# Проверка статистических гипотез

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

29 сентября 2013 г.

## Содержание



## 1 Тест Шапиро–Уилка

Нулевая гипотеза $H_0$ теста Шапиро–Уилка заключается в том, что случайная величина, выборка `x` которой известна, распределена по нормальному закону. Альтернативная гипотеза $H_1$ заключается в том, закон распределения не является нормальным.

Для выполнения теста Шапиро-Уилка предназначена функция `shapiro.test(x)`, принимающая на вход выбокру `x` объема не меньше 3 и не больше 5000. Функция возвращает список со следующими компонентами:

- `statistic` – значение статистики теста, которую принято обозначать буквой $W$;
- `p.value` – апроксимация p-value для полученного значения статистики;
- `method` – строка с названием теста;
- `data.name` – имя переменной, содержащей выборку, которая была переданна функции `shapiro.test` в качестве аргумента.

Следует отметить, что несмотря на то, что возвращаемое значение является списком, его вывод на экран обрабатывается особым образом, позволяя более компактно представить данные о выполненном тесте.

Рассмотрим несколько примеров. Применим тест Шапиро-Уилка к выборкам из различных распределений, сгенерированным с помощью стандартных функций языка R.

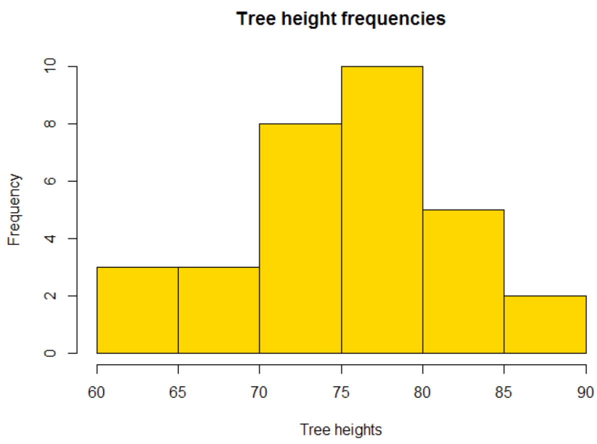


Рис. 1:

```
> 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

> 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
```

В данном случае при уровне значимости, например, $\alpha = 0.05$ для выборки, сгенерированной функцией `rnorm`, гипотеза $H_0$ должна быть принята (так как p-value $> \alpha$), а для выборки, полученной с помощью `runif`, $H_0$ следует отклонить, приняв альтернативную гипотезу.

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.

```
> hist(trees[, "Height"], col = "gold", xlab = "Tree heights",
    main = "Tree height frequencies")
> shapiro.test(trees[, "Height"])

        Shapiro-Wilk normality test

data:  trees[, "Height"]
W = 0.9655, p-value = 0.4034
```

Таким образом, при уровне значимости, например, $\alpha = 0.05$ гипотезу $H_0$ о нормальности распределения принимаем. Гистограмма высот деревьев из рассматриваемого набора данных представлена на рис. 1.

## 2 Тест Колмогорова–Смирнова

Тест Колмогорова–Смирнова предназначен для проверки гипотез об отличии интегральных функций распределения $F_1$ и $F_2$ двух случайных величин на основании их выборок (в данном случае говорят о двухвыборочный тесте Колмогорова–Смирнова) или выборки одной случайной величины и аналитически заданной интегральной функции для другой (одновыборочный тест). При этом нулевая гипотеза может формулироваться, как «$F_1 = F_2$», «$F_1 \leq F_2$» или «$F_1 \geq F_2$». При этом альтернативная гипотеза прямо противоположна нулевой.

Для выполнения теста Колмогорова–Смирнова предназначена функция `ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)`, где

- `x` – выборка значений первой случайной величины;
- `y` – выборка значений второй случайной величины для двухвыборочного теста, либо имя или сама интегральная функция распределения второй случайной величины для одновыборочного теста;
- `...` – параметры распределения второй случайной величины для одновыборочного теста;
- `alternative` – строка, определяющая тип альтернативной (и косвенно нулевой) гипотезы. Возможны значения: `"two.sided"` соответствует $H_0$ :«$F_1 = F_2$», $H_1$ :«$F_1 \neq F_2$»; `"less"` – $H_0$ :«$F_1 \geq F_2$», $H_1$ :«$F_1 < F_2$»; `"greater"` – $H_0$ :«$F_1 \leq F_2$», $H_1$ :«$F_1 > F_2$»;
- `exact` – логическое значение, обозначающее требуется ли точное вычисление p-value или достаточно его апроксимации, либо `NULL`. В случае `exact = NULL` решение о точном вычислении p-value принимается автоматически на основе данных, на которых производится тест.

Функция возвращает список со следующими компонентами:

- `statistic` – значение статистики теста, которую принято обозначать буквами $D$, $D^-$, $D^+$ в зависимости от нулевой гипотезы;
- `p.value` – p-value (точное значение или апроксимация) для полученного значения статистики;
- `method` – строка с названием теста;
- `data.name` – имена переменных, содержащих выборки, которые были переданны функции `ks.test` в качестве аргументов `x` и `y`;
- `alternative` – строка с описанием альтернативной гипотезы;

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"`. Проведем двухвыборочные тесты для всех возможных пар `"x"`, `"y"` и `"z"`, а также одновыборочные тесты для сравнения распределения с равномерным.

```
> ks.test(randu$x, randu$y)

        Two-sample Kolmogorov-Smirnov test

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

Warning message:
In ks.test(randu$x, randu$y) :
  p-value will be approximate in the presence of ties
> ks.test(randu$x, randu$z)

        Two-sample Kolmogorov-Smirnov test

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

> ks.test(randu$y, randu$z)

        Two-sample Kolmogorov-Smirnov test

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

> ks.test(randu$x, punif)

        One-sample Kolmogorov-Smirnov test

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

> ks.test(randu$y, punif)

        One-sample Kolmogorov-Smirnov test

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

> ks.test(randu$z, punif)

        One-sample Kolmogorov-Smirnov test

data:  randu$z
D = 0.0455, p-value = 0.3782
alternative hypothesis: two-sided
```

В качестве самостоятельного задания предлагается визуализировать точки из данного набора в трехмерии (для этого, например, может быть использована функция `plot3d` из пакета `rgl`).

## 3 t-тест Стьюдента

Одновыборочный t-тест предназначен для проверки равенства математического ожидания нормально распределенной случайной величины (для которой известна лишь выборка) некоторому заданному значению в предположении, что дисперсия не известна. Двухвыборочный тест служит для сравнения математических ожиданий нормально распределенных случайных величин в предположении, что их дисперсии равны, хотя и не известны. Таким образом, нулевая гипотеза формулируется как «$E(X) = \mu$» или «$E(X) - E(Y) = \mu$».

Для выполнения различных вариантов t-теста Стьюдента предназначена функция `t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)`, где

- `x` – выборка значений первой случайной величины;
- `y` – выборка значений второй случайной величины для двухвыборочного теста, либо `NULL`;
- `alternative` – строка, определяющая тип альтернативной гипотезы. Возможны значения: `"two.sided"` соответствует $H_1$ : «$E(X) \neq \mu$» или «$E(X)-E(Y) \neq \mu$»; `"less"` – $H_1$ : «$E(X) < \mu$» или «$E(X)-E(Y) < \mu$»; `"greater"` – $H_1$ :«$E(X) > \mu$» или «$E(X) - E(Y) > \mu$»;
- `paired` – логическое значение, обозначающее являются ли выборки в двухвыборочном тесте независимыми (`paired = FALSE`) или нет (`paired = TRUE`). Примером, когда выборки являются зависимыми является исследование одних и тех же объектов с некоторой разницей во времени (например, измерение артериального давления у одних и тех же людей до и после приема лекарст).
- `var.equal` – логическое значение, указывающее являются ли дисперсии двух рассматриваемых случайных величин одинаковыми (хотя и неизвестными) или различными;
- `conf.level` – уровень доверительного интервала для математического ожидания первой случайной величины или разности мат. ожиданий первой и второй случайных величин, который будет вычислен при выполнении теста.

Функция возвращает список со следующими компонентами:

- `statistic` – значение статистики теста, которую принято обозначать буквой $t$;
- `parameter` – количество степеней свободы $t$-статистики;
- `p.value` – p-value для полученного значения статистики;
- `method` – строка с названием теста;
- `data.name` – имена переменных, содержащих выборки, которые были переданны функции `t.test` в качестве аргументов `x` и `y`;

- `alternative` – строка с описанием альтернативной гипотезы;
- `conf.int` – доверительный интервал;
- `estimate` – оцененные по выборкам значения мат. ожиданий;
- `null.value` – значение параметра `mu`, использовавшееся для теста.

В качестве простого примера рассмотрим выборки из нормальных распределений с одинаковыми дисперсиями и различными мат. ожиданиями. Применим t-тест, полагая в качестве альтернативной гипотезы неравенство мат. ожиданий.

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

Как можно видеть, в данном случае при уровне значимсоти $\alpha = 0.1$, следует принять нулевую гипотезу о равенсте мат. ожиданий. Изменим теперь альтернативную гипотезу. Пусть альтернативная гипотеза постулирует, что мат. ожидание первой случайной величины меньше мат. ожидания второй.

```
> 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 $\alpha = 0.1$ нулевую гипотезу следует отвергнуть и принять альтернативную.

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"` – разность измеренной скорости света и значения 299000 км/с. Примем во внимание только последний столбец. Предположим, что измеренная скорость света имеет нормальное распределение (проверьте это с помощью рассмотренных ранее статистических тестов). Сформулируем нулевую гипотезу: «математическое ожидание генеральной совокупности равно 299792.458 км/с» (принятое в настоящее время значение скорости света).

```
> t.test(morley$Speed, mu = 792.458)

        One Sample t-test

data:  morley$Speed
t = 7.5866, df = 99, p-value = 1.824e-11
alternative hypothesis: true mean is not equal to 792.458
95 percent confidence interval:
 836.7226 868.0774
sample estimates:
mean of x
    852.4
```

Из полученных результатов делаем вывод, что при уровне значимости $\alpha = 0.05$ нулевую гипотезу следует отклонить и принять альтернативную.

Также 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. Проверим нулевую гипотезу о равенстве средней продолжительности сна при приеме двух типов снотворного при различных альтернативных гипотезах.

```
> group1 = sleep[sleep$group == 1, "extra"]
> group2 = sleep[sleep$group == 2, "extra"]
> boxplot(sleep[sleep$group == 1, "extra"], sleep[sleep$group
    == 2, "extra"], col = c("gold", "dodgerblue"), xlab =
    "group", ylab = "extra")
> t.test(group1, group2)

        Welch Two Sample t-test

data:  group1 and group2
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal
    to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean of x mean of y
     0.75      2.33

> t.test(group1, group2, alternative = "less")

        Welch Two Sample t-test

data:  group1 and group2
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 of x mean of y
```

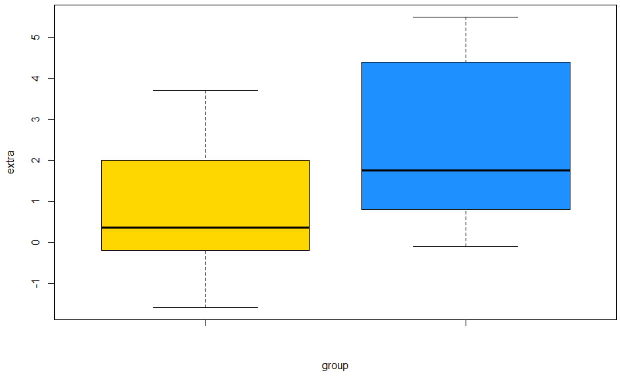


Рис. 2:

```
      0.75       2.33
```

При уровне значимости $\alpha = 0.05$ нулевую гипотезу следует принять в первом случае и отвергнуть во втором. Таким образом, на содержательном языке можно сделать вывод, что второе снотворное приводит к большему увеличению продолжительности сна, чем первое. Визуализация представлена на рис. 2.

# 4 F-тест Фишера

Тест Фишера предназначен для проверки соотношения (в том числе и равенства) дисперсий двух нормально распределенных случайных величин, т.е. $H_0$ : «$\sigma^2(X)/\sigma^2(Y) = r$». Для выполнения данного статистического теста предназначена функция `var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, ...)`, где

- `x` – выборка значений первой случайной величины;
- `y` – выборка значений второй случайной величины;
- `ratio` – величина предполагаемого отношения дисперсий $r$;
- `alternative` – строка, определяющая тип альтернативной гипотезы. Возможны значения: `"two.sided"` соответствует $H_1$ : «$\sigma^2(X)/\sigma^2(Y) \neq r$»; `"less"` – $H_1$ : «$\sigma^2(X)/\sigma^2(Y) < r$»; `"greater"` – $H_1$ :«$\sigma^2(X)/\sigma^2(Y) > r$»;
- `conf.level` – уровень доверительного интервала отношения дисперсии первой случайной величины ко второй, который будет вычислен при выполнении теста.

Функция возвращает список со следующими компонентами:

- `statistic` – значение статистики теста, которую принято обозначать буквой $F$;
- `parameter` – количество степеней свободы $F$-статистики;

- `p.value` – p-value для полученного значения статистики;
- `method` – строка с названием теста;
- `data.name` – имена переменных, содержащих выборки и переданные функции в качестве аргументов `x` и `y`;
- `alternative` – строка с описанием альтернативной гипотезы;
- `conf.int` – доверительный интервал;
- `estimate` – оцененное значение отношения дисперсий;
- `null.value` – значение параметра `ratio`, использовавшееся для теста.

В качестве модельного примера рассмотрим две выборки из нормального распределения и одинаковыми дисперсиями, но разными мат. ожиданиями и проверим с помощью F-теста Фишера гипотезу о том, что отношения дисперсий равно единице.

```
> set.seed(0)
> x = rnorm(50, mean = 0, sd = 2)
> y = rnorm(50, mean = 10, sd = 2)
> var.test(x, y)

        F test to compare two variances

data:  x and y
F = 1.1432, num df = 49, denom df = 49, p-value = 0.6414
alternative hypothesis: true ratio of variances is not equal to
    1
95 percent confidence interval:
 0.648724 2.014488
sample estimates:
ratio of variances
          1.143174
```

При уровне значимости $\alpha = 0.05$ нулевую гипотезу следует принять.

Вернемся к примеру Стьюдента с данными о влиянии разных снотворных на продолжительность сна. Теперь сравним дисперсии: рассмотрим альтернативные гипотезы о том, что отношение дисперсий не равно единице и, затем, что это отношение меньше единицы.

```
> group1 = sleep[sleep$group == 1, "extra"]
> group2 = sleep[sleep$group == 2, "extra"]
> var.test(group1, group2)

        F test to compare two variances

data:  group1 and group2
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

> var.test(group1, group2, alternative = "less")

```

```
        F test to compare two variances

data:  group1 and group2
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
```

В обоих случаях при уровне значимости $\alpha = 0.05$ следует принять нулевую гипотезу.

# 5 Критерий согласия $\chi^2$ Пирсона

Критерий согласия $\chi^2$ Пирсона применяется для проверки гипотезы о том, что случайная величина имеет заданное распределение, т.е. $H_0$ : «Случайная величина $X$ имеет интегральную функцию распределения $F(x)$», а также гипотезы о независимости двух признаков, т.е. $H_0$ : «$X$ независит от $Y$». Для выполнения данного статистического теста предназначена функция `chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)`, где

- `x` – числовой вектор, представляющий собой статистический ряд[1], фактор, содержащий выборку значений случайной величины $X$, или матрица сопряженности признаков[2];
- `y` – если `x` является фактором, то `y` также должен быть фактором такой же длины, содержа выборку значений случайной величины $Y$, таким образом, $(x_i, y_i)$ – реализация двумерной дискретной случайной величины. По `x` и `y` вычисляется таблица сопряженности. В остальных случаях данный параметр игнорируется;
- `p` – вектор, содержащий вероятности попадания в интервалы разбиений значений случайных величин. Должен иметь такую же длину, что и `x`;
- `rescale.p` – логическое значение, определяющее будет ли для вычисления p-value использоваться метод Монте–Карло (`rescale.p = TRUE`), или $\chi^2$ распределение;
- `B` – количество испытаний в методе Монте–Карло;
- `correct` – логическое значение, указывающее, требуется ли применять непрерывную коррекцию для $2 \times 2$ матриц.

[1] Множество допустимых значений $[a, b]$ непрерывной случайной величины разбивается на $n$ непересекающихся интервалов $(a_i, b_i]$, $i = 1, 2, \dots, n$, для каждого из которых по имеющейся выборке подсчитывается частота $p_i$ попадания в него. Набор $p_1, p_2, \dots, p_n$ называется статистическим рядом.

[2] Таблица сопряженности признаков является обобщением статистического ряда на случай двух случайных величин. Множество значений каждой из них разбивается на непересекающиеся интервалы. Элемент таблицы сопряженности признаков $p_{i,j}$ содержит вычисленное по выборке значение частоты одновременного попадания случайной величины $X$ в $i$-й интервал и случайной величины $Y$ в $j$-й. Для дискретной случайной величины рассматриваются не интервалы, а значения, которые она принимает.

Таким образом, если задана (явно или неявно) таблица сопряженности, то выполняется тест с нулевой гипотезой «случайные величины независимы» и альтернативной «случайные величины зависимы», в остальных случаях проверяется нулевая гипотеза «$X$ имеет заданное распределение», где распределение определяется вектором `p`.

Функция возвращает список со следующими компонентами:

- `statistic` – значение $\chi^2$-статистики;
- `parameter` – количество степеней свободы $\chi^2$-статистики. Равно `NA`, если для отыскания p-value использовался метод Монте–Карло;
- `p.value` – p-value для полученного значения статистики;
- `method` – строка с названием теста;
- `data.name` – имена переменных, содержащих данные для теста;
- `observed` – число точек, попавших в $i$-й интервал группировки. Равно `x`, если `x` вектор, а `y` не используется;
- `expected` – теоретическое число точек (в предположении выполнения нулевой гипотезы), попавших в $i$-й интервал группировки;
- `residuals` – остатки Пирсона: $\frac{observed-expected}{\sqrt{expected}}$.

Рассмотрим классический пример с бросанием монеты. Бюффон бросал монету 4040 раз, при этом герб выпал 2048 раз. Используя критерий согласия $\chi^2$, проверим, что монета симметрична. Итак, нулевая 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
```

При уровне значимости $\alpha = 0.05$ нулевую гипотезу в данном случае следует принять.

Рассмотрим пример на проверку независимости двух случайных величин, если известны выборки $x$ и $y$. Соответствующие пары компонентов векторов $x$ и $y$ требуется рассматривать как реализации двумерной случайной величины $(X, Y)$.

```
> set.seed(0)
> x = rnorm(100)
> y = rnorm(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:
In chisq.test(x, y) : Chi-squared approximation may be incorrect
```

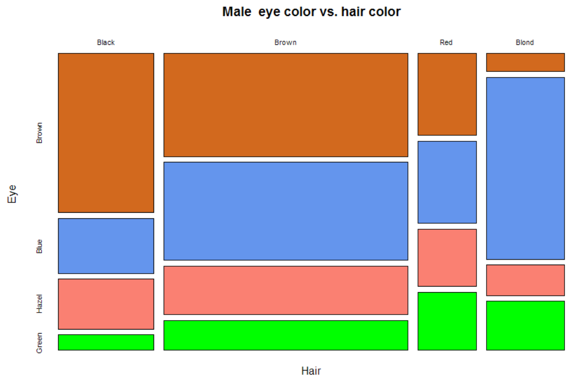


Рис. 3:

Пусть, например, был выбран уровень значимости $\alpha = 0.05$, тогда нулевую гипотезу о независимости случайных признаков принимаем.

Рассмотрим другой пример на проверку независимости двух случайных величин. Массив `HairEyeColor` из библиотеки `datasets` содержит информацию о поле, цвете волос и глаз у 592 студентов. По сути, данный массив представляет собой две (одна для мужчин, другая для женщин) таблицы зависимости цвета глаз от цвета волос. Каждая ячейка этих таблиц содержит количество человек с заданными признаками. Визуализируем данные и проверим нулевую гипотезу о том, что у мужчин цвет глаз не зависит от цвета волос.

```
> mosaicplot(HairEyeColor[, , "Male"], col  = c("chocolate",
    "cornflowerblue",  "salmon", "green"), main = "Male  eye
    color vs. hair color")
> chisq.test(HairEyeColor[, , "Male"], simulate.p.value = TRUE)

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

data:  HairEyeColor[, , "Male"]
X-squared = 41.2803, df = NA, p-value = 0.0004998
```

При уровне значимости $\alpha = 0.05$ гипотезу следует отклонить и признаки считать зависимыми. Построенная мозаичная диаграмма приведена на рис. 3.

## 6 Задания к лабораторной работе

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 $\sqrt{k}$, и 200 реализаций случайной вели-

чины, распределенной по закону $\chi^2$ 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. Пусть $area\_log = \log_{10}(\log_{10}(area))$, $population\_log = \log_{10}(\log_{10}(population))$.

   Методом наименьших квадратов постройте функцию $f(\cdot)$, моделирующую зависимость $population\_log$ от $area\_log$ с помощью линейной функции $population\_log = f(area\_log) = \beta_0 + \beta_1 area\_log$, т.е. подберите коэффициенты $\beta_0$ и $\beta_1$. Используя тест Колмогорова-Смирнова, проверьте гипотезу о том, что `population_log` и `f(area_log)` относятся к одному непрерывному распределению. Уровень значимости $\alpha = 0.05$.

4. Используя критерий $\chi^2$ проверьте 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.