УДК 615.036.2

ВИЗУАЛИЗАЦИЯ БИОМЕДИЦИНСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОЙ СРЕДЫ R

© 2018 г. [1]В. Л. Егошин, [2]С. В. Иванов, [3]Н. В. Саввина, [4]С. Б. Калмаханов, [3,5]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

В статье представлены основные принципы работы программной среды R в применении к визуализации исследовательских данных. Описаны основные типы диаграмм, алгоритмы их создания, особенности работы с графиками различных видов в применении к различным типам данных.

Ключевые слова: статистических анализ данных, визуализация, диаграммы, R

VISUALIZATION OF BIOMEDICAL DATA USING R

[1]V. L. Egoshin, [2]S. V. Ivanov, [3]N. V. Savvina, [4]S.B. Kalmakhanov, [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

The paper presents basic principles of using R software for visualization of biomedical research data. Basic types of graphs and algorithms for graph creation are presented. Specification of using different graph types in implementation of different data types is described.

Key words: statistical analysis, visualization, graphs, R

Библиографическая ссылка:

Егошин В. Л., Иванов С. В., Саввина Н. В., Калмаханов С. Б., Гржибовский А. М. Визуализация исследовательских данных с использованием программной среды R // Экология человека. 2018. № 8. С. 52–64.

Egoshin V. L., Ivanov S. V., Savvina N. V., Kalmakhanov S. B., Grjibovski A. M. Visualization of Biomedical Data Using R. *Ekologiya cheloveka* [Human Ecology]. 2018, 8, pp. 52-64.

Визуализация является важной составной частью анализа данных независимо от вида проводимого исследования [6]. Американский математик John W. Tukey, известный своими работами в области анализа данных, разработавший концепцию разведочного анализа данных («Exploratory Data Analysis») и создавший диаграмму «boxplot» («ящик с усами»), считал, что «величайшей ценностью графика является возможность увидеть то, чего мы не ожидали увидеть» [14]. По мнению другого известного специалиста в области анализа данных John Chambers, «нет статистического метода более мощного, чем хорошо подобранный график» [7].

На этапе первоначального знакомства с данными, собственно и называемого разведочным анализом, визуализация может и должна предшествовать выполнению статистических тестов и созданию моделей. На данном этапе результаты визуальной оценки могут вести к продолжению подготовительной работы с данными, например к их трансформации для более эффективного проведения анализа.

Именно визуализация позволяет всецело рассмотреть данные, и исследователь должен владеть умением делать данные видимыми и уметь видеть их. Визуализация данных органично дополняет и делает воспринимаемой читателем описательную статистику исследовательских данных [1].

Среди компьютерных программ, использующихся для анализа и визуализации данных, именно программная среда R известна своими широкими графическими возможностями.

Nathan Yau назвал R излюбленным инструментом для создания информационной графики, «... если вы примете особенности R, перед вами раскроются большие возможности. Вы сможете делать графику типографского качества... и полюбите гибкость R» [4].

Для визуализации в R могут использоваться возможности базовых и дополнительных пакетов. Так, пакет «ggplot2», являющийся частью пакета «tidyverse», представляет собой возможность визуализации как «Grammar of Graphics». В данной концепции создание графиков рассматривается через добавление слоев, каждый их которых выполняет свои функции: представление данных и эстетическое отображение, включающее переменные по осям x и y, цвет, размер

и форму выводимых в поле графика объектов, сами геометрические объекты, статистические трансформации, представление шкал, координатных систем, фасетирование (разделение графиков на несколько субграфиков, создаваемых по одному принципу, для подгрупп этой же переменной/переменных), создание тем диаграмм [16].

Выбор геометрического объекта в «ggplot2» может быть выполнен по схеме, представленной в табл. 1.

Таблица 1

Выбор геометрического объекта для визуализации данных в R

Переменные	Геометрический объект в «ggplot2»	Применяется для изучения
Одна непрерывная переменная	geom_histogram, geom_density	Распределение непрерывной переменной
Одна дискретная переменная	geom_bar	Распределение частот дискретной переменной
Две переменные: x – непрерывная, y – непрерывная	geom_point, geom_jitter, geom_smooth, geom_text	Связь двух непрерывных переменных
Две переменные: x – дискретная, y – непрерывная	geom_boxplot, geom_violin	Распределение непрерывной переменной в подгруппах
Две переменные: x – дискретная, y – дискретная	geom_count, geom_jitter	Распределение частот двух дискретных переменных
Три переменные	geom_tile	Связь трех переменных
Показатели в разные промежутки времени	geom_line	Временные ряды

Разумеется, возможный спектр геометрических объектов не исчерпывается теми, которые указаны в данной таблице.

В зависимости от цели визуализации используемые средства могут отличаться. Так, при выполнении разведочного анализа данных акцент делается на геометрические объекты, отображение подгрупп данных, в то время как оформлению графика уделяется меньшее внимание. Напротив, при подготовке текста публикации, научного труда или презентации в график будут вводиться элементы коммуникационного характера: изменение шкал, подписи на графиках, выбор цветового решения и т. п. [11].

Для демонстрации приемов работы с пакетом «ggplot2» будет использоваться случайная выборка (с небольшими изменениями значений) Архангельского областного регистра родов [5].

Для визуализации данных в R наряду с «ggplot2» используются и другие пакеты – «tidyverse», «ggthemes», «RColorBrewer», «grid», «gridExtra».

Разумеется, перед проведение визуализации данных они должны быть импортированы в систему. Для импорта файла с данными в формате .sav потребуется функция из пакета «foreign», как показано на рис. 1 (листинг 1).

Листинг 1

```
library(foreign)
df <- read.spss("Simulated_sample.sav", to.data.
frame = TRUE)
# в набор данных будет добавлен столбец
df <-df %>%
mutate(M_age_group =factor(cut(Maternal_age,
breaks =c(14, 18, 30 , 40, 50),
labels =c('<18', '18-30', '31-40', '>40'))))
```

Рис. 1. Импорт данных (листинг 1)

Графическое представление непрерывных переменных

Для изучения распределения непрерывной переменной традиционно наиболее часто используется гистограмма. При этом особое значение в гистограмме имеет такой параметр, как «binwidth» (ширина «контейнера»), определяющий объединение близких значений изучаемого показателя.

Сравните гистограмму со значением параметра binwidth = 100 (рис. 2 – листинг 2) и binwidth = 200 (рис. 3 – листинг 3). На рисунках наглядно определяются различия между двумя данными гистограммами.

Листинг 2

```
ggplot(df, aes(BIrthweight)) +
geom_histogram(binwidth =100,
colour ='grey50')
# binwidth задает размер "контейнера" (bin)
```

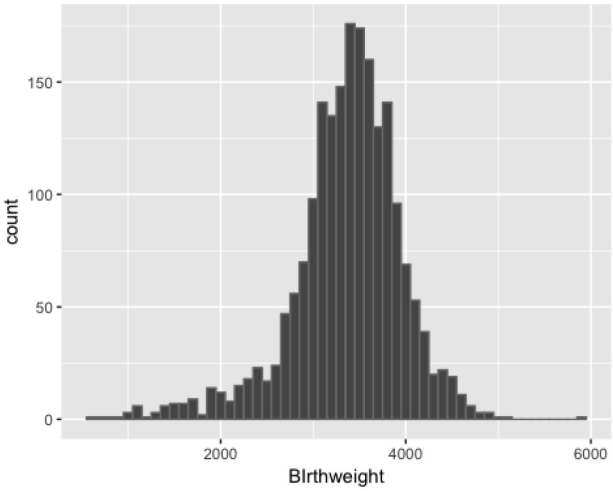


Рис. 2. Гистограмма со значением параметра binwidth = 100 (листинг 2)

Для улучшения восприятия гистограммы может быть использован параметр «colour» (*color, col*), который определяет цвет границы «контейнера», и параметр «fill», который определяет «заливку» площади. Название по оси x берется из названия переменной (столбца), название по оси y определяется самой функцией. Тема *theme_minimal()* удаляет задний фон и линии осей на диаграмме.

Дополнительная информация об изучаемой непрерывной переменной может быть получена при визуализации разделения данных на подмножества в соответствии с категориальными переменными (пол, возрастная группа, уровень образование и т. п.). Данное действие может быть выполнено путем присвоения названий категориальных переменных таким

Листинг 3

```
ggplot(df, aes(BIrthweight)) +
geom_histogram(binwidth =200, colour ='grey50',
fill ='white') +
theme_minimal()
```

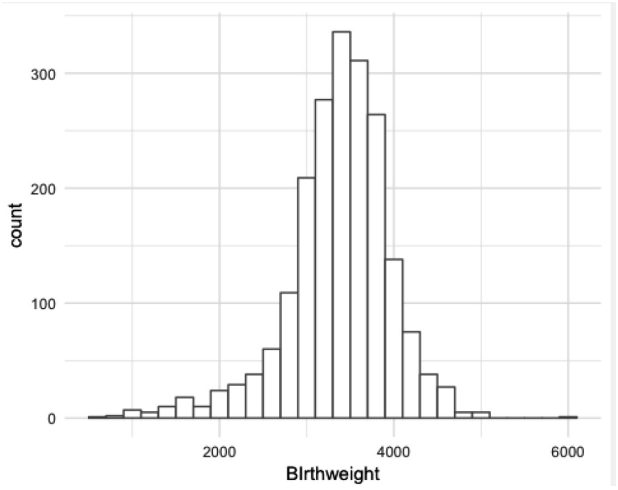


Рис. 3. Гистограмма со значением параметра binwidth = 200 (листинг 3)

параметрам, как «colour», «fill», «shape», «size», «linetype», или при проведении фасетирования. Так, листинг 4 (рис. 4) демонстрирует использование параметра «fill» для выделения на графике значений, соответствующих полу новорожденного.

Листинг 4

```
df %>%
ggplot(aes(BIrthweight, fill = Infant_sex)) +
geom_histogram(binwidth =100, colour ='grey50') +
scale_fill_grey() +
theme_grey() +
labs(x ='Birth Weiglt',
title ='Birth Weight',
subtitle ='divided by Age Group',
caption ='From Regional Register',
fill ='Infant sex')
```

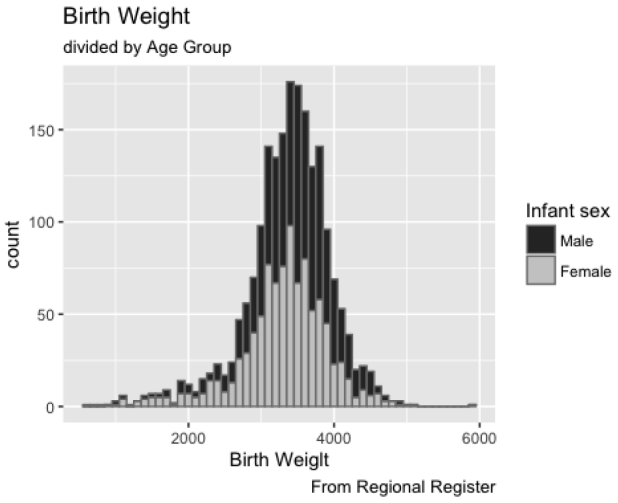


Рис. 4. Использование параметра «fill» для разделения значений по гендерному признаку (листинг 4)

Функция «labs» позволяет вывести на рисунок подписи к графику: x — название переменной по оси x, y — название переменной по оси y, «title» — заголовок, «subtitle» — подзаголовок, «caption» — надпись, которая часто используется для указания источника данных. В используемом примере функция «fill» задает название для легенды графика.

Листинг 5 (рис. 5) и листинг 6 (рис. 6) — примеры использования фасетирования — разделения графического представления данных по различным качественным признакам.

Листинг 5

```
df %>%
ggplot(aes(BIrthweight, fill = Infant_sex)) +
geom_histogram(binwidth =100, colour ='grey50') +
facet_wrap( ~Year_of_birth, ncol =1) +
scale_fill_grey() +
theme_bw() +
labs(x ='Вес в граммах',
y ='',
title ='Вес ребёнка',
subtitle ='Распределение по годам и полу ребенка',
caption ='Данные областного регистра',
fill ='Пол ребёнка')
```

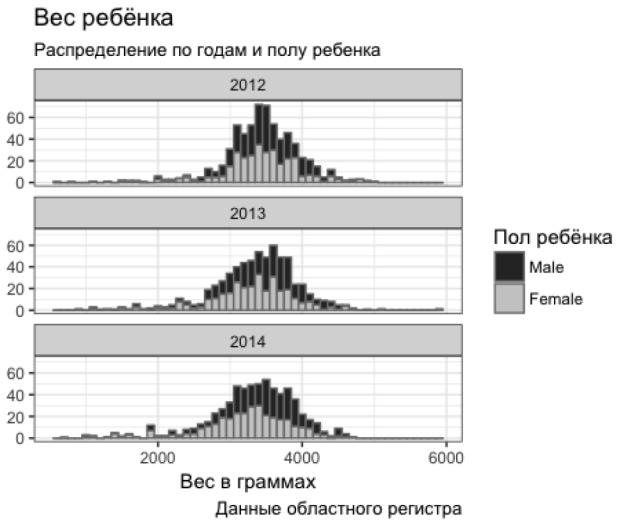


Рис. 5. Фасетирование по признакам — год и пол ребенка (листинг 5)

Листинг 6

```
df %>%
filter(M_age_group %in%c('18-30', '31-40')) %>%
ggplot(aes(BIrthweight)) +
geom_histogram(binwidth =100, colour ='grey50') +
facet_grid(Year_of_birth ~M_age_group) +
scale_fill_grey() +
theme_bw() +
labs(x ='Вес в граммах',
y ='',
title ='Вес ребёнка у матерей 18-30 и 31-40
лет',
subtitle ='Распределение по годам и возрастной
группе матери',
caption ='Данные областного регистра')
```

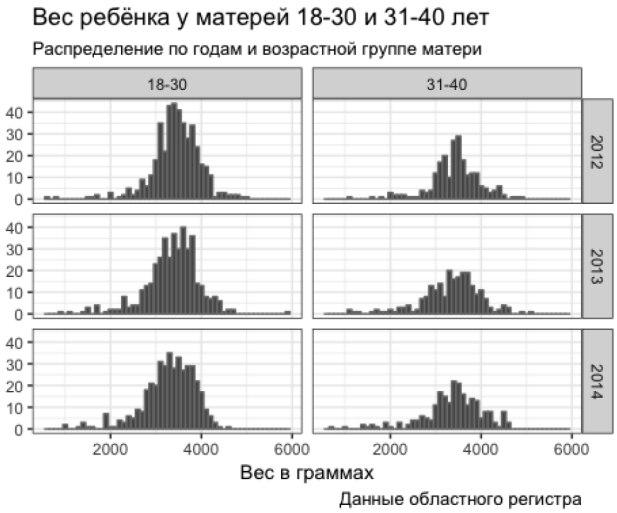


Рис. 6. Фасетирование по признакам — год и возрастная группа матерей (листинг 6)

Помимо гистограммы для представления непрерывных переменных может быть использована диаграмма плотности. Она подобна гистограмме, но если в гистограмме высота «контейнера» (значение y на диаграмме) равна количеству случаев с величиной значения, входящей в данный промежуток, то в диаграмме плотности значение y определяется как «PDF» («probability density function») — рис. 7 (листинг 7).

Листинг 7

```
ggplot(df, aes(BIrthweight)) +
geom_density(aes())
```

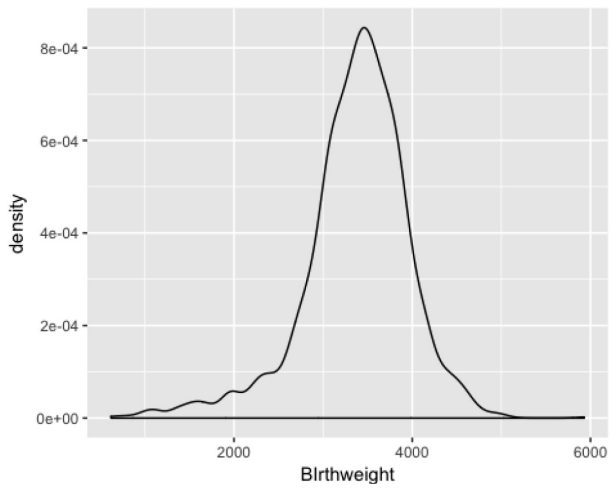


Рис. 7. Пример диаграммы плотности (листинг 7)

Пакет «gridExtra» позволяет выводить несколько диаграмм как один рисунок. Для этого необходимо создать графики, сохранить их в переменные, а потом «объединить» в один график функцией данного пакета «grid.arrange».

Выведем на одном рис. 8 три диаграммы: гистограмму, диаграмму плотности и наложение диаграммы плотности на гистограмму (листинг 8).

Листинг 8

```
library(gridExtra)
gg1 <-ggplot(df, aes(y = ..density..,
BIrthweight)) +
geom_histogram(binwidth =200, fill ='white',
colour ='grey50') +
labs(x ='', y ='', title ='Histogram') +
theme_bw()

gg2 <-ggplot(df, aes(y = ..density..,
BIrthweight)) +
geom_density(aes(y = ..density..)) +
labs(x ='', y ='', title ='Density plot') +
theme_bw()

gg3 <-ggplot(df, aes(y = ..density..,
BIrthweight)) +
geom_histogram(binwidth =200, fill ='white',
colour ='grey50') +
geom_density(aes(y = ..density..)) +
labs(x ='', y ='', title ='Histogram & Density
plot') +
theme_bw()

grid.arrange(gg1, gg2, gg3, nrow =1)
```

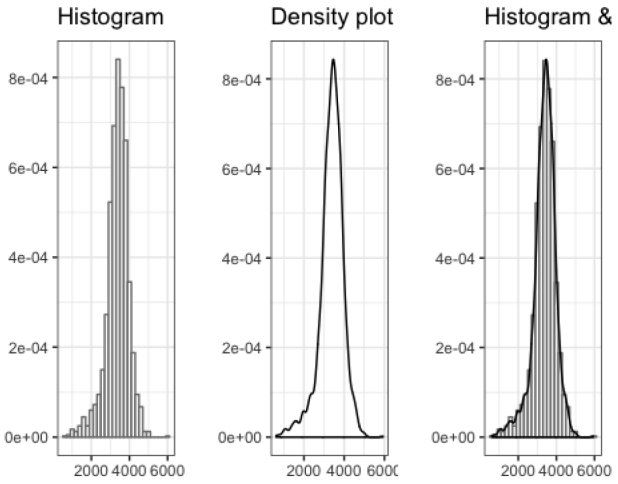


Рис. 8. Использование функции «grid.arrange» для объединения нескольких графиков на одном рисунке (листинг 8)

При использовании диаграмм плотности их также можно разделять в соответствии со значениями категориальных переменных, используя заливку (рис. 9 — листинг 9) или изменяя линию контура (цвет, форма), как это показано на рис. 10 (листинг 10). При этом параметр «scale_fill_grey» используется для вывода графика в черно-белых тонах. Легенда размещается в поле графика, позиция легенды задается как вектор значений x, y в относительных величинах.

Листинг 9

```
df %>%select(BIrthweight, Delivery_type) %>%
na.omit() %>%
ggplot(aes(BIrthweight, fill = Delivery_type)) +
geom_density() +
scale_fill_grey() +
theme_bw() +
theme(legend.position =c(.8, .8))
```

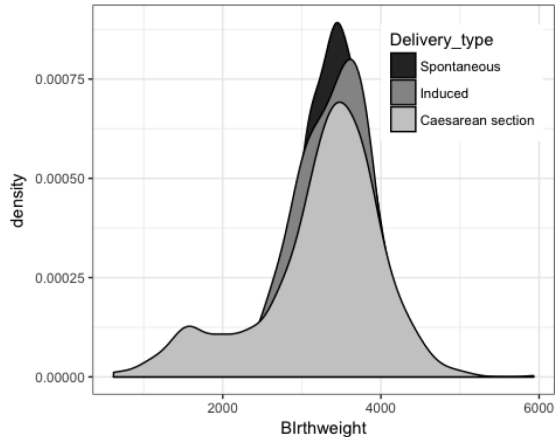


Рис. 9. Разделение диграммы плотности с использованием заливки (листинг 9)

Листинг 10

```
df %>%select(BIrthweight, Delivery_type) %>%
na.omit() %>%
ggplot(aes(BIrthweight, linetype = Delivery_type)) +
geom_density() +
theme_bw() +
theme(legend.position =c(.8, .8))
```

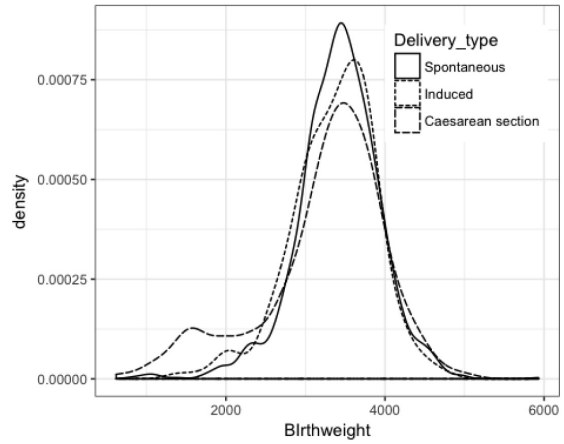


Рис. 10. Разделение диаграммы плотности с использованием линии контура (листинг 10)

Следует отметить, что выделение значений переменной «Delivery_type» на графике достигнуто выбором линий разного типа, и для такого решения предпочтительным будет использование цвета линии.

Помимо вышеописанных гистограммы и диаграммы распределения для визуализации непрерывных переменных могут быть использованы диаграмма «boxplot» («ящик с усами») и скрипичная диаграмма («violin plot»). Данные виды диаграмм рекомендуются к использованию совместно с категориальными (дискретными) переменными, хотя их можно использовать и самостоятельно [8]. Пример формирования данных диаграмм представлен на рис. 11 (листинг 11).

Листинг 11

```
gb <-ggplot(df, aes(x =1, y = BIrthweight)) +
geom_boxplot() +
theme_minimal() +
labs(x ='', y ='', title ='Boxplot')

gv <-ggplot(df, aes(x =1, y = BIrthweight)) +
geom_violin(draw_quantiles =c(.25, .5, .75)) +
theme_minimal() +
labs(x ='', y ='', title ='Violin plot')

grid.arrange(gb, gv, nrow =1)
```

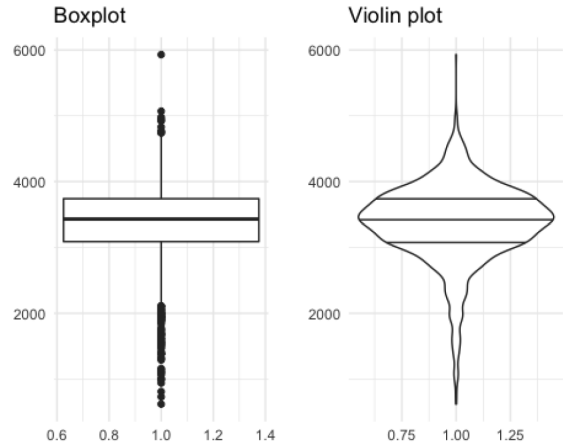


Рис. 11. Примеры диаграммы «boxplot» и скрипичной диаграммы (листинг 11)

Аргумент draw_quantiles = c(0.25, 0.5, 0.75) в скрипичной диаграмме позволяет отобразить на графике разделение на задаваемые квантили. В коробочной диаграмме использование аргумента varwidth = TRUE позволяет делать ширину «коробки» на диаграмме прямо пропорциональной количеству наблюдений. При этом строки с показателями «None», «Unknown» столбца Education не использованы в графике в связи с малым количеством значений в подгруппах (рис. 12, листинг 12).

Листинг 12

```
df %>%
select(Education, BIrthweight) %>%
na.omit() %>%
filter(Education != 'None'&Education != 'Unknown'
) %>%
ggplot(aes(Education, BIrthweight)) +
geom_boxplot(varwidth =TRUE) +
coord_flip() +# поворот оси на 90 градусов
theme_minimal()
```

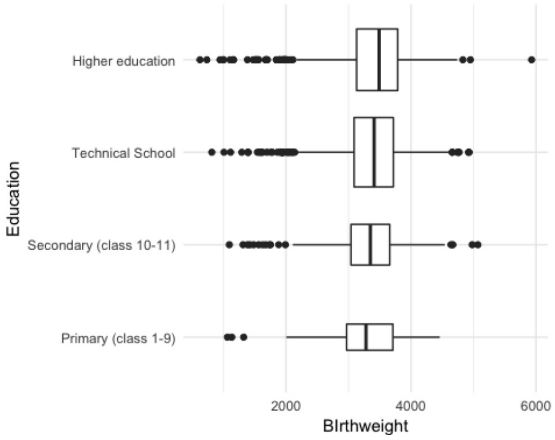


Рис. 12. Разделение диаграммы «boxplot» (листинг 12)

Скрипичная диаграмма аналогична квартильной, при этом она может предоставить больше информации о распределение непрерывной переменной. На рис. 13 (листинг 13) показана связь срока окончания беременности, уровня образования и возрастной группы.

Листинг 13

```
df %>%
select(Education, Gestational_age, M_age_
group) %>%
na.omit() %>%
filter(Education != 'None'&Education != 'Unknown'
) %>%
ggplot(aes(Education, Gestational_age)) +
geom_violin(draw_quantiles =c(0.25, 0.5, 0.75)) +
coord_flip() +
facet_wrap(~M_age_group) +
theme_minimal()
```

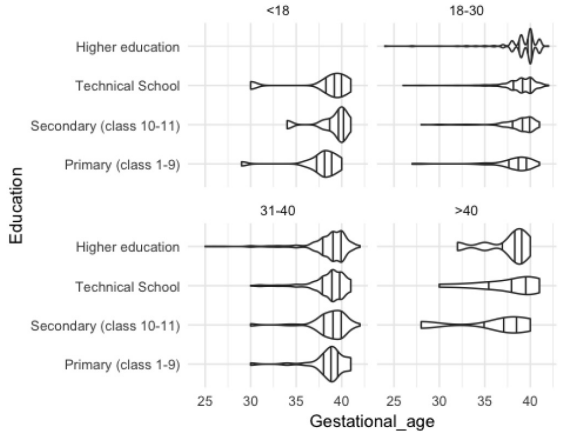


Рис. 13. Разделение скрипичной диаграммы (листинг 13)

Графическое представление дискретных переменных

Для изучения распределения частот дискретной переменной наиболее предпочтительным является использование столбиковых диаграмм.

Столбиковые диаграммы наглядны, просты для понимания, в данных диаграммах высота столбца (значение по оси y) равна количеству значений признака, т. е. частоты встречаемости в выборке или популяции (рис. 14, листинг 14).

Как и в приведенных выше вариантах представления данных, в отношении столбиковой диаграммы также возможно разделение на подгруппы в соответствии с другими категориальными переменными, проводимое методом позиционирования. Для этого используется параметр «position» со значениями «stack» (применяется по умолчанию – рис. 15, листинг 15), «dodge» (рис. 16, листинг 16), «fill» (рис. 17, листинг 17). При этом столбиковая диаграмма со значением параметра position = "fill" создает диаграмму, в которой учитываются пропорциональные отношения.

Листинг 14

```
ggplot(df, aes(Gestational_age)) +
geom_bar()
```

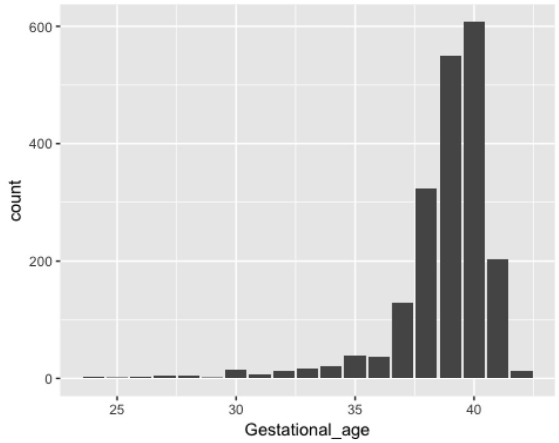


Рис. 14. Столбиковая диаграмма (листинг 14)

Листинг 15

```
df %>%
filter(!is.na(Delivery_type)) %>%
ggplot(aes(Gestational_age, fill = Delivery_type))
+
geom_bar(position ='stack') +
scale_fill_grey() +
theme_minimal() +
theme(legend.position =c(.2, .8)) +
ggtitle('geom_bar(position = "stack")')
```

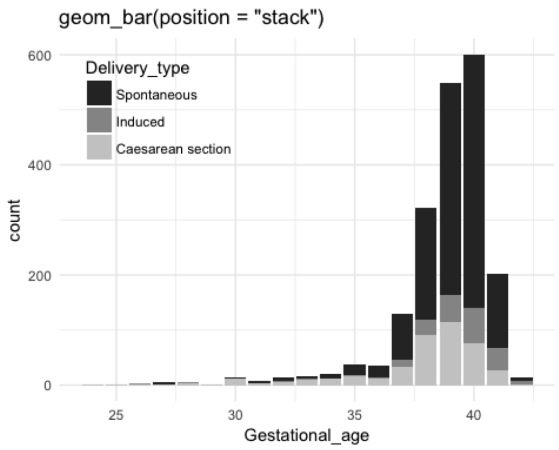


Рис. 15. Разделение столбиковой диаграммы со значениями «stack» по умолчанию (листинг 15)

Листинг 16

```
df %>%
filter(!is.na(Delivery_type)) %>%
ggplot(aes(Gestational_age, fill = Delivery_type))
+
geom_bar(position ='dodge') +
scale_fill_grey() +
theme_minimal() +
theme(legend.position =c(.2, .8)) +
ggtitle('geom_bar(position = "dodge")')
```

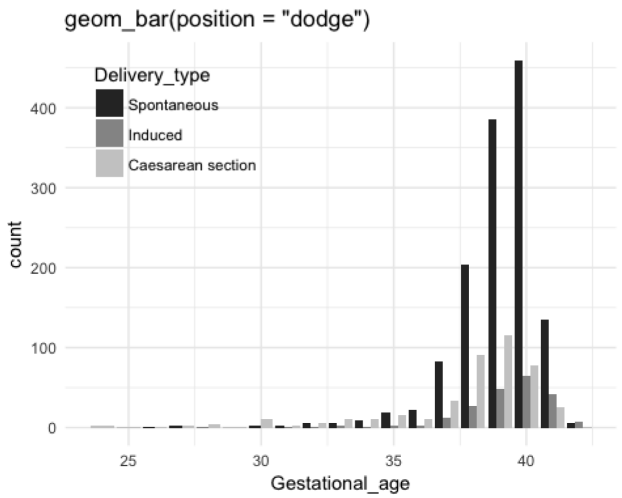


Рис. 16. Разделение столбиковой диаграммы с использованием параметра «dodge» (листинг 16)

Листинг 17

```
df %>%
filter(!is.na(Delivery_type)) %>%
ggplot(aes(Gestational_age, fill = Delivery_type))
+
geom_bar(position =’fill’) +
scale_fill_grey() +
theme_minimal() +
ggtitle(‘geom_bar(position = “fill”)’)
```

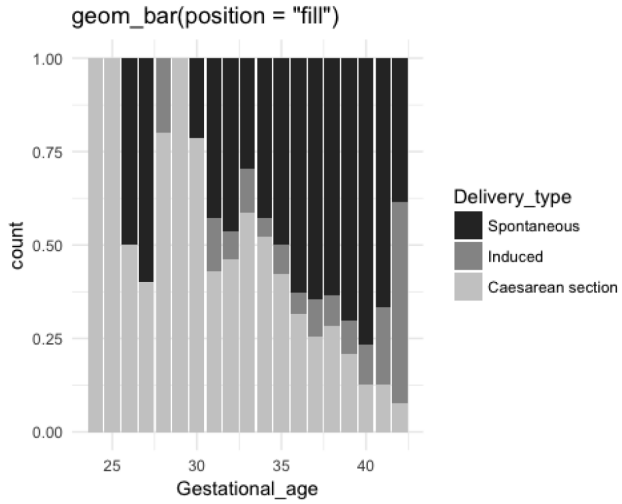


Рис. 17. Разделение столбиковой диаграммы с использованием параметра «fill» (листинг 17)

Столбиковая диаграмма может быть использована для отображения не только непрерывных, но и дискретных числовых данных. Для создания такой диаграммы первоначально необходимо создать таблицу данных из таблицы частоты значений числового признака, и только затем создается столбиковая диаграмма (рис. 18, листинг 18).

Листинг 18

```
# таблица частоты значений
table(df$Maternal_age)
## 15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32
## 1 3 8 19 30 55 54 83 100 118 148 140 137
126 119 150 123 114
## 33 34 35 36 37 38 39 40 41 42 43 44 46
## 89 87 66 69 51 37 28 20 12 8 3 1 1
# таблица данных на основе таблицы частоты зна-
чений, первые четыре строки
as.data.frame(table(df$Maternal_age))[1:4, ]
## Var1 Freq
## 1 15 1
## 2 16 3
## 3 17 8
## 4 18 19
# создание графика
as.data.frame(table(df$Maternal_age)) %>%
ggplot(aes(x = Var1, Freq)) +
geom_bar(stat =‘identity’)
```

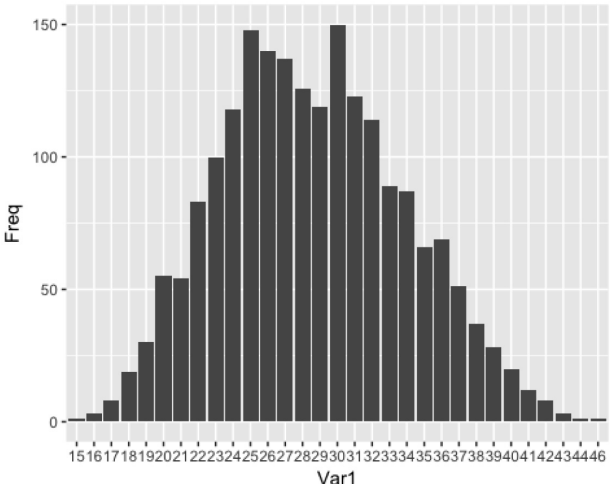


Рис. 18. Создание столбиковой диаграммы для дискретных числовых данных (листинг 18)

Как и в предыдущих примерах, подгруппы наблюдения можно визуализировать по-отдельности с использованием функции позиционирования, разделение также проводится по категориальному (качественному) признаку (рис. 19, листинг 19).

Листинг 19

```
df %>%
select(Maternal_age, Marital_status) %>%
na.omit() %>%
table() %>%as.data.frame() %>%
ggplot(aes(x = Maternal_age, y = Freq, fill =
Marital_status)) +
geom_bar(stat =’identity’) +
scale_fill_grey(start = .1, end = .9) +#определя-
ет черно-белую шкалу для графика
theme_minimal() +
theme(legend.position =c(.85, .8)) +
theme(axis.text.x =element_text(angle =90)) +#
изменение расположения значений оси x
labs(x =’Возраст матери’, y =’’, fill =’Семейное
положение’,
title =’Возраст матери и семейное положение’)
```

Рис. 19. Разделение столбиковой диаграммы для дискретных числовых данных (листинг 19)

Столбиковые диаграммы также могут быть использованы для демонстрации средних значений и показателей вариабельности. Данные диаграммы, представленные на рис. 20 (листинг 20) и рис. 21 (листинг 21), полезны при визуальной оценке достоверности различий.

Листинг 20

```
# mean & sd
df %>%
select(M_age_group, Gestational_age) %>%
na.omit() %>%
group_by(M_age_group) %>%
summarise(GestAgeMean =mean(Gestational_age),
GestAgeSD =sd(Gestational_age)) %>%
ggplot(aes(M_age_group, y = GestAgeMean, fill =
M_age_group)) +
geom_bar(stat ='identity', position ='dodge') +
# определение размера отклонений
geom_errorbar(aes(ymin = GestAgeMean -GestAgeSD,
ymax = GestAgeMean +GestAgeSD),
position ='dodge', width = .25) +
theme(legend.position ='none') +
scale_fill_grey(start = .4, end = .7) +
labs(x ='Maternal Age Group',
y ='',
title ='Gestation Age',
subtitle =paste('Mean', '\U00B1', '1 standard
deviation'))
```

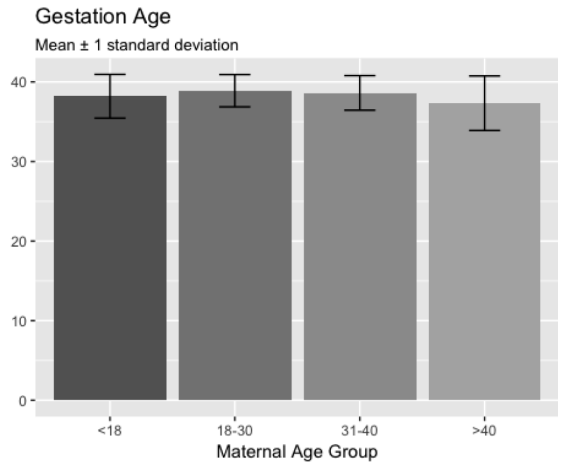


Рис. 20. Создание столбиковой диаграммы с представлением среднего арифметического и стандартного отклонения (листинг 20)

Листинг 21

```
# mean & standard error
# функция для расчета стандартной ошибки
se95 <-function(x) {
se95 <-qt(.975, df =length(x)-1)*sd(x)/
sqrt(length(x))
se95
}

# mean & se
df %>%
select(M_age_group, Gestational_age) %>%
na.omit() %>%
group_by(M_age_group) %>%
summarise(GestAgeMean =mean(Gestational_age),
GestAgeSE =se95(Gestational_age)) %>%
ggplot(aes(M_age_group, y = GestAgeMean, fill =
M_age_group)) +
geom_bar(stat ='identity', position ='dodge') +
geom_errorbar(aes(ymin = GestAgeMean -GestAgeSE,
ymax = GestAgeMean +GestAgeSE),
position ='dodge', width = .25) +
labs(x ='Maternal Age Group',
y ='',
title ='Gestation Age',
subtitle =paste('Mean', '\U00B1', '1 standard
error')) +
scale_fill_grey(start = .4, end = .7) +
theme(legend.position ='none')
```

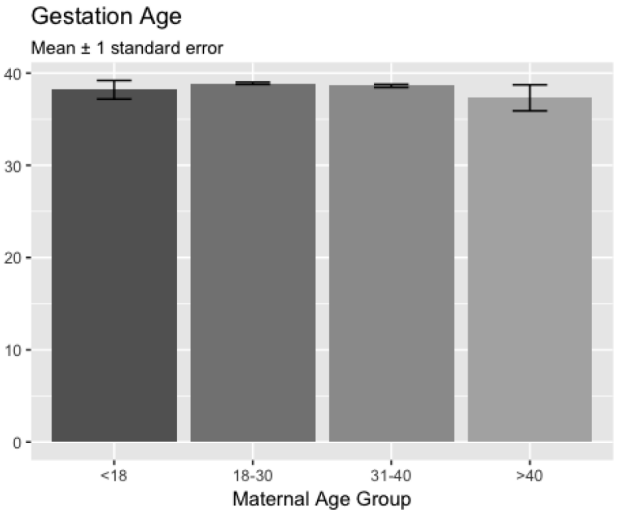


Рис. 21. Создание столбиковой диаграммы с представлением среднего арифметического и ошибки среднего (листинг 21)

Визуализация является необходимым средством для изучения связей между переменными. Так, для визуальной оценки взаимосвязей между двумя непрерывными переменными обычно используются точечные диаграммы (скаттерограммы), представленные на рис. 22 (листинг 22).

Листинг 22

```
ggplot(df, aes(Paternal_age, Maternal_age)) +geom_
point()
```

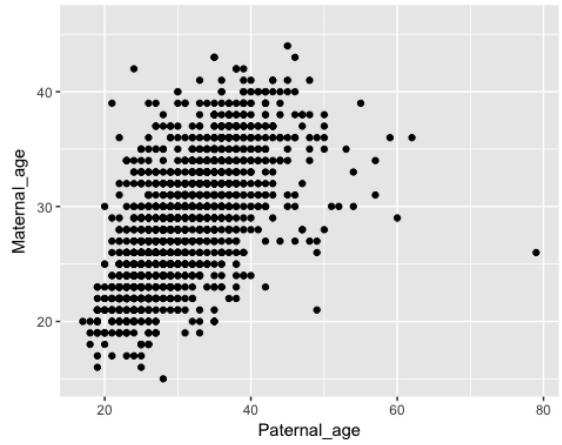


Рис. 22. Создание точечной диаграммы (листинг 22)

На поле графика в правой части можно увидеть значение возраста отца около 80 лет, посчитаем это значение за «выброс» (атипичное наблюдение) и отфильтруем данные. Для этого добавим к графику линию тренда, используя функцию «geom_smooth» (рис. 23, листинг 23).

Листинг 23

```
df %>%filter(Paternal_age <65) %>%
ggplot(aes(Paternal_age, Maternal_age)) +
geom_point() +geom_smooth(method ='lm')
```

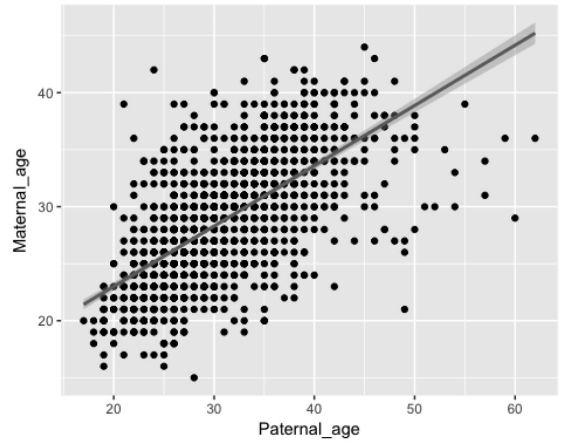


Рис. 23. Добавление линии тренда к точечной диаграмме (листинг 23)

При использовании точечной диаграммы также возможно провести ее разделение на подгруппы с использованием таких параметров, как цвет и форма маркера (рис. 24, листинг 24). В данном листинге заданные значения в слое *scale_y_continuous(breaks = seq(35, 45, 2), limits = c(35, 45))* определили шкалу оси y и диапазон данных.

Следует отметить, что недостатком точечных диаграмм является совпадение значений у точек на диаграмме при большом количестве наблюдений.

Листинг 24

```
df %>%
filter(Maternal_age >35) %>%
ggplot(aes(Paternal_age, Maternal_age, shape =
Infant_sex)) +
geom_point(size =3) +
scale_colour_grey() +
scale_y_continuous(breaks =seq(35, 45, 2), limits
=c(35, 45)) +
theme(legend.position =c(.8, .8))
```

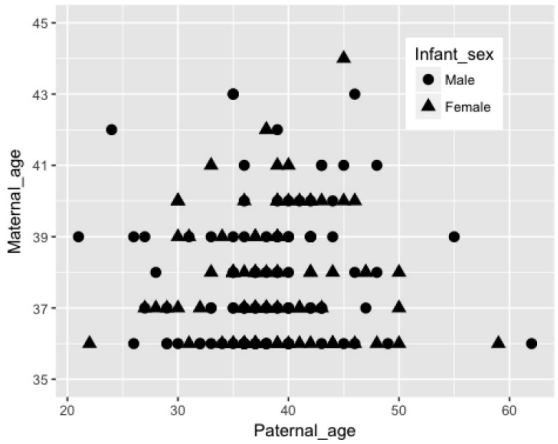


Рис. 24. Разделение точечной диаграммы с использованием формы маркера (листинг 24)

Для показа реального количества значений, приходящегося на пересечение значений по осям графика, используются геометрические образы, например с использованием функции «geom_count», как это показано на рис. 25 (листинг 25).

Листинг 25

```
df %>%
filter(Paternal_age <65 ) %>%
ggplot(aes(Paternal_age, Maternal_age)) +
geom_count() +
geom_smooth(method ='lm')
```

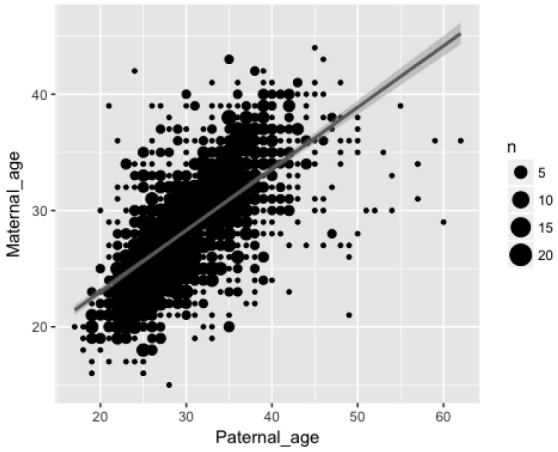


Рис. 25. Использование функции «geom_count» при создании точечной диаграммы (листинг 25)

Для визуальной оценки взаимосвязей между двумя непрерывными переменными также могут быть использованы пузырьковые диаграммы (рис. 26, листинг 26). В данных диаграммах первая переменная — значения по оси x, вторая переменная — значения по оси y, третья переменная — специальный параметр «size».

Листинг 26

```
df %>%
select(Gestational_age, Maternal_age) %>%
na.omit() %>%
filter(Gestational_age >30&Maternal_age >20) %>%
group_by(Gestational_age, Maternal_age) %>%
summarise(n =n()) %>%
ggplot(aes(Gestational_age, Maternal_age, size =
n)) +
geom_point(color ='grey50') +
scale_x_continuous(breaks =30:42, limits =c(31,
42)) +
scale_y_continuous(breaks =20:42, limits =c(21,
42)) +
theme_minimal() +
labs(x ='Gestational Age',
y ='Maternal Age',
title ='Count')
```

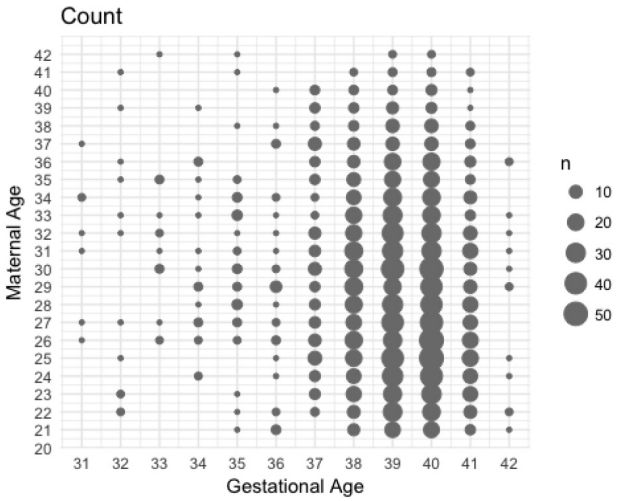


Рис. 26. Создание пузырьковой диаграммы (листинг 26)

Сходные результаты достигаются при использовании вафельной диаграммы (рис. 27, листинг 27). Размер значения в ячейке такой диаграммы определяется по её цвету или оттенку серого цвета.

Листинг 27

```
df %>%
select(Gestational_age, Maternal_age) %>%
na.omit() %>%
group_by(Gestational_age, Maternal_age) %>%
summarise(n =n()) %>%
ggplot(aes(Gestational_age, Maternal_age, fill =
n)) +
geom_tile(color ='grey50') +
scale_fill_gradient(low ='lightgrey', high
='black') +
theme_minimal() +
labs(x ='Gestational Age',
y ='Maternal Age',
title ='Count')
```

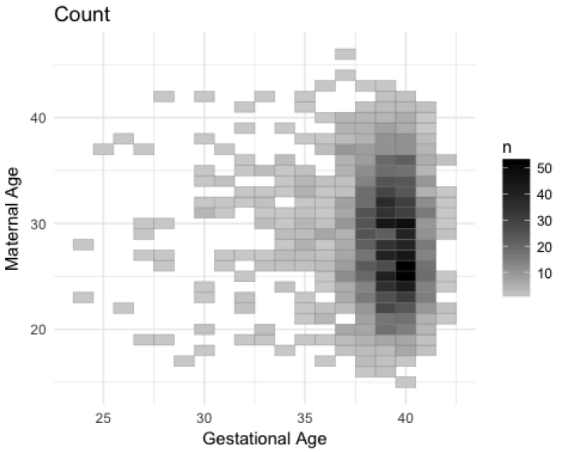


Рис. 27. Создание вафельной диаграммы (листинг 27)

Графическое представление временных рядов

Графическое представление временных рядов как последовательности значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, является одним из способов анализа временных рядов и прогнозирования (экстраполирования) на основе данного анализа.

Как известно, для создания подобного графика необходимы две переменные: по оси x — переменная времени, по оси y — значения показателя в определенный момент времени (рис. 28, листинг 28).

Листинг 28

```
df %>%
# группировка данных
group_by(Year_of_birth) %>%
# вычисление показателя
summarise(Anemia_Level =mean(Anemia)) %>%
# график
ggplot(aes(x = Year_of_birth,
y = Anemia_Level)) +
geom_point() +
geom_line() +
scale_x_continuous(breaks =2012:2014) +
theme_minimal()
```

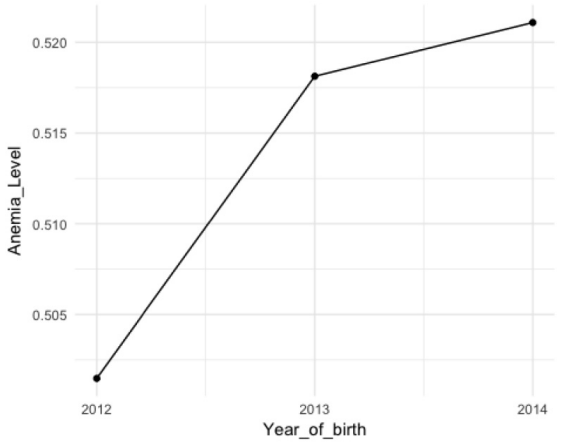


Рис. 28. Создание временного ряда (листинг 28)

Путем использования параметра «group» можно разделить значения временного ряда на подгруппы и

тем самым отследить изменения показателя в подгруппах (рис. 29, листинг 29).

Листинг 29

```
# Динамика количества родовых дефектов по годам
с учетом вида родоразрешения
# использование table для получения показателя
tt <- table(df$Birth_defect)
tt
##
## no yes
## 1911 74
tt[2]
## yes
## 74
df %>%
# выбор столбцов для изучения
select(Delivery_type, Year_of_birth, Birth_
defect) %>%
# удаление NA
drop_na() %>%
# группировка
group_by(Year_of_birth, Delivery_type) %>%
# расчет показателя – количества дефектов
summarise(Birth_defect_Count =table(Birth_defect)
[2]) %>%
# график
ggplot(aes(x = Year_of_birth,
y = Birth_defect_Count,
group = Delivery_type,
linetype = Delivery_type)) +
geom_point(size =3) +
geom_line(size =1) +
theme_minimal() +
scale_x_continuous(breaks =2012:2014) +
labs(x ='Year', y ='', title ='Count of Birth
Defects')
```

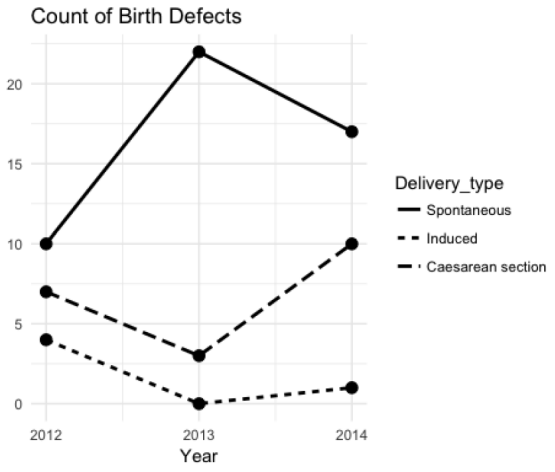


Рис. 29. Разделение временного ряда (листинг 29)

Создание графиков с надписями на русском языке

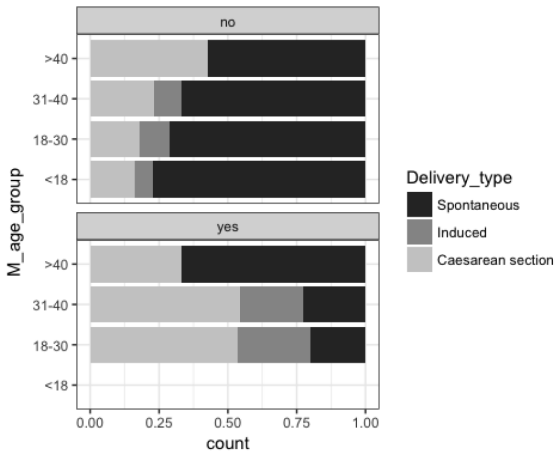
R является программной средой, предполагающей использование английского языка, поэтому при создании графика названия переменных также будут выводиться на английском языке. Данный факт не имеет особого значения на этапе разведочного анализа данных, но для представления графика русскоязычной аудитории все названия диаграммы (заголовок, подзаголовок, названия осей и переменных на осях, названия легенды и выведенных переменных и прочее) можно вывести и на русском языке. К сожалению, это потребует увеличения размера кода, но это условие является неизбежным, если требуется русификация графика.

На рис. 30 (листинг 30) представлен процесс русификации диаграммы.

Листинг 30

```
# создание таблицы данных для графика
df_pre <-df %>%
mutate(M_age_group =factor(cut(Maternal_age,
breaks =c(14, 18, 30 , 40, 50),
labels =c('<18', '18-30', '31-40', '>40')))) %>%
select(Delivery_type,
M_age_group,
Preeclampsia) %>%
na.omit() %>%
mutate(Preeclampsia =factor(Preeclampsia, levels
=c(0, 1),
labels =c('no', 'yes')))

# график без надписей на русском
ggplot(df_pre, aes(M_age_group, fill = Delivery_
type)) +
geom_bar(position ='fill') +
scale_fill_grey() +
coord_flip() +
theme_bw() +
facet_wrap(~Preeclampsia, ncol =1)
```




```
# График с надписями
# создание названий групп при фасетировании
tt <-table(df$Preeclampsia)
tt # Случаи преэклампсии
##
## 0 1
## 1930 70
facet_labels <-c(yes =paste('Преэклампсия',
as.character(tt[2]), 'случаев'),
no =paste('Не было преэклампсии',
as.character(tt[1]), 'случаев'))

# График с надписями
ggplot(df_pre, aes(M_age_group, fill = Delivery_
type)) +
geom_bar(position ='fill') +
# поворот графика на 90 градусов
coord_flip() +
theme_bw() +
facet_wrap(~Preeclampsia, ncol =1,
labeller=labeller(Preeclampsia = facet_labels)) +
# названия осей, заголовка, подзаголовка, сопро-
водительной подписи
```

```
labs(x ='Возрастная группа матери',
y ='Количество, %',
title ='Виды родоразрешения',
subtitle ='При преэклампсии в разных возрастных
группах',
caption ='По данным областного регистра') +
# работа с легендой: выбор цвета для переменных,
название легенды и переменных
scale_fill_manual(values =c('grey20', 'grey50',
'grey90'),
name ='Вид родоразрешения',
labels =c('Спонтанное',
'Индуцированное',
'Кесарево сечение')) +
# создание названий для переменной x (при пово-
роте графика ось не изменяется)
scale_x_discrete(labels =c('младше 18','от 18 до
30 лет',
'от 30 до 40 лет',
'Старше 40 лет')) +
# перевод значений по оси y в проценты
scale_y_continuous(labels = scales::percent)
```

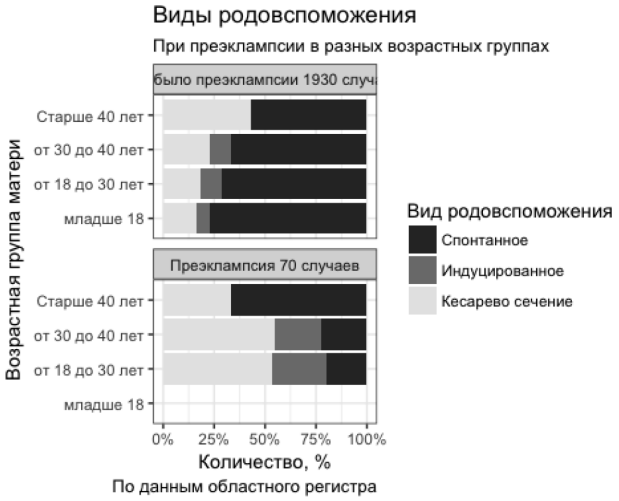


Рис. 30. Русификация диаграммы (листинг 30)

Заметим, что содержание графика не изменилось при добавлении надписей, первый график информативен для исследователя, второй — больше подходит для публикации.

Сохранение графиков

Созданные графики могут быть сохранены для дальнейшего использования с помощью функции ggsave('filename'), где «filename» — название файла вводится в кавычках с расширением. Графики могут быть сохранены в pdf, png, jpeg и других графических форматах.

Дополнительная информация по работе с пакетом «ggplot2» может быть получена при изучении технической документации, пособий и доступных интернет-ресурсов [2, 3, 9, 12, 13, 15], содержащих большое количество примеров использования данного пакета.

В заключении хотелось бы сказать, что рассматриваемые в данной работе графики представлены исключительно с целью показать возможности R, поэтому не стоит искать в наших примерах глубокую научную составляющую. С научными публикациями с использованием всего Архангельского областного регистра родов можно ознакомиться в международных базах данных Web of Science, Scopus, e-library и др.

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. *Нейтан Яу.* Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами / пер. с англ. Кировой С. М.: Манн, Иванов и Фербер, 2013, 352 с.

5. *Усынина А. А., Одланд И. 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. *Chambers J., Cleveland W., Beat Kleiner B., Tukey P.* Graphical methods for Data Analysis. Wadsworth, 1983.

8. Data Visualization Cheat Sheet. URL: https://www.rstudio.com/resources/cheatsheets/ (дата обращения: 18.05.2018).

9. ggplot2 Reference. URL: http://ggplot2.tidyverse.org/reference/ (дата обращения: 18.05.2018).

10. Graphs with ggplot2. URL: http://www.cookbook-r.com/Graphs/ (дата обращения: 18.05.2018).

11. *Grolemund G., Wickham H.* R for data science. URL: http://r4ds.had.co.nz (дата обращения: 18.05.2018).

12. How to make any plot in ggplot2. URL: http://r-statistics.co/ggplot2-Tutorial-With-R.html (дата обращения: 18.05.2018).

13. R: анализ и визуализация данных. URL: http://r-analytics.blogspot.ru (дата обращения: 18.05.2018).

14. *Tukey J. W.* Exploratory data analysis. Reading, PA: Addison-Wesley, 1977.

15. Tutorials. URL: http://flowingdata.com/category/tutorials/ (дата обращения: 18.05.2018).

16. *Wickham H.* ggplot2: Elegant Graphics for Data Analysis. New York, Springer, 2009.

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. Kabacoff 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 vizualizatsiya dannykh s pomoshch'yu R* [Data statistical analysis using R]. Moscow, 2015, 496 p.

4. Nathan Yau. *Iskusstvo vizualizatsii v biznese. Kak predstavit' slognuyu informatsiyu prostimi obrazami* [Art of visualisation in bisiness. How to present difficult information

by simple pictures], per. s angl. Kirova S. Moscow, Mann, Ivanov i Ferber Publ., 2013, 352 p.

5. Usynina A. A., Odland J. Ø., 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. Chambers J., Cleveland W., Beat Kleiner B., Tukey P. *Graphical methods for Data Analysis*. Wadsworth, 1983.

8. Data Visualization Cheat Sheet. Available at: https://www.rstudio.com/resources/cheatsheets/ (accessed: 18.05.2018).

9. ggplot2 Reference. Available at: http://ggplot2.tidyverse.org/reference/ (accessed:18.05.2018).

10. Graphs with ggplot2. Available at: http://www.cookbook-r.com/Graphs/ (accessed:18.05.2018).

11. Grolemund G.,Wickham H. R for data science. Available at: http://r4ds.had.co.nz (accessed: 18.05.2018).

12. How to make any plot in ggplot2. Available at: http://r-statistics.co/ggplot2-Tutorial-With-R.html (accessed:18.05.2018).

13. R: Analysis and visualization of data. Available at: http://r-analytics.blogspot.ru (accessed:18.05.2018).

14. Tukey J.W. *Exploratory data analysis*. Reading, PA, Addison-Wesley, 1977.

15. Tutorials. Available at: http://flowingdata.com/category/tutorials/ (accessed:18.05.2018).

16. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, Springer. 2009.

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 проспект, д. 51, офис 1252

Тел.: +79214717053

E-mail: Andrej.Grjibovski@gmail.com