 1.00000000 0.1883494 0.06348339

Sex 0.18834936 1.0000000 0.15387567

Age 0.06348339 0.1538757 1.00000000

$p.value

pH Sex Age

pH 0.0000000 0.2938576 0.7256056

Sex 0.2938576 0.0000000 0.3925658

Age 0.7256056 0.3925658 0.0000000

$statistic

pH Sex Age

pH 0.0000000 1.0677962 0.3541749

Sex 1.0677962 0.0000000 0.8670701

Age 0.3541749 0.8670701 0.0000000

Вывод: P-value > 0.05 для всех коэффициентов – нулевая гипотеза не отвергается, значит, коэффициенты корреляции незначимы.

На основе результатов корреляционного анализа можно сделать вывод о том, что на уровень кислотности ткани мозга обратно пропорционально влияет возраст участника исследования.

# **Заключение**

По результатам исследования можно сделать вывод, что на уровень кислотности ткани мозга влияет возраст человека – чем старше человек, тем меньше водородный показатель, то есть человек менее тревожен. Пол человека не влияет на уровень кислотности. Таким образом, первая гипотеза, выдвинутая в постановке исследовательской задачи, не отвергается, а вторая – не принимается. Гипотезы были оценены на основании парных и частных коэффициентов корреляции, то есть тестноты взаимосвязи соответствующих переменных. Для оценки третьей гипотезы о влиянии этноса недостаточно данных, так как в выборке представлены преимущественно записи о людях кавказского этноса.

Стоит отметить, что выборка была маленькой и результаты могут существенно отличаться для большего набора данных. Достигнутое решение сильно ограничено и нет уверенности, что гипотезы будут применятся на практике, так как не были приняты во внимание многие важные факторы, которые влияют на кислотность ткани мозга, такие как история болезни, наследственность, принимаемые до момента смерти лекарства и препараты. Также нет данных о людях младше 18 и старше 90 лет.

Однако именно эти недостатки могут послужить дальнейшему развитию исследования. Была обработана лишь часть информации, предложенной учёными НАН США. Таким образом, можно ввести новые характеристики человека и факторы влияния на водородный показатель и, возможно, найти другие зависимости, или добавить информации о генах в мозге человека и исследовать, к примеру, как гены изменяются с возрастом.