КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Институт фундаментальной медицины и биологии

Кафедра биохимии, биотехнологии и фармакологии

АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА ПРОГРАММИРОВАНИЯ R

Часть 1

Учебно-методическое пособие

Казань - 2024

УДК 51-76; 573
ББК 28

Печатается по решению Учебно-методической комиссии Института фундаментальной медицины и биологии КФУ Протокол № 8 от 15 мая 2024г

Рецензенты:

Кандидат биологических наук, доцент О.С. Козлова

Доктор биологических наук, профессор Т.В. Багаева

Доктор биологических наук, зав. кафедрой генетики Каюмов А.Р.

Акберова Н.И.
Анализ данных с использованием языка программирования R.
Часть1./Н.И.Акберова. – Казань: Казанский федеральный университет, 2024. – 50 с.

В учебно-методическом пособии систематизированы базовые методы статистической обработки данных, продемонстрировано применение методов биостатистики для анализа и визуализации результатов эксперимента с использованием свободной программной среды вычислений R и графической оболочки Rstudio. В Часть 1 включены методы разведочного анализа данных, их визуализации, проверки распределения на нормальность, описания количественных признаков, параметрические и непараметрические критерии для их сравнения.

Каждый раздел пособия снабжён примерами реализации статистических методов на языке R, а также вопросами и заданиями для самостоятельного решения.

Пособие предназначено для использования в курсах магистратуры «Компьютерные технологии в биологии» и «Биостатистика», а также для широкого круга студентов и аспирантов медико-биологического профиля при подготовке курсовых и выпускных квалификационных работ и проведении научных исследований.

СОДЕРЖАНИЕ

Введение. Краткий обзор основных методов биостатистики

Все изучаемые в биологии и медицине признаки можно разделить на количественные и качественные (факториальные). В зависимости от формата признаков используются определенные статистические подходы и методы для их описания и сравнения. На схеме (Рис.1) представлены основные статистические критерии, которые применяются при анализе биологических и медицинских данных, в зависимости от задач исследования. Схема может служить определителем для выбора корректного статистического критерия.

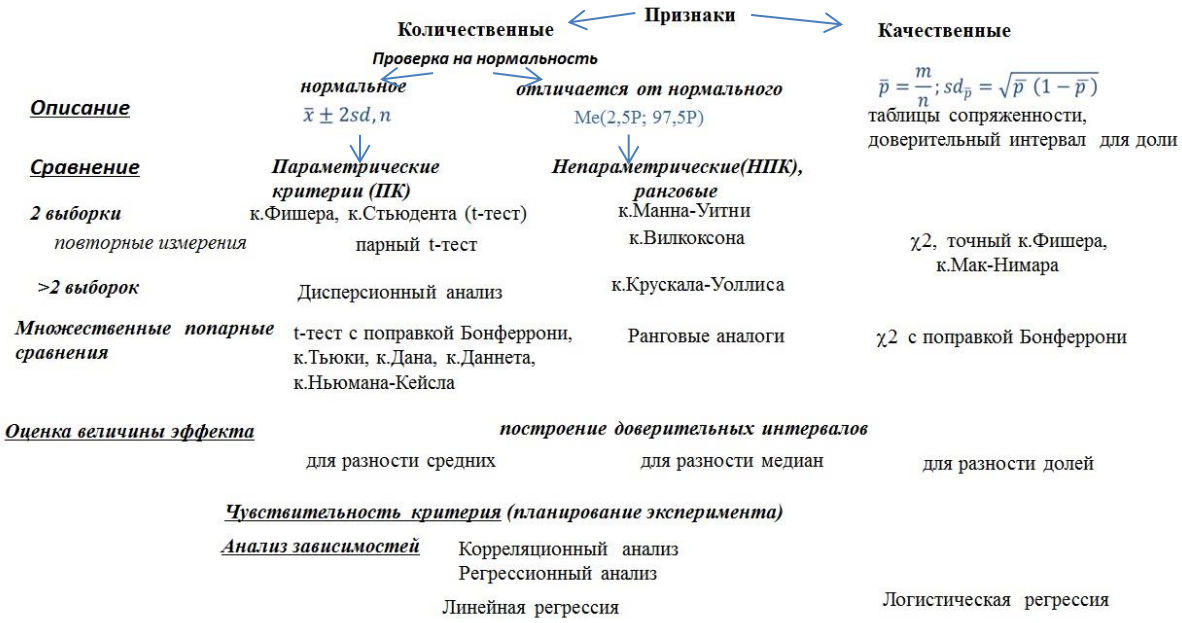

Рис. 1. Схема-определитель статистических методов

Так, если вы работаете с количественными признаками, прежде всего вы должны выяснить, можно ли использовать модель нормального распределения для его описания и параметрические критерии для сравнения вариантов этого признака. Если распределение изучаемого количественного признака нормально, то для описания и сравнения используются параметры нормального распределения, такие как среднее, дисперсия, стандартное отклонение. Если распределение отличается от нормального, для описания в качестве меры центра распределения применяют медиану, а вариабельность признака описывают с помощью перцентилей, для сравнения вариантов такого признака используют соответствующие ранговые методы в зависимости от количества сравниваемых групп и дизайна эксперимента.

В случае работы с качественными данными для их описания используют построение доверительного интервала для истинной доли, часто представляют

качественные данные в виде таблиц сопряженности, с которыми работают основные критерии сравнения качественных данных в различных группах.

Для оценки величины различия значений признака в группах сравнения используют доверительные интервалы, обычно строят 95%-ные доверительные интервалы (ДИ). В случае нормально распределенного количественного признака это 95%-ный ДИ для разности средних, для признака, распределение которого отличается от нормального, это 95%-ный ДИ для разности медиан, в случае качественного признака строят 95%-ный ДИ для разности долей.

Тема1. Количественные данные. Проверка распределения на нормальность

Графические методы

Для определения характера выборочного распределения количественного признака прежде всего строят графики, обычно используют гистограммы и боксплоты (коробчатые графики, «ящики с усами»).

С этой целью гистограммы целесообразно применять для выборок большого объема.

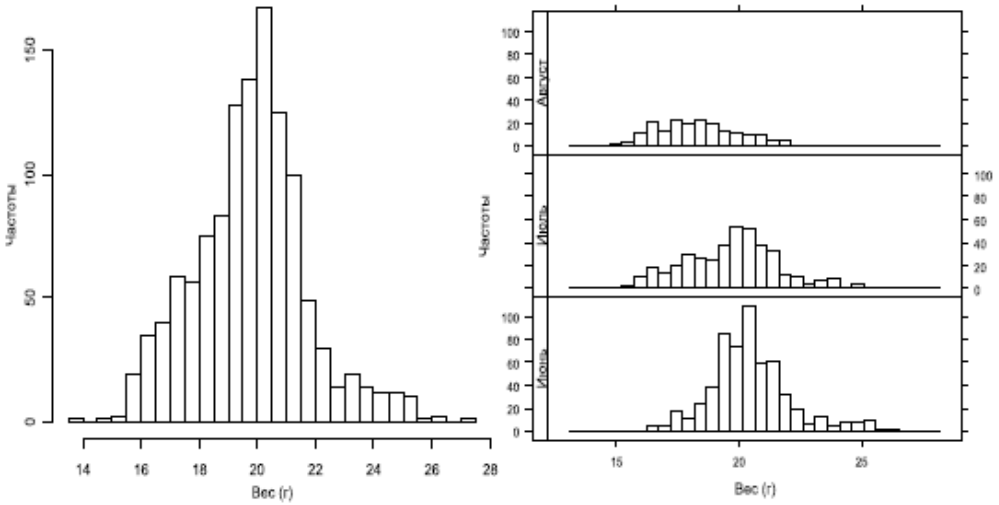

Рис. 2. Гистограммы веса воробьев

Например, распределение веса 1193 воробьев (Zuur et al., 2010) явно асимметрично (слева), а по гистограммам, построенным для выборок веса 50 воробьев (справа), сделать определенный вывод трудно (Рис.2).

Боксплоты активно применяются для выборок разного объема. В виде боксплотов удобно представлять на одном графике варианты количественного признака, и это позволяет сформулировать гипотезы и на

основании симметричности боксплотов (или ее отсутствия) выбрать корректные статистические критерии для их проверки.

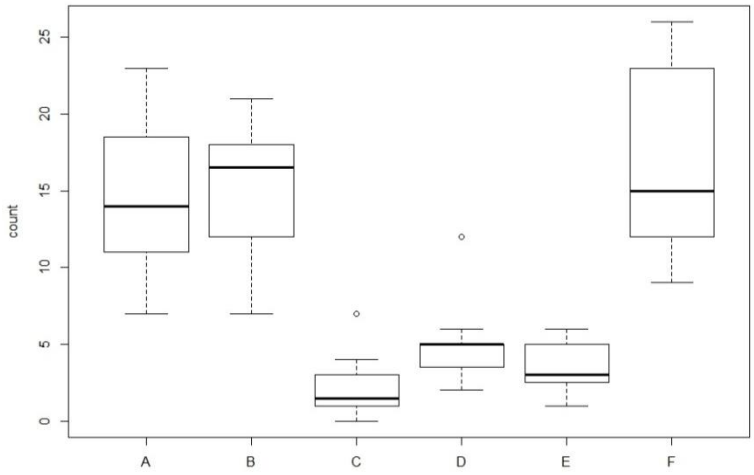

Рис. 3. Примеры боксплотов

На графике представлены боксплоты для количества насекомых при обработке различными инсектицидами, в каждой выборке по 12 наблюдений (Рис.3).

Кроме гистограмм и боксплотов для обоснования выбора модели нормального распределения для описания и сравнения изучаемого признака можно строить графики квантилей.

Гистограммы

В языке R имеется большое количество инструментов для визуализации выборочного распределения количественного признака. В базовой версии R есть функция **hist(x)**, основным аргументом которой являются значения признака.

```
> hist(InsectSprays$count)
```

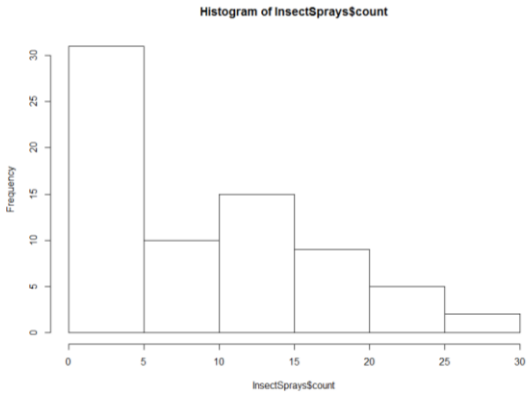

Рис. 4. Гистограмма распределения количества насекомых

На рисунке 4 представлена гистограмма распределения количества насекомых после воздействия шести инсектицидных спреев из таблицы данных InsectSprays (всего 72 наблюдения), по виду которой понятно, что применять модель нормального распределения для работы с этим признаком нельзя.

```
> hist(InsectSprays$count[InsectSprays$spray=="E"])
```

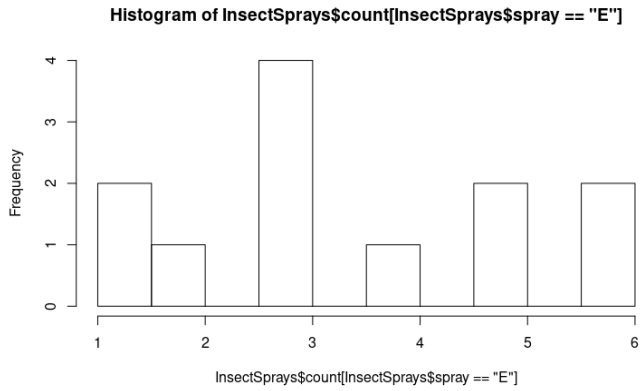

Рис. 5. Гистограмма распределения количества насекомых после спрея E

В случае, когда гистограмма построена для меньшей выборки, на графике 5 показано распределение количества насекомых после воздействия спрея E (12 наблюдений), сложнее сделать однозначный вывод о характере распределения и, соответственно, выбрать адекватную модель для описания и сравнения признака.

Боксплоты

Боксплоты, в отличие от гистограмм, позволяют более определенно говорить о характере распределения количественного признака при малых выборках. В базовой версии R есть функция **boxplot(x)**, основным аргументом которой являются значения признака

```
> boxplot(InsectSprays$count)
```

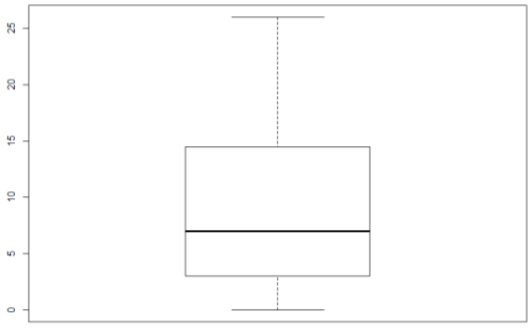

Рис. 6. Боксплот распределения количества насекомых

На рисунке 6 представлен боксплот для количества насекомых после воздействия шести инсектицидных спреев из таблицы данных InsectSprays (всего 72 наблюдения), видна явная асимметрия, медиана сильно смещена вниз, усы разного размера, очевидно, что модель нормального распределения не годится в данном случае.

```
> boxplot(InsectSprays$count[InsectSprays$spray=="E"])
```

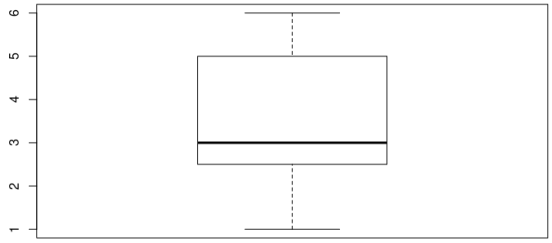

Рис. 7. Боксплот распределения количества насекомых после спрея E

Боксплот дает больше информации о распределении и в случае меньшей выборки. Так, на боксплоте 7 количества насекомых после воздействия спрея E (12 наблюдений) асимметрия распределения такая же явная, как в случае 72 наблюдений.

Графики квантилей (q-q plots, quantile-quantile plots)

Графики квантилей показывают совпадение (или несовпадение) квантилей изучаемого количественного признака с теоретическими для нормального распределения с таким же средним и стандартным отклонением. Квантиль-квантильный график без доверительных огибающих легко построить с помощью базовых функций **qqnorm()** и **qqline**(), задавая в качестве аргумента значения признака.

```
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length)
```

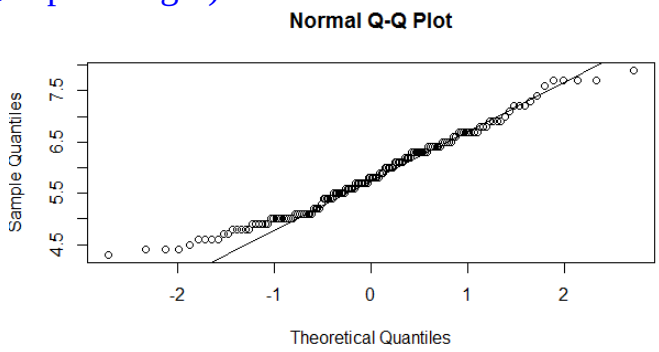

Рис. 8. График квантилей длины лепестков цветков ирисов

Так, для Sepal.Length из датафрейма iris не все значения совпадают с линией квантилей для нормального распределения с такими же, как у Sepal.Length, средним и стандартным отклонением (Рис.8)

Функция **qqPlot()** пакета **car** строит 95%-ную область для квантилей выборки такого же объема из нормального распределения с такими же как у Sepal.Length средним и стандартным отклонением

```
> library(car)
> qqPlot(iris$Sepal.Length, dist= "norm", col=palette()[1] , pch=19,
xlab="Квантили нормального распределения",
ylab="Наблюдаемые квантили",
main="Сравнение квантилей эмпирического и нормального
распределений")
```

(Жирным выделены аргументы, достаточные для построения графика квантилей с доверительной областью.)

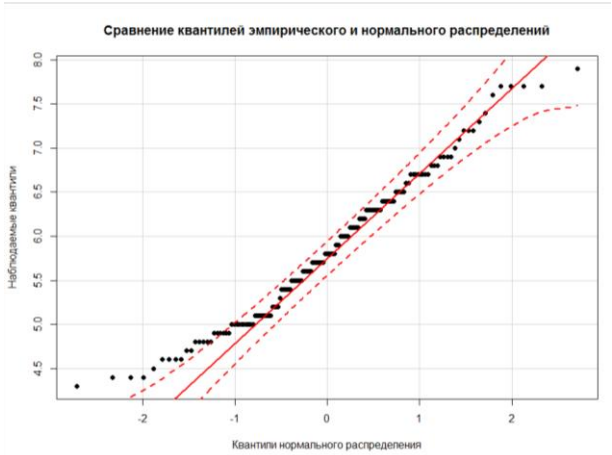

Рис. 9. График квантилей длины лепестков цветков ирисов с 95%-ной доверительной областью для квантилей нормального распределения

Теперь очевидно, что меньшие значения признака выходят за границы области квантилей для нормального распределения, что говорит о том, что нельзя применять эту модель для описания и сравнения признака Sepal.Length из датафрейма iris (Рис.9)

Графики функции плотности распределения

Для анализа распределений очень удобны функции **sm.density()** и **sm.density.compare()** из пакета **sm**, которые строят графики, на которых показаны выборочное распределение изучаемого количественного признака (тонкая черная линия), доверительная область для значений для нормально распределенного признака с таким же средним и стандартным отклонением

(показана голубым цветом), кроме того, на оси X насечками показаны все значения выборки.

```
> library(sm)
> sm.density(iris$Sepal.Length, model = "Normal",
xlab=" iris$Sepal.Length", ylab="Функция плотности распределения")
```

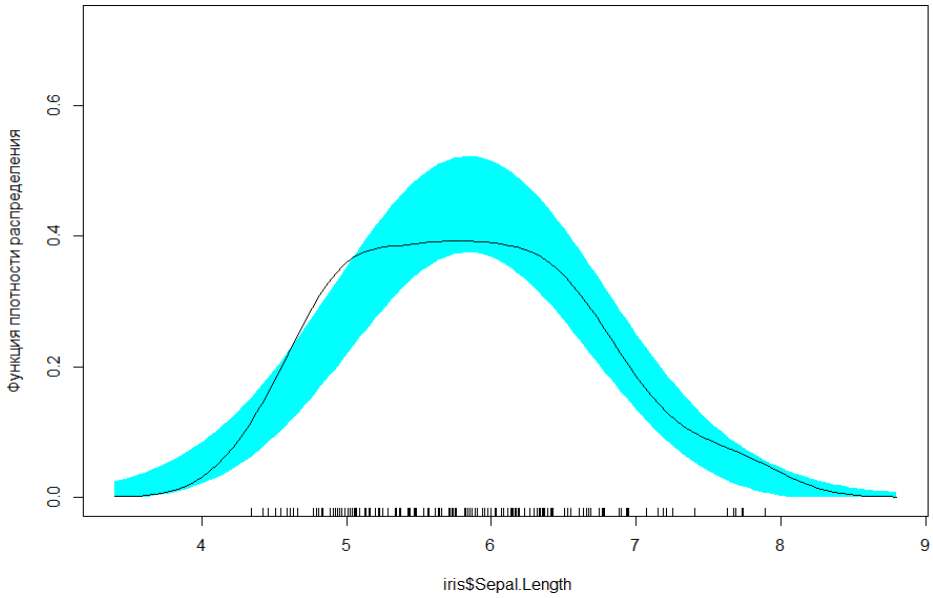

Рис. 10. График функции плотности распределения длины лепестков цветков ирисов с 95%-ной доверительной областью для квантилей нормального распределения

На графике 10 представлена функция плотности распределения длины лепестков ирисов (Sepal.Length) из таблицы данных iris, линия выборочного распределения выходит за рамки доверительной области для нормального распределения, поэтому нельзя использовать параметры нормального распределения для описания этого признака и параметрические критерии для сравнения его вариантов

Задание. Постройте гистограмму, боксплот, график квантилей с доверительной областью для количества насекомых (count) при обработке спреем F из таблицы данных InsectSprays, сделайте вывод о применимости модели нормального распределения

Задание. Постройте график функции плотности распределения ширины лепестков ирисов (Sepal.Width) из таблицы данных iris, сделайте вывод о применимости модели нормального распределения

Формальные методы

Кроме графических способов проверки нормальности распределения количественного признака существуют критерии значимости. Нулевую гипотезу можно сформулировать так: "анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение". Если получаемая при помощи того или иного теста вероятность ошибки p оказывается меньше некоторого заранее принятого уровня значимости (обычно, 0.05), нулевая гипотеза отклоняется.

Практически все такие критерии реализованы в R: базовая функция **shapiro.test(),** при помощи которой можно выполнить широко используемый тест Шапиро-Уилка, функции из пакета **nortest**, реализующие другие распространенные тесты на нормальность - **ad.test()** - тест Андерсона-Дарлинга, **cvm.test()** - тест Крамера фон Мизеса, **lillie.test()** - тест Колмогорова-Смирнова в модификации Лиллиефорса, **sf.test()** - тест Шапиро-Франсия. Во всех функциях аргументом являются значения исследуемого признака.

Ниже показаны отчеты о проведенных тестах для длины лепестков цветков ирисов (Sepal.Length). Во всех тестах p-value меньше 0.05, следует отклонить гипотезу о нормальности распределения этого признака. Следует отметить, что наиболее чувствительными являются тесты Шапиро-Уилка, Шапиро-Франсия, Андерсона-Дарлинга, менее чувствительными - тест Крамера фон Мизеса и тест Колмогорова-Смирнова в модификации Лиллиефорса.

```
> shapiro.test(iris$Sepal.Length)

        Shapiro-Wilk normality test

data:  iris$Sepal.Length
W = 0.97609, p-value = 0.01018

> library(nortest)
> ad.test(iris$Sepal.Length)

        Anderson-Darling normality test

data:  iris$Sepal.Length
A = 0.8892, p-value = 0.02251
```

```
> cvm.test(iris$Sepal.Length)

        Cramer-von Mises normality test

data:  iris$Sepal.Length
W = 0.1274, p-value = 0.04706

> lillie.test(iris$Sepal.Length)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  iris$Sepal.Length
D = 0.088654, p-value = 0.005788

> sf.test(iris$Sepal.Length)

        Shapiro-Francia normality test

data:  iris$Sepal.Length
W = 0.97961, p-value = 0.02621
```

Задание. Проведите тесты на нормальность количества насекомых из таблицы данных InsectSprays

Задание. Постройте гистограмму, боксплот, график квантилей с доверительной областью, проведите тесты на нормальность для количества насекомых (count) при обработке спреем A из таблицы данных InsectSprays.

Анализ однородности данных

Рассмотренные методы проверки нормальности распределения количественного признака можно использовать для разведочного анализа причин асимметрии в его распределении, что позволяет выявить неоднородность значений в выборке.

Рассмотрим на примере анализа длины чашелистиков цветков ирисов (Petal.Length) из датафрейма iris.

```
> hist(iris$Petal.Length)
```

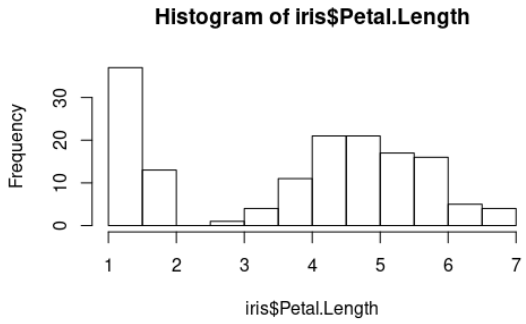

Рис. 11. Гистограмма распределения длины чашелистиков цветков ирисов

На гистограмме 11 видим бимодальность в распределении изучаемого признака

```
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length)
```

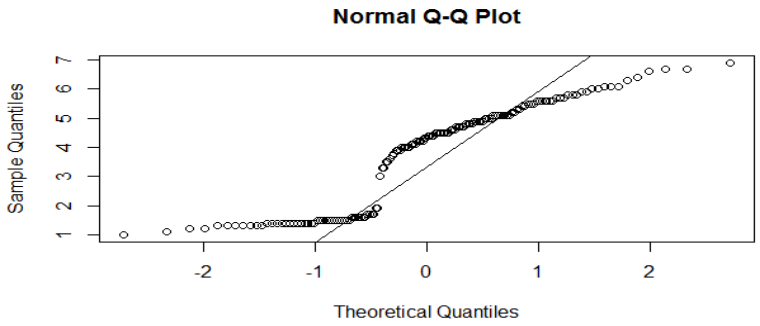

Рис. 12. График квантилей длины чашелистиков цветков ирисов

```
> library(car)
> qqPlot(iris$Petal.Length, dist= "norm", col=palette()[1] , pch=19,
xlab="Квантили нормального распределения",
ylab="Наблюдаемые квантили",
main="Сравнение квантилей эмпирического и нормального
распределений")
```

На графиках квантилей 12 и 13 видим явное несоответствие наблюдаемых квантилей и квантилей нормального распределения, при этом показано четкое разделение значений на 2 группы.

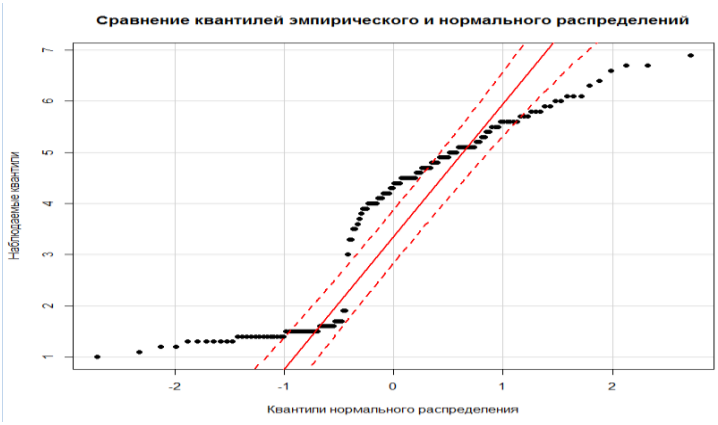

Рис. 13. График квантилей длины чашелистиков цветков ирисов с доверительной областью

Построим графики функции плотности распределения:

```
> library(sm)
> sm.density(iris$Petal.Length, model = "Normal", xlab="iris$Petal,Length",
ylab="Функция плотности распределения")
```

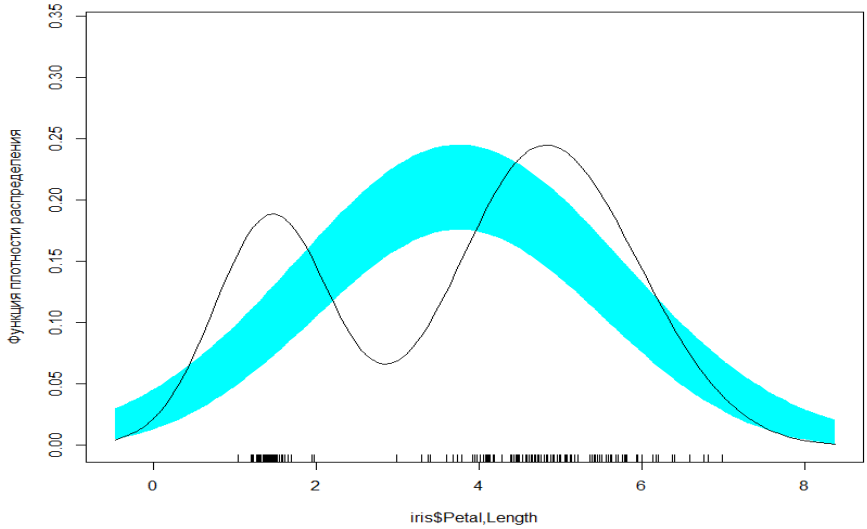

Рис. 14. График функции плотности распределения длины чашелистиков цветков ирисов с 95%ной доверительной областью

График функции плотности распределения длины чашелистиков цветков ирисов (Рис.14) наглядно демонстрирует «несогласие» распределения Petal.Length с нормальным распределением с таким же средним и стандартным отклонением, при этом наблюдаем две группы значений, которые, похоже, нормально распределены, но взяты из распределений с разными средними и стандартными отклонениями.

Не связана ли такая бимодальность с тем, что в выборке собраны цветки разных видов ирисов, т.е. с признаком Species?

```
>boxplot(iris$Petal.Length~iris$Species)
```

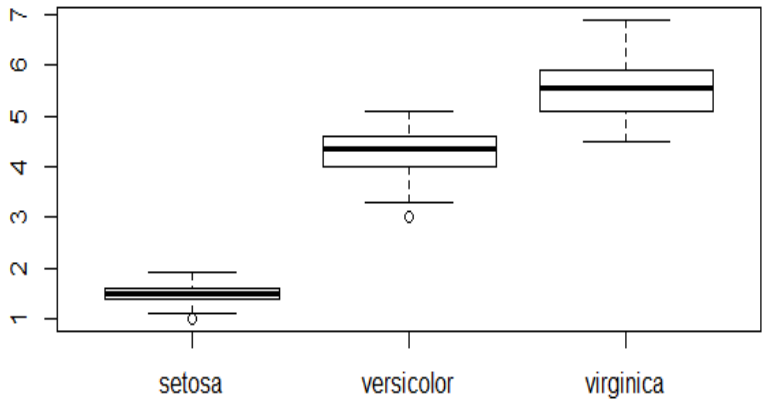

Рис. 15. Боксплоты длины чашелистиков цветков трех видов ирисов

График боксплотов для трех видов ирисов (Рис.15) показывает, что у цветков сорта setosa длина чашелистиков меньше, чем у двух других сортов.

Проверим нормальность распределения в двух выявленных группах. В объект setosa сохраним значения длины чашелистиков этого сорта:

```
> setosa<-subset(iris, Species=="setosa")$Petal.Length
> sm.density(setosa, model = "Normal", xlab="setosa_Petal.Length",
ylab="Функция плотности распределения")
```

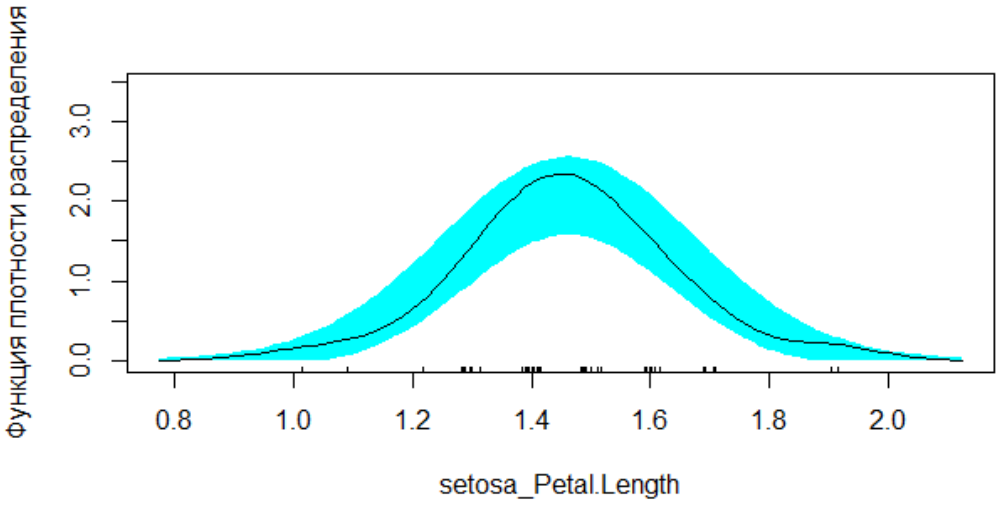

Рис. 16. График функции плотности распределения длины чашелистиков цветков ирисов вида setosa с 95%ной доверительной областью

График функции плотности распределения (Рис.16) наглядно демонстрирует нормальность распределения длины чашелистиков у цветков ирисов сорта setosa , такой же вывод следует из отчета по критерию Шапиро-Уилка (p-value> 0.05)

```
> shapiro.test(setosa)

        Shapiro-Wilk normality test

data:  setosa
W = 0.95498, p-value = 0.05481
```

В объект ver_vir сохраним значения длины чашелистиков видов versicolor и virginica:

```
> versicolor<-subset(iris, Species=="versicolor")$Petal.Length
> virginica<-subset(iris, Species=="virginica")$Petal.Length
> ver_vir<-c(versicolor,virginica)
```

Построим график квантилей:

```
> qqnorm(ver_vir)
> qqline(ver_vir)
```

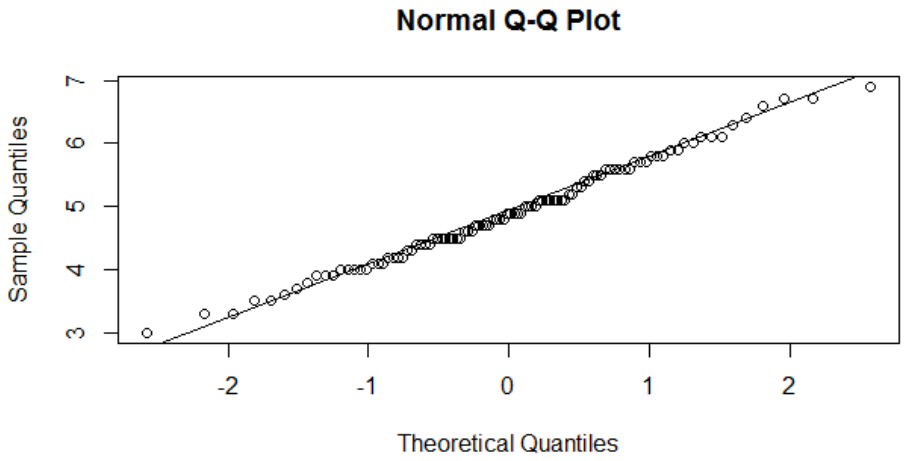

Рис. 17. Квантильный график распределения длины чашелистиков цветков ирисов видов versicolor и virginica

Построим график функции плотности распределения:

```
> sm.density(ver_vir, model = "Normal",
xlab="versicolor+virginica_Petal.Length",
ylab="Функция плотности распределения")
```

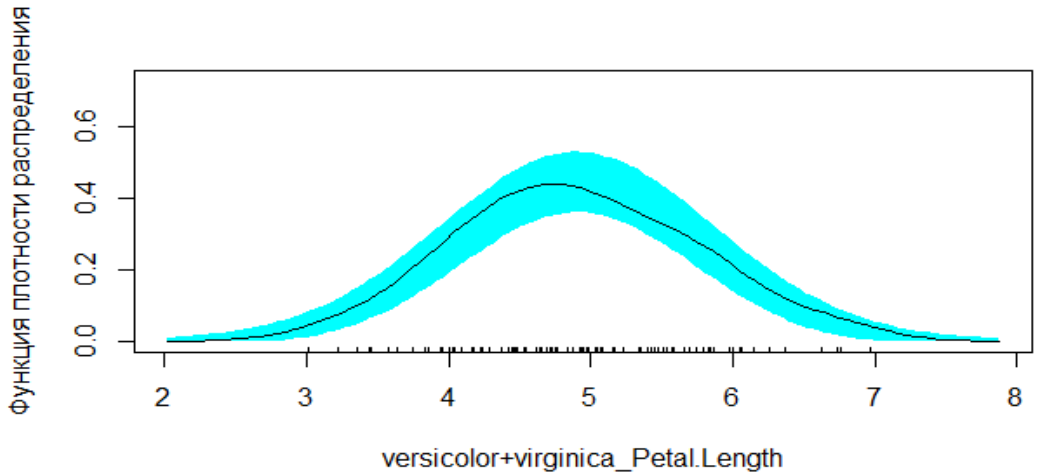

Рис. 18. График функции распределения длины чашелистиков цветков ирисов видов versicolor и virginica

Оба графика (Рис.17, 18) наглядно демонстрируют нормальность распределения значений длины чашелистиков сортов versicolor и virginica. Критерий Шапиро-Уилка подтверждает вывод о нормальности распределения значений в этой группе, p-value > 0.05

```
> shapiro.test(ver_vir)

        Shapiro-Wilk normality test

data:  ver_vir
W = 0.99099, p-value = 0.7445
```

Вернемся к гистограмме длины чашелистиков цветков ирисов и покажем на ней , что один пик соответствует значениям вида setosa (серый цвет), другой – объединенной группе цветков видов versicolor и virginica (синий цвет) (Рис.19)

```
> hist(iris$Petal.Length, breaks=50, freq=F)
> hist(setosa, breaks=8, freq=F, col="grey", add=T )
> hist(ver_vir, breaks=50, freq=F, col="blue", add=T )
```

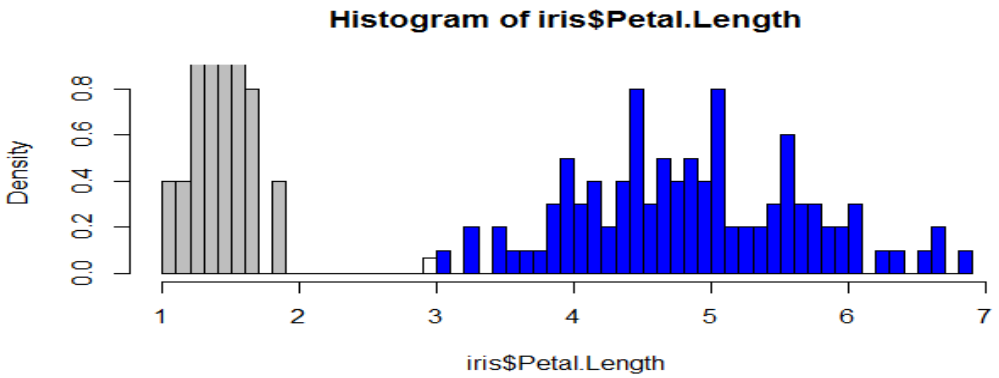

Рис. 19. Гистограмма распределения длины чашелистиков цветков ирисов трех видов

```
> boxplot(iris$Petal.Length,setosa,ver_vir)
> legend("top",c("1-iris,2-setosa,3-ver+vir"))
```

Рис.20. Боксплоты длины чашелистиков цветков ирисов трех видов (1), вида setosa (2), видов versicolor+virginica(3)

Представленные на графике 20 боксплоты показывают, что признак длина чашелистиков распределен нормально, а асимметрия боксплота 1 связана с неоднородностью значений в выборке, в которой объединены значения, принадлежащие двум распределениям с разными средними и стандартными отклонениями

Задание. Проведите проверку на нормальность распределения ширины чашелистиков цветков ирисов (Petal.Width) из таблицы данных iris. Наблюдается ли асимметрия в распределении этого признака, с чем она может быть связана?

Тема 2. Количественные данные. Описание и визуализация

Универсальным способом графического представления выборочных значений для количественных признаков является боксплот, коробчатый график или ящик с усами, в котором жирная линия, проходящая внутри ящика (прямоугольной области), - медиана. Нижняя граница ящика 1 квартиль, верхняя - 3 квартиль, ровно половина всех наблюдений (центральная половина упорядоченной выборки) находится внутри числового диапазона, ограниченного ящиком. Высота ящика называется межквартильным (интерквартильным) размахом.

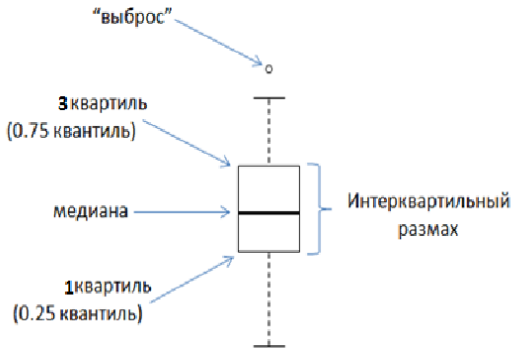

Рис. 21. Боксплот

Концы усов показывают наиболее удалённые от верхней и нижней границ ящика значения, расстояние между которыми и границами ящика не превышает полутора межквартильных размахов. Отдельно стоящие точки на графике – это «выбросы». Так называют экстремально отстоящие от медианы значения, то есть те значения, расстояние между которыми и границами ящика (то есть 1м и 3м квартилями) превышает полтора межквартильных размаха.

Охарактеризовать количественный признак означает определить центральную тенденцию и описать разброс. Оценкой центра распределения служат среднее ($\bar{X}$, mean) и медиана (Me, median), для описания и визуализации вариабельности используют стандартное отклонение (sd, standard deviation), и квантили (квартили и перцентили).

Если количественный признак имеет нормальное распределение, для его полноценного описания и представления, в т.ч. в научных публикациях, используют

$$\bar{x} \pm 2sd, n$$

где n –объем выборки. Таким образом описывается вариабельность признака с 95%ной достоверностью. Указание объема наблюдений позволит

использовать параметрические критерии значимости для сравнения этих данных с результатами других исследований.

Если распределение признака отличается от нормального, то использовать эту модель нельзя. В этом случае используют квантили. Мерой центра распределения будет медиана, для того, чтобы охарактеризовать с 95%ной достоверностью вариабельность такого признака, распределение которого будет асимметричным, используют значения 2.5 перцентиля и 97.5 перцентиля:

Me(2,5P; 97,5P)

Для примера правильного описания количественных признаков используем доступную для анализа базу данных системы Behavioral Risk Factor Surveillance System (BRFSS), в рамках которой проводится ежегодный телефонный опрос 350 000 человек в США. BRFSS предназначена для выявления факторов риска для здоровья среди взрослого населения и выявления новых тенденций. Респондентам задают вопросы об их диете и еженедельной физической активности, статуса ВИЧ - инфицирования, об употреблении табака, а также об уровне медицинского обеспечения. На веб-сайте BRFSS (http:// www.cdc.gov/brfss) содержится полное описание исследования, в том числе научно-исследовательских вопросов, которые мотивируют этот проект. Мы сосредоточимся на случайной выборке в 20000 человек из опроса BRFSS, проведенного в 2000 году, таблице данных cdc, которая расположена на сайте http://www.openintro.org/stat/data/cdc.R. В датафрейме cdc представлены переменные: genhlth, exerany, hlthplan, smoke100, height, weight, wtdesire, age, и gender. Каждая из этих переменных соответствует вопросу, который был задан в опросе. Так, 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 физические упражнения (1) или нет (0). Переменная hlthplan указывает, имеет ли респондент ту или иную форму медицинского страхования (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). Загрузим датафрейм cdc в рабочее пространство RSudio:

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

Будем работать с ответами первых 150 респондентов:

```
> cdc150<-cdc[1:150, ]
```

Посмотрим сводку по полученной таблице данных:

```
> summary(cdc150)
```

```
     genhlth        exerany          hlthplan         smoke100
excellent:36   Min.   :0.0000   Min.   :0.0000   Min.   :0.00
very good:55   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.00
good     :41   Median :1.0000   Median :1.0000   Median :0.00
fair     :14   Mean   :0.7667   Mean   :0.8867   Mean   :0.42
poor     : 4   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.00
               Max.   :1.0000   Max.   :1.0000   Max.   :1.00
    height          weight         wtdesire          age        gender
Min.   :59.00   Min.   :100.0   Min.   : 99.0   Min.   :18.00   m:76
1st Qu.:64.00   1st Qu.:138.5   1st Qu.:130.0   1st Qu.:32.00   f:74
Median :67.00   Median :165.5   Median :150.0   Median :43.00
Mean   :67.22   Mean   :166.2   Mean   :153.2   Mean   :43.93
3rd Qu.:70.00   3rd Qu.:190.0   3rd Qu.:175.0   3rd Qu.:53.00
Max.   :77.00   Max.   :280.0   Max.   :260.0   Max.   :87.00
```

Построим боксплоты для роста, веса, желаемого веса и возраста респондентов, среди которых 76 мужчин и 74 женщины:

```
> boxplot(cdc150[ , 5:8])
```

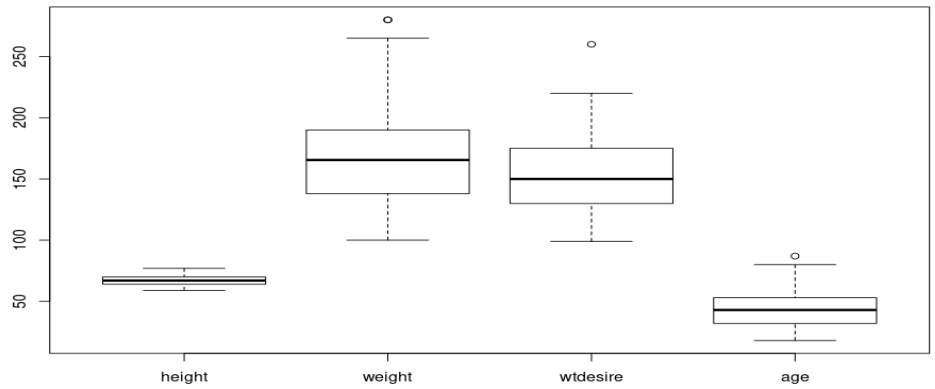

Рис. 22. Боксплоты количественных признаков из набора данных cdc150

Глядя на график 22, можно предположить, что из 4-х признаков, представленных на графике, только рост может быть распределен нормально. Проверим это:

```
> boxplot(cdc150$height)
```

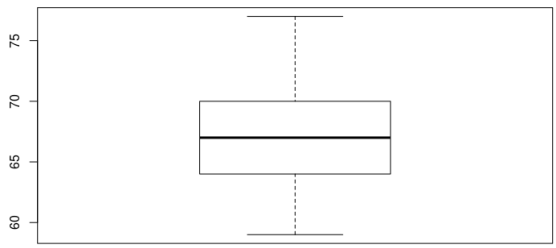

Рис. 23. Боксплот для роста респондентов из набора данных cdc150

В боксплоте 23 медиана расположена посередине ящика, т.е. равна среднему, но усы все-таки отличаются. Построим график функции плотности распределения:

```
> library(sm)
> sm.density(cdc150$height,model="norm")
```

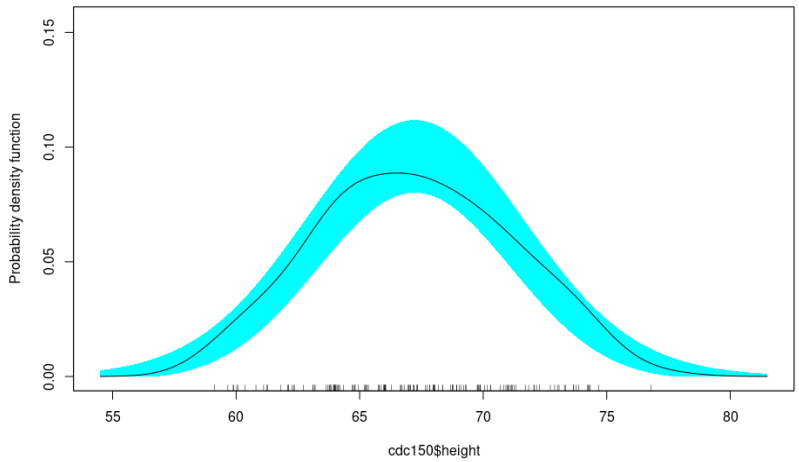

Рис. 24. График функции плотности распределения роста респондентов из набора данных cdc150

Линия для выборочного распределения находится в области для нормального распределения (Рис. 24).

Проверим нормальность распределения по росту с помощью критерия Шапиро-Уилка:

```
> shapiro.test(cdc150$height)

        Shapiro-Wilk normality test

data:  cdc150$height
W = 0.98225, p-value = 0.05003
```

Гипотезу о нормальности оставляем в силе, p-value не меньше 0.05, для описания роста респондентов будем использовать параметры нормального распределения – среднее и стандартное отклонение.

```
> mean(cdc150$height)
[1] 67.22
> sd(cdc150$height)
[1] 3.873624
> 2*sd(cdc150$height)
[1] 7.747249
```

Таким образом, рост респондентов следует представить как 67.22±7.75 дюйма

```
> mean(cdc150$height)-2*sd(cdc150$height)
[1] 59.47275
> mean(cdc150$height)+2*sd(cdc150$height)
[1] 74.96725
```

При этом с 95%ной достоверностью можно утверждать, что респонденты не ниже 59.47 дюймов, но не выше примерно 75 дюймов.

Опишем вес респондентов.

```
> boxplot(cdc150$weight)
```

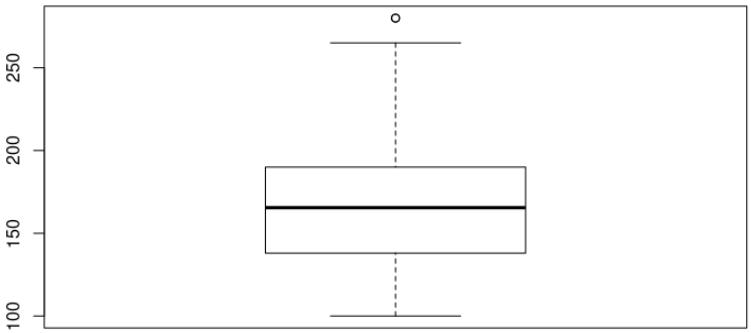

Рис. 25. Боксплот для веса респондентов из набора данных cdc150

Боксплот явно асимметричен (Рис. 25), построим график функции плотности распределения веса:

```
> sm.density(cdc150$weight,model="norm")
```

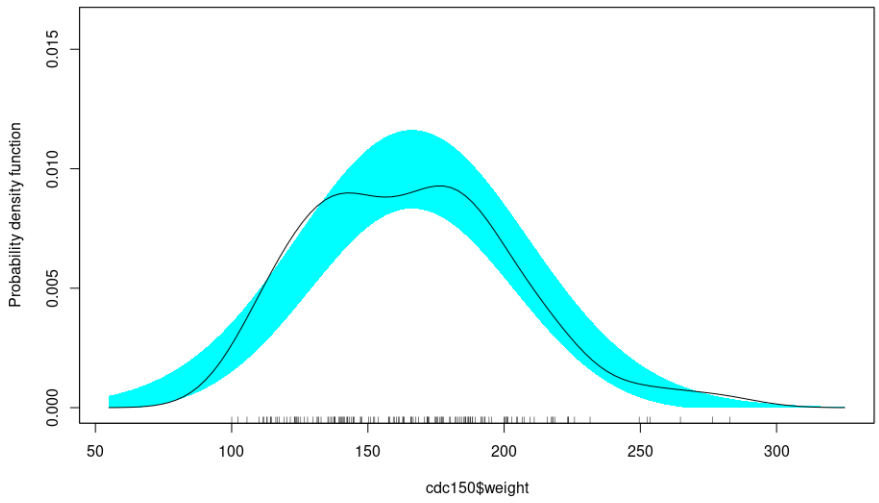

Рис. 26. График функции плотности распределения веса респондентов из набора данных cdc150

Выборочная линия функция плотности вероятности выходит за границы доверительной области для нормального распределения, и вид ее отличается от нормального (Рис.26).

```
> shapiro.test(cdc150$weight)

	Shapiro-Wilk normality test

data:  cdc150$weight
W = 0.96863, p-value = 0.001665
```

В тесте Шапиро-Уилка нулевая гипотеза отвергается, p-value < 0.05.

Вывод: нельзя использовать параметры нормального распределения. Считаем медиану и перцентили.

```
> median(cdc150$weight)
[1] 165.5
> quantile(cdc150$weight, c(0.025, 0.975))
  2.5%   97.5% 
109.175 250.000 
```

Представляем вес респондентов как 165.5 (109.18; 250), т.е. с 95%ной достоверностью можно утверждать, что вес респондентов не меньше 109.18 фунтов, но не больше 250 фунтов.

Задание. Оцените вариабельность возраста респондентов-мужчин из базы данных cdc

Задание. Опишите вес цыплят из датафрейма chickwts

Тема 3. Параметрические критерии сравнения количественных признаков

Для сравнительного анализа данных используются критерии значимости, которые разделяют на параметрические и непараметрические, двухвыборочные и для сравнения более двух выборок, предназначенные для сравнения количественных и качественных признаков и т.д. И все они строятся по общей схеме. Сначала формулируется нулевая гипотеза H_0 об отсутствии различий между группами. Затем рассчитывается фактическое значение Ф (в разных критериях используются различные формулы). Фактическое значение сравнивается с критическим К, рассчитанным для случая, когда справедлива нулевая гипотеза. Если Ф < К, нулевая гипотеза остается в силе. Если Ф > К, нулевая гипотеза отвергается, и указывается уровень значимости (p-value), при котором нулевая гипотеза отвергается.

Рассмотрим параметрические критерии, в которых используются параметры нормального распределения, такие как среднее, дисперсия, поэтому они предназначены только для сравнения выборок нормально распределенных признаков.

Двухвыборочные параметрические критерии

F критерий Фишера

Предназначен для сравнения дисперсий

$H_0: \sigma_1^2=\sigma_2^2$

$F = s_1^2/s_2^2, \quad s_1^2>s_2^2$

$F_{кр}(\alpha, df_1, df_2)$

$F \geq F_{кр}$ ~~H_0~~: $\sigma_1^2 \neq \sigma_2^2$ $F < F_{кр}$ $H_0: \sigma_1^2=\sigma_2^2$

p-value <0.05 отклоняем $\mathbf{H_0}$

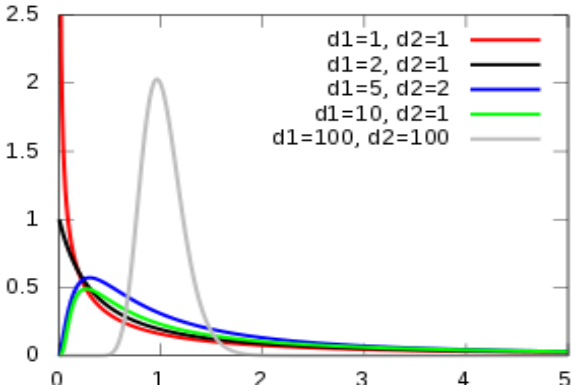

Рис.27. Графики F распределения

В a сохраним 10 случайных значений из нормального распределения со средним 5 и стандартным отклонением 2, в b - 15 случайных значений из нормального распределения со средним 5 и стандартным отклонением 6.

```
> a<-rnorm(10, mean=5,sd=2)
> b<-rnorm(15, mean=5,sd=6)
```

Построим боксплоты (Рис.28):

```
> boxplot(a,b)
```

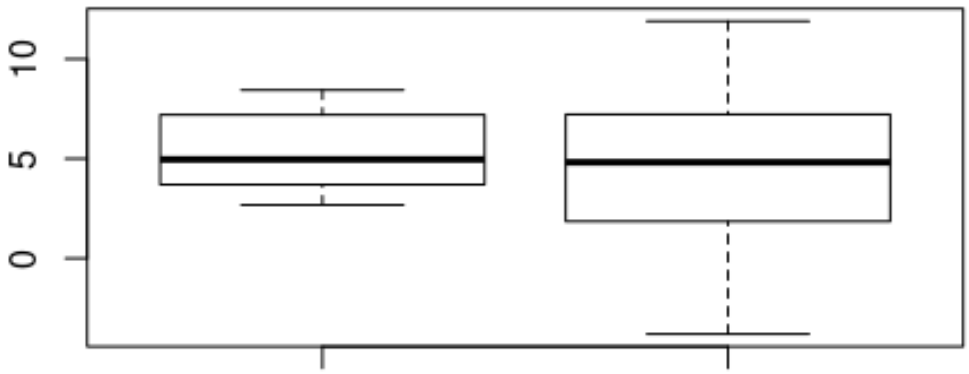

Рис. 28. Боксплоты для двух нормально распределенных выборок

Сравним дисперсии выборок a и b с помощью критерия Фишера

```
> var.test(a,b)

        F test to compare two variances

data:  a and b
F = 0.21376, num df = 9, denom df = 14, p-value = 0.0252
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.0666058 0.8118428
sample estimates:
ratio of variances
          0.213758
```

Поскольку p-value<0.05, делаем вывод о значимом отличии дисперсий выборок (p-value = 0.0252).

Создадим еще одну выборку b1, сохранив 15 случайных значений из нормального распределения со средним 4.5 и стандартным отклонением 2.5:

```
> b1<-rnorm(15, mean=4.5,sd=2.5)
```

Построим боксплоты для выборок a и b1 (Рис.29) :

```
> boxplot(a,b1)
```

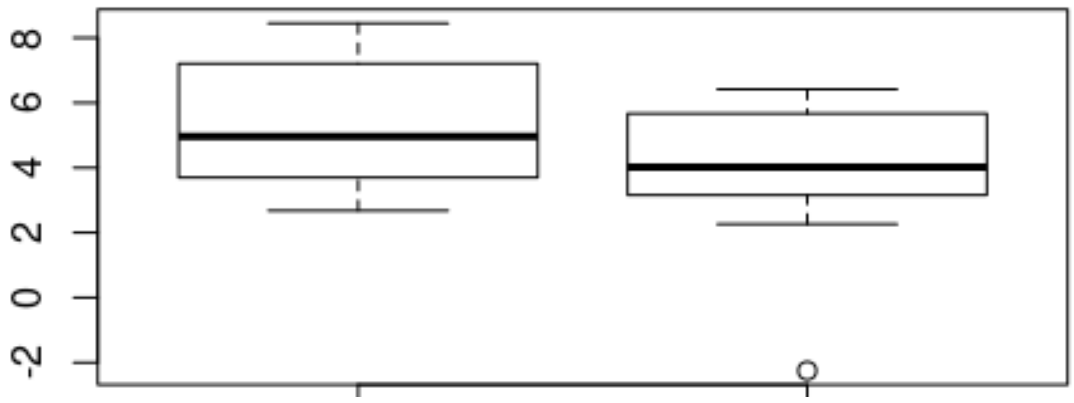

Рис.29. Боксплоты для двух нормально распределенных выборок с одинаковыми дисперсиями

Сравним дисперсии выборок a и b1:

```
> var.test(a,b1)
```

F test to compare two variances

data: a and b1
F = 0.8253, num df = 9, denom df = 14, p-value = 0.7916
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.257158 3.134440
sample estimates:
ratio of variances
0.8252973

Дисперсии однородны (p-value = 0.79), теперь можно применять к. Стьюдента

Задание. Сравните вариабельность возраста женщин и мужчин из базы данных cdc150

Критерий Стюдента, t-тест

Предназначен для сравнения средних двух независимых выборок

$H_0: \mu_1=\mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}}$$

$t_{кр}(\alpha, df)$ $df=n_1+n_2-2$

$t \geq t_{кр}$ ~~$H_0: \mu_1\neq\mu_2$~~ $t < t_{кр}$ $H_0:\mu_1=\mu_2$

p-value <0.05 отклоняем H_0

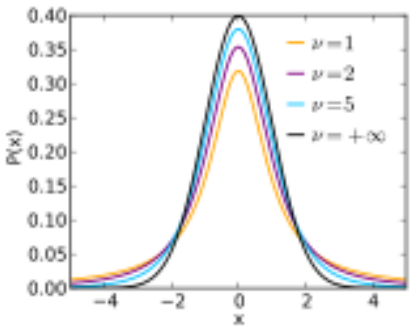

Рис. 30. График t-распределения

Выше мы определили, используя критерий Фишера, что дисперсии выборок a и b1 однородны, теперь можно сравнить средние:

> t.test(a,b1)

Welch Two Sample t-test
data: a and b1
t = 1.6222, df = 20.721, p-value = 0.1199
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3880188 3.1300571
sample estimates:
mean of x mean of y
 5.309846 3.938827

В R t-тест реализован в модификации Уэлча, более чувствительной при сравнении средних выборок разных объемов. Поскольку p-value > 0.05, нулевая гипотеза остается в силе, делаем заключение об отсутствии различий между выборками a и b1.

В a1 из нормального распределения со средним 7 и стандартным отклон ением 4 сохраним 10 случайных значений и построим боксплоты для выборо к a1 и b1 (Рис.31):

```
> a1<-rnorm(10, mean=7,sd=4)
> boxplot(a1,b1)
```

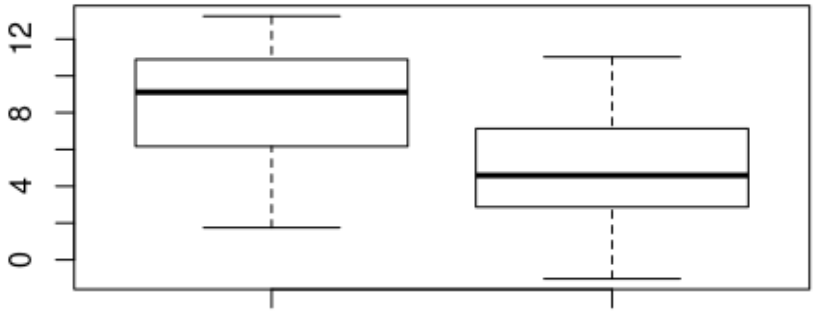

Рис. 31. Боксплоты для двух нормально распределенных выборок с одинаковыми дисперсиями

Сравним дисперсии выборок a1 и b1, используем критерий Фишера

```
> var.test(a1,b1)
```

F test to compare two variances

data: a1 and b1
F = 1.2263, num df = 9, denom df = 14, p-value = 0.7064
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.3820967 4.6572891
sample estimates:
ratio of variances
1.226263

Поскольку p-value > 0.05, дисперсии выборок a1 и b1 однородны, сравн им средние этих выборок:

```
> t.test(a1,b1)
```

Welch Two Sample t-test

data: a1 and b1
t = 2.5367, df = 18.022, **p-value = 0.02065**
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:

0.619002 6.584199
sample estimates:
mean of x mean of y
8.603192 5.001591

Нулевую гипотезу о равенстве средних выборок a1 и b1 отвергаем на уровне значимости 0.02

var.test() и **t.test()** работают также с функциями:

Рассмотрим на примере признака желаемый вес (wtdesire) у женщин и м ужчин из таблицы данных cdc150.

Построим боксплот для этого признака, разделив на группы по полу (Рис. 32):

```
> boxplot(wtdesire~gender, data=cdc150)
```

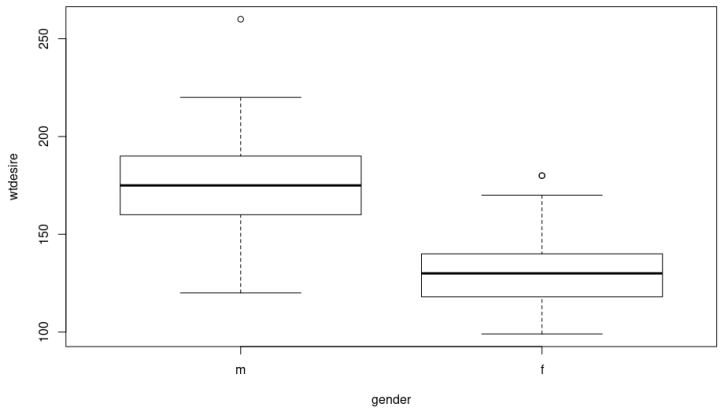

Рис. 32. Боксплоты желаемого веса у мужчин и женщин из cdc150

```
> aggregate(wtdesire~gender, data=cdc150, mean)
  gender wtdesire
1      m 175.0132
2      f 130.8243
 > aggregate(wtdesire~gender, data=cdc150, median)
  gender wtdesire
1      m      175
2      f      130
```

Проверим однородность дисперсий желаемого веса у женщин и мужчин

```
> var.test(wtdesire~gender, data=cdc150)
```

F test to compare two variances

data: wtdesire by gender
F = 1.5708, num df = 75, denom df = 73, p-value = 0.05428
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:

0.9916609 2.4846580
sample estimates:
ratio of variances
1.570825

Дисперсии однородны, проведем сравнение средних, применим t-тест:

```
> t.test(wtdesire~gender, data=cdc150)

        Welch Two Sample t-test

data:  wtdesire by gender
t = 12.796, df = 142.53, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 37.36240 51.01526
sample estimates:
mean in group m mean in group f
      175.0132        130.8243
```

Нулевую гипотезу отвергаем, желаемый вес (wtdesire) у женщин и мужчин значимо отличается (p-value < 2.2e-16)

Для визуализации групповых распределений (Рис.34) удобно использовать **sm.density.compare**():

```
> sm.density.compare(cdc150$wtdesire, cdc150$gender, model = "none")
```

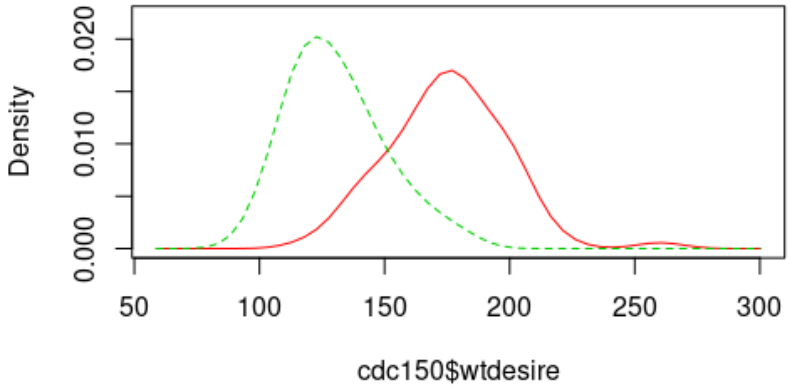

Рис.33. Графики функции плотности распределения желаемого веса мужчин и женщин из набора данных cdc150

Задание. Сравните возраст женщин и мужчин из базы данных cdc150

Задание. Сравните длину чашелистиков цветков ирисов сорта setosa с длиной чашелистиков цветков сортов versicolor и virginica

Парный t-тест (повторные измерения, до-после)

Одна из целей правильно спланированного эксперимента — свести к минимуму все источники вариаций, кроме экспериментального эффекта, но невозможно устранить вариацию, присущую биологическому признаку. В таких случаях лучшим дизайном эксперимента будет измерение значений изучаемого признака до воздействия фактора и после, т.е. каждый субъект или объект измеряется дважды, в результате чего получаются пары наблюдений. В таких случаях для сравнения средних значений используется парный t-тест, и анализируют выборку различий в парах.

$H_0: \mu_1=\mu_2$

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

$t_{кр}(\alpha, df)$ $df=n-1$

$t \geq t_{кр}$ ~~H_0~~: $\mu_1 \neq \mu_2$ $t < t_{кр}$ $H_0: \mu_1=\mu_2$

p-value <0.05 отклоняем H_0

В качестве примера проанализируем индекс желтизны маховых перьев птиц.

Загрузим таблицу данных в рабочее пространство:

```
> Input = (" Bird   Typical  Odd
A      -0.255   -0.324
B      -0.213   -0.185
C      -0.190   -0.299
D      -0.185   -0.144
E      -0.045   -0.027
F      -0.025   -0.039
G      -0.015   -0.264
H       0.003   -0.077
I       0.015   -0.017
J       0.020   -0.169
K       0.023   -0.096
L       0.040   -0.330
M       0.040   -0.346
N       0.050   -0.191
O       0.055   -0.128
P       0.058   -0.182 ")
 > b = read.table(textConnection(Input),header=TRUE)
```

Построим боксплоты (Рис.34):

```
> boxplot(b$Typical, b$Odd)
```

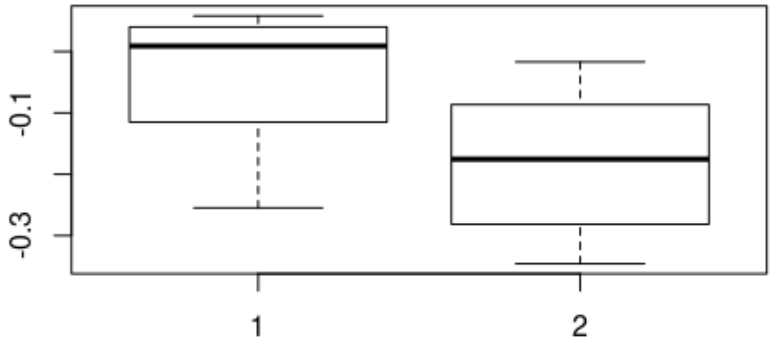

Рис. 34. Боксплоты индекса желтизны маховых перьев (1-Typical, 2- Odd)

Проверим признак на нормальность распределения:

```
> shapiro.test(c(b$Typical, b$Odd))

        Shapiro-Wilk normality test

data:  c(b$Typical, b$Odd)
W = 0.92062, p-value = 0.02161
```

Критерий Шапиро-Уилка показывает, что распределение изучаемого признака имеет отличия от нормального распределения. Добавим в таблицу новый столбец, в котором сохраним разницу в индексе желтизны для каждой птицы, построим боксплот для разницы (Рис.35):

```
> b$d<-b$Typical- b$Odd
> boxplot(b$d)
```

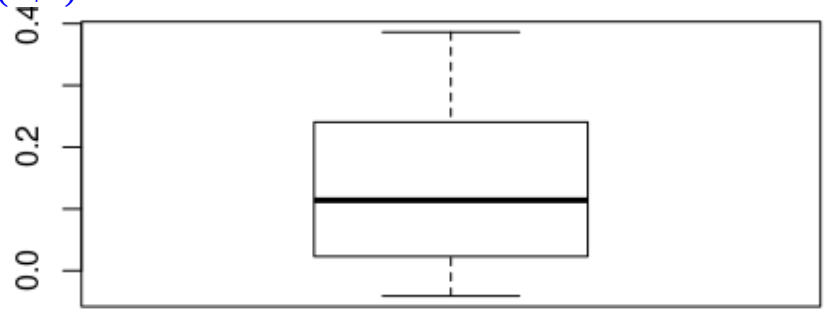

Рис. 35. Боксплот разницы в парах значений индекса желтизны маховых перьев

Проверим нормальность распределения разницы с помощью критерия Шапиро-Уилка

```
> shapiro.test(b$d)

        Shapiro-Wilk normality test
data:  b$d
W = 0.93987, p-value = 0.3474
```

С помощью критерия Шапиро-Уилка подтвердили нормальность распределения парного различия. Построим коробчатую диаграмму с линиями, связывающими парные измерения:

```
> typic<-b$Typical
> odd<-b$Odd
> pd <- paired(typic, odd)
> plot(pd, type = "profile") + theme_bw()
```

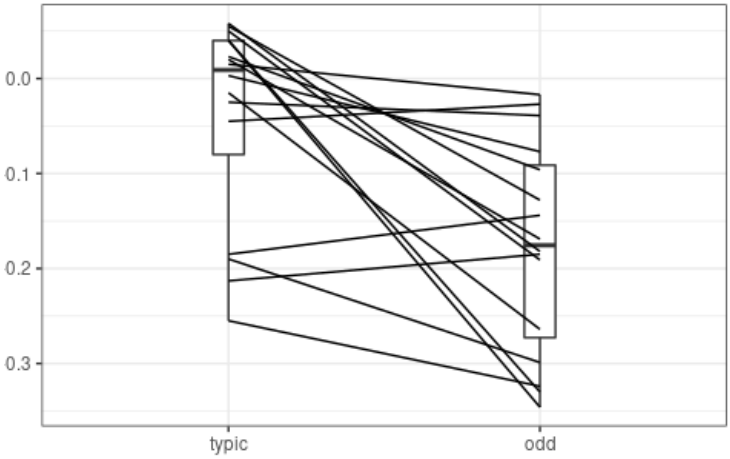

Рис. 36. Боксплоты значений индекса желтизны маховых перьев с линиями, связывающими пары измерений

В 12ти парах видим снижение показателя (Рис.36)

Проведем парный t-тест, используя функцию **t.test()** и придав аргументу **paired** логическое значение **TRUE** :

```
> t.test(b$Typical, b$Odd, paired = TRUE)
```

```
        Paired t-test

data:  b$Typical and b$Odd
t = 4.0647, df = 15, p-value = 0.001017
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.06521848 0.20903152
sample estimates:
mean of the differences
             0.137125
```

Выявили значимое различие в индексе желтизны маховых перьев (**p-value = 0.001017**)

Задание. При изучении скорости ферментативной реакции студенты делили аликвоты экстракта цветной капусты пополам, к одной половине добавляли малонат, получили следующие результаты:

student	a	b	c	d	e	f	g	h	i	j	k	l	
no_malonate	0.026	0.052	0.045	0.09	0.012	0.084	0.006	0.078	0.016	0.005	0.12	0.055	0.05
malonate	0.026	0.047	0.005	0.088	0.02	0.087	0.007	0.074	0.015	0.001	0.08	0.035	0.05

Является ли малонат ингибитором ферментативной реакции?

Сравнение более двух выборок нормально распределенного признака
Дисперсионный анализ

Для примера разберем ширину лепестков цветков ирисов трех видов ирисов:

Проверим этот признак на нормальность:

```
> shapiro.test(iris$Sepal.Width)

        Shapiro-Wilk normality test

data:  iris$Sepal.Width
W = 0.98492, p-value = 0.1012
```

Построим боксплоты (Рис.37):

```
> boxplot(Sepal.Width~Species, data=iris)
```

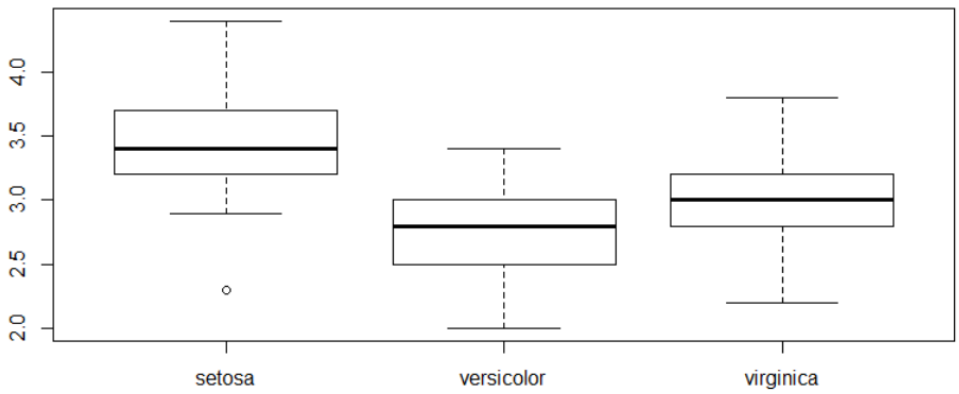

Рис. 37. Боксплоты ширины лепестков цветков ирисов трех видов

Подсчитаем выборочные средние:

```
> aggregate(Sepal.Width~Species, data=iris, mean)
     Species Sepal.Width
1     setosa       3.428
2 versicolor       2.770
3  virginica       2.974
```

Представим на одном графике значения в группах и групповые средние (Рис.38):

```
> stripchart(iris$Sepal.Width~iris$Species, xlab="Sepal.Width",ylab="Species")
> means<-tapply(iris$Sepal.Width,iris$Species, mean)
> stripchart(means, method="stack", col="red", at=2.5,add=T)
```

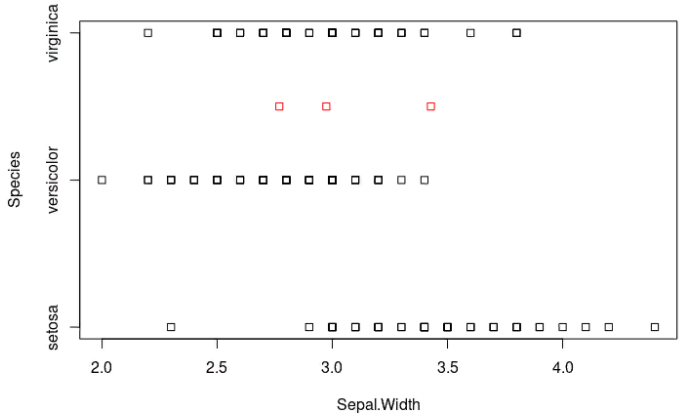

Рис. 38. Значения ширины лепестков цветков ирисов трех видов и групповые средние (показаны красными квадратами)

Суть дисперсионного анализа состоит в том, что мы можем оценить общую дисперсию ширины лепестков цветков трех видов ирисов двумя способами: на основе внутригрупповой вариации и на основе вариации групповых средних (на графике средние значения ширины лепестков для каждого сорта ирисов показаны красными квадратами). Сравнение двух оценок общей дисперсии, межгрупповой (факториальной) и внутригрупповой (остаточной), покажет, зависит ли ширина лепестков от сорта ирисов .

В R имеется много функций для проведения дисперсионного анализа, используем базовую функцию **aov(),** в которой дисперсионный комплекс задается формулой.

```
> aov(Sepal.Width~Species, data=iris)

Call:
   aov(formula = Sepal.Width ~ Species, data = iris)

Terms:
                  Species Residuals
Sum of Squares  11.34493  16.96200
Deg. of Freedom         2       147
Residual standard error: 0.3396877
Estimated effects may be unbalanced
```

Заполним таблицу дисперсионного анализа:

```
> summary(aov(Sepal.Width~Species, data=iris))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  11.35   5.672   49.16 <2e-16 ***
Residuals   147  16.96   0.115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

В таблице Df- число степеней свободы, Sum Sq -вариация, Mean Sq -дис персия, F - фактическое значение F, value Pr(>F) - уровень значимости, при котором отвергается нулевая гипотеза о равенстве групповых средних.

Уровень значимости <2e-16 меньше 0.05, это означает, что хотя бы одна группа в дисперсионном комплексе отличается от остальных, что ширина лепестков цветков зависит от вида ирисов. Для того, чтобы определить, какая группа от какой отличается значимо, надо провести попарные множест венные сравнения.

Задание. Зависит ли урожай (вес сухих растений) от типа обработки растений? Используйте данные таблицы PlantGrowth

Задание. Связана ли летальность (status) с возрастом (age) пациентов? Используйте данные таблицы данных Melanoma из пакета MASS

Множественное попарное сравнение

Для того, чтобы выяснить, какие группы значимо отличаются, после дисперсионного анализа используют методы попарного сравнения с учетом эффекта множественности сравнений: критерий Стьюдента с поправкой Бонферрони, тест Тьюки, тесты Данна, Даннета, Ньюмана-Кейсла.

В базовой версии R имеется функция **TukeyHSD()**, которая сочетается с функцией **aov()**, и кроме поправленного на множественность сравнений уровня значимости (p adj) для каждой пары строит 95%ные доверительные интервалы для разности средних, показывает среднее различие в паре (diff), нижнюю (lwr) и верхнюю (upr) границы доверительного интервала.

```
> TukeyHSD(aov(Sepal.Width~Species, data=iris))
```

```
 Tukey multiple comparisons of means
   95% family-wise confidence level

Fit: aov(formula = Sepal.Width ~ Species, data = iris)
```

```
$Species
                          diff         lwr        upr     p adj
versicolor-setosa    -0.658 -0.81885528 -0.4971447 0.0000000
virginica-setosa     -0.454 -0.61485528 -0.2931447 0.0000000
virginica-versicolor  0.204  0.04314472  0.3648553 0.0087802
```

Поскольку поправленное на множественность сравнений p adj во всех трех парах меньше 0.05, с 95%-ной достоверностью утверждаем, что цветки трех видов ирисов (setosa, versicolor, virginica) отличаются друг от друга по ширине лепестков.

Задание. Как влияет тип обработки растений на урожай? Используйте данные таблицы PlantGrowth.

Тема 4. Непараметрические критерии сравнения количественных признаков

При сравнении вариантов количественных признаков, распределение которых отличается от нормального, некорректно использовать параметрические критерии, поскольку они опираются на параметры нормального распределения. В таких случаях применяют непараметрические критерии, в которых для расчетов используются не значения количественного признака, а их ранги. Для сравнения двух групп используют критерий Уилкоксона-Манна-Уитни, если сравниваются больше двух групп, применяют критерий Крускала-Уоллиса.

Сравнение двух независимых выборок

Разберем ранговые методы на примере данных первых 150 респондентов мониторинга состояния здоровья (датафрейм cdc150).

Проанализируем вес женщин и мужчин. Сначала посторим боксплоты веса мужчин и женщин (Рис. 39):

```
> boxplot(weight~gender,data=cdc150)
```

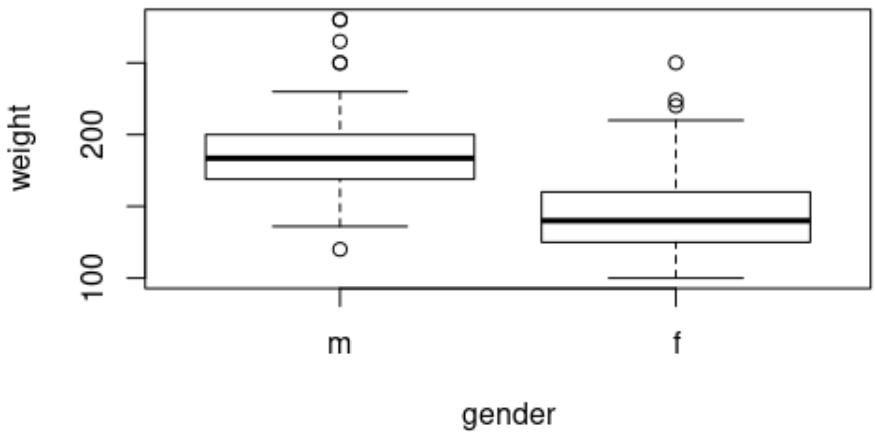

Рис. 39. Вес мужчин и женщин из датафрейма cdc150

Создадим таблицу с оценкой выборочных медиан веса мужчин и женщин:

```
> aggregate(weight~gender,data=cdc150,median)
  gender weight
1      m  183.5
2      f  140.0
```

Проверим нормальность распределения веса с помощью критерия Шапиро-Уилка:

```
> shapiro.test(cdc150$weight)

        Shapiro-Wilk normality test

data:  cdc150$weight
W = 0.96863, p-value = 0.001665
```

Поскольку уровень значимости p-value = 0.001665 меньше 0.05, распределение веса не согласуется с нормальным распределением, не можем использовать параметрический t-тест. Применим его ранговый аналог критерий Манна-Уитни (или критерий ранговых сумм Вилкоксона).

В R этот критерий реализован в базовой функции **wilcox.test():**

```
> wilcox.test(weight~gender,data=cdc150)

        Wilcoxon rank sum test with continuity correction
data:  weight by gender
W = 4616, p-value = 1.176e-11
alternative hypothesis: true location shift is not equal to 0
```

Уровень значимости меньше 0.05, вес мужчин и женщин отличается (p-value=1.176e-11)

Задание. Сравните вес курящих и некурящих мужчин из базы данных cdc150

Задание. Есть ли различия в весе женщин, делающих и не делающих зарядку, из базы данных cdc150?

Сравнение двух выборок (повторные измерения, до-после)

Для сравнения двух измерений для признаков, распределение которых отличается от нормального, существует ранговый аналог парного t-теста – тест Вилкоксона (или критерий знаковых рангов Вилкоксона), в основе которого лежит анализ выборки разницы рангов в парах измерений. В R используется функция для сравнения рангов двух выборок **wilcox.test()** , и по аналогии с парным t-тестом, добавляется аргумент **paired**, которому присваивается логическое значение TRUE

Разберем пример из исследования мечехвостов. Выясним, изменилось ли их количество в 2012 г по сравнению с 2011 г., подсчет производили на одних и тех же пляжах:

```
> Input = ("
Beach               Year.2011    Year.2012
'Bennetts Pier'      35282        21814
'Big Stone'         359350        83500
'Broadkill'          45705        13290
'Cape Henlopen'      49005        30150
'Fortescue'          68978       125190
'Fowler'              8700         4620
'Gandys'             18780        88926
'Higbees'            13622         1205
'Highs'              24936        29800
'Kimbles'            17620        53640
'Kitts Hummock'     117360        68400
'Norburys Landing'  102425        74552
'North Bowers'       59566        36790
'North Cape May'     32610         4350
'Pickering'         137250       110550
'Pierces Point'      38003        43435
'Primehook'         101300        20580
'Reeds'              62179        81503
'Slaughter'         203070        53940
'South Bowers'      135309        87055
'South CSL'         150656       112266
'Ted Harvey'        115090        90670
'Townbank'           44022        21942
'Villas'             56260        32140
'Woodland'             125         1260")
```

Сохраним данные в датафрейм crab:

```
> crab = read.table(textConnection(Input),header=TRUE)
```

Построим боксплоты количества мечехвостов для 2011 и 2012:

```
> boxplot(crab$Year.2011,crab$Year.2012)
```

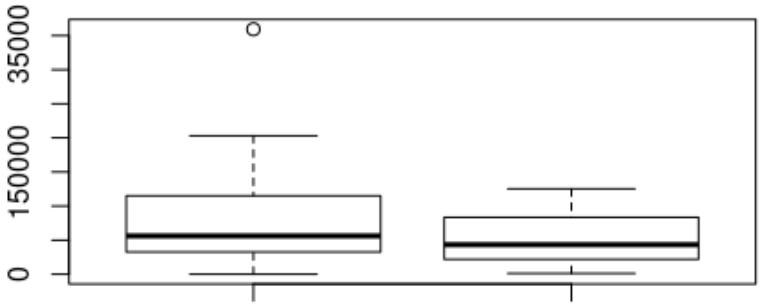

Рис. 40. Количество мечехвостов, собранных в 2011г. (слева) и 2012г.(справа)

Похоже, распределение количества мечехвостов асимметрично (Рис.40). Посчитаем разницу количества мечехвостов для каждого пляжа, для выборки различий построим боксплот:

```
> crab$d<-crab$Year.2011-crab$Year.2012
> boxplot(crab$d)
```

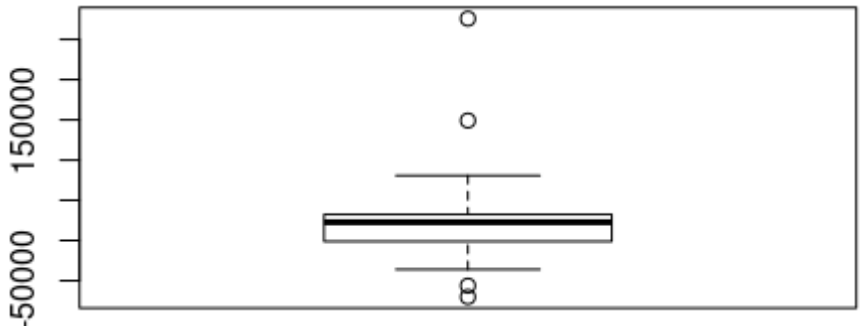

Рис. 41. Боксплот для разницы количества мечехвостов, собранных в 2011г. и 2012г

На боксплоте (Рис. 41) очевидна асимметрия разницы в количестве. Проверим нормальность, применив критерий Шапиро-Уилка:

```
> shapiro.test(crab$d)

        Shapiro-Wilk normality test

data:  crab$d
W = 0.76725, p-value = 6.778e-05
```

Уровень значимости меньше 0.05, распределение признака отличается от нормального. Используем ранговый двухвыборочный критерий для повторных измерений:

```
> wilcox.test(crab$Year.2011,crab$Year.2012 , paired = T)
```

```
        Wilcoxon signed rank test

data:  crab$Year.2011 and crab$Year.2012
V = 249, p-value = 0.01874
alternative hypothesis: true location shift is not equal to 0
```

Уровень значимости меньше 0.05, делаем вывод о значимом различии количества мечехвостов в 2011 и 2012 гг.

Рассмотрим еще один пример парных измерений длины маховых перьев у птиц. Данные сохраняем во фрейме D:

```
> Input = ("Bird    Feather   Length
A       Typical   -0.255
B       Typical   -0.213
C       Typical   -0.19
D       Typical   -0.185
E       Typical   -0.045
F       Typical   -0.025
G       Typical   -0.015
H       Typical    0.003
I       Typical    0.015
J       Typical    0.02
K       Typical    0.023
L       Typical    0.04
M       Typical    0.04
N       Typical    0.05
O       Typical    0.055
P       Typical    0.058
A       Odd       -0.324
B       Odd       -0.185
C       Odd       -0.299
D       Odd       -0.144
E       Odd       -0.027
F       Odd       -0.039
G       Odd       -0.264
H       Odd       -0.077
I       Odd       -0.017
J       Odd       -0.169
K       Odd       -0.096
L       Odd       -0.33
M       Odd       -0.346
N       Odd       -0.191
O       Odd       -0.128
P       Odd       -0.182")
   > D = read.table(textConnection(Input),header=TRUE)
```

Построим график боксплотов длины типичных и добавочных маховых п
ерьев и покажем пары измерений у каждой птицы:

```
> library(PairedData)
> typic<-D$Length[D$Feather=="Typical"]
> odd<-D$Length[D$Feather=="Odd"]
> pd <- paired(typic, odd)
> plot(pd, type = "profile") + theme_bw()
```

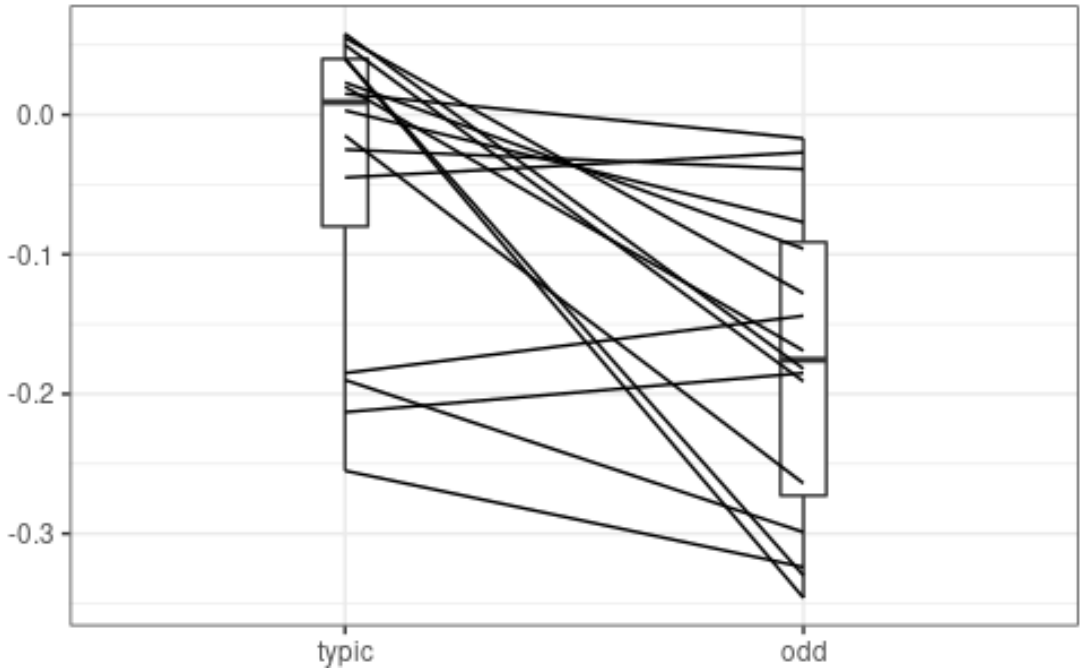

Рис. 42. Боксплоты длины типичных и добавочных маховых перьев

Очевидна асимметрия распределения признака (Рис. 42), проведем проверку распределения длины маховых перьев на нормальность:

```
> shapiro.test(D$Length)
        Shapiro-Wilk normality test
data:  D$Length
W = 0.92062, p-value = 0.02161
```

Уровень значимости меньше 0.05, распределение признака отличается от нормального. Используем ранговый двухвыборочный критерий для парных измерений:

```
> wilcox.test(Length ~ Feather, data=D,  paired=TRUE)
Wilcoxon signed rank test
data:  Length by Feather
V = 10, p-value = 0.001312
alternative hypothesis: true location shift is not equal to 0
```

Уровень значимости меньше 0.05, делаем вывод о значимом различии в длине маховых перьев (p-value = 0.001312)

Задание. В датафрейме Tobacco из пакета PairedData представлены количества поражений, вызванных двумя вирусными препаратами, инокулированными в две половинки каждого табачного листа. Отличаются ли эти вирусные препараты по способности поражать листья табака?

Непараметрическое сравнение более двух выборок

В случае, когда распределение исследуемого не подчиняется нормальному закону, и сравниваются 3 или больше выборок, используют ранговый аналог дисперсионного анализа – критерий Крускала-Уоллиса. В R есть базовая функция **kruskal.test()**, в которой данные задаются в виде формулы.

Разберем на примере анализа длины лепестков (Sepal.Length) цветков трех видов ирисов из встроенной таблицы данных iris.

Построим боксплоты (Рис. 43) и сведем в таблицу медианы длины лепестков цветков ирисов видов setosa, versicolor, virginica :

```
> boxplot(Sepal.Length~Species,data=iris)
```

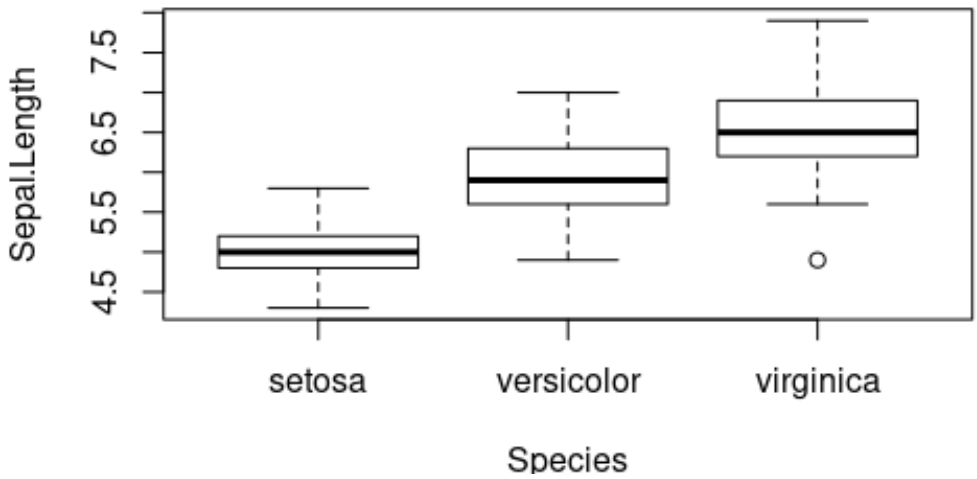

Рис. 43. Боксплоты длины лепесков цветков трех видов ирисов

```
> aggregate(Sepal.Length~Species,data=iris, median)
   Species Sepal.Length
1     setosa          5.0
2 versicolor          5.9
3  virginica          6.5
```

Проверим на нормальность с помощь критерия Шапиро-Уилка:

```
> shapiro.test(iris$Sepal.Length)

        Shapiro-Wilk normality test

data:  iris$Sepal.Length
W = 0.97609, p-value = 0.01018
```

Распределение признака длина лепестков отличается от нормального, сравниваем 3 группы - применим ранговый критерий Крускала-Уоллиса:

```
> kruskal.test(Sepal.Length~Species,data=iris)

        Kruskal-Wallis rank sum test

data:  Sepal.Length by Species
Kruskal-Wallis chi-squared = 96.937, df = 2, p-value < 2.2e-16
```

Рассчитанный уровень значимости меньше 0.05, нулевую гипотезу об отсутствии различий отвергаем, хотя бы одна группа отличается от остальных (p-value < 2.2e-16)

Для выявления, какие именно группы отличаются, необходимо провести попарное множественное сравнение.

Ранговые методы множественного попарного сравнения

Если критерий Крускала-Уоллиса является значимым, следует провести апостериорный анализ, чтобы определить, какие группы отличаются друг от друга. В R есть базовая функция **pairwise.wilcox.test()**, которая выполняет ранговое попарное сравнение с поправкой на множественность сравнений. Выполним расчет для длины лепестков цветков ирисов трех видов :

```
> pairwise.wilcox.test(iris$Sepal.Length,iris$Species,p.adjust.method = "bonf")

        Pairwise comparisons using Wilcoxon rank sum test

data:  iris$Sepal.Length and iris$Species

           setosa  versicolor
versicolor 2.5e-13 -
virginica  < 2e-16 1.8e-06

P value adjustment method: bonferroni
```

Рассчитанные с поправкой Бонферрони p-value для всех пар меньше 0.05, длины лепестков цветков ирисов видов setosa, versicolor, virginica значимо отличаются друг от друга.

Тест Данна для множественных сравнений

В пакетах R реализованы многие ранговые критерии для множественных сравнений. Вероятно, самым популярным тестом для этого является тест Данна, который выполняется с помощью функции **dunnTest()** в пакете **FSA**. Корректировку значений p можно внести с помощью разных методов. Зар (2010) утверждает, что тест Данна подходит для групп с неодинаковым количеством наблюдений.

```
> library(FSA)
> dunnTest(Sepal.Length~Species,data=iris,method="bh")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Benjamini-Hochberg method.

              Comparison         Z      P.unadj        P.adj
1    setosa - versicolor -6.106326 1.019504e-09 1.529257e-09
2     setosa - virginica -9.741785 2.000099e-22 6.000296e-22
3 versicolor - virginica -3.635459 2.774866e-04 2.774866e-04
```

Рассчитанные с поправкой Бенжамини-Хокберга значения p (**P.adj**) для всех пар меньше 0.05, длины лепестков цветков ирисов сорта setosa, versicolor, virginica значимо отличаются друг от друга.

Задание. Проанализируйте длину чашелистиков (Petal.Length) цветков ирисов сорта setosa, versicolor, virginica из встроенной таблицы iris

Задание. Зависит ли урожай (вес сухих растений) от типа обработки растений? Используйте данные таблицы PlantGrowth

ЗАКЛЮЧЕНИЕ

Таким образом, в Части1 разобраны методы разведочного анализа данных, их визуализации, проверки распределения количественных признаков на нормальность, методы их описания в зависимости от модели распределения, параметрические и непараметрические критерии для их сравнения.

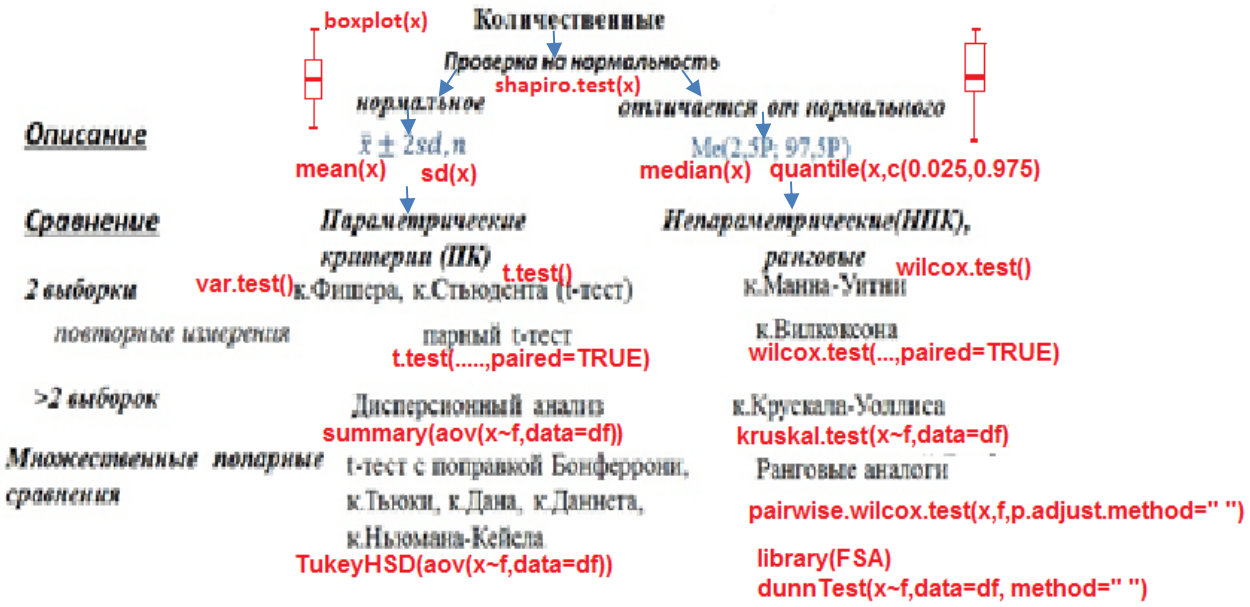

Рис. 44. Схема-определитель статистических методов анализа количественных признаков с функциями языка R

На рисунке (Рис. 44) представлен фрагмент схемы-определителя статистических методов для анализа количественных признаков, дополненный функциями для их визуализации, описания, сравнения на языке R.

КОНТРОЛЬНЫЕ ЗАДАЧИ

1. Используя набор анатомических данных домашних кошек cats из пакета MASS, сравните **а)** вес тела (`Bwt`), **б)** вес сердца (`Hwt`) у котов и кошек

2. Определите, изменился ли вес девочек после лечения анорексии, используйте набор данных Anorexia из пакета PairedData

3. В наборе данных BloodLead из пакета PairedData представлены сопоставленные парные данные, соответствующие уровням свинца в крови для 33 детей, родители которых работали на фабрике, связанной со свинцом, и 33 контрольных детей из их района. Сравните уровень свинца в крови детей этих двух групп.

4. Синдром Кушинга — гипертоническое расстройство, связанное с избыточной секрецией кортизола надпочечниками. Наблюдения в наборе данных Cushings из пакета MASS представляют собой скорость выведения с мочой двух стероидных метаболитов. Определите, **а)** отличается ли уровень экскреции тетрагидрокортизона у больных с разным типом синдрома, **б)** отличается ли уровень экскреции прегнанетриола у больных с разным типом синдрома. **в)** Сравните уровни экскреции надпочечниками тетрагидрокортизона и прегнанетриола у больных с синдромом Кушинга

5. В базе данных Melanoma из пакета MASS содержатся данные о 205 пациентах со злокачественной меланомой в Дании. Сравните летальность (`time`) **а)** у мужчин и женщин, **б)** в группах пациентов различного статуса. **в)** Опишите признак `thickness` и сравните толщину опухоли у мужчин и женщин

6. Датафрейм Rabbit из пакета MASS представляет набор данных, полученных при исследовании артериального давления у кроликов. Пять кроликов были исследованы дважды, после лечения физиологическим раствором (контроль) и после лечения антагонистом 5-HT_3 MDL 72222. После каждого лечения внутривенно вводили возрастающие дозы фенилбигуанида с 10-минутными интервалами и измеряли давлени**е.** Сравните изменение давления **(**`BPchange`**)** после лечения (`Treatment`) физиологическим раствором (`Control`) и после лечения MDL 72222(`MDL`)

7. Используя встроенный в R набор данных chickwts, определить, влияет ли тип кормления цыплят на их вес

8. Используя встроенный в R набор данных CO2, определить, влияет ли **а)** предварительное охлаждение (Treatment) и **б)** место происхождения (Type) на поглощения двуокиси углерода растением Echinochloa crus-galli (ежовник обыкновенный)

9. Используя команду source (http://www.openintro.org/stat/data/cdc.R), загрузите в рабочее пространство базу данных cdc. **а)** Определите, отличается ли вес (weight) в зависимости от состояния здоровья (genhlth) респондентов. **б)** Сравните рост мужчин и женщин, используйте коробчатый график для визуализации **в)** Добавьте новую переменную индекс массы тела (BMI) во фрейме данных cdc. BMI является соотношением веса и высоты и может быть рассчитан по формуле BMI=weight/height2×703 (703-приблизительный коэффициент для перевода имперских единиц (дюймов и фунтов) в метрические (метры и килограммы)). Определите, отличается ли BMI в зависимости от состояния здоровья (genhlth) респондентов? **г)** Сравните индекс массы тела (BMI) мужчин и женщин **д)** Сравните индексы массы тела (BMI) у респондентов-женщин старше 40 и моложе 40 лет.

10. В наборе данных Barley из пакета PairedData представлены 12 парных данных, соответствующих урожайности ячменя Glabron и Velvet, выращенного на разных фермах. Используя эти данные, **а)** опишите урожайность ячменя и представьте графически, **б)** сравните урожайность двух сортов ячменя

11. В наборе данных crabs из пакета MASS представлены результаты морфологических измерений крабов Leptograpsus variegatus, собранных в Австралии. Сравните а) размер лобной части (`FL`) у крабов голубого и оранжевого цвета, б) длину панциря (`CL`) у самцов крабов голубого цвета, в) ширину панциря (`CW`) у самок крабов оранжевого цвета, г) размер лобной части (`FL`) у самцов и самок оранжевого цвета.

12. В датафрейме SMBassWB из пакета FSA представлены данные о росте мелкоротого окуня (Micropterus dolomieu), пойманного в Вест-Беарскин-Лейк, штат Миннесота. Сравните а) возраст рыбы при вылове (agecap), б) общую длина пойманной рыбы (lencap) в зависимости от вида снасти, используемой для ловли рыбы (gear)

13. Используя набор данных Animals из пакета MASS, в котором представлены средняя масса мозга и тела 28 видов наземных животных, описать и представить графически а) массу мозга для представленных в датафрейме млекопитающих, б) массу тела приматов.

14. Популяция женщин в возрасте не менее 21 года, индейцев Пима, живущих недалеко от Финикса, штат Аризона, была проверена на диабет в соответствии с критериями Всемирной организации здравоохранения, результаты представлены в наборе данных из пакета. Сравните **а)** индекс массы тела (bmi),**б)** концентрацию глюкозы в плазме при пероральном тесте на толерантность к глюкозе(glu), **в)** диастолическое давление (bp**), г)** возраст (age) в зависимости от наличия и отсутствия диабета (type)

15. В рамках исследования факторов риска, связанных с низкой массой тела ребенка при рождении, в Медицинском центре Бэйстейт, Спрингфилд, Массачусетс были собраны данные, доступные в датафрейме birthwt в пакете MASS. Сравните **а)** вес ребенка при рождении от курящих и некурящих матерей, **б)** вес ребенка при рождении от матерей возрастных групп не старше и старше 23 лет

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

1. Н.И. Акберова, О.С. Козлова: Основы анализа данных и программирования в R: учебно-методическое пособие – Казань: Альянс, 2017. – 33 с.
2. А.Б. Шипунов, Е.М. Балдин, П.А. Волкова и др.: Наглядная статистика. Используем R! ДМК Пресс, 2014, 298 с. (ISBN: 978-5-97060-094-8)
3. 2. Роберт И. Кабаков: R в действии. Анализ и визуализация данных на языке R М.: ДМК Пресс, 2014. – 588 с. (ISBN 978-5-947060-077-1)
4. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р.: Введение в статистическое обучение с примерами на языке R. (ISBN: 978-5-97060-293-5)
5. Мастицкий С.Э., Шитиков В.К.: Статистический анализ и визуализация данных с помощью R (ISBN: 978-5-97060-301-7)
6. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R Тольятти, Лондон, 2017, 351 с.
7. Митин И.В., Русаков В.С. Анализ и обработка экспериментальных данных. М. издательство НЭВЦ ФИПТ, 1998, 48 с.
8. Волкова П. А., Шипунов А. Б. Статистическая обработка данных в учебно-исследовательских работах. М.: Форум, 2012, 96 с.
9. Буховец А.Г., Москалев П.В., Богатова В.П., Бирючинская Т.Я. Статистический анализ данных в системе R. Учебное пособие. – Воронеж: ВГАУ, 2010.
10. Волкова П. А., Шипунов А. Б. Статистическая обработка данных в учебно-исследовательских работах. М.: Экопресс, 2008.
11. Гланц С. Медико-биологическая статистика. Пер. с англ. — М., Практика, 1998.
12. The R Project for Statistical Computing. Режим доступа https://www.r-project.org/
13. Rstudio. Take control of your R code. Режим доступа https://www.rstudio.com/products/rstudio/
14. R: Анализ и визуализация данных. Режим доступа http://r-analytics.blogspot.ru/