УДК 619.25

РАСЧЕТ ПОКАЗАТЕЛЕЙ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОЙ СРЕДЫ R

© 2018 г. **[1]В. Л. Егошин, [2]С. В. Иванов, [3]Н. В. Саввина, [4]Г. Ж. Капанова, [3-6]А. М. Гржибовский**

[1]Павлодарский филиал Государственного медицинского университета г. Семей, г. Павлодар, Казахстан; [2]Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова, г. Санкт-Петербург; [3]Северо-Восточный федеральный университет им. М. К. Аммосова, г. Якутск; [4]Казахский Национальный Университет им. аль-Фараби, г. Алматы, Казахстан; [5]Северный государственный медицинский университет, г. Архангельск; [6]Западно-Казахстанский государственный медицинский университет им. Марата Оспанова, г. Актобе, Казахстан

В статье представлены основные алгоритмы работы в программной среде R для расчета показателей описательной статистики исследовательских данных, в том числе с применением специализированных пакетов. Приведены детализированные примеры использования функций R для описания и визуализации количественных и категориальных переменных.

Ключевые слова: описательная статистика, количественные данные, категориальные данные, визуализация, R

DESCRIPTIVE STATISTICS USING R

[1]V. L. Egoshin, [2]S. V. Ivanov, [3]N. V. Savvina, [4]G. Zh. Kapanova, [3-5]A. M. Grjibovski

[1]Semey State Medical University, Pavlodar Campus, Pavlodar, Kazakhstan; [2]I. P. Pavlov First St. Petersburg State Medical University, St. Petersburg, Russia; [3]North-Eastern Federal University, Yakutsk, Russia; [4]Al-Farabi Kazakh National University, Almaty, Kazakhstan; [5]Northern State Medical University, Arkhangelsk, Russia; [6]West Kazakhstan Marat Ospanov State Medical University, Aktobe, Kazakhstan

The article presents basic algorithms of R software using for calculating descriptive statistics of biomedical data including the use specialized packages. Detailed examples of the use of R functions for description and visualization of quantitative and categorical data are given.

Key words: descriptive statistics, continuous data, categorical data, visualization, R

Библиографическая ссылка:

Егошин В. Л., Иванов С. В., Саввина Н. В., Капанова Г. Ж., Гржибовский А. М. Расчёт показателей статистики с использованием программной среды R // Экология человека. 2018. № 9. С. 55–64.

Egoshin V. L., Ivanov S. V., Savvina N. V., Kapanova G. Zh., Grjibovski A. M. Descriptive Statistics Using R. *Ekologiya cheloveka* [Human Ecology]. 2018, 9, pp. 55-64.

В биомедицинских исследованиях анализ данных, полученных в их результате независимо от типа исследования, всегда начинается с описательной статистики – одного из разделов статистической науки, в рамках которого изучаются методы описания и представления основных свойств данных [5]. Для этого используются общепринятые способы описания результатов исследования – они могут быть представлены в виде суммарных показателей, а также в табличном и графическом виде.

Описательная статистика при анализе данных предполагает изучение следующих характеристик выборки [1, 15]:

– для непрерывных числовых переменных – оценку центральных тенденций и показателей распределения;

– для дискретных и категориальных переменных – оценку частотного распределения наблюдений.

Программная среда R обладает большими возможностями для изучения показателей описательной статистики, позволяя не только вычислять их числовые значения, но и обеспечивать эффективную визуализацию данных. Методы описательной статистики в R достаточно подробно изложены в руководствах [2, 3, 7–10], помимо которых также существует много интернет-ресурсов [11–15], освещающих прикладные вопросы применения R при анализе данных.

В программной среде R для проведения описательной статистики используются как базовые функции R, так и функции дополнительных пакетов.

Для демонстрации возможностей R в описательной статистике будут использованы данные Архангельского областного регистра родов [5]. Загрузка используемых пакетов и таблицы данных в систему R представлена на рис. 1 (листинг 1).

Листинг 1

```
library(tidyverse)
library(DescTools)
# импорт данных из файла формата sav (подготов-
лен в программе SPSS)
df <- foreign::read.spss(«Simulated_sample.sav”,
to.data.frame = TRUE)
# Примечание
# запись типа «название пакета»::»название функ-
```

```
ции» (например, foreign::read.spss) позволяет ис-
полнять функцию, не загружая сам пакет
#
# добавление новых переменных
df <- df %>%
  mutate(lowBirthWeight = factor(ifelse(BIrthweight
< 2500, 'yes', 'no')),
  Infant_sex = as.factor(as.character(Infant_sex)),
  Preeclampsia = factor(Preeclampsia, levels =
c(0, 1),
  labels = c('no', 'yes')))

df_s <- df %>%
  select(Maternal_age, Gestational_age,
BIrthweight,
  Marital_status, Delivery_type, Infant_sex)

df_t <- df %>%
  select(lowBirthWeight, Infant_sex, Delivery_type,
  Preeclampsia, Birth_defect)
```

Рис. 1. Загрузка пакетов «tidyverse», «DescTools» и таблицы данных в R (листинг 1)

Как было сказано выше, параметры описательной статистики и подходы к анализу для непрерывных данных (к ним относятся, например, возраст, уровень гемоглобина, масса тела) и для дискретных данных (пол, уровень образования, наличие определенного заболевания) принципиально различаются.

Далее в статье будут приведены алгоритмы описательной статистики для обоих типов данных биомедицинских исследований.

Описательная статистика для непрерывных данных

Основным математическим показателем оценки центральной тенденции для непрерывных данных является среднее арифметическое. В R для расчета среднего арифметического применяется команда *mean(x, trim = 0, na.rm = FALSE, ...)*. В данной команде используются следующие параметры функции:

– *x* (здесь и далее) – объект, для которого вычисляется показатель (например, числовой вектор);

– *trim* – доля наблюдений (от 0 до 0,5), которые будут удалены при расчете с каждого конца отсортированного вектора;

– по умолчанию *na.rm = FALSE*, для того, чтобы *NA* не учитывались при вычислении, параметр должен быть установлен *na.rm = TRUE* (специальные значения *«NA» – «Not applicable»* – кодирует отсутствующие значения в вариационных рядах и таблицах данных, так как в R просто пропуски значений не допускаются);

– многоточие в синтаксисе команды указывает на возможность использования и других параметров.

По аналогии со средним арифметическим для расчета взвешенной средней используется команда *weighted.mean(x, w, ..., na.rm = FALSE)*, для расчета медианы – *median(x, na.rm = FALSE, ...)*.

Далее приведем пример определения средних значений для числового вектора. Следует отметить, что вектор в R – это вариационный ряд, и он может быть создан пользователем. Числовая переменная из таблицы данных также может быть представлена в виде вектора, при расчете необходимо учитывать возможность наличия «NA» (недоступных значений). Создание числового вектора и расчет основных показателей описательной статистики представлены на рис. 2 (листинг 2).

Листинг 2

```
# создание числового вектора
v1 <- c(4, 5, 8, 3, 4, 5, 6, 8, 3)
mean(v1) # вычисление среднего арифметического
## [1] 5.111111
sort(v1) # отсортированный по возрастанию число-
вой вектор
## [1] 3 3 4 4 5 5 6 8 8
median(v1) # медиана числового вектора
## [1] 5
# числовой вектор из таблицы данных
mean(df$Paternal_age) # определение среднего
арифметического без использования дополнительных
параметров функции
## [1] NA
sum(is.na(df$Paternal_age)) # определение количе-
ства NA в столбце таблицы
## [1] 253
# определение среднего арифметического с учетом
NA
mean(df$Paternal_age, na.rm = TRUE)
## [1] 30.74585
# функция range выводит минимальное и максималь-
ное значения
range(df$Paternal_age, na.rm = TRUE)
## [1] 17 79
# определение среднего арифметического с учетом
NA среди 90% "центральных" значений (удалены по
5% с каждого конца вектора)
mean(df$Paternal_age, trim = .05, na.rm = TRUE)
## [1] 30.41894
# определение медианы
median(df$Paternal_age, na.rm = TRUE)
## [1] 30
```

Рис. 2. Создание числового вектора и использование функций описательной статистики (листинг 2)

Следует отметить, что в некоторых случаях значения переменной в векторе вводятся как бинарные («0» и «1»), а сама переменная при этом может быть не обозначена как категориальная. В этом случае использование функции mean() позволяет вычислить долю показателей, кодированных как «1». Пример такого вычисления приведен на рис. 3 (листинг 3).

Листинг 3

```
# создание числового вектора со значениями 0, 1
set.seed(12) # для воспроизводимости
# создание числового вектора и вывод значений
(v2 <- sample(0:1, 12, replace = TRUE))
## [1] 0 1 1 0 0 0 0 1 0 0 0 1
mean(v2)
## [1] 0.3333333
# пример из используемой таблицы данных
str(df$Anemia) # структура переменной Anemia та-
блицы данных df
## num [1:2000] 1 1 0 0 1 1 0 0 1 1 ...
mean(df$Anemia)
## [1] 0.5135
```

Рис. 3. Расчет доли значений «1» в векторе с бинарной переменной (листинг 3)

В отличие от среднего арифметического взвешенная средняя используется для вычисления среднего арифметического из сгруппированной таблицы по формуле $\frac{\sum x \cdot w}{\sum w}$, где x — вектор значений, а w — вектор весов.

Например, при регистрации продолжительности врачебного приема зафиксированы следующие значения: 5, 10, 15, 30, 45 минут. Соответствующее количество случаев для каждого варианта времени приема составило соответственно 10, 20, 15, 5, 2. В данном случае взвешенная средняя рассчитывается для того, чтобы определить средняя время приема — рис. 4 (листинг 4).

Листинг 4

```
time_v <- c(5, 10, 15, 30, 45)
count_v <- c(10, 20, 15, 5, 2)
weighted.mean(time_v, count_v)
## [1] 13.75
```

Рис. 4. Расчет взвешенного среднего (листинг 4)

Помимо оценки центральной тенденции для непрерывных данных огромное значение имеют и показатели распределения.

Оценка распределения данных может быть выполнена уже в процессе их визуализации. Наиболее известным является «колоколообразное» распределение непрерывных данных (нормальное распределение, распределение Гаусса), но в биомедицинских исследованиях часто могут встречаться и другие типы распределений, отличающиеся от нормального и требующие иных подходов к анализу. Для оценки распределения непрерывных данных используются гистограммы, диаграммы плотности, коробочные (ящичные, «ящик с усами»), скрипичные диаграммы, примеры которых приведены на рис. 5.

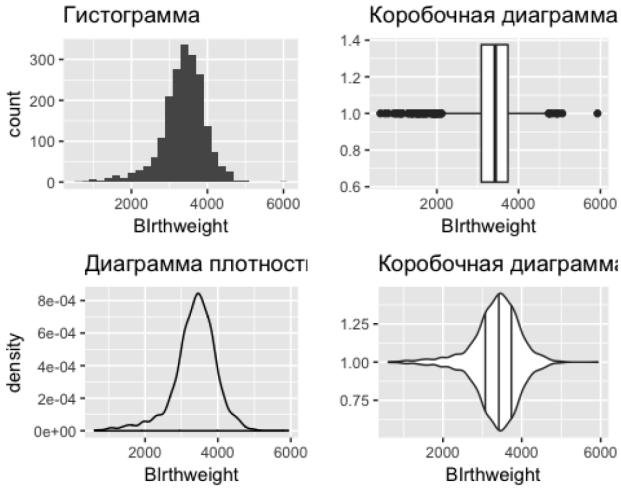


Рис. 5. Примеры визуализации непрерывных данных

В R показатели распределения непрерывных данных могут быть вычислены с использованием следующих функций: min(x), max(x), range(x), var(x), sd(x), mad(x), quantile(x, probs = , na.rm = FALSE), IQR(x). Информация о минимальном и максимальном значении числового вектора может быть получена при использовании функций min( ), max( ), range( ) — рис. 6 (листинг 5).

Листинг 5

```
min(df$BIrthweight, na.rm = TRUE)
## [1] 620
max(df$BIrthweight, na.rm = TRUE)
## [1] 5930
range(df$BIrthweight, na.rm = TRUE)
## [1] 620 5930
```

Рис. 6. Расчет минимальных и максимальных значений для вектора (листинг 5)

Важными характеристиками распределения числовой переменной являются дисперсия, стандартное отклонение, квантили. Функция «mad», определяющая абсолютное отклонение от медианы, используется для оценки распределения несколько реже.

Параметр «probs» в функции «quantile» позволяет создать желаемые значения квантилей:

- для вывода квартилей: *probs = seq(.25, .75, .25)*;
- для вывода квинтилей: *probs = seq(.2, .8, .2)*;
- для вывода децилей: *probs = seq(.1, .9, .1)*;
- для вывода процентилей: *probs = seq(.01, .99, .01)*.

При этом формат функции будет следующим: *seq(from = , to = , by =)*. Название параметров можно не указывать, но порядок ввода функции необходимо соблюдать. Для вывода квантилей можно создать вектор квантильных значений. Например, для вывода квартилей нужно указать следующие значения: *probs = c(.25, .5, .75)*.

Функция «IRQ» позволяет определить межквартильное расстояние — расстояние между первым и третьим квартилем.

Пример использования данных функций представлен на рис. 7 (листинг 6).

Листинг 6

```
# определение дисперсии
var(df$BIrthweight, na.rm = TRUE)
## [1] 354075.6
# определение стандартного отклонения
sd(df$BIrthweight, na.rm = TRUE)
## [1] 595.0425
# абсолютное отклонение медианы
mad(df$BIrthweight, na.rm = TRUE)
## [1] 489.258
# определение квартилей
quantile(df$BIrthweight, probs = c(.25, .5, .75),
na.rm = TRUE)
## 25% 50% 75%
## 3090 3430 3740
# определение квартилей – код с указанием названий параметров функции seq
quantile(df$BIrthweight, probs = seq(from = .25,
to = .75, by = .25), na.rm = TRUE)
## 25% 50% 75%
## 3090 3430 3740
# определение квинтилей
quantile(df$BIrthweight, probs = seq(.2, .8, .2),
na.rm = TRUE)
## 20% 40% 60% 80%
## 3010.0 3305.2 3540.0 3810.0
```

```
# определение децилей
quantile(df$BIrthweight, probs = seq(.1, .9, .1),
na.rm = TRUE)
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
## 2695.0 3010.0 3160.0 3305.2 3430.0 3540.0
3660.0 3810.0 4002.0
# определение процентилей с сохранением в векто-
ре, вывод первых пяти значений
percentil <- quantile(df$BIrthweight, probs =
seq(.01, .99, .01), na.rm = TRUE)
percentil[1:5]
## 1% 2% 3% 4% 5%
## 1390.00 1689.20 1949.64 2110.00 2260.00
# определение межквартильного расстояния
IQR(df$BIrthweight, na.rm = TRUE)
## [1] 650
```

Рис. 7. Расчет показателей распределения вектора (листинг 6)

Помимо вышеуказанных алгоритмов описательная статистика непрерывных данных может проводиться и с использованием специального пакета. Данный пакет позиционируется его автором Andri Signorell как набор инструментов для описательной статистики и разведочного анализа данных, и в него включены многие функции, отсутствующие в базовых пакетах.

Одним из преимуществ использования пакета «DescTools» является то, что эти функции позволяют не только рассчитать сами показатели описательной статистики, но и провести их интервальную оценку — определить доверительные интервалы — рис. 8 (листинг 7).

Листинг 7

```
# среднее арифметическое и доверительный интервал
MeanCI(df$BIrthweight, na.rm = TRUE)
## mean lwr.ci upr.ci
## 3370.836 3344.735 3396.937
# величина доверительного интервала может быть
определена пользователем, по умолчанию использу-
ется значение 0.95
MeanCI(df$BIrthweight, na.rm = TRUE, conf.level
= .9)
## mean lwr.ci upr.ci
## 3370.836 3348.935 3392.737
MeanCI(df$BIrthweight, na.rm = TRUE, conf.level =
.99)
## mean lwr.ci upr.ci
## 3370.836 3336.522 3405.150
# медиана и доверительный интервал
MedianCI(df$BIrthweight, na.rm = TRUE)
## median lwr.ci upr.ci
## 3430 3410 3450
# дисперсия и доверительный интервал
VarCI(df$BIrthweight, na.rm = TRUE)
## var lwr.ci upr.ci
## 354075.6 333105.5 377100.2
# коэффициент вариации и доверительный интервал
CoefVar(df$BIrthweight, na.rm = TRUE)
## [1] 0.1765267
# коэффициент асимметрии (skewness) и доверитель-
ный интервал
Skew(df$BIrthweight, na.rm = TRUE,
  conf.level = .95, ci.type = 'basic')
## skew lwr.ci upr.ci
## -0.8908888 -1.0939665 -0.7123040
# коэффициент эксцесса (kurtosis) и доверительный
интервал

Kurt(df$BIrthweight, na.rm = TRUE,
  conf.level = .95, ci.type = 'basic')
## kurt lwr.ci upr.ci
## 2.350422 1.801657 2.885762
# определение стандартной ошибки
MeanSE(df$BIrthweight, na.rm = TRUE)
## [1] 13.30888
```

Рис. 8. Расчет показателей описательной статистики с использованием пакета «DescTools» (листинг 7)

Среднее значение и доверительный интервал (как и другие показатели) могут быть получены сразу для всех числовых переменных таблицы данных с использованием функций «apply» и «sapply» — рис. 9 (листинг 8).

Листинг 8

```
# определение показателей с использованием функ-
ций apply, sapply
# sapply(df, is.numeric) выбирает только числовые
переменные
options(digits = 3)
apply(df_s[sapply(df_s, is.numeric)], 2, MeanCI,
na.rm = TRUE)
## Maternal_age Gestational_age BIrthweight
## mean 28.6 38.8 3371
## lwr.ci 28.4 38.7 3345
## upr.ci 28.8 38.9 3397
```

Рис. 9. Использование функций «apply» и «sapply» для анализа непрерывных переменных (листинг 8)

Группировка данных и вычисление суммарных показателей также могут быть выполнены с использованием пакета «tidyverse» — рис. 10 (листинг 9).

Листинг 9

```
dsm <- df %>%
  select(BIrthweight,
  Smoking_before_pregnancy,
  Smoking_during_pregnancy) %>%
  drop_na() %>%
  group_by(Smoking_before_pregnancy,
  Smoking_during_pregnancy) %>%
  summarise(mean_BW = MeanCI(BIrthweight)[1],
  n_count = n(),
  ci95_low = MeanCI(BIrthweight)[2],
  ci95_upper = MeanCI(BIrthweight)[3]) %>%
  ungroup() %>%
  mutate(Smoking_BD = paste('Before', Smoking_
before_pregnancy,
  'During', Smoking_during_pregnancy)) %>%
  select(Smoking_BD, mean_BW, ci95_low, ci95_
upper, n_count)

# результаты группировки и получения итоговых
значений
# функция pander из пакета pander позволяет вы-
вести данные более "аккуратным" образом

pander::pander(dsm)
```

Smoking_BD	mean_BW	ci95_low	ci95_upper	n_count
Before no During no	3396	3366	3426	1486
Before no During yes	3820	3312	4328	2
Before yes During no	3580	3410	3750	41
Before yes During yes	3211	3142	3280	299

```
# визуализация

ggplot(dsm, aes(Smoking_BD, y = mean_BW)) +
  geom_bar(stat = 'identity', color = 'grey50',
fill = 'white') +
  geom_errorbar(aes(ymin = ci95_low,
  ymax = ci95_upper), width = .6) +
  geom_text(aes(label = n_count, y = 2000),
  size = 4, color = 'grey20') +
  scale_x_discrete(labels = c('Не курят', 'Стали
курить \n во время \n беременности', 'Перестали
курить \n во время \n беременности', 'Продолжали
курить \n во время \n беременности')) +
  labs(x = '', y = 'Значение',
  title = 'Средний вес новорожденного и 95% ДИ',
  subtitle = 'указано количество случаев')
```

Рис. 10. Использование пакета «tidyverse» (листинг 9)

Описательная статистика для категориальных данных

В качестве показателя центральной тенденции используется мода, функция для ее расчета («Mode») присутствует в пакете «DescTools», она же используется для вычисления данного показателя и среди числовых значений — рис. 11 (листинг 10).

Листинг 10

```
Mode(df$Delivery_type)
## [1] "Spontaneous"
Mode(df$Gestational_age)
## [1] 40
# вычисление моды у всех категориальных перемен-
ных в таблице данных
# !sapply(df, is.numeric) позволяет отобрать не-
числовые переменные

apply(df_s[!sapply(df_s, is.numeric)], 2, Mode)
## Marital_status Delivery_type Infant_sex
## "Married" "Spontaneous" "Male"
```

Рис. 11. Расчет моды (листинг 10)

Помимо количественной характеристики мода может быть отражена графически. С помощью данной функции вычисляются значения распределения переменной, далее создается упорядоченная столбиковая диаграмма — рис. 12 (листинг 11).

Листинг 11

```
# функция
barplot_order <- function(x) {
  as.data.frame(table(na.omit(x))) %>%
  mutate(Var1 = factor(Var1, levels =
  Var1[order(Freq, decreasing = TRUE)])) %>%
  ggplot(aes(x = Var1, y = Freq)) +
  geom_bar(stat = 'identity') +
  geom_text(aes(label = Freq, y = Freq/2),
  colour = 'white', size = 3) +
  xlab('') + theme_minimal()
}

# Пример применения
barplot_order(df$Marital_status)
```

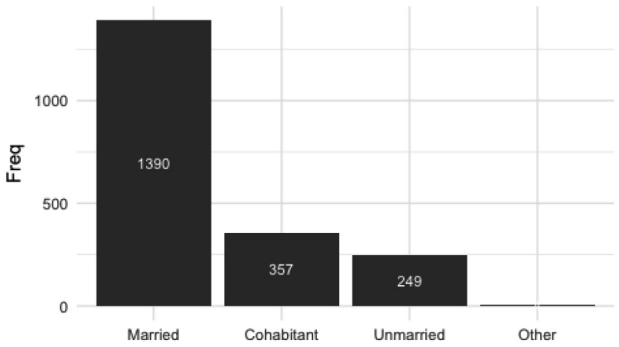


Рис. 12. Графическое представление моды (листинг 11)

Для категориальных переменных основным параметром описательной статистики является частота встречаемости данных категорий (например, доля мужчин и женщин в выборке, доля в выборке пациентов разных возрастных категорий, доля неблагоприятных исходов лечения и т. п.).

Базовыми функциями для оценки частотного распределения являются следующие:

— *table()* — создает таблицу сопряженности («contingency table», «crosstab»), причем порядок указания переменных в функции определяет макет таблицы;

— *addmargins()* — функция, добавляющая итоговые значения по столбцам и рядам;

— *prop.table()* — функция, определяющая доли значений;

— *ftable()* — функция, позволяющая выводить таблицы и включающая три и более переменные в более удобном для восприятия виде.

Применение функций с визуализацией данных для оценки частотного распределения для одной категориальной переменной продемонстрировано на рис. 13 (листинг 12).

Листинг 12

```
# одна переменная
ttable1 <- table(df_t$lowBirthWeight)

ttable1
##
##  no  yes
## 1835  138
addmargins(ttable1) # добавлена сумма
##
##  no  yes  Sum
```

```
## 1835 138 1973
prop.table(ttable1) # данные представлены в виде
долей
##
## no yes
## 0.93 0.07
```

Рис. 13. Применение функций с визуализацией данных для оценки частотного распределения для одной категориальной переменной (листинг 12)

В составе пакета «DescTools» присутствует функция *Freq()*, расширяющая возможности функции *table()* для одной переменной. Результатом выполнения этой функции является таблица данных, включающая в качестве столбцов перечень значений изучаемой категориальной переменной, количество значений («freq»), долю значения («perc»), кумулятивную сумму значений («cumfreq»), кумулятивную сумму долей («perc»). Порядок вывода значений переменной определяется аргументом функции *ord* – рис. 14 (листинг 13).

Листинг 13

```
Freq(df$Marital_status)
## level freq perc cumfreq cumperc
## 1 Unmarried 249 12.4% 249 12.4%
## 2 Married 1’390 69.5% 1’639 81.9%
## 3 Cohabitant 357 17.8% 1’996 99.8%
## 4 Other 4 0.2% 2’000 100.0%
Freq(df$Marital_status, ord = ‘desc’) # сортировка в порядке убывания
## level freq perc cumfreq cumperc
## 1 Married 1’390 69.5% 1’390 69.5%
## 2 Cohabitant 357 17.8% 1’747 87.3%
## 3 Unmarried 249 12.4% 1’996 99.8%
## 4 Other 4 0.2% 2’000 100.0%
Freq(df$Marital_status, ord = ‘asc’) # сортировка в порядке возрастания
## level freq perc cumfreq cumperc
## 1 Other 4 0.2% 4 0.2%
## 2 Unmarried 249 12.4% 253 12.7%
## 3 Cohabitant 357 17.8% 610 30.5%
## 4 Married 1’390 69.5% 2’000 100.0%
```

Рис. 14. Применение функций для оценки частотного распределения одной категориальной переменной (листинг 13)

Использование функций «apply» и «sapply» позволяет выполнить частотный анализ всех категориальных переменных в таблице данных – рис. 15 (листинг 14).

Листинг 14

```
apply(df_s[!sapply(df_s, is.numeric)], 2, Freq,
ord = ‘desc’)
## $Marital_status
## level freq perc cumfreq cumperc
## 1 Married 1’390 69.5% 1’390 69.5%
## 2 Cohabitant 357 17.8% 1’747 87.3%
## 3 Unmarried 249 12.4% 1’996 99.8%
## 4 Other 4 0.2% 2’000 100.0%
##
## $Delivery_type
## level freq perc cumfreq cumperc
## 1 Spontaneous 1’351 67.9% 1’351 67.9%
## 2 Caesarean section 423 21.3% 1’774 89.2%
## 3 Induced 215 10.8% 1’989 100.0%
##
## $Infant_sex
## level freq perc cumfreq cumperc
## 1 Male 1’033 51.6% 1’033 51.6%
## 2 Female 967 48.4% 2’000 100.0%
```

Рис. 15. Использование функций «apply» и «sapply» для анализа категориальных переменных (листинг 14)

Показатели частотного распределения можно также получить, используя функции пакета «tidyverse» – рис. 16 (листинг 15).

Листинг 15

```
df %>%
  select(Infant_sex, Birth_defect, lowBirthWeight,
  Preeclampsia, Delivery_type) %>%
  drop_na() %>%
  group_by(Preeclampsia, Birth_defect,
  Delivery_type) %>%
  summarise(n_count = n()) %>%
  ungroup() %>%
  mutate(portion = n_count/sum(n_count))
## # A tibble: 11 x 5
## Preeclampsia Birth_defect Delivery_type n_
count portion
## <fct> <fct> <fct> <int> <dbl>
## 1 no no Spontaneous 1276 0.647
## 2 no no Induced 194 0.0983
## 3 no no Caesarean section 365 0.185
## 4 no yes Spontaneous 49 0.0248
## 5 no yes Induced 2 0.00101
## 6 no yes Caesarean section 18 0.00912
## 7 yes no Spontaneous 15 0.00760
## 8 yes no Induced 14 0.00710
## 9 yes no Caesarean section 35 0.0177
## 10 yes yes Induced 3 0.00152
## 11 yes yes Caesarean section 2 0.00101
```

Рис. 16. Расчет показателей частотного распределения с использованием пакета «tidyverse» (листинг 15)

Как и для непрерывных переменных, для категориальных выборочных данных целесообразно проводить не только точечную, но и интервальную оценку – вычислять доверительный интервал. Пакет «DescTools» включает в себя функцию «BinomCI», позволяющую рассчитывать доверительные интервалы для биномиальных пропорций с использованием большинства популярных методов (Wald, Wilson, Agresti-Coull, Jeffreys, Clopper-Pearson и др.), причем необходимый метод расчета задается пользователем. Синтаксис данной функции: *BinomCI(x, n, conf.level = 0.95, method = c(“wilson”, “wald”, “agresti-coull”, “jeffreys”, “modified wilson”, “modified jeffreys”, “clopper-pearson”, “arcsine”, “logit”, “witting”, “pratt”), rand = 123)*, где x – количество «успехов» (значений переменной, для которой проводится оценка), а n – общее количество наблюдений. Пример использования данной функции приведен на рис. 17 (листинг 16).

Листинг 16

```
# например, необходимо вычислить 95% доверительный интервал для доли мальчиков (556) среди новорожденных (1050).

BinomCI(556, 1050)
```

```
## est lwr.ci upr.ci
## [1,] 0.53 0.499 0.56
# при работе с таблицей данных можно использовать функцию table

(t1 <- with(df, table(Infant_sex)))
## Infant_sex
## Female Male
## 967 1033
sum(t1)
## [1] 2000
t1[2]
## Male
## 1033
BinomCI(t1[2], sum(t1), method = 'wald')
## est lwr.ci upr.ci
## [1,] 0.516 0.495 0.538
```

Рис. 17. Использование функции «BinomCI» (листинг 16).

Для расчета доверительных интервалов также может быть создана пользовательская функция, включающая в себя функцию «BinomCI».

В пользовательской функции *ci_pr()* на входе указывается вектор (столбец таблицы данных), величина доверительного интервала (по умолчанию — 0,95) и метод расчета (по умолчанию — Wald). В результате использования функции рассчитываются значение доли, нижний и верхний уровни для выбранного уровня доверительной вероятности — рис. 18 (листинг 17).

Листинг 17

```
ci_pr <- function(x, conf_i = .95, methd =
'wald') {
 t1 <- table(x)
 m <- matrix(rep(0, 3 * length(t1)), ncol = 3,
 dimnames = list(names(t1),
 c('portion','low_level', 'upper_level')))
 for(i in 1 : length(t1)) {
 p_res <- BinomCI(t1[i], sum(t1), conf.level =
conf_i,
 method = methd)
 for(j in 1:3) m[i, j] <- p_res[j]
 }
 return(m)
}

# Примеры использования функции
ci_pr(df$Infant_sex)
## portion low_level upper_level
## Female 0.483 0.462 0.505
## Male 0.516 0.495 0.538
ci_pr(df$Education)
## portion low_level upper_level
## None 0.000501 0.000000 0.00148
## Primary (class 1-9) 0.063095 0.052431 0.07376
## Secondary (class 10-11) 0.150726 0.135034
0.16642
## Technical School 0.437656 0.415898 0.45941
## Higher education 0.345518 0.324662 0.36637
## Unknown 0.002504 0.000312 0.00470
```

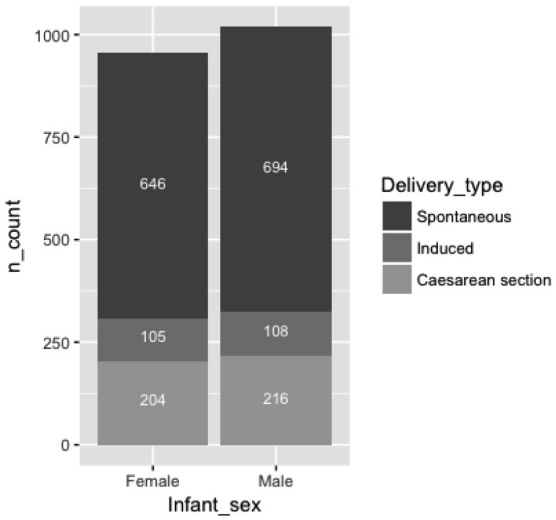
Рис. 18. Использование функции ci_pr() (листинг 17)

Стандартная задача при анализе результатов биомедицинских исследований — изучение частотного распределения двух категориальных переменных. С этой целью создается таблица сопряженности, которая является средством представления совместного распределения переменных, предназначенным для исследования связи между ними. Таблица сопряженности — наиболее универсальное средство изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения. Пример работы с таблицами сопряженности представлен на рис. 19 (листинг 18).

Листинг 18

```
# две переменных
ttable2 <- with(df_t, table(Delivery_type,
Infant_sex))

# визуализация
df_t %>%
 group_by(Infant_sex, Delivery_type) %>%
 summarise(n_count = n()) %>%
 arrange(Infant_sex, desc(Delivery_type)) %>%
 mutate(y_pos = cumsum(n_count) - .5 * n_
count) %>%
 ggplot(aes(x = Infant_sex, y = n_count,
 fill = Delivery_type)) +
 geom_bar(stat = 'identity') +
 geom_text(aes(label = n_count, y = y_pos),
 colour = 'white', size = 3, vjust = .2) +
 scale_fill_grey(start = .4, end = .7) +
 coord_fixed(ratio = 1/300)
```




```
ttable2
## Infant_sex
## Delivery_type Female Male
## Spontaneous 646 694
## Induced 105 108
## Caesarean section 204 216
# добавлены суммы по рядам и столбцам
addmargins(ttable2)
## Infant_sex
## Delivery_type Female Male Sum
## Spontaneous 646 694 1340
## Induced 105 108 213
## Caesarean section 204 216 420
## Sum 955 1018 1973
# добавлены суммы по столбцам
addmargins(ttable2, 1)
## Infant_sex
## Delivery_type Female Male
## Spontaneous 646 694
## Induced 105 108
## Caesarean section 204 216
## Sum 955 1018
# добавлены суммы по рядам
```

```
addmargins(ttable2, 2)
## Infant_sex
## Delivery_type Female Male Sum
## Spontaneous 646 694 1340
## Induced 105 108 213
## Caesarean section 204 216 420
# значения в долях
prop.table(ttable2)
## Infant_sex
## Delivery_type Female Male
## Spontaneous 0.3274 0.3517
## Induced 0.0532 0.0547
## Caesarean section 0.1034 0.1095
# добавление суммарных значений для долей
addmargins(prop.table(ttable2))
## Infant_sex
## Delivery_type Female Male Sum
## Spontaneous 0.3274 0.3517 0.6792
## Induced 0.0532 0.0547 0.1080
## Caesarean section 0.1034 0.1095 0.2129
## Sum 0.4840 0.5160 1.0000
# значения в долях по рядам
prop.table(ttable2, 1)
## Infant_sex
## Delivery_type Female Male
## Spontaneous 0.482 0.518
## Induced 0.493 0.507
## Caesarean section 0.486 0.514
# значения в долях по столбцам
prop.table(ttable2, 2)
## Infant_sex
## Delivery_type Female Male
## Spontaneous 0.676 0.682
## Induced 0.110 0.106
## Caesarean section 0.214 0.212
# визуализация
df_t %>%
 group_by(Infant_sex, Delivery_type) %>%
 summarise(n_count = n()) %>%
 mutate(ssum = sum(n_count),
 prop_n = round(n_count/ssum,3),
 lab_name = paste0(as.character(prop_n *
100),'%')) %>%
 ggplot(aes(x = Infant_sex, y = n_count, fill =
Delivery_type)) +
 geom_bar(position = 'fill', stat = 'identity') +
 geom_text(aes(label = lab_name, y = prop_n),
 vjust = 1.8, position = 'stack', size = 3,
 colour = 'white') +
 scale_y_continuous(labels = scales::percent) +
 scale_fill_grey(start = .2, end = .7) +
 coord_fixed(ratio = 3)
```

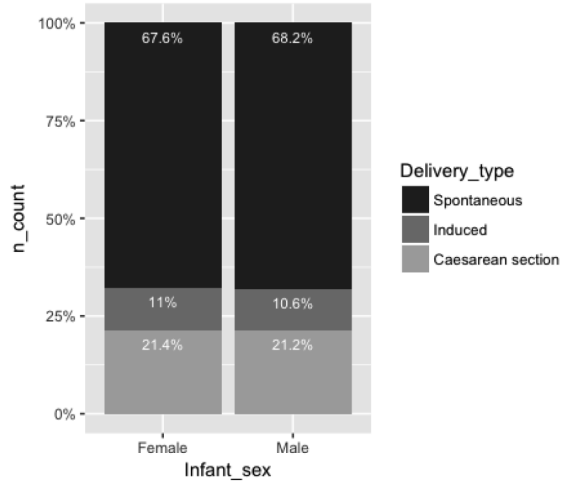


Рис. 19. Описательная статистика и визуализация данных в таблице сопряженности (листинг 18)

Имеющиеся в R функции *table()*, *prop.table()*, *addmargins()* позволяют также изучать взаимодействие трех и более переменных — рис. 20 (листинг 19).

```
Листинг 19
# три переменных
ttable3 <- with(df_t, table(Infant_sex, Delivery_
type,
 lowBirthWeight))

ftable(ttable3)
## lowBirthWeight no yes
## Infant_sex Delivery_type
## Female Spontaneous 610 36
## Induced 96 9
## Caesarean section 165 39
## Male Spontaneous 670 24
## Induced 106 2
## Caesarean section 188 28
ftable(prop.table(ttable3))
## lowBirthWeight no yes
## Infant_sex Delivery_type
## Female Spontaneous 0.30917 0.01825
## Induced 0.04866 0.00456
## Caesarean section 0.08363 0.01977
## Male Spontaneous 0.33958 0.01216
## Induced 0.05373 0.00101
## Caesarean section 0.09529 0.01419
# четыре переменных

ttable4 <- with(df_t, table(Birth_defect, Infant_
sex,
 Delivery_type, lowBirthWeight))
ftable(ttable4)
## lowBirthWeight no yes
## Birth_defect Infant_sex Delivery_type
## no Female Spontaneous 596 34
## Induced 94 9
## Caesarean section 155 37
## Male Spontaneous 639 22
## Induced 103 2
## Caesarean section 184 24
## yes Female Spontaneous 14 2
## Induced 2 0
## Caesarean section 10 2
## Male Spontaneous 31 2
## Induced 3 0
## Caesarean section 4 4
```

Рис. 20. Оценка таблицы сопряженности с тремя переменными (листинг 19)

Как и в случае с непрерывными переменными, функция «summary» позволяет получить сразу большой набор показателей описательной статистики и для категориальных переменных. Если для непрерывных данных с использованием этой функции рассчитываются среднее арифметическое, медиана, первый и третий квартили, минимальное и максимальное значение, то для категориальных данных оценивается частотное распределение. При этом указывается количество «NA» в переменных — рис. 21 (листинг 20).

```
Листинг 20
options(digits = 3)
summary(df_s)
## Maternal_age Gestational_age BIrthweight
Marital_status
## Min. :15.0 Min. :24.0 Min. : 620 Unmarried
: 249
```

```
## 1st Qu.:25.0 1st Qu.:38.0 1st Qu.:3090
Married :1390
## Median :28.0 Median :39.0 Median :3430
Cohabitant: 357
## Mean :28.6 Mean :38.8 Mean :3371 Other : 4
## 3rd Qu.:32.0 3rd Qu.:40.0 3rd Qu.:3740
## Max. :46.0 Max. :42.0 Max. :5930
## NA's :12 NA's :1
## Delivery_type Infant_sex
## Spontaneous :1351 Female: 967
## Induced : 215 Male :1033
## Caesarean section: 423
## NA's : 11
##
```

Рис. 21. Использование функции «summary» для оценки таблицы сопряженности (листинг 20)

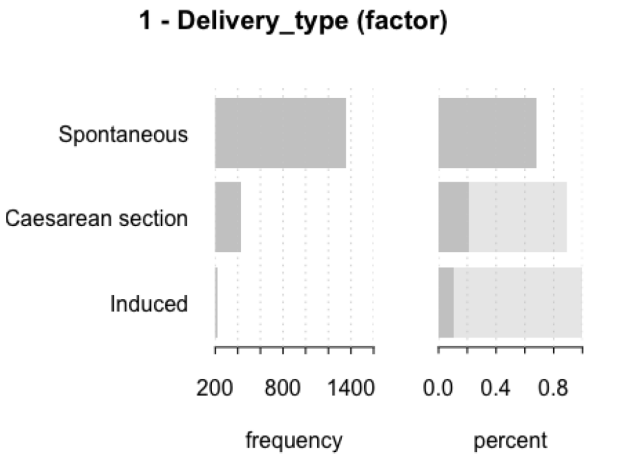
Выполнить описательную статистику данных можно и с помощью функции «Desc» пакета «DescTools». Данная функция позволяет выводить суммарные показатели и диаграммы — рис. 22 (листинг 21).

В результате использования данной функции выводятся значения частоты признака переменной в числах и процентах, а также кумулятивные показатели. Графики при этом представлены полосковыми диаграммами.

Заметим, что в отличие от категориальных переменных для числовых показателей в результате использования этой функции выводятся значения среднего арифметического и 95 % доверительного интервала, стандартное отклонение, коэффициент вариации, показатели асимметрии, эксцесса, квантили, пять высших и пять низших значений. При этом отображаются гистограмма с кривой плотности, коробочная диаграмма и кумулятивная диаграмма.

Листинг 21

```
Desc(df_s[, c('Delivery_type', 'Maternal_age')])
## ------------------------------------------------
------------------------
## Describe df_s[, c("Delivery_type", "Maternal_
age")] (data.frame):
##
## data.frame: 2000 obs. of 2 variables
##
## Nr ColName Class NAs Levels
## 1 Delivery_type factor 11 (0.5%) (3):
1-Spontaneous,
## 2-Induced, 3-Caesarean
## section
## 2 Maternal_age numeric .
##
##
## ------------------------------------------------
------------------------
## 1 - Delivery_type (factor)
##
## length n NAs unique levels dupes
## 2'000 1'989 11 3 3 y
## 99.5% 0.5%
##
## level freq perc cumfreq cumperc
## 1 Spontaneous 1'351 67.9% 1'351 67.9%
## 2 Caesarean section 423 21.3% 1'774 89.2%
## 3 Induced 215 10.8% 1'989 100.0%
```




```
/2018-06-22
## ------------------------------------------------
------------------------
## 2 - Maternal_age (numeric)
##
## length n NAs unique 0s mean meanCI
## 2'000 2'000 0 31 0 28.60 28.37
## 100.0% 0.0% 0.0% 28.83
##
## .05 .10 .25 median .75 .90 .95
## 20.00 22.00 25.00 28.00 32.00 36.00 38.00
##
## range sd vcoef mad IQR skew kurt
## 31.00 5.31 0.19 5.93 7.00 0.21 -0.46
##
## lowest : 15.0, 16.0 (3), 17.0 (8), 18.0
(19), 19.0 (30)
## highest: 41.0 (12), 42.0 (8), 43.0 (3),
44.0, 46.0
```

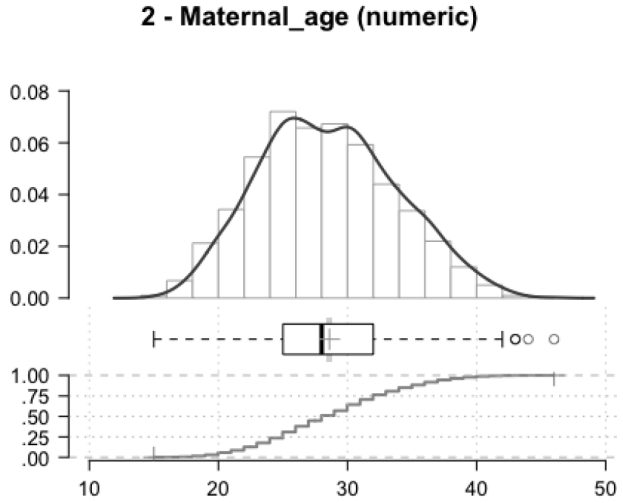


Рис. 22. Использование функции «Desc» для расчета показателей описательной статистики и визуализации категориальных и непрерывных данных (листинг 21)

В заключение перечислим возможности некоторых дополнительных пакетов для получения показателей описательной статистики в среде R, так как они позволяют рассчитывать большее количество показателей описательной статистики, в том числе и в подгруппах. Примером такого пакета является пакет «DescTools», который был представлен выше.

Другой пакет – «psych» – был создан для работы в экспериментальной психологии. Для описательной статистики в нем используются функции «describe» и «describeBy», позволяющие выводить большее количество показателей описательной статистики. Функция «describeBy» при этом дает возможность выводить данные в подгруппах, разделение на которые проводится по значениям категориальных переменных, при этом допускается группировка по нескольким категориальным переменным.

В другом пакете «broom» используется функция «tidy», работающая аналогично функции «describe» пакета «psych».

Функция «CreateTableOne» пакета «tableone» дает возможность вывести показатели описательной статистики для переменных любого типа, включенных в таблицу данных. При этом возможна стратификация на подгруппы при использовании параметра strata. По умолчанию сравнительная оценка проводится с применением непараметрических тестов.

Дополнительная информация по применению методов описательной статистики в среде R может быть получена при изучении технической документации, пособий и доступных интернет-ресурсов, содержащих большое количество примеров использования программной среды R для этой цели.

Список литературы

1. *Гржибовский А. М., Унгуряну Т. Н., Горбатова М. А.* Описательная статистика с использованием пакетов статистических программ SPSS и STATA // Наркология. 2017. № 4. С. 36–51.

2. *Кабаков Р. И.* R в действии. Анализ и визуализация данных в программе R / пер. с англ. П. А. Волковой. М.: ДМК Пресс, 2014. 588 с.

3. *Мастицкий С. Э., Шитиков В. К.* Статистический анализ и визуализация данных с помощью R. М.: ДМК Пресс, 2015. 496 с.

4. R: анализ и визуализация данных. URL: http://r-analytics.blogspot.ru. (дата обращения: 18.06.1018).

5. *Усынина А. А., Одланд Й. О., Пылаева Ж. А., Пастбина И. М., Гржибовский А. М.* Регистр родов Архангельской области как важный информационный ресурс для науки и практического здравоохранения // Экология человека. 2017. № 2. С. 58–64.

6. *Холматова К. К., Харькова О. А., Гржибовский А. М.* Классификация научных исследований в здравоохранении // Экология человека. 2016. № 1. С. 57–64.

7. *Bilder R. C., Loughin T. M.* Analysis of Categorical Data using R. CRC Press, 2015.

8. *Crawley M. J.* Statistics. An Introduction using R. 2nd ed. Wiley, 2015.

9. *Crawley M. J.* The R Book. 2nd ed. Wiley, 2013.

10. *Dalgaard P.* Introductory Statistics with R. 2nd ed. Springer, 2008.

11. Function of DescTool. URL: https://www.rdocumentation.org/packages/DescTools/versions/0.99.19 (дата обращения 18.06.2018).

12. *Grolemund G., Wickham H.* R for data science. URL: http://r4ds.had.co.nz (дата обращения: 18.06.2018).

13. R - Mean, Median and Mode. URL: https://www.tutorialspoint.com/r/r_mean_median_mode.htm (дата обращения: 18.06.2018).

14. RStudio Cheat Sheet. URL: https://www.rstudio.com/resources/cheatsheets/ (дата обращения: 18.06.2018).

15. *Stewart A.* Basic Statistics and Epidemiology. Practical Guide, 4 edition. CRC Press, 2016.

References

1. Grjibovski A. M., Unguryanu T. N., Gorbatova M. A. Descriptive statistics using SPSS and STATA software. *Narkologiya* [Narcology]. 2017, 4, pp. 36-51. [In Russian]

2. Kabakov R. I. *R v deystvii. Analiz i vizualizaciya dannyh v programme R* [R in action: data analysis and visualization using R software], per. s angl. P. A. Volkova. Moscow, 2014, 588 p.

3. Mastickiy S. E., Shitikov V. K. *Statisticheskiy analiz i vizualizaciya dannyh s pomoshch'yu R* [Data statistical analysis using R]. Moscow, 2015, 496 p.

4. R: analysis and data visualization. Available at: http://r-analytics.blogspot.ru (accessed: 18.06.2018)

5. Usynina A. A., Odland Jon Øyvind, Pylaeva Zh. A., Pastbina I. M., Grjibovski A. M. Arkhangelsk County Birth Registry as an Inportant Source of Information for Research and Healthcare. *Ekologiya cheloveka* [Human Ecology]. 2017, 2, pp. 58-64. [In Russian]

6. Kholmatova K. K., Kharkova O. A., Grjibovski A. M. Types of Research in Health Sciences. *Ekologiya cheloveka* [Human Ecology]. 2016, 1, pp. 57-64. [In Russian]

7. Bilder R. C., Loughin T. M. *Analysis of Categorical Data using R*. CRC Press, 2015.

8. Crawley M. J. *Statistics. An Introduction using R*. 2nd rd. Wiley, 2015.

9. Crawley M. J. *The R Book*. 2nd ed. Wiley, 2013.

10. Dalgaard P. *Introductory Statistics with R*. 2nd ed. Springer, 2008.

11. Function of DescTool. Available at: https://www.rdocumentation.org/packages/DescTools/versions/0.99.19 (accessed: 21.06.2018)

12. Grolemund G., Wickham H. R for data science. Available at: http://r4ds.had.co.nz (accessed: 18.06.2018)

13. R - Mean, Median and Mode. Available at: https://www.tutorialspoint.com/r/r_mean_median_mode.htm (accessed: 18.06.2018)

14. RStudio Cheat Sheet. Available at: https://www.rstudio.com/resources/cheatsheets/ (accessed: 18.06.2018)

15. Stewart A. *Basic Statistics and Epidemiology*. Practical Guide, 4 edition. CRC Press, 2016.

Контактная информация:

Гржибовский Андрей Мечиславович – доктор медицины, заведующий ЦНИЛ Северного государственного медицинского университета, г. Архангельск; профессор Северо-Восточного федерального университета, г. Якутск; почетный доктор Международного казахско-турецкого университета, г. Туркестан (Казахстан); почетный профессор Государственного медицинского университета г. Семей (Казахстан)

Адрес: 163000 г. Архангельск, Троицкий пр., д. 51

E-mail: Andrej.Grjibovski@gmail.com