КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ФУНДАМЕНТАЛЬНОЙ МЕДИЦИНЫ И БИОЛОГИ
Кафедра биохимии и биотехнологии

Н.И.АКБЕРОВА

Краткое введение в R и RStudio

Учебно-методическое пособие

Казань – 2014

Целью этих практических занятий является познакомить вас с R и RStudio, которые вы будете использовать на протяжении всего курса и изучить статистические понятия, рассмотренные в лекциях, а также научиться анализировать реальные данные и делать обоснованные выводы.

R - это название языка программирования и RStudio является удобный интерфейс.

Сегодня мы начинаем с фундаментальных строительных блоков R и RStudio: интерфейс, чтения данных и основных команд.

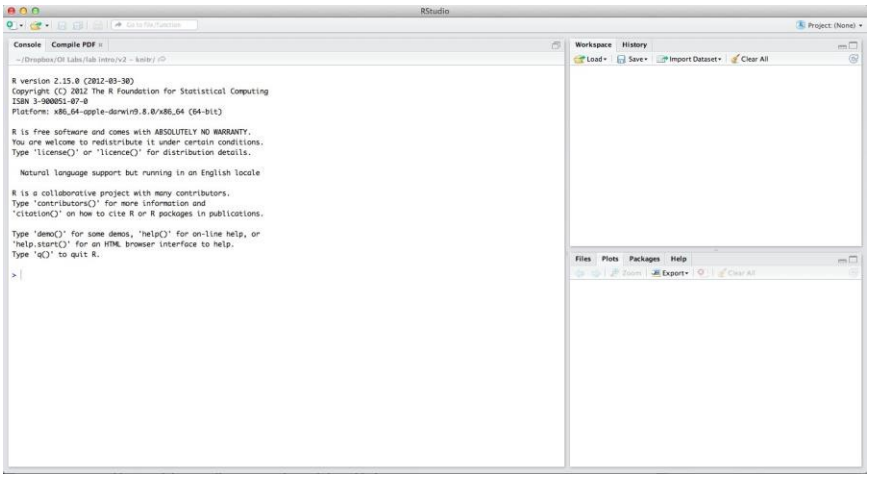


Рис.1. Интерфейс программы RStudio

Панель в правом верхнем углу содержит рабочее пространство, а также историю команд, введенных ранее. Любые графики, которые вы генерируете, будут отображаться в панели в правом нижнем углу. Панель слева называется консоль. Каждый раз вы запустите RStudio, она будет иметь тот же текст в верхней части консоли, говоря вам версию R, с которой вы работаете. Ниже находится приглашение (командная строка). Как следует из названия, это, по сути, запрос команды, где происходит взаимодействие с R. Эти команды и их синтаксис развивались в течение десятилетий (буквально) и в настоящее время обеспечивают то, что многие пользователи считают, довольно естественный

способ доступа к данным, их организации, описания и вызова статистических выкладок.

Для начала введите в командную строку R (т.е. сразу после > в консоли). Можно сделать ввод вручную или скопируйте и вставьте из этого документа:

```
source("http://www.openintro.org/stat/data/present.R")
```

Данные сохраняются во фрейме данных под названием present.

Эта команда инструктирует R получить доступ к сайту OpenIntro и взять данные: количество мальчиков и девочек, родившихся в США в разные годы. Вы должны видеть, что рабочая зона в правом верхнем углу окна RStudio теперь отображает набор данных с именем present, в котором 63 наблюдения на 3 переменных. При взаимодействии с R вы будете создавать серию объектов. Иногда вы загрузите их, как сейчас, а иногда вы будете создавать их сами, как продукт вычислений или некоторого выполненного анализа. Обратите внимание, что вы обращаетесь к данным из Интернета, эта команда (и весь анализ) будет работать везде, где есть доступ к сети Интернет.

Данные Present

Настоящий массив данных содержит количества рождений в США мальчиков и девочек с 1940 по 2002 год. Мы можем взглянуть на данные, введя в консоль:

```
present
```

Вы должны увидеть четыре колонки чисел, каждая строка которых представляет определенный год: первый элемент в каждой строке - номер строки (им можно

воспользоваться, чтобы получить доступ к данным из отдельных лет), вторым является год, а третий и четвертый - количества рожденных в этом году мальчиков и девочек, соответственно. Воспользуйтесь полосой прокрутки на правой стороне окна консоли для изучения полного набора данных.

Обратите внимание, что номера строк в первом столбце не являются частью настоящего массива данных. R добавляет их, чтобы помочь вам сделать визуальное сравнение. Можно рассматривать их как индекс. На самом деле, обычно сравнение с электронной таблицей может быть полезным. R сохранил данные в своего рода таблицу, называемую фреймом (a *data frame)*.

Вы можете узнать размерность фрейма, введя:

```
dim(present)
```

Эта команда должна вывести [1] 63, указывая, что в фрейме есть 63 строки и 3 колонки. Вы можете увидеть имена этих столбцов (или переменных), введя:

```
names(present)
```

Вопрос 1 [выбрать ответ] Сколько переменных в этом фрейме?

- 2
- 3
- 4
- 63
- 2002

Задание Какие годы включены в этот фрейм? *Подсказка:* Для ответа посмотрите переменную year.

Вы видите, что фрейм данных содержит колонки year, boys и girls. На этом этапе вы можете заметить, что команды в R похожи на математические функции, и вызов R команд означает выполнение функции с некоторым числом аргументов. Команды dim и names , например, имеют единственный аргумент - название фрейма данных.

Одним из преимуществ RStudio является встроенный «просмотрщик» (вьюер) данных. Нажмите на название **present** в правом верхнем окне, где перечислены объекты в рабочей области. После этого появится альтернативный показ граф в верхнем левом окне. Вы можете закрыть просмотр данных, нажав на «x» в верхнем левом углу.

Небольшое исследование

Давайте начнем изучать данные более внимательно. Мы можем получить отдельно доступ к данным в одной колонке фрейма данных, используя команду:

```
present$boys
```

Эта команда показывает количество только мальчиков по годам.

Вопрос 2 [выбрать ответ] Какую команду нужно использовать, чтобы в результате получать количество девочек по годам ?

(a) present$boys
(b) present$girls
(c) girls
(d) present[girls]
(e) $girls

Обратите внимание, что R напечатал эти данные в другом формате. Когда мы смотрели на полный фрейм данных, мы увидели 63 строк, по одному значению

в каждой строке. Теперь данные больше не структурированы в виде таблицы с другими переменными, поэтому они отображаются один за другим. Объекты, которые печатаются в этом случае, называются векторами; они представляют собой набор цифр. R добавил числа в скобках [ ] в левой части распечатки, чтобы указать местоположение в пределах вектора. Например, за [1] следует 1211684, что указывает на то, что 1211684 является первым элементом в векторе. И если [43] начинает строчку, то это означает - первое число в этой строке будет представлять 43-ю запись в векторе.

R имеет несколько мощных функций для создания графики. Мы можем создать простой график количества ежегодно рождаемых девочек с помощью команды:

```
plot(x = present$year, y = present$girls)
```

По умолчанию R создает точечный график (a scatterplot), рисуя кружок для каждой пары x, y. График сам по себе должен появиться на вкладке "Графики "(“Plots”) в нижней правой панели RStudio. Обратите внимание, что команда снова выглядит как функция, на этот раз с двумя аргументами, разделенными запятой. Первый аргумент в функции графика задает переменную для оси абсцисс x, а второй - на оси ординат y. Если мы хотим, чтобы точки были соединены в линию, мы можем добавить третий аргумент, буква " l " для линии:

```
plot(x = present$year, y = present$girls, type = "l")
```

Вопрос 3 [выбрать ответ] Существует ли наметившаяся тенденция в количестве девочек, рожденных за эти годы? Как бы вы ее описали?

(a) Нет никакой тенденции в количестве девочек, рожденных с 1940 по 2002.
(b) Существует первоначальное увеличение кличества рожденных девочек, пик которого приходится к 1960 г. После 1960 г. наблюдается снижение числа рожденных девочек, но оно снова начинает расти в начале 1970-х. Общей тенденцией является увеличение числа девочек, рожденных в США начиная с 1940-х годов.
(c) Существует первоначальное увеличение числа рожденных девочек, которое достигает максимума к 1960 г., а затем после 1960 года количество рожденных девочек уменьшается.
(d) число рожденных девочек уменьшилось с течением времени.
(e) Существует первоначальное увеличение числа рожденных девочек, которое осталось на уровне 1960 г. и не изменялось с тех пор.

Как же узнать, что можно было добавить третий аргумент. Для удобства пользователей все функции в R хорошо задокументированы. Чтобы прочитать, что делает функция и узнать аргументы, которые доступны, просто наберите в вопросительный знак, за которым следует имя функции. Например, попробуйте следующее:

```
?plot
```

Обратите внимание, что файл справки заменяет график в нижней правой панели. Вы можете переключаться между графиком и файлом помощи с использованием вкладок в верхней части этой панели.

Теперь предположим, что мы хотим построить график общего количества рождений. Чтобы вычислить это, мы могли бы использовать R как калькулятор. Мы можем ввести математическое выражение

```
1211684 + 1148715
```

чтобы увидеть общее количество рождений в 1940 году. Мы могли бы повторить это каждый раз для каждого года, но есть более быстрый способ. Если мы сложим вектора рождения для мальчиков и девочек, R вычисляет все суммы одновременно:

```
present$boys + present$girls
```

Вы увидите 63 числа, каждый из которых представляет собой сумму количества мальчиков и девочек, рожденных в каждый год. Можете проверить некоторые из них и убедиться, что они правильные. Поэтому, мы можем сделать график общего количества родившихся по годам с помощью команды:

```
plot(present$year, present$boys + present$girls, type = "l")
```

На этот раз, обратите внимание, мы не использовали имена первых двух аргументов. Мы можем сделать это потому, что файл справки показывает, что по умолчанию для plot первым аргументом является переменная x, а вторым аргументом - переменная y.

Вопрос 4 [выбрать ответ] В каком году мы видим наибольшее число рождений в США?

(a) 1961
(b) 1960
(c) 1957
(d) 1991

Подсказка: Вы можете обратиться к файлам помощи или к ссылке (http:// cran.r-project.org/ doc/ contrib/ Short-refcard.pdf) чтобы найти полезные команды. Например, вы можете

посмотреть функцию which.max.

Подобно тому, как мы рассчитали сумму новорожденных мальчиков и девочек, мы можем вычислить отношение числа мальчиков к числу девочек, рожденных в 1940 году, введя:

```
1211684/1148715
```

или мы можем использовать полные вектора:

```
present$boys/present$girls
```

Доля новорожденных мальчиков:

```
1211684/(1211684 + 1148715)
```

или это может также вычисляться одновременно для всех лет:

```
present$boys/(present$boys + present$girls)
```

Обратите внимание, что в R как с калькулятором, вы должны отдавать себе отчет в порядке операций. Здесь мы хотим разделить количество мальчиков на общее число новорожденных, поэтому мы должны использовать скобки. Без них R будет сначала делать деление, затем сложение, и долю вы не получите

Вопрос 5 [TRUE / FALSE] Теперь сделайте график доли мальчиков с течением времени, и на основе графика определите, является ли следующее утверждение истинным или ложным (TRUE или FALSE): доля мальчиков, родившихся в США, уменьшилась с течением времени.

- TRUE
- FALSE

Совет: с помощью клавиш со стрелками вверх и вниз вы можете просматривать предыдущие команды вашей так называемой истории команд. Вы также можете получить доступ к ней, нажав на вкладке ***history*** *в верхней правой панели. Это позволит вам сэкономить много времени в будущем.*

Наконец, в дополнение к простым математическим операторам, таким как вычитание и деление, вы можете попросить R делать сравнения, такие как больше, >, меньше, <, и равенства, ==. Например, мы можем спросить, превосходят ли мальчики девочек по рождаемости в каждом году с помощью выражения:

```
present$boys > present$girls
```

Эта команда возвращает 63 значения либо ИСТИНА (TRUE), если в этом году было больше мальчиков, чем девочек, или ЛОЖЬ (FALSE), если в этом году это не так. Этот вывод показывает данные иного рода, чем мы рассматривали до сих пор. Во фрейме данных present значения численные (год, количество мальчиков и девочек). Здесь мы попросили R создать логические данные, где значения являются ИСТИНОЙ или ЛОЖЬЮ. В общем, анализ данных будет включать в себя различные виды типов данных, и одной из причин для использования R является то, что он способен представлять и выполнять операции вычисления со

многими из них.

Вопрос 6 [выбрать] Что из ниже следующего правда?

(a) Каждый год рождается больше девочек, чем мальчиков.
(b) Каждый год больше новорожденных мальчиков, чем девочек.
(c) В половине лет рождается больше мальчиков, в другой половине – больше девочек.

Теперь попробуйте ответить на следующие вопросы без дополнительных указаний по кодированию, только основываясь на том, что вы узнали до сих пор.

Вопрос 7 [выбрать] Сделайте график, который отображает соотношение мальчик-к-девочке за каждый год. Что вы видите?

(a) Похоже, нет никакой тенденции в соотношении мальчик-к-девочке с 1940 по 2002.
(b) Существует первоначально увеличение соотношения мальчик-к-девочка, который достигает максимума около 1960 г. После 1960 г. наблюдается снижение в соотношении мальчик-к-девочке, но начинает увеличиваться в середине 1970-х годов.
(c) Существует первоначальное уменьшение соотношения мальчик-к-девочке, а затем увеличение в период между 1960 г. и 1970 г. , а затем снижение.
(d) Соотношение мальчик-к-девочке все время увеличивалось.
(e) Существует начальное снижение отношения мальчик-к-девочке до уровня около 1960 г. и с тех пор далее остается неизменным

Вопрос 8 [выбрать] Рассчитайте абсолютные различия между количеством мальчиков и девочек, родившихся в каждом году, и определите, в каком году была самая большая абсолютная разница в количествах новорожденных девочек и мальчиков?

(a) 1963
(b) 1946
(c) 2002
(d) 1940

Эти данные взяты из отчета Центров по контролю за заболеваниями (*http:// www.cdc.gov/ nchs/ data/ nvsr/ nvsr53/ nvsr53 20.pdf*). Вы можете посмотреть его, если вы хотите прочитать больше об анализе соотношения полов при рождении в США.

Это было краткое введение в R и RStudio. Очень полезно просматривать веб-сайты для R *http:// www.r-project.org* и RStudio *http:// rstudio.org*, если вы заинтересованы в получении дополнительной информации или хотите получить данные для дополнительной практики по адресу: *http:// openintro.org*

Данные

Существует определение статистики как области знаний, которая фокусируется на превращении информации в знания. Первым шагом в этом процессе является обобщение и описание «сырой» информации - данных. На этих практических занятиях попробуем разобраться в вопросах здравоохранении путем создания простых графических и численных обобщений наборов данных, собранных Центрами по контролю и профилактике заболеваний (Centers for Disease Control and Prevention, CDC). Поскольку это достаточно большая база данных, мы ее используем для получения необходимых навыков обработки данных и их подмножеств.

С чего следует начинать

Система Behavioral Risk Factor Surveillance System (BRFSS) является ежегодным телефонным опросом 350 000 человек в США. Как следует из

названия, BRFSS предназначена для выявления факторов риска для здоровья среди взрослого населения и выявления новых тенденций. Например, респондентам задают вопросы об их диете и еженедельной физической активности, статуса ВИЧ-инфицирования, об употреблении табака, а также уровня по медицинскому обеспечению. На веб-сайте BRFSS (*http:// www.cdc.gov/ brfss*) содержится полное описание исследования, в том числе научно-исследовательских вопросов, которые мотивируют этот проект, и много интересных результатов, полученных из данных BRFSS.

Мы сосредоточимся на случайной выборке в 20000 человек из опроса BRFSS, проведенного в 2000 году. Хотя в этой базе данных более 200 переменных, мы будем работать с небольшой группой переменных различных типов.

Начнем с загрузки набор данных 20000 наблюдений в рабочее пространство R. После запуска RStudio, введите следующую команду:

```
source("http://www.openintro.org/stat/data/cdc.R")
```

База данных cdc, которая появляется в рабочей области представляет собой *матрицу данных*, в ней в каждой строке представлено наблюдение для одного респондента, и каждый столбец представляет переменную. В языке R это формат данных называется *фрейм данных*, этот термин будем использовать на протяжении всего курса.

Для просмотра имен переменных, наберите команду:

```
names(cdc)
```

Вы должны получить следующие имена переменных: genhlth, exerany, hlthplan, smoke100, height, weight, wtdesire, age, и gender. Каждая из этих переменных соответствует вопросу, который был задан в опросе. Например, для genhlth, респондентам было предложено оценить их общее состояние здоровья, отвечая либо отлично, очень хорошо, хорошо, удовлетворительно или плохо (excellent, very good, good, fair or poor). Переменная exerany указывает, делал ли респондент в прошлом месяце (1) или нет (0) физические упражнения.

Переменная hlthplan указывает, имеет ли респондент медицинское страхование в той или иной форме (1) или нет (0). Переменная smoke100 указывает, выкурил ли респондент по крайней мере 100 сигарет в своей жизни. Другие переменные записывают рост респондента в дюймах (height), вес в фунтах(weight), а также их желаемый вес (wtdesire) , возраст (age) и пол (gender).

Вопрос 1 [выбрать] Сколько наблюдений и переменных в этом фрейме данных?

- 9 наблюдения; 20000 переменных
- 10 наблюдений; 20000 переменных
- 20000 наблюдений; 9 переменных
- 20000 наблюдений; 10 переменных

Вопрос 2 [выбрать] Какого типа переменная genhlth?

- количественная, непрерывная
- количественная, дискретная
- качественная (непорядковая)
- качественная (порядковая)

Вопрос 3 [выбрать] Какого типа переменная weight?

- количественная, непрерывная
- количественная, дискретная
- качественная (непорядковая)
- качественная (порядковая)

Вопрос 4 [выбрать] Какого типа переменная smoke100?

- количественная, непрерывная
- количественная, дискретная
- качественная (непорядковая)
- качественная (порядковая)

Мы можем взглянуть на первые несколько записей (строк) наших данных с помощью команды:

```
head(cdc)
```

Иточно так же мы можем посмотреть на несколько последних записей, введя:

```
tail(cdc)
```

Вы также можете посмотреть на *весь* фрейм данных сразу, введя его название в консоль, но это может быть неразумным. Мы знаем, что cdc имеет 20000 строк, и чтение всего набора данных будет перегружать ваш экран. Лучше бросить быстрые взгляды на небольшие фрагменты с использованием head, tail или методов выделения подмножества, с которыми вы позже познакомитесь

Сводки и таблицы

Анкета BRFSS это огромная сокровищница информации. Первым шагом в любом анализе является необходимость первести всю эту информацию в несколько сводных таблиц статистических данных и в графики. В качестве простого примера, функция summary возвращает численное резюме: минимальное значение, первый квартиль, медиану, среднее, второй квартиль, и максимальное значение

Для того, чтобы получить это для переменной weight , введите:

```
summary(cdc$weight)
```

Получим в результате:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   68.0   140.0   165.0   169.7   190.0   500.0 
```

R можно использовать как калькулятор. Если вы хотите вычислить вероятное отклонение для веса респондентов , посмотрев на вывод команды summary выше, введите:

```
190 - 140
```

R имеет также встроенные функции для вычисления параметров сводной статистики по одному. Например, чтобы вычислить среднее, медиану и дисперсию веса (weight), введите :

```
mean(cdc$weight)
var(cdc$weight)
median(cdc$weight)
```

Эти характеристики очень осмыслены для описания количественных переменных, вроде веса, но как быть с качественными (категорийными) данными? В таком случае обычно рассматривают частоту встречаемости каждого варианта признака или относительное распределение частот. Функция table делает это путем подсчета, сколько раз был дан ответ каждого вида. Например, чтобы увидеть количество людей, которые выкурили 100 сигарет, введите:

```
table(cdc$smoke100)
```

или можно посмотреть на относительное распределение частот, введя:

```
table(cdc$smoke100)/20000
```

Обратите внимание, как R автоматически делит все записи в таблице на 20000 в этой команде. Это похоже на то, что мы видели, когда умножали или делили вектор чисел: R применяет это действие по всем записям в векторах. Мы убедились, что это работает и для таблиц. Далее, по таблице мы делаем диаграмму, поставив table внутри команды barplot:

```
barplot(table(cdc$smoke100))
```

Обратите внимание! Мы вычислили таблицу cdc$smoke100, а затем сразу же обратились к графической функции, barplot. Это важная идея: команды R могут быть вложенными. Вы можете разбить это на два этапа, введя следующее:

```
smoke = table(cdc$smoke100)
barplot(smoke)
```

В этом случае мы создаем новый объект, таблицу, с именем smoke (содержание которого мы можем видеть, введя smoke в консоль), а затем используем его в качестве входных данных для barplot.

Вопрос 5 [выбрать] Создайте сводную таблицу для gender. Сколько мужчин в этой выборке?

(a) 4657
(b) 6972
(c) 9569
(d) 10431
(e) 20000

Вопрос 6 [выбрать] Сосчитайте относительное распределение частот для genhlth. Какую долю составляют находящиеся в отличном здоровье? Выберите самый близкий ответ.

(a) 0.15
(b) 0.23
(c) 0.30
(d) 0.51

Команда table может быть использована для табулирования любого количества переменных. Например, чтобы изучить количество курящих и некурящих для каждого пола , мы можем использовать следующее:

```
gender_smokers = table(cdc$gender, cdc$smoke100)
```

Сначала создаем новый объект, таблицу с именем gender smokers (содержание которой можно увидеть, введя gender smokers в консоли). Напомним, что 1 указывает, что респондент выкурил по крайней мере 100 сигарет. Строки относятся к полу. Для создания мозаичной диаграммы по этой таблице, мы должны ввести следующую команду:

```
mosaicplot(gender_smokers)
```

Мы могли бы сделать это в два этапа путем сохранения таблицы в одной строке и применяя mosaicplot в следующей (см. пример с table/barplot выше).

Вопрос 7 [Выбрать] Что можно выявить о привычке к курению среди мужчин и женщин по мозаичной диаграмме гендерными ?

(a) Мозаичная диаграмма показывает, что мужчины более склонны к курению, чем женщины.
(b) Мозаичная диаграмма показывает, что мужчины менее склонны к курению, чем женщины
(c) Мозаичная диаграмма показывает, что мужчины и женщины имеют одинаковую склонность к курению.

Вставка: Как R думает о данных

Мы уже упомянули, что R хранит данные в виде фреймов данных, которые можно представить себе как таблицы. Каждая строка представляет отдельное наблюдение (различные респонденты), а каждый столбец представляет собой разные переменные (первый - genhlth, второй - exerany и так далее). Мы видим размер фрейма данных рядом с именем объекта в рабочей области или можем ввести в консоль:

```
dim(cdc)
```

Теперь, если мы хотим получить доступ к подмножеству полного фрейма данных, мы можем использовать обозначения строк и столбцов. Например, чтобы увидеть шестую переменную для пятьсот шестьдесят седьмого респондента, используется формат

```
cdc[567, 6]
```

который означает, что мы хотим увидеть элемент нашего набора данных, который находится в 567-м ряду (имеется в виду пятьсот шестьдесят седьмой человек или наблюдение) и в шестом столбце (в данном случае это weight). Мы

знаем, что weight является шестой переменной, потому что это шестой элемент в списке имен переменных

```
names(cdc)
```

Для того, чтобы увидеть вес первых 10-ти респондентов , нужно ввести:

```
cdc[1:10, 6]
```

В этом выражении, мы попросили вывести данные только для строк в диапазоне от 1 до 10. R использует ":" чтобы создать диапазон значений, так 1:10 расширяется до 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. В этом можно убедиться, введя

```
1:10
```

Наконец, если мы хотим все данные для первых 10-ти респондентов, вводим:

```
cdc[1:10, ]
```

Оставляя индекс или диапазон пустым (т.е. ничего между запятой и квадратными скобками не набрано), мы получаем все столбцы. Это немного противоречит интуиции. В R, как правило, мы опускаем номер столбца, чтобы увидеть все столбцы фрейма данных. Точно так же, если оставить индекс или диапазон для строк пустым, мы получим все наблюдения, (а не только пятьсот шестьдесят седьмой, или строки с 1 по 10) Попробуйте ввести следующее (чтобы вывести на экран вес для всех 20000 респондентов):

```
cdc[, 6]
```

Напомним, что столбец 6 представляет вес респондентов, поэтому команда выше вернула все веса в наборе данных. Альтернативный метод доступа к данным о весе это обращение к имени переменной. Ранее мы набрали names(cdc), чтобы увидеть все переменные, содержащиеся в фрейме данных cdc. Мы можем использовать любое из имен переменных, чтобы выбрать записи из фрейма:

```
cdc$weight
```

Знак $ говорит R найти во фрейме данных cdc столбец, называемый weight. Поскольку это один вектор, мы можем видеть его подмножества с помощью

всего одного индекса в квадратных скобках. Мы получим вес для пятьсот шестьдесят седьмого респондента, введя

```
cdc$weight[567]
```

Сходным образом, получим вес 10-ти первых респондентов, введя:

```
cdc$weight[1:10]
```

Команда выше возвращает тот же результат, что и cdc[1:10,6] команда. Обе нотации (обозначения строк и столбцов и обозначение со знаком доллара) широко используются, что вы выберете использовать, зависит от ваших личных предпочтений.

Еще немного о подмножествах

Часто оказывается полезным извлекать из множества данных все случаи, которые имеют определенные специфические характеристики. Этого добиваются с помощью кондиционирующих команд (команд с заданием условий). Сначала рассмотрим выражения типа

```
cdc$gender == "m"
```

или

```
cdc$age > 30
```

Эти команды производят значений TRUE и FALSE,. Существует одно значение для каждого респондента, где TRUE указывает, что человек был мужчиной (для первой команды) или старше 30 (вторая команда).

Предположим, мы хотим извлечь выборке данных только для мужчин, или просто для тех, кому за 30 лет, для этого мы можем использовать функцию R subset. Например, команда

```
mdata = subset(cdc, cdc$gender == "m")
```

создаст новый набор данных с именем mdata, содержащий только мужчин из фрейма данных cdc . (Обратите внимание на двойной знак равенства!). Вы можете взглянуть на первые несколько рядов, как обычно:

```
head(mdata)
```

Этот новый фрейм данных содержит все те же переменные, но только строки для мужчин. Таким образом, мы можем поделить данные на основе значений одной или нескольких переменных.

Можно задавать условия, используя знаки **&** и |. Амперсанд **&** читается как "и", так что

```
m_and_over30 = subset(cdc, cdc$gender == "m" & cdc$age > 30)
```

выберет данные для мужчин старше 30 лет.

Знак |обозначает “или”, поэтому

```
m_or_over30 = subset(cdc, cdc$gender == "m" | cdc$age > 30)
```

выберет данные для мужчин или всех респондентов старше 30 лет (чем эта группа может быт интересна на вскидку сказать сложно, сейчас важна техническая сторона). В принципе, вы можете использовать столько "и" и "или", сколько вам нравится для перехода к подмножеству.

Упражнеие Создайте новый объект с именем under23 and smoke , который содержит все наблюдения для респондентов в возрасте младше 23 лет, которые выкурили по крайней мере 100 сигарет в своей жизни. Запишите команду, что вы использовали для создания нового объекта в качестве ответа.

Вопрос 8 [выбрать] Сколько наблюдений в созданном объекте under23 and smoke?

(a) 620
(b) 771
(c) 7,244
(d) 10,436
(e) 17,591

Количественные данные

Имея в руках инструменты выделения подмножеств, возвращаемся к задаче, с которой начали: сделать основные краткие сводки из анкет BRFSS. Мы уже рассмотрели качественные признаки smoke и gender, теперь сосредоточимся на количественных признаках. Для визуализации количественных показателей используют 2 основных метода: рисуют диаграммы box plots и histograms. Мы можем построить box plot («ящик с усами») для одной переменной, используя команду:

```
boxplot(cdc$height)
```

Вы можете сравнить расположение компонентов коробки со сводной статистикой

```
summary(cdc$height)
```

Убедитесь, что медиана и верхние и нижние квартили, представленные в summary, совпадают с таковыми в графике. Цель boxplot заключается в предоставлении переменной с целью сравнения по нескольким категориям. Так что мы можем, например, сравнивать рост мужчин и женщин с помощью

```
boxplot(cdc$height ~ cdc$gender)
```

В этом выражении новое обозначение. Символ ~ может быть прочитан "по сравнению с" или "в зависимости от". Так мы просим R построить boxplot для height, где группы определяются по полу (рост мужчин и рост женщин).

Далее рассмотрим новую переменную, которая не показывается непосредственно в этом фрейме данных: индекс массы тела (BMI). BMI является соотношением веса и роста и может быть рассчитана как:

$$BMI = weight\ (lb) / height\ (in)^2 \times 703$$

где 703 - приблизительный коэффициент для изменения единиц от метрических (метров и килограммов) в имперских (дюймов и фунтов)

В следующих двух строчках сначала создается новый объект с именем bmi , затем строится box plots для этого признака для групп, определяемых переменной cdc$genhlth.

```
bmi = (cdc$weight/cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```

Обратите внимание, что арифметические действия в первой строке применяются ко всем 20000 наблюдений во фрейме данных cdc. То есть для каждого из 20 000 участников, мы берем их вес, делим на квадрат их роста, а затем умножаем на 703. В результате получаем 20000 значений BMI, по одному для каждого респондента. Это еще одна из причин, почему любят R: он позволяет выполнять вычисления, используя очень простые выражения.

Вопрос 9 [выбрать] Что является ложью (*false)* на основании boxplot BMI vs. general health?

(a) Медиана индекса массы тела примерно 25 для всех категорий общего состояния здоровья, но есть небольшое увеличение медианы BMI по мере снижения общего состояния здоровья (от excellent до poor).
(b) IQR несколько увеличивается по мере снижения общего состояния здоровья (от excellent до poor).
(c) Среди людей с отменным здоровьем (excellent) есть некоторые с необычно низкими BMI по сравнению с остальной частью группы.
(d) Распределения BMI каждой группы по состоянию здоровья скошены влево.

Упражнение Выберите другую качественную переменную из базы данных и посмотрите, как она соотносится с BMI. Объясните, почему вы думаете, что она будет иметь отношение к BMI

Наконец, давайте сделаем несколько гистограмм. Мы можем посмотреть на гистограмму для возраста респондентов с помощью команды:

```
hist(cdc$age)
```

Гистограммы, как правило, очень хороший способ увидеть форму распределения по признаку, форма может меняться в зависимости от того, как данные распределяются между различными «бункерами». Вы можете контролировать количество «бункеров», добавив аргумент к команде. В следующих двух строках, мы сначала сделаем гистограмму по умолчанию для bmi, затем аргументом 50 breaks

```
hist(bmi)
hist(bmi, breaks = 50)
```

Обратите внимание, что вы можете переключаться между графиками, нажав вперед и назад стрелки в правом нижнем окне RStudio, чуть выше графиков. Сравните эти две гистограммы.

Упражнение Используйте функцию plot создайте график (a scatterplot) веса относительно желаемого веса.

Вопрос 10 [выбрать] Что справедливо для связи переменных вес и желаемый вес?

(a) слабая отрицательная линейная
(b) слабая положительная линейная
(c) сильная положительная линейная
(d) сильная отрицательная линейная

Заключение

Проделанный анализ данных позволил нам достаточно хорошо продвинуться в анализе информации из анкет BRFSS. Мы нашли интересную связь между курением и полом, что-то можем сказать о взаимосвязи между оценкой людьми их общего состояния здоровья и их BMI. Мы также получили существенные вычислительные средства – научились рассчитывать статистические сводки данных, отбирать подмножества и строить графики. Все эти методы будут использоваться на протяжении всего курса.

Контрольные задания

Задание 1

American Community Survey предоставляет скачиваемые данные из различных обследований общества в Соединенных Штатах. С помощью команды **download.file ()** скачать данные из опроса о жилье в штате Айдахо в 2006 г. с сайта :

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv

Загрузите эти данные в R. Книга кодирования, описывающая имена переменных находится по адресу:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf

Сколько категорий стоимостью $ 1 млн или больше?

- 25
- 47
- 24
- 53

Задание 2

Используйте данные из задания 1. Рассмотрим переменную FES. Какой из принципов "аккуратных данных" (tidy data) нарушаются в этой переменной?

- «Аккуратные данные» имеют одно наблюдение для каждой строки
- Числовые значения в «аккуратные данных» не могут представлять категории.
- «Аккуратные данные» содержат одну переменную в каждом столбце.
- Каждая переменная в «аккуратных данных» была преобразована для интерпретации.

Задание 3

Скачать Excel таблицу из данных Natural Gas Aquisition Program по адресу:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx

(original data source: http://catalog.data.gov/dataset/natural-gas-acquisition-program)

Прочитайте строки 18-23 и столбцы 7-15 в R и присвойте результат переменной с именем **dat**

Чему равно значение выражения sum(dat$Zip*dat$Ext,na.rm=T)

- 338924
- 184585
- NA
- 36534720

Задание 4

Прочитайте XML данные о ресторанах г.Балтимора с сайта:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml

Сколько ресторанов имеют zipcode 21231?

- 181
- 127
- 156
- 100

Задание 5

Скачайте данные опроса 2006 г. о жилье для штата Айдахо с помощью команды **download.file ()** по адресу:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv

Используя команду **fread()** загрузите данные в R, назовите объект **DT**

Что из перечисленного ниже является самым быстрым способом для расчета средних значений переменной **pwgtp15** для мужчин и женщин с использованием пакета **data.table**?

- tapply(DT$pwgtp15,DT$SEX,mean)
- mean(DT[DT$SEX==1,]$pwgtp15); mean(DT[DT$SEX==2,]$pwgtp15)
- DT[,mean(pwgtp15),by=SEX]

- rowMeans(DT)[DT$SEX==1]; rowMeans(DT)[DT$SEX==2]
- sapply(split(DT$pwgtp15,DT$SEX),mean)
- mean(DT$pwgtp15,by=DT$SEX)

Задание 6

Рассмотрим приведенную ниже таблицу с описанием набора данных лиц, которые добровольно зарегистрировались в государственной школе.

Имя	Год рождения	Телефон	Кол-во сестер(братьев)	Годовой доход
Анна	1975	8929223	0	66000
Федор	1984	8629237	3	22500
…	…	…	…	…

Какие переменные качественные, а какие - количественные?

- Все качественные, количественных нет.
- **качественные:** Имя, Телефон
количественные: Год рождения, Кол-во сестер(братьев), Годовой доход.
- **качественные:** Имя, Телефон , Кол-во сестер(братьев),
количественные: Год рождения, Годовой доход.
- **качественные:** Имя
количественые: Год рождения, Телефон , Кол-во сестер (братьев), Годовой доход.

Задание 7

В социальном обследовании, проводимом ежегодно в Соединенных Штатах, спрашивается, сколько друзей у людей (number of friends) и как они оценивают свой уровень счастья (very happy, pretty happy, not too happy). Для того чтобы оценить связь между этими двумя переменными исследователь вычисляет среднее количество друзей для людей, которые классифицировали себя как очень

счастлив, довольно счастлив, и не слишком счастлив. Какие переменные независимые? зависимые ?

- **независимые:**very happy, pretty happy, not too happy
зависимые: number of friends
- **независимые:**number of friends
зависимые: very happy, pretty happy, not too happy
- **независимые:**happiness level (categorical with 3 levels)
зависимые: number of friends
- **независимые:**number of friends
зависимые: happiness level (categorical with 3 levels)

Задание 8

В исследовании, опубликованном в 2011 PNAS USA, 120 пожилых мужчин и женщин (средний возраст около 65 лет), которые добровольно согласились участвовать в этом исследовании, были случайным образом распределены в две группы. В первой группе добровольцы ходили по дорожке в парке три раза в неделю; в другой - делали множество менее аэробных упражнений, в том числе йогу и тренировки с отягощением. Через год сканирование мозга показало, что у «пешеходов»гиппокамп (часть мозга, отвечающая за формирование воспоминаний) увеличился в объеме в среднем примерно на 2%; в другой группе объем гиппокампа снизился на 1,4%. Что из перечисленного ниже ложно?

- Причинная связь между ходьбой и расширения гиппокампа могут быть выведены на основе этих результатов.
- Независмой переменной является тип упражнений, а зависимой переменной является изменение объема гиппокампа.
- Результаты этого исследования могут быть обобщены на всех пожилых.

Задание 9

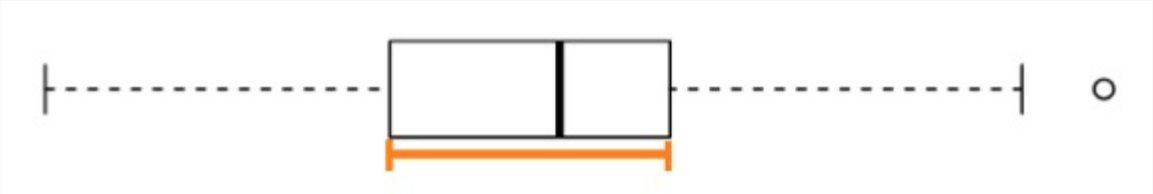
Что из ниже перечисленного является одним из четырех принципов экспериментального дизайна?

- стратификация
- контроль

- клстеризация
- нет ответа

Задание 10

Что из нижеперечисленного является шириной коробки («ящика») - показано оранжевым на рисунке?



- стандартное отклонение
- медиана
- диапазон
- IQR (межквартильный диапазон)
- среднее

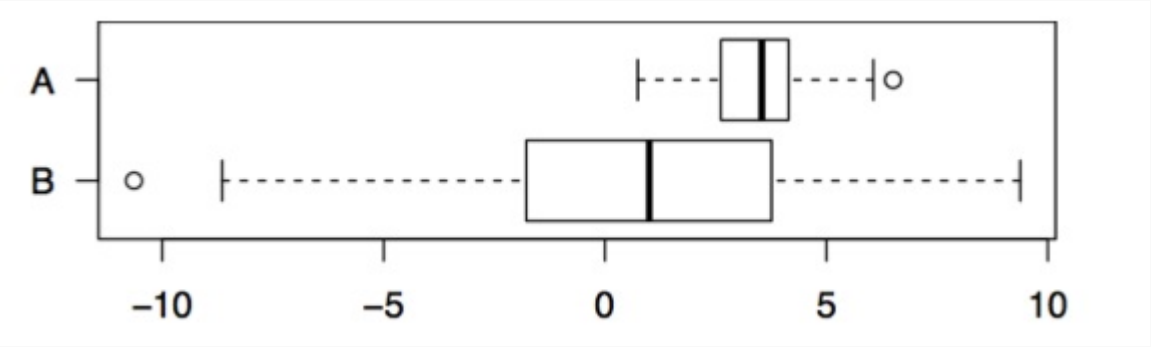
Задание 11

Правда или ложь: Статистика среднее / медиана (среднее разделить на медиану) может использоваться в качестве меры асимметрии (или правой, или левой). Если эта статистика меньше 1, распределение, скорее всего, скошено влево.

- ложь
- правда

Задание 12

Два распределения (A и B), показаны на графике ниже. Что из нижеперечисленного не подтверждается графиком?



- Оба распределения примерно симметричны
- Оба распределения унимодальны
- Медиана А превышает медиану В
- вариабельность в В больше, чем в А.

Задание 13

В одном американском городе был проведен опрос о жилье, чтобы определить цену типичного дома в городе, в котором проживает в основном средний класс, но есть очень дорогой пригород. Средняя стоимость дома в этом городе примерно $ 650 000. Что из нижеперечисленного является наиболее вероятно правдой?

- Большинство домов в этом городе стоят более $ 650 000
- Большинство домов в этом городе стоят меньше, чем $ 650 000.
- Есть примерно одинаковое количество домов в этом городе, которые стоят более $ 650 000, и менее этой суммы
- Чтобы ответить на этот вопрос нужно знать стандартное отклонение

Задание 14

Довольно часто рыбу маркируют ошибочно в супермаркетах и даже в ресторанах. В приведенной ниже таблице показаны результаты исследования, где случайная выборка из 156 рыб, приготовленных для продажи, были генетически протестированы. Исследователи классифицируют каждый образец как маркированный правильно или ошибочно. Какая часть копченой рыбы в пробе

была неправильно маркирована? Выберите самый близкий ответ.

	Копченная рыба	Некопченная рыба	Всего
Неправильно маркирована	28	11	39
Правильно маркирована	8	109	117
Всего	36	120	156

- 78%
- 72%
- 18%
- 9%
- 28%

Задание 15

Профессора регулярно дают две версии экзамена, при этом Профессор также может предоставить сводную статистику по каждой версии. Предположим, приводится следующее краткое изложение:

	Кол-во	Среднее	Медиана	Стандартное отклонение
Версия 1	53	65,4	72	16
Версия 1	65	66,5	71	17

Студент, который сдавал Вариант А говорит, что он должен получить дополнительное балл, потому что его экзамен был сложнее, о чем свидетельствует более низки средний балл для версии А, как показано в таблице. Хороший ли это аргумент? Выберите лучший ответ ниже.

- Мы должны знать форму распределения для каждой версии, чтобы определить, действует ли этот аргумент

- Мы должны знать минимум и максимум для каждой версии, чтобы определить, действует ли этот аргумент.
- Да. Разница в средних экзаменационных оценках означает, что существует разница в сложности между версиями.
- Нет. Средние баллы близки при рассмотрении разброса распределений. Разница может быть случайной.
- Нет. Медиана версии A выше