**Проверка гипотез в R**

Статистика позволяет оценить какие-то стохастические процессы, которые происходят в мире. Центральное понятие статистики — генеральная совокупность, множество всех элементов какой-либо группы, параметр которой мы хотим оценить:

* все жители РБ при оценке роста;
* все возможные тексты писателя (реальные и потенциальные) при оценке частоты встречаемости каких-либо слов, гласных, сочетаний букв;
* все возможные курсы валют при попытке оценить курс валюты завтра;
* и т. д.

Весь статистический анализ строится на основе предположений о свойствах генеральной совокупности и некоторой выборки из генеральной совокупности. Так если мы не можем взять всю генеральную совокупность и оценить ее параметр θ (матожидание — средний рост, доля встречаемости гласных в текстах писателя и т. д.), то мы берем случайную выборку из генеральной совокупности и оцениваем параметр выборки θ̂ и делаем предположения о том, как параметр может быть устроен в генеральной совокупности. Если выборка, которой мы располагаем, содержит в себе генеральную совокупность, то нужда оценить некоторый параметр, казалось бы, отпадает (в таком случае задача переходит в область теории вероятностей):

* Какая доля слов «не» в корпусе текстов Пушкина?

Однако бывают задачи, которые даже обладая генеральной совокупностью, можно переформулировать в статистические:

* Какая доля слов «не» будет в свежеобнаруженном тексте Пушкина длины n?
* Исследователь каждый год ездит на остров Суматра и обнаруживает каждый год несколько неизвестных науке видов ящериц. С каждый годом он обнаруживает неизвестные науке виды ящериц все реже и реже. Можем ли мы оценить сколько ящериц неизвестного вида исследователь найдет в этом году?

При проверке статистических гипотез вы формулируете гипотезу о параметре генеральной совокупности (вашу нулевую гипотезу, H0). Затем вы создаете выборку из этой генеральной совокупности и вычисляете статистику, которая применяется для формирования заключений об этом параметре. Допуская, что нулевая гипотеза справедлива, вы вычисляете вероятность получения для генеральной совокупности наблюдаемого в выборке или большего значения статистики. Если вероятность достаточно мала, вы отвергаете нулевую гипотезу в пользу противоположной (ее называют альтернативной, или исследовательской, H1).

**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 должен быть не меньше 3 и не больше 5000.

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.

Пример 1. Нормальное распределение.

*> set.seed(0)*

*> shapiro.test(rnorm(100, mean = 2, sd = 5))*

*Shapiro-Wilk normality test*

*data: rnorm(100, mean = 2, sd = 5)*

*W = 0.98957, p-value = 0.6303*

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

Пример 2. 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 = 0.95606, p-value = 0.002126*

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

Пример 3. Таблица данных trees из библиотеки datasets содержит замеры диаметра, высоты и объёма вишнёвых деревьев. Проверим гипотезу о том, что высоты деревьев распределены по нормальному закону.

*> colnames(trees)*

*[1] "Girth" "Height" "Volume"*

*> x <- trees[, "Height"]*

*> shapiro.test(x)*

*Shapiro-Wilk normality test*

*data: x*

*W = 0.96545, p-value = 0.4034*

При уровне значимости, например, α = 0.05, гипотезу о нормальности распределения принимаем.

**Критерий Колмогорова–Смирнова**

Общий вид функции:  
ks.test(x, y, ..., alternative = c(“two.sided”, “less”, “greater”), exact = NULL)

Тест Колмогорова–Смирнова для одной или двух выборок. Аргументы:

* x — вектор, содержащий выборку.
* y — вектор, содержащий вторую выборку, или символьная строка с именем распределения.
* ... — параметры распределения.
* alternative — символьный аргумент, обозначающих тип альтернативной гипотезы. Принимает одно из следующих значений: “two.sided” (по умолчанию), “less” или “greater”.
* exact — NULL или логическое значение, обозначающее требуется ли точное вычисление p-value. 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 принадлежат одному и тому же непрерывному распределению.
* Если y — символьная переменная (имя непрерывного распределения), то выполняется одновыборочный тест Колмогорова–Смирнова, проверяющий нулевую гипотезу о том, что x принадлежит заданному распределению.
* Возможные значения “two.sided”, “less” и “greater” параметра alternative определяют альтернативную гипотезу, заключающуюся в том, что эмпирическая функция распределения выборки x не совпадает с теоретической (“two.sided”), не больше ее (“less”) или не меньше ее (“greater”).

Точное значение p-value не вычисляется в двувыборочном тесте, если alternative = “less” или alternative = “greater”. 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 — символьный аргумент — описание альтернативной гипотезы;
* method — символьный аргумент — название используемого метода;
* data.name — название массива данных, подвергнутых тесту.

*> colnames(randu)*

*[1] "x" "y" "z"*

*> nrow(randu)*

*[1] 400*

*> attach(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:*

*In ks.test(x, y) : p-value will be approximate in the presence of ties*

*> 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(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 = 0.055524, p-value = 0.1697*

*alternative hypothesis: two-sided*

*> detach(randu)*

**Критерий согласия χ2 (хи-квадрат) Пирсона**

Общий вид функции:

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 χ2 Пирсона для простых гипотез и тест на проверку независимости признаков.

Аргументы функции:

* 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 нормируется 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 χ2-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 сопряжённости признаков). Т. е. i-я компонента вектора x содержит количество элементов выборки, попавших в i-й интервал группировки. В этом случае выполняется тест на проверку соответствия (согласия) выборки заданным вероятностям p. Таким образом, основная (нулевая) 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 сопряжённости признаков и выполняется тест на проверку их независимости.

Таким образом, если аргумент y не задан, то функция chisq.test работает со статистическим рядом или таблицей сопряжённости, а не непосредственно с самой выборкой.

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

Функция chisq.test возвращает список, состоящий из следующих полей:

* statistics — значение χ2-статистики Пирсона.
* parameter — число степеней свободы распределения 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 c 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.
* expected — теоретическое число точек (в предположении выполнения гипотезы), попадающих в i-й интервал группировки.
* residuals — остатки Пирсона:

Пример 1.

Рассмотрим классический пример с бросанием монеты. Бюффон бросал монету 4040 раз, при этом герб выпал 2048 раз.

Используя критерий согласия χ2, проверим, что монета симметрична. Итак, основная гипотеза заключается в том, что вероятность выпадения герба равна p1 = 1/2, вероятность выпадения решки — p2 = 1/2.

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

*Chi-squared test for given probabilities*

*data: c(2048, 1992)*

*X-squared = 0.77624, df = 1, p-value = 0.3783*

Пусть, например, был выбран уровень значимости α = 0.05. Тогда, так как α < p-value, гипотезу принимаем.

Пример 2.

Таблица HairEyeColor из библиотеки datasets содержит информацию о поле, цвете волос и глаз у 592 студентов. Элементы таблицы — количество человек из данной группы.

Проверим гипотезу о том, что для мужчин цвет глаз не зависит от цвета волос.

*> male <- HairEyeColor[, , "Male"]*

*> male*

*Eye*

*Hair Brown Blue Hazel Green*

*Black 32 11 10 3*

*Brown 53 50 25 15*

*Red 10 10 7 7*

*Blond 3 30 5 8*

*> mosaicplot(male, col = c("brown", "blue", "red", "green"))*

*> chisq.test(male, simulate.p.value = TRUE)*

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

*data: male*

*X-squared = 41.28, df = NA, p-value = 0.0004998*

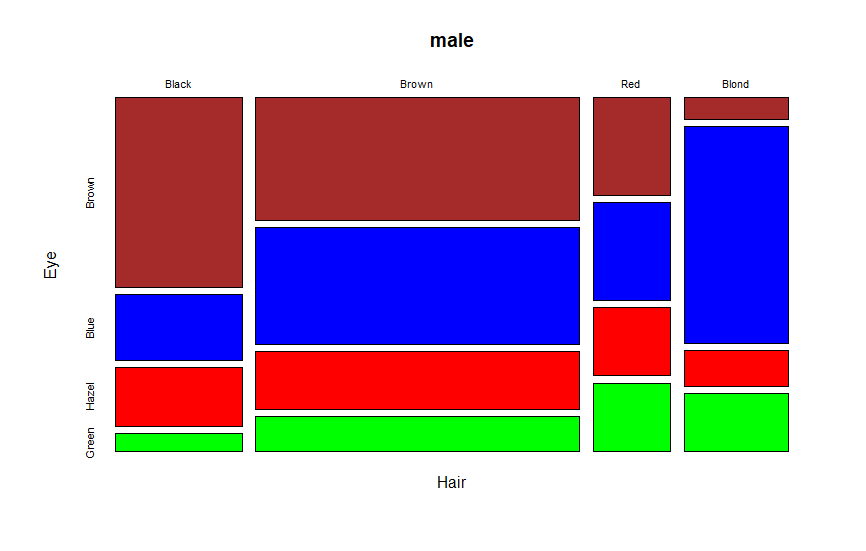


Рисунок 1 — Диаграмма цвета волос и глаз у мужчин

При уровне значимости, например, α = 0.05, гипотезу следует отклонить и признаки считать зависимыми.

Проведём аналогичное исследование для женщин.

*> female <- HairEyeColor[, , "Female"]*

*> female*

*Eye*

*Hair Brown Blue Hazel Green*

*Black 36 9 5 2*

*Brown 66 34 29 14*

*Red 16 7 7 7*

*Blond 4 64 5 8*

*> mosaicplot(female, col = c("brown", "blue", "red", "green"))*

*> chisq.test(female, simulate.p.value = TRUE)*

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

*data: female*

*X-squared = 106.66, df = NA, p-value = 0.0004998*

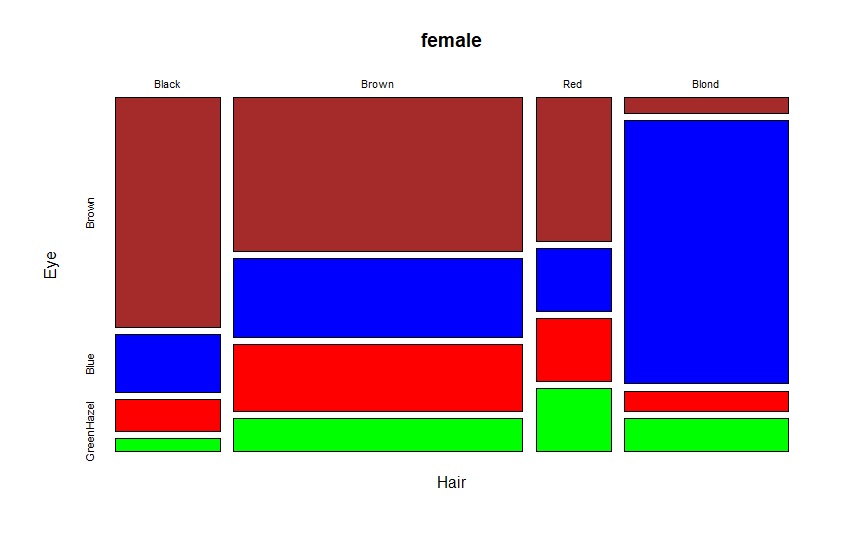


Рисунок 2 — Диаграмма цвета волос и глаз у женщин

При уровне значимости α = 0.05 гипотезу о независимости следует отклонить и признаки считать зависимыми.

**t-тест Стьюдента**

Общий вид функции:

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, ...)

Выполняется одно- или двувыборочный тест Стьюдента. Одновыборочный t-тест предназначен для проверки равенства среднего значения выборки из нормально распределённой генеральной совокупности некоторому заданному значению в предположении, что дисперсия не известна.

Двувыборочный тест служит для сравнения двух средних значений выборок из нормально распределённых генеральных совокупностей в предположении, что их дисперсии равны, хотя и не известны.

Аргументы:

* x — числовой вектор, содержащий элементы первой выборки.
* y — числовой вектор, содержащий элементы второй выборки.
* alternative — символьный аргумент, определяющий тип альтернатив-ной гипотезы. Возможные значения: “two.sided” — средние значения не равны (по умолчанию), “less” или “greater”.
* exact — либо NULL, либо логический аргумент. Отвечает за 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 объединённой 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.
* subset — вектор, определяющий используемое подмножество наблюдений.
* na.action — функция, которая вызывается, как только в данных встретилось 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 объединённая выборка. По умолчанию var.equal = FALSE и отклонение оценивается отдельно для каждой выборки. При этом происходит надлежащая корректировка числа степеней свободы.

Объект, возвращаемый функцией t.test, — 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.

Пример.

Рассмотрим данные об измерениях скорости света, полученные А. А. Майкельсоном и Э. У. Морли во время 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/c). 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).

*> light <- morley[, "Speed"]*

*> 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*

Найдено значение статистики, 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 α = 0.05. Так как p-value < α, гипотезу отклоняем.

**F-тест Фишера**

Функции:

var.test(x, y, ratio = 1, alternative = c(“two.sided”, “less”, “greater”), conf.level = 0.95, ...)

var.test(formula, data, subset, na.action, ...)

Выполняется F-тест Фишера для проверки на равенство стандартных отклонений двух нормально распределённых 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.side” (по умолчанию), “less”, “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.
* subset — вектор, определяющий используемое подмножество наблюдений.
* na.action — функция, которая вызывается, как только в данных встретилось значение NA.

Нулевая гипотеза постулирует, что отношение стандартных отклонений генеральных совокупностей, из которых выбраны x и y соответственно, равно отношению ratio.

Объект, возвращаемый функцией var.test, это список со следующими полями:

* data — используемые выборки (названия переменных, которым присвоены значения выборок).
* F — значение F-статистики Фишера.
* num df — число степеней свободы.

Пример.

Рассмотрим две выборки из разных нормальных генеральных совокупностей. Проверим, что генеральные совокупности имеют одинаковое стандартное отклонение.

*> 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*