МАШИННОЕ ОБУЧЕНИЕ
ЛАБОРАТОРНЫЙ ПРАКТИКУМ
(предварительный вариант)

Н.Ю. Золотых, А.Н. Половинкин, С.Н. Чернышова

9 октября 2007 г.

1. *Практическое машинное обучение*

1.1. Проверка статистических гипотез

1.1.1. Тест Шапиро–Уилка

Функция

```
shapiro.test(x)
```

выполняет тест Шапиро–Уилка. Нуль-гипотеза заключается в том, что случайная величина, выборка `x` которой известна, распределена по нормальному закону. Объем выборки должне быть не меньше `3` и не больше `5000`.

Рассмотрим несколько примеров. Протестируем стандартные датчики распределений. Начнем с нормального распределения.

```
> set.seed(0)
> shapiro.test(rnorm(100, mean = 2, sd = 5))

        Shapiro-Wilk normality test

data:  rnorm(100, mean = 2, sd = 5)
W = 0.9896, p-value = 0.6303
```

При уровне значимости, например, $\alpha = 0.05$ гипотеза должна быть принята, так как p-value $> \alpha$.

Теперь рассмотрим равномерное распределение.

```
> set.seed(0)
> shapiro.test(runif(100, min = -10, max = 10))

        Shapiro-Wilk normality test

data:  runif(100, min = -10, max = 10)
W = 0.9561, p-value = 0.002126
```

При уровне значимости, например, $\alpha = 0.05$ гипотеза должна быть отвергнута, так как p-value $< \alpha$.

Рассмотрим еще один пример. Фрейм данных `trees` из библиотеки `datasets` содержит замеры диаметра, высоты и объема вишневых деревьев. Проверим гипотезу о том, что высоты деревьев распределены по нормальному закону.

```
> colnames(trees)
[1] "Girth"  "Height" "Volume"
> x <- trees[, "Height"]
```

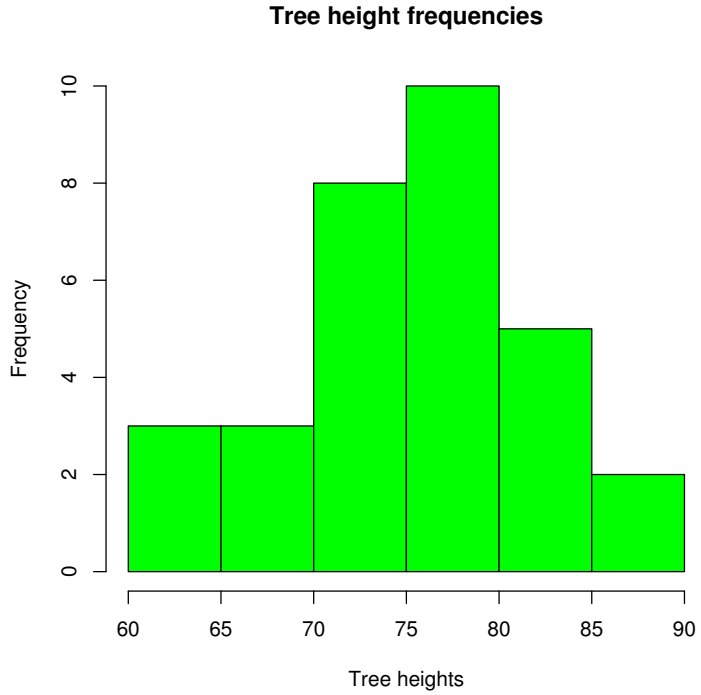

Рис. 1.1. Гистограмма представляет частоты высот вишневых деревьев

```
> hist(x, col = "green", xlab = "Tree heights",
    main = "Tree height frequencies")
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9655, p-value = 0.4034
```

При уровне значимости, например, $\alpha = 0.05$ гипотезу о нормальности распределения принимаем. Гистограмма представлена на рис. 1.1.

Объект, возвращаемый функцией `shapiro.test`, — это список со следующими полями:

`statistics` — значение статистики Шапиро–Уилка,

`p.value` — p-value,

`method` — строка `"Shapiro-Wilk normality test"`,

`data.name` — строка, содержащее имя данных, подвергнутых тесту (в предыдущем примере `x`).

1.1.2. Критерий Колмогорова–Смирнова

Описание:

```
ks.test(x, y, ...,
    alternative = c("two.sided", "less", "greater"),
    exact = NULL)
```

Одно- или двувыборочный тест Колмогорова–Смирнова.

`x` — вектор, содержащий выборку. `y` — вектор, содержащий вторую выборку, или символьная строка с именем распределения. `...` — параметры распределения. `alternative` — одно из следующих значений: `"two.sided"` (по умолчанию), `"less"`, or `"greater"`, обозначающих тип альтернативной гипотезы. См. детали ниже. `exact` — `NULL` или логическое значение, обозначающее требуется ли точное вычисление p-value. См. детали ниже. Не используется в двувыборочном тесте, если `alternative = "less"` или `alternative = "greater"`.

Детали: Если `y` — числовой вектор, то выполняется двувыборочный тест Колмогорова–Смирнова, проверяющий нуль-гипотезу о том, что x и y выбраны из одного непрерывного распределения.

Если `y` — строка, содержащая имя непрерывного распределения, то выпоняется одновыборочный тест Колмогорова–Смирнова, проверяющий нуль-гипотезу о том, что x имеет заданное непрерывное распределение.

Возможные значения `"two.sided"`, `"less"` и `"greater"` параметра `alternative` определяют альтернативную гипотезу, которая может заключаться в том, что интегральная функция распределения выборки `x` не совпадает с гипотетической (`"two.sided"`), не больше ее (`"less"`) или не меньше ее (`"greater"`).

Точное p-value не вычисляется в двувыборочном тесте, если `alternative = "less"` или `alternative = "greater"`. Если `exact = NULL` (по умолчанию), то точное p-value вычисляется, если объем выборки меньше `100` в одновыборочном тесте или если объемы выборок меньше `10000` в двувыборочном тесте.

В одновыборочном тесте параметры гипотетического распределения должны быть известны точно, а не вычисляться по выборке `x`. Вариант теста Колмогорова–Смирнова с оценкой параметров не поддерживается.

Объект, возвращаемый функцией `ks.test`, — это список со следующими полями:

`statistics` — значение статистики Колмогорова–Смирнова,

`p.value` — p-value,

`alternative` — символьная строка с описанием альтернативной гипотезы,

`method` — символьная строка с названием используемого метода,

`data.name` — строка, содержащее имя данных, подвергнутых тесту.

Рассмотрим пример. Фрейм данных `randu` из библиотеки `datasets` содержит 400 троек псевдо-случайных чисел из интервала $[0, 1]$, последовательно выдаваемых (печально) известной функцией `RANDU`, имеющейся в компиляторе VAX FORTRAN под операционной системой VMS 1.5. Значения записаны в матрицу с тремя столбцами, называемыми именами `x`, `y`, `z`.

```
> colnames(randu)
[1] "x" "y" "z"
```

```
> nrow(randu)
[1] 400
> attach(randu)
> detach(randu)
> ks.test(x, y)

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.085, p-value = 0.1111
alternative hypothesis: two-sided

Warning message:
cannot compute correct p-values with ties in: ks.test(x, y)
> ks.test(x, z)

        Two-sample Kolmogorov-Smirnov test

data:  x and z
D = 0.0875, p-value = 0.09353
alternative hypothesis: two-sided

> ks.test(y, z)

        Two-sample Kolmogorov-Smirnov test

data:  y and z
D = 0.0475, p-value = 0.7576
alternative hypothesis: two-sided

> ks.test(x, punif)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0555, p-value = 0.1697
alternative hypothesis: two-sided

> ks.test(y, punif)

        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.0357, p-value = 0.6876
alternative hypothesis: two-sided
```

```
> ks.test(z, punif)

        One-sample Kolmogorov-Smirnov test

data:  z
D = 0.0455, p-value = 0.3782
alternative hypothesis: two-sided
> detach(randu)
```

1.1.3. t-тест Стьюдента

Описание:

```
t.test(x, ...)
t.test(x, y = NULL,
    alternative = c("two.sided", "less", "greater"),
    mu = 0, paired = FALSE, var.equal = FALSE,
    conf.level = 0.95, ...)
t.test(formula, data, subset, na.action, ...)
```

Выполняет одно- или двувыборочный t-тест. Одновыборочный t-тест предназначен для проверки равенства среднего значения нормально распределенной генеральной совокупности некоторому заданному значению в предположении, что дисперсия не известна. Двувыборочный тест служит для сравнения двух средних значений из нормально распределенных генеральных совокупностей в предположении, что их дисперсии равны, хотя и не известны.

Аргументы:

`x` — вектор, содержащий данные (первая выборка).

`y` — вектор, содержащий данные (вторая выборка) — опционально.

`alternative` — одно из следующих значений: `"two.sided"` (по умолчанию), `"less"`, or `"greater"`, обозначающих тип альтернативной гипотезы.

`exact` — `NULL` или логическое значение, обозначающее требуется ли точное вычисление p-value. Не используется в двувыборочном тесте, если `alternative = "less"` или `alternative = "greater"`.

`mu` — математическое ожидание или разность математических ожиданий, если задано две выборки.

`paired` — логическое значение, указывающее, требуется ли выполнить спаренный t-тест.

`var.equal` — логическое значение, считающее, считать ли разбросы равными. Если `TRUE`, то вычисляется разброс для объединенной выборки.

`conf.level` — доверительный уровень.

`formula` — формула вида `lhs ~ rhs`, где `lhs` — числовой вектор, а `rhs` — фактор с двумя классами. Только для двувыборочного теста.

`data` — матрица или фрейм данных, из которых берутся данные (опционально).

`subset` — вектор, определяющий используемое подмножество наблюдений.

`na.action` — функция, которая вызывается, как только в данных встрети-

лось значение NA.

Детали: Если y и formula не заданы, то выполняется одновыборочный тест, проверяющий, что выборка x имеет среднее, равное 0.

Если paired = TRUE, то должны быть определены и иметь одинаковую длину векторы x и y.

Значения NA и NaN из данных удаляются (если paired = TRUE, то при этом удаляется соответствующее значение из второй выборки).

Если var.equal = TRUE, то для оценки отклонения используется объединенная выборка. По умолчанию var.equal = FALSE и отклонение оценивается отдельно для каждой выборки. При этом происходит надлежащяя корректировка числа степеней свободы.

Возвращаемое значение: Объект, возвращаемый функцией t.test, — это список со следующими полями:

statistics — значение t-статистики Стьюдента,

parameter — число степеней свободы,

p.value — p-value,

conf.int — доверительный интервал для математического ожидания,

estimate — оценка математического ожидания для одновыборочного теста или разности математических ожиданий для двувыборочного теста,

null.value — предполагаемое математическое ожидание или разность предполагаемых математических ожиданий для двувыборочного теста (входной параметр mu),

alternative — символьная строка с описанием альтернативной гипотезы,

method — символьная строка с названием используемой модификации метода,

data.name — строка, содержащее имя (имена) данных, подвергнутых тесту.

Примеры: В качестве простых примеров сравним математические ожидания у двух выборок, полученных с помощью функции rnorm:

```
> set.seed(0)
> x <- rnorm(100, mean = 0, sd = 4)
> y <- rnorm(100, mean = 1, sd = 4)
```

Итак, x и y — две выборки из нормальных генеральных совокупностей с одинаковым среднеквадратическим отклонением, но разными средними. Применим к выборкам t-тест, полагая, что в качестве нуль-гипотезы постулируется, что генеральные совокупности имеют одинаковое среднее. Альтернативная гипотеза пусть утверждает, что средние не равны.

```
> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = -1.3896, df = 196.428, p-value = 0.1662
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -1.7590435  0.3048036
sample estimates:
mean of x mean of y
0.0906738 0.8177938
```

Дадим разъяснение полученным результатам. Найдено значение статистики `t`, число степеней свободы `df`, величина p-value. Указаны границы 95% доверительного интервала для разности мат. ожиданий первого и второго распределений. Приведены оценки математических ожиданий для каждого распределения. Пусть уровень значимости равен $\alpha = 0.1$. Так как p-value $> \alpha$, то гипотезу о том, что математические ожидания у распределений равны, принимаем.

Изменим теперь альтернативную гипотезу. Пусть альтернативная гипотеза постулирует, что вторая генеральная совокупность имеет бóльшее математическое ожидание, чем первая:

```
> t.test(x, y, alternative = "less")

        Welch Two Sample t-test

data:  x and y
t = -1.3896, df = 196.428, p-value = 0.08311
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1376404
sample estimates:
mean of x mean of y
0.0906738 0.8177938
```

Теперь при уровне значимости $\alpha = 0.1$, так как p-value $< \alpha$, то нуль-гипотезу отклоняем и принимаем альтернативную гипотезу.

В качестве еще одного примера рассмотрим данные об измерениях скорости света, полученные А.А. Майкельсоном и Э.У. Морли во время знаменитого эксперимента 1887 г. Данные содержатся во фрейме `morley` библиотеки `datasets`. Фрейм содержит три столбца: `Expt` — номер эксперимента (от 1 до 5), `Run` — номер испытания (каждый эксперимент состоял из 20 испытаний), `Speed` — скорость света минус 299000 (в км/с). Примем во внимание только последний столбец. Предположим, что генеральная совокупность (замеры скорости света) имеет нормальное распределение. Сформулируем нуль-гипотезу: математическое ожидание генеральной совокупности равно 299792.458 (принятое в настоящее время значение скорости света).

```
> t.test(light - 792.458)

        One Sample t-test

data:  light - 792.458
```

```
t = 7.5866, df = 99, p-value = 1.824e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 44.26459 75.61941
sample estimates:
mean of x
   59.942
```

Дадим разъяснение полученным результатам. Найдено значение статистики `t`, число степеней свободы `df`, величина p-value. Указаны границы 95% доверительного интервала для оценки мат. ожидания выборки `light` $-$ 792.458. Приведены оценки математических ожиданий для каждой группы. Пусть уровень значимости равен $\alpha = 0.05$. Так как p-value $< \alpha$, гипотезу отклоняем.

Рассмотрим классический пример Стьюдента. Фрейм данных `sleep` из библиотеки `datasets` содержит информацию об увеличении продолжительности сна у 10 пациентов, которым давали два типа снотворного. У фрейма `sleep` 20 строк и два столбца с именами `extra` и `group`. В первом столбце содержатся числовые значения, равные увеличению продолжительности сна (в часах) после принятия снотворного, во втором столбце — тип снотворного (`factor`).

```
> colnames(sleep)
[1] "extra" "group"
> nrow(sleep)
[1] 20
> plot(extra ~ group, data = sleep, col = c("gold", "coral"))
```

Графический результат представлен на рис. 1.2.

В качестве нуль-гипотезы возьмем утверждение, что данные из первой и второй группы имеют одинаковое среднее (при условии, что обе выборки удовлетворябт нормальному закону распределения). Альтернативная гипотеза будет заключаться в том, что средние не совпадают.

```
> t.test(extra ~ group, data = sleep)
       Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.0794
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

Дадим разъяснение полученным результатам. Найдено значение статистики `t`, число степеней свободы `df`, величина p-value. Указаны границы 95% доверительного интервала для оценки разности мат. ожиданий. Приведены оценки математических ожиданий для каждой группы.

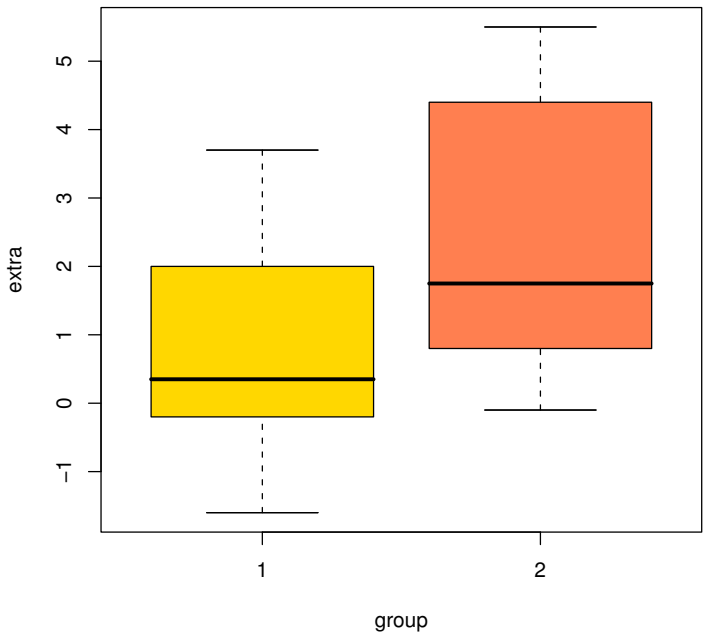

Рис. 1.2. Влияние двух типов снотворного на изменение продолжительности сна

Пусть выбран уровень значимости $\alpha = 0.05$. Мы видим, что $\alpha <$ p-value, таким образом, гипотезу о том, что данные имеют разные распределения принять нельзя.

Изменим альтернативную гипотезу и посмотрим, как это повлияет на p-value. Пусть теперь альтернативная гипотеза постулирует, что мат. ожидание второго распределения больше мат. ожидания первого распределения.

```
> t.test(extra ~ group, data = sleep, alternative = "less")
         Welch Two Sample t-test
 data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.0397
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1066185
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

Теперь при том же уровне значимости $\alpha = 0.05$ мы должны отклонить нуль-гипотезу и принять альтернативную гипотезу. На содержательном языке альтернативная гипотеза говорит о том, что второе снотворное приводит к боль-

шему увеличению продолжительности сна, чем первое.

1.1.4. F-тест Фишера

```
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"),
         conf.level = 0.95, ...)
var.test(formula, data, subset, na.action, ...)
```

Выполняет F-тест Фишера для проверки на равенство стандартных отклонений двух нормально распределенных генеральных совокупностей.

Аргументы:

`x`, `y` — числовые векторы, содержащие выборки из разных генеральных совокупностей, или линейные модели (возвращаемые функцией `lm`).

`ratio` — предполагаемая величина отношения стандартных отклонений в первой и второй генеральных совокупностях.

`alternative` — одно из следующих значений: `"two.sided"` (по умолчанию), `"less"`, or `"greater"`, обозначающих тип альтернативной гипотезы.

`conf.level` — доверительный уровень для возвращаемого доверительного интервала.

`formula` — формула вида `lhs ~ rhs`, где `lhs` — числовой вектор, а `rhs` — фактор с двумя классами.

`data` — матрица или фрейм данных, из которых берутся данные для `formula` (опционально).

`subset` — вектор, определяющий используемое подмножество наблюдений.

`na.action` — функция, которая вызывается, как только в данных встретилось значение `NA`.

Детали: Нуль-гипотеза постулирует, что отношение стандартных отклонений генеральных совокупностей, из которых выбраны x и y соответственно, равна отношению *ratio*.

Возвращаемое значение: Объект, возвращаемый функцией `var.test`, — это список со следующими полями:

`statistics` — значение F-статистики Фишера,

`parameter` — число степеней свободы,

`p.value` — p-value,

`conf.int` — доверительный интервал для математического ожидания,

`estimate` — оценка математического ожидания для одновыборочного теста или разности математических ожиданий для двувыборочного теста,

`null.value` — предполагаемое математическое ожидание или разность предполагаемых математических ожиданий для двувыборочного теста (входной параметр `mu`),

`alternative` — символьная строка с описанием альтернативной гипотезы,

`method` — символьная строка с названием используемой модификации метода,

`data.name` — строка, содержащее имя (имена) данных, подвергнутых тесту.

Примеры: Рассмотрим две выборки из разных нормальных генеральных совокупностей:

```
> x <- rnorm(50, mean = 0, sd = 2)
> y <- rnorm(50, mean = 10, sd = 2)
```

Проверим, что генеральные совокупности имеют одинаковое стандартное отклонение.

```
> var.test(x, y)

        F test to compare two variances

data:  x and y
F = 0.7353, num df = 49, denom df = 49, p-value = 0.2852
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4172619 1.2957270
sample estimates:
ratio of variances
         0.7352942
```

Пусть, например, уровень значимости равен $\alpha = 0.05$. Так как $\alpha <$ p-value, то гипотезу о равенстве стандартных отклонений принимаем.

Вернемся к примеру Стьюдента с данными о влиянии разных снотворных на продолжительность сна. Теперь сравним стандартные отклонения. Вначале в качестве альтернативной возмем гипотезу (по умолчанию), что отношение стандартных отклонений не равно 1:

```
> var.test(extra ~ group, data = sleep)

        F test to compare two variances

data:  extra by group
F = 0.7983, num df = 9, denom df = 9, p-value = 0.7427
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.198297 3.214123
sample estimates:
ratio of variances
         0.7983426
```

Пусть, например, уровень значимости равен $\alpha = 0.1$. Так как $\alpha <$ p-value, то гипотезу о равенстве стандартных отклонений принимаем.

Теперь в качестве альтернативной возьмем гипотезу, что отношение отклонений меньше 1:

```
> var.test(extra ~ group, data = sleep, alternative = "less")
```

```
        F test to compare two variances

data:  extra by group
F = 0.7983, num df = 9, denom df = 9, p-value = 0.3714
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.000000 2.537846
sample estimates:
ratio of variances
         0.7983426
```

При уровне значимости $\alpha = 0.1$ гипотезу о равенстве стандартных отклонений снова принимаем.

1.1.5. Критерий согласия χ^2 Пирсона

Описание:

```
chisq.test(x, y = NULL, correct = TRUE,
    p = rep(1/length(x), length(x)), rescale.p = FALSE,
    simulate.p.value = FALSE, B = 2000)
```

Функция реализует критерий согласия χ^2 Пирсона для простых гипотез и тест на проверку независимости признаков.

Аргументы:

`x` — вектор или матрица.

`y` — вектор. Игнорируется, если `x` — матрица.

`correct` — логическое значение, указывающее, требуется ли применять непрерывную коррекцию для 2×2 матриц.

`p` — вектор, содержащий вероятности. Должен иметь такую же длину, что и `x`.

`rescale.p` — логическое значение. Если `TRUE`, то `p` при необходимости масштабируется, так, чтобы сумма его компонентов была равна 1.

`simulate.p.value` — логическое значение. Если `TRUE`, то p-value вычисляется с помощью метода Монте–Карло, в противном случае используется χ^2 распределение.

`B` — количество испытаний в методе Монте–Карло.

Детали: Если `x` — вектор или матрица с одним столбцом или одной строкой, а вектор y не задан, то `x` рассматривается как статистический ряд (одномерная таблица сопряженности признаков), т. е. i-я компонента вектора `x` содержит количество точек попавших в i-й интервал группировки. В этом случае выполняется тест на проверку согласия данных заданным вероятностям p. Таким образом, нуль-гипотеза заключается в том, что вероятность попадания в i-й интервал группировки равна i-й компоненте вектора `p`. По умолчанию, задаются равные вероятности.

Если `x` — матрица не менее чем с 2 строками и 2 столбцами, то `x` рассматривается как двумерная таблица сопряженности признаков и выполняется тест на проверку их независимости.

Итак, обратим внимание, что если аргумент `y` не задан, то функция `chisq.test` работает со статистическим рядом или таблицей сопряженности, а не с выборкой.

Если `x` и `y` числовые векторы или векторы факторов одной и той же длины (числовые векторы конвертируются в векторы факторов), то соответствующие пары их компонентов рассматриваются как реализации двумерной случайной величины (X, Y) и выполняется тест на проверку независимости признаков X и Y.

Возвращаемое значение: Объект, возвращаемый функцией `chisq.test`, — это список со следующими полями:

`statistics` — значение χ^2-статистики Пирсона,

`parameter` — число степеней свободы распределения χ^2; равно `NA`, если для отыскания p-value использовался метод Монте–Карло,

`p.value` — p-value,

`method` — символьная строка с названием используемой модификации теста, а также указанием того, использовались ли непрерывная коррекция и метод Монте–Карло,

`data.name` — строка, содержащее имя (имена) данных, подвергнутых тесту,

`observed` — число точек, попавших в i-й интервал группировки (равно `x` на входе, если `x` — вектор, а `y` не используется),

`expected` — теоретическое число точек (в предположении выполнения гипотезы), попавших в i-й интервал группировки,

`residuals` — остатки Пирсона:

$$\frac{\texttt{observed} - \texttt{expected}}{\sqrt{\texttt{expected}}}.$$

Примеры: Рассмотрим классический пример с бросанием монеты. Бюффон бросал монету 4040 раз, при этом герб выпал 2048 раз. Используя критерий согласия χ^2, проверим, что монета симметрична. Итак, нуль-гипотеза заключается в том, что вероятность выпадания герба равна $p_1 = 1/2$, вероятность выпадания решки — $p_2 = 1/2$.

```
> chisq.test(c(2048, 1992))

        Chi-squared test for given probabilities

data:  c(2048, 1992)
X-squared = 0.7762, df = 1, p-value = 0.3783
```

Пусть, например, был выбран уровень значимости $\alpha = 0.05$. Так как $\alpha <$ p-value, то гипотезу принимаем.

Рассмотрим простой пример на проверку независимости двух генеральных совокупностей, если известны выборки `x`, `y`. Соответствующие пары компонентов векторов `x`, `y` требуется рассматривать как реализации двумерной случайной величины (X, Y).

```
> set.seed(0)
```

```
> x <- rnorm(100)
> y <- runif(100)
> chisq.test(x, y)

        Pearson's Chi-squared test

data:  x and y
X-squared = 9900, df = 9801, p-value = 0.239

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(x, y)
```

Пусть, например, был выбран уровень значимости $\alpha = 0.05$. Так как $\alpha <$ p-value, то гипотезу о независимости случайных признаков принимаем.

Рассмотрим пример на проверку независимости двух случайных величин. Таблица `HairEyeColor` из библиотеки `datasets` содержит информацию о поле, цвете волос и глаз у 592 студентов. Таблица имеет 3 размерности:

`"Hair"`: `HairEyeColor["Black", ,]`, `HairEyeColor["Brown", ,]`, `HairEyeColor["Red", ,]`, `HairEyeColor["Blond", ,]`,

`"Eye"`: `HairEyeColor[, "Brown",]`, `HairEyeColor[, "Blue",]`, `HairEyeColor[, "Hazel",]`, `HairEyeColor[, "Green",]`,

`"Sex"`: `HairEyeColor[, , "Male"]`, `HairEyeColor[, , "Female"]`.

Элементы таблицы — количество человек из данной группы. Проверим нуль-гипотезу о том, что общее цвет глаз мужчин не зависит от цвета волос мужчин. Во-первых, построим таблицу сопряженности признаков:

```
> men <- HairEyeColor[, , "Male"]
> men
       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    38   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8
> mosaicplot(men,
     col = c("chocolate", "cornflowerblue", "salmon", "green"),
     main = "Male eye color vs. hair color")
> chisq.test(men, simulate.p.value = TRUE)

        Pearson's Chi-squared test with simulated p-value
        (based on 2000 replicates)

data:  men
X-squared = 42.1633, df = NA, p-value = 0.0004998
```

При уровне значимости, например, $\alpha = 0.05$, гипотезу следует отклонить и признаки считать зависимыми. Построенная мозаичная диаграмма приведена на рис. 1.3. Аналогичные результаты можно получит и для женщин.

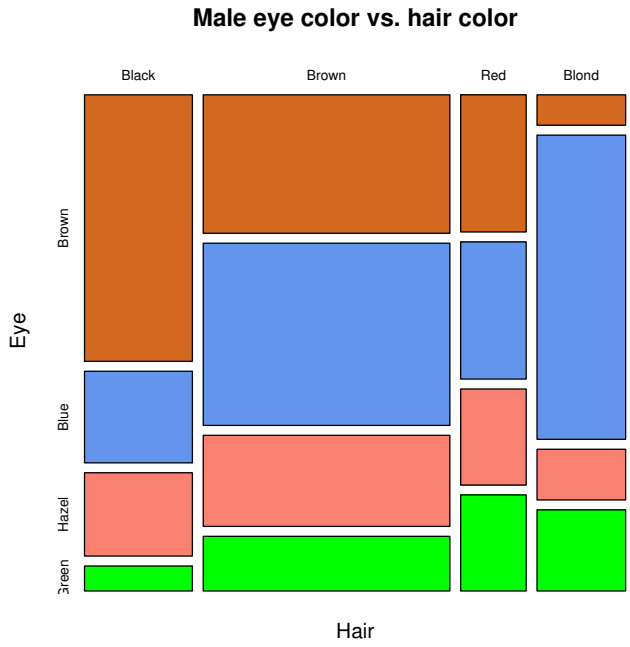

Рис. 1.3. Зависимость цвета глаз от цвета волос у мужчин

1.1.6. Задания для лабораторной работы

1) Используя тест Шапиро-Уилка, проверьте, являются ли нормально распределенными характеристики цветов ириса (фрейм данных `iris`). Уровень значимости $\alpha = 0.05$.

2) Для `k=10,15,20,25,30` сгенерируйте 200 реализаций нормальной распределенной случайной величины с мат. ожиданием, равным `k`, и стандартным отклонением, равным $\sqrt{2k}$, и 200 реализаций случайной величины, распределенной по закону χ^2 с `k` степенями свободы. Используя тест Колмогорова-Смирнова, проверьте гипотезу о том, что данные выборки относятся к одному непрерывному распределению. Уровень значимости $\alpha = 0.05$.

3) Загрузите таблицу из файла `allcountries.txt`, содержащую информацию о населении, площади и ряде других характеристик современных государств. Выберите из таблицы те страны, для которых доступна информация о населении и площади (нет отсутствующих значений `NA`) и площадь больше `10.0`. Пусть `area_log = log10(log10(area))`, `population_log = log10(log10(population))`. Постройте линейную регрессию (используя функцию `lm`) для зависимости `population_log` от

`area_log`. Используя тест Колмогорова-Смирнова, проверьте гипотезу о том, что `population_log` и `f(area_log)`, где `f()` — построенная регрессионная функция, относятся к одному непрерывному распределению. Уровень значимости $\alpha = 0.05$.

4) Используя критерий χ^2 проверьте нуль-гипотезу, состоящую в том, что цвет глаз женщин не зависит от цвета волос (на фрейме данных `HairEyeColor`).

5) Загрузите таблицу из файла `readingspeed.txt`, которая содержит информацию о скорости чтения у детей в зависимости от применяемой методики обучения (`DRA` — direct reading activities, `SC` — standart curriculum). Используя t-тест, проверьте гипотезу о том, что среднее время чтения для обеих методик совпадает (используйте разные альтернативные гипотезы). Объясните полученные результаты.