КЛАССИЧЕСКИЕ МЕТОДЫ И КРИТЕРИИ СТАТИСТИКИ

1. **Тест Шапиро-Уилка**

Функция shapiro.test(x) выполняет тест Шапиро-Уилка. Нулевая 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(x), - это список со следующими полями:

* statistics – значение статистики Шапиро-Уилка;
* p.value – полученный уровень значимости;
* method – строка, Shapiro-Wilk normality test;
* data.name - строка содержащая имя данных, подвергнутых тесту.

Пример 1.

>x1<-rnorm(100,2,5)

>shapiro.test(x1)

Shapiro-Wilk normality test

data: x1

W = 0.97847, p-value = 0.101

При уровне значимости 0,05 гипотеза должна быть принята, т.к. p-value = 0.101>0.05.

Пример 2.

>shapiro.test(runif(100,min=-10, max=10))

Shapiro-Wilk normality test

data: runif(100, min = -10, max = 10)

W = 0.96068, p-value = 0.004499

При уровне значимости 0,05 гипотеза о нормальности распределения должна быть отвергнута, т.к. p-value = 0.004499<0.05.

1. **Критерий Колмогорова-Смирнова** для одной или двух выборок

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

Аргументы:

* x – вектор, содержащий выборку;
* y -вектор, содержащий вторую выборку, или символьная строка с именем распределения;
* … - параметры распределения;
* alternative – символьный аргумент, обозначающий 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 нулевую гипотезу о том, х 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, х принадлежит заданному распределению.

Пример 3.

> ks.test(y, z)

Two-sample Kolmogorov-Smirnov test

data: y and z

D = 0.0475, p-value = 0.7576

alternative hypothesis: two-sided

Так как, p-value = 0.7576>0.05, нулевую гипотезу о принадлежности двух выборок y и z к одному распределению принимаем.

Пример 4.

> ks.test(z, punif)

One-sample Kolmogorov-Smirnov test

data: z

D = 0.0455, p-value = 0.3782

alternative hypothesis: two-sided

Так как, p-value = 0.3782>0.05, нулевую гипотезу о принадлежности выборки z к равномерному распределению принимаем.

1. **Критерий согласия хи-квадрат Пирсона используется** для проверки нулевой гипотезы о независимости признаков.

>chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE,

simulate.p.value = FALSE, B = 2000)

Аргументы функции:

* x – вектор или матрица;
* y – вектор. Игнорируется, если х – матрица;
* сorrect – логическое значение, указывающее, требуется ли применять непрерывную коррекцию для 2×2 матриц;
* р –вектор, содержащий вероятность. Должен иметь такую же длину, что и х;
* rescale.p – логическое значение. Если TRUE, то р при необходимости нормируются 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.

Пример 5.

>chisq.test(x, y)

Pearson’s Chi-squared test

data: x and y

X-squared = 9900, df = 9801, p-value = 0.239

Так как, p-value = 0.239>0.05, то гипотезу о независимости случайных признаков можно принять.

1. **t-тест Стьюдента**

**Критерий Стьюдента (t)** применяется для проверки нулевой гипотезы о равенстве средних значений двух совокупностей, хотя существует также и одновыборочная модификация этого метода.

Основные допущения, на которых основан критерий Стьюдента:

* сравниваемые выборки должны происходить из нормально распределенных совокупностей;
* дисперсии сравниваемых генеральных совокупностей должны быть равны.

Кроме того, в своей исходной форме, t-критерий предполагает независимость сравниваемых выборок.

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, ...)

Одновыборочный t-тест предназначен для проверки равенства среднего значения выборки из нормально распределенной генеральной совокупности в предположении, что дисперсия не известна.

Двувыборочный тест служит для сравнения двух средних значений выборок из нормально распределенных генеральных совокупносей 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 альтернативной гипотезы. Возможны значения: ”two.sided”- средние 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”;
* и другие.

Пример 6.

> 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

Найдено значение t-статистики, число степеней свободы df, величина p-value. Указаны границы 95% доверительного интервала для разности математических ожиданий распределений первой и второй выборки. Приведены оценки математических ожиданий распределений для каждого распределения. Так как p-value = 0.1662>0.05, то гипотезу о том, что средние двух выборок равны, принимаем.

**Одновыборочный t-критерий**

Этот вариант критерия Стьюдента служит для проверки нулевой ***гипотезы о равенстве среднего значения (m1) генеральной совокупности***, из которой была взята выборка, некоторому априори известному значению (m0): H0: m1= m0.

Предположим, у нас имеются данные по суточному потреблению энергии, поступающей с пищей (кДж/сутки), для 11 женщин: пример заимствован из книги (Altman, 1981):

d.intake <- c(5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770)

Среднее значение для этих 11 наблюдений составляет:

mean(d.intake)

[1] 6753.6

Зададимся вопросом: отличается ли это выборочное среднее значение от установленной нормы в 7725 кДж/сутки? Разница между нашим выборочным значением и этим нормативом довольно прилична: 7725 - 6753.6 = 971.4. Но насколько эта разница статистически значима с учетом уровня вариации приведенных выше 11 значений? Ответить на этот вопрос поможет одновыборочный t-тест. Как и другие варианты t-теста, одновыборочный тест Стьюдента выполняется в R при помощи функции t.test():

t.test(d.intake, mu = 7725)

One Sample t-test

data: d.intake

t = -2.8208, df = 10, p-value = 0.01814

alternative hypothesis: true mean is not equal to 7725 95 percent confidence interval:

5986.348 7520.925

sample estimates:

mean of x

6753.636

Видим, что для имеющихся выборочных данных t-критерий составляет -2.821 при 10 степенях свободы (df). Вероятность получить такое (либо большее) значение t при условии, что проверяемая нулевая гипотеза верна, оказалась весьма мала: p-value =0.01814 (во всяком случае, это меньше 5%). Следовательно, можно отклонить проверяемую нулевую гипотезу о равенстве выборочного среднего значения нормативу и принять альтернативную гипотезу (alternative hypothesis: true mean is not equal to 7725). Принимая это предположение, рискуем ошибиться с вероятностью менее 5%.

Помимо t-критерия, числа степеней свободы, р-значения и выборочного среднего (sample estimates: mean of x), программа рассчитала также 95%-ный доверительный интервал (95 percent confidence interval) для истинной разницы между выборочным средним значением суточного потребления энергии и нормативом. Если бы повторили аналогичный тест много раз для разных групп из 11 женщин, то в 95% случаев эта разница оказалась бы в диапазоне от 5986.3 до 7520.9

кДж/сутки.

**Сравнение двух независимых выборок**

При сравнении двух выборок проверяемая нулевая гипотеза состоит в том, что обе эти выборки происходят из нормально распределенных генеральных совокупностей с одинаковыми средними значениями: H0: m1= m2.

Рассмотрим пример о суточном расходе энергии (expend) у худощавых женщин (lean) и женщин с избыточным весом (obese), приведенный в книге П. Дальгаарда (Dalgaard, 2008). Данные из этого примера (подробнее см. ?energy) входят в состав пакета ISwR, сопровождающего эту книгу:

library(ISwR)

data(energy)

attach(energy)

head(energy)

expend stature

1 9.21 obese

2 7.53 lean

3 7.48 lean

4 8.08 lean

5 8.09 lean

6 10.15 lean

Соответствующие средние значения потребления энергии в рассматриваемых группах пациенток можно найти с использованием знакомой нам функции tapply():

tapply(expend, stature, mean)

lean obese

8.07 10.30

Вопрос заключается в том, различаются ли эти средние значения статистически? Проверим гипотезу об отсутствии разницы при помощи t-теста:

t.test(expend ~ stature)

Welch Two Sample t-test

data: expend by stature

t = -3.8555, df = 15.919, p-value = 0.001411

alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval:

-3.459167 -1.004081

sample estimates:

mean in group lean mean in group obese

8.066154 10.297778

Обратите внимание на использование знака ~ в вызове функции t.test(). Это стандартный для R способ записи формул, описывающих связь между переменными. В нашем случае выражение expend ~ stature можно расшифровать как "зависимость суточного потребления энергии (expend) от статуса пациентки (stature)".

Согласно величине полученного р-значения (p-value = 0.001411), средний уровень потребления энергии у женщин из рассматриваемых весовых групп статистически значимо различается. При этом истинная разница между средними значениями с вероятностью 95% находится в диапазоне от -3.5 до -1.0 (см. 95 percent confidence interval).

Следует подчеркнуть, что при выполнении двухвыборочного t-теста функция R по умолчанию принимает, что дисперсии сравниваемых совокупностей не равны, и, как следствие, выполняет t-тест в модификации Уэлча. Мы можем изменить такое поведение программы, воспользовавшись аргументом var.equal = TRUE: (от variance – дисперсия, и equal – равный):

t.test(expend ~ stature, var.equal = TRUE)

Two Sample t-test

data: expend by stature

t = -3.9456, df = 20, p-value = 0.000799

alternative hypothesis: true difference in means is not equal

to 0

95 percent confidence interval:

-3.411451 -1.051796

sample estimates:

mean in group lean mean in group obese

8.066154 10.297778

Р-значение стало еще меньше, и мы так же, как и после теста в модификации Уэлча, можем сделать вывод о наличии существенной разницы групповых средних.

Однако такое совпадение выводов будет иметь место не всегда и, следовательно, на разницу между групповыми дисперсиями (или ее отсутствие) следует обращать серьезное внимание при выборе и интерпретации того или иного варианта t-теста.

**Сравнение двух зависимых выборок**

Зависимыми, или парными, являются две выборки, содержащие результаты измерений какого-либо количественного признака, выполненных на одних и тех же объектах. Во многих исследованиях определенный отклик измеряется у одних и тех же объектов до и после экспериментального воздействия. При такой схеме эксперимента исследователь более точно оценивает эффект воздействия именно потому, что прослеживает его фактически у каждого уникального объекта.

Нас интересуют свойства выборки, составленной из разностей значений признака у одних и тех же объектов, а точнее – "истинная средняя разность" как результат экспериментального воздействия (обозначим его δ). Если верна нулевая гипотеза H0: δ = 0, утверждающая, *что средняя разность δ между парами реализаций случайных величин статистически значимо не отличается от нуля*, *то нет оснований предполагать, что эффект воздействия имеет место*.

Возьмем другой пример о суточном потреблении энергии, измеренном уже у одних и тех же 11 женщин до и после определенного цикла:

data(intake) # из пакета ISwR

attach(intake)

head(intake)

pre post

1 5260 3910

2 5470 4220

3 5640 3885

4 6180 5160

5 6390 5645

Индивидуальные разности потребления энергии у этих женщин составляют:

post - pre

[1] -1350 -1250 -1755 -1020 -745 -1835 -1540 -1540

[9] -725 -1330 -1435

Усреднив эти индивидуальные разницы, получим

mean(post - pre)

[1] -1320.5

Задача заключается в том, чтобы оценить, насколько статистически значимо эта средняя разность отличается от нуля. Применим парный критерий Стьюдента (обратите внимание на использование аргумента paired = TRUE):

t.test(pre, post, paired = TRUE)

Paired t-test

data: pre and post

t = 11.9414, df = 10, p-value = 3.059e-07

alternative hypothesis: true difference in means is not equal

to 0

95 percent confidence interval:

1074.072 1566.838

sample estimates:

mean of the differences

1320.455

Как видим, рассчитанное программой р-значение оказалось намного меньше 0.05, что позволяет нам сделать заключение о наличии существенной разницы в потреблении энергии у исследованных женщин до и после. Истинная величина эффекта (в абсолютном выражении) с вероятностью 95% находится в интервале от 1074.1 до 1566.8 кДж/сутки.

**Использование рангового критерия Уилкоксона-Манна-Уитни**

Одно из важных условий корректного применения критерия Стьюдента состоит в том, что анализируемые выборки должны принадлежать нормально распределенным генеральным совокупностям. В случаях, когда это условие не выполняется, вместо критерия Стьюдента следует использовать его непараметрический аналог – критерий Уилкоксона (Wilcoxon rank test). Здесь необходимо сразу пояснить, что создатели системы R под названием "критерий Уилкоксона" (или "тест Уилкоксона") объединяют как метод, предложенный собственно Ф. Уилкоксоном (Wilcoxon) в 1945 г., так и опубликованный несколько позднее (1947 г.) метод Манна-Уитни. Первый из этих методов обычно используется для сравнения двух парных выборок, тогда как второй предназначен для сравнения двух независимых выборок.

**Одновыборочный критерий Уилкоксона**

Этот вариант критерия (Wilcoxon signed rank test) служит для проверки нулевой гипотезы о том, что анализируемая выборка происходит из симметрично распределенной генеральной совокупности с центром в точке µ0.

Обратимся к данным о суточном потреблении энергии у 11 женщин и выясним, имеются ли отличия от нормативного значения 7725 кДж/сутки:

d.intake <- c(5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770)

Для выполнения теста Уилкоксона в системе R используется функция wilcox.test():

wilcox.test(d.intake, mu = 7725)

Wilcoxon signed rank test with continuity correction

data: d.intake

V = 8, p-value = 0.0293

alternative hypothesis: true location is not equal to 7725

Warning message:

In wilcox.test.default(d.intake, mu = 7725) :

cannot compute exact p-value with ties

Видим, что, p-value = 0.0293 не превышает 0.05, это позволяет отклонить нулевую гипотезу о том, что суточное потребление энергии у обследованных 11 женщин не отличается от принятой нормы. Обратите внимание на выданное программой предупреждение о том, что полученное значение вероятности р не является точным из-за наличия в данных значений с одинаковыми рангами (Warning message... cannot compute exact p-value with ties). Проблема расчета точных р-значений при наличии повторяющихся значений в данных характерна для статистических методов, основанных на рангах, и критерий Уилкоксона здесь, увы, не исключение. При наличии повторяющихся наблюдений р-значение рассчитывается путем аппроксимации распределения критерия Уилкоксона нормальным распределением.

**Сравнение двух независимых выборок**

Если сравниваемые выборки являются независимыми (аргумент paired =FALSE), то мы имеем дело с критерием Уилкоксона, который в англоязычной литературе называют Wilcoxon rank sum test . Проверяемая с его помощью нулевая гипотеза состоит в том, что *центры распределений, из которых происходят сравниваемые выборки, смещены относительно друг друга на величину µ* (например, µ = 0).

Используем рассмотренный ранее пример о суточном расходе энергии (expend) у худощавых женщин (lean) и женщин с избыточным весом (obese):

data(energy) # из пакета ISwR

attach(energy)

str(energy)

'data.frame': 22 obs. of 2 variables:

$ expend : num 9.21 7.53 7.48 8.08 8.09 ...

$ stature: Factor w/ 2 levels "lean","obese": 2 1 1 1 1 1 1 1 1 1 ...

Проверим гипотезу об отсутствии разницы в потреблении энергии у женщин из этих двух групп при помощи критерия Уилкоксона для независимых выборок:

wilcox.test(expend ~ stature, paired = FALSE)

Wilcoxon rank sum test with continuity correction

data: expend by stature

W = 12, p-value = 0.002122

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(x = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, :

cannot compute exact p-value with ties

Согласно полученному р-значению (p-value = 0.002122), потребление энергии у женщин из рассматриваемых весовых групп статистически значимо различается.

**Сравнение двух зависимых выборок**

Сейчас для нас более важен тот факт, что обе сравниваемые выборки происходят из ненормально распределенных генеральных совокупностей. Это дает нам весомые основания выполнить сравнение при помощи парного рангового критерия Уилкоксона.

Как и в парном тесте Стьюдента, находят разницу между всеми имеющимися парными выборочными наблюдениями с целью проверить нулевую гипотезу о том, что медиана полученных разностей равна нулю (либо какому-либо другому, отличному от нуля значению). Здесь (псевдо)-медианой распределения F называют медиану распределения (u + v)/2, где u и v являются независимыми переменными, каждая из которых имеет распределение F. Если распределение F симметрично, псевдомедиана и медиана совпадают (подробнее см. ?wilcox.test).

Используем рассмотренный ранее пример о суточном потреблении энергии, измеренном у одних и тех же 11 женщин до и после:

data(intake) # из пакета ISwR

attach(intake)

Сравнить два периода по потреблению энергии при помощи критерия Уилкоксона можно следующим образом (обратите внимание на использование аргумента paired =TRUE):

wilcox.test(pre, post, paired = TRUE)

Wilcoxon signed rank test with continuity correction

data: pre and post

V = 66, p-value = 0.00384

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(pre, post, paired = T) :

cannot compute exact p-value with ties

Как видим, рассчитанное программой р-значение оказалось меньше 0.05, что позволяет нам сделать заключение о наличии статистически значимой разницы в потреблении энергии у исследованных женщин до и после. (Для сравнения: р-значение, полученное при помощи критерия Стьюдента было << 0.001). Мы можем оценить доверительный интервал, в котором с определенной вероятностью находится истинная величина эффекта, воспользовавшись аргументом conf.int (вероятность задается при помощи аргумента conf.level; по умолчанию рассчитывается 95%-ный доверительный интервал):

wilcox.test(pre, post, paired = TRUE, conf.int = TRUE)

Wilcoxon signed rank test with continuity correction

data: pre and post

V = 66, p-value = 0.00384

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

1037.5 1582.5

sample estimates:

(pseudo)median

1341.332

Warning messages:

1: In wilcox.test.default(pre, post, paired = TRUE, conf.int = TRUE):

cannot compute exact p-value with ties

2: In wilcox.test.default(pre, post, paired = TRUE, conf.int = TRUE):

cannot compute exact confidence interval with ties

Видим, что истинная разность уровней потребленной энергии с вероятностью 95% находится в интервале от 1037.5 до 1581.5 кДж/сутки. Из-за наличия повторяющихся наблюдений, расчет точных доверительных пределов оказался невозможным. Псевдомедиана ((pseudo)median) индивидуальных разностей между парными значениями потребления энергии была оценена в 1341.3 кДж/сутки.

Важно отметить одно из ограничений критерия Уилкоксона для двух выборок (зависимых или независимых): если общее количество наблюдений не превышает 6, то обнаружить разницу между выборками с уровнем ошибки в 5% просто невозможно.

**Задание**

1. Откройте файл данных mtcars, проверьте данные mpg на нормальность распределения (всю совокупность значение и по категориям переменной vs).
2. Проверьте, существует ли зависимость переменной mpg по категориям переменной vs.
3. Существует ли разница в значениях переменной mpg категории vs?
4. Открыть таблицу данных **trees** из библиотеки **datasets**, содержащую замеры диаметра, высоты и объема вишневых деревьев (datasets::trees или View(trees)
5. Выведите имена столбцов таблицы **trees**.
6. С помощью теста Шапиро-Уилка проверьте на нормальность каждый столбец таблицы. Сделайте выводы.
7. С помощью критерия согласия Пирсона проверьте гипотезы о независимости переменных. Сделайте выводы.
8. Откройте фрейм данных **randu** из библиотеки **datasets**,содержащий 400 троек псевдослучайных чисел из интервала [1;0]. Значения записаны в матрицу с тремя столбцами, называемыми именами x, y, z.
9. С помощью двувыборочного теста Колмогорова-Смирнова проверьте гипотезы о том, что **x, y, z** принадлежат одному и тому же непрерывному распределению. Объясните полученные результаты.

С помощью теста согласия Колмогорова-Смирнова проверьте гипотезы о том, что **x** принадлежит к нормальному виду распределения, а y – к равномерному распределению.

1. Откройте таблицу данных ldeaths (datasets::ldeaths).
2. Проверьте гипотезы о равенстве средних значений смертности по годам из таблицы ldeaths с помощью t-теста Стьюдента и о независимости данных с помощью критерия Пирсона.
3. Откройте таблицу данных **HairEyeColor** из библиотеки **datasets**, содержащуюинформацию о поле, цвете волос и глаз у 592 студентов.
4. Проверьте гипотезу о том, что для мужчин цвет глаз не зависит от цвета волос. Для этого сначала постройте таблицу сопряженных признаков по данным для мужчин (**male**). Затем, с помощью критерия Пирсона проверьте гипотезу. Сделайте вывод.
5. Проведите аналогичное исследование для женщин (**female**). Проанализируйте полученные результаты.
6. Постройте мозаичные диаграммы зависимости цвета волос и глаз для мужчин и для женщин (по таблицам сопряженных признаков **male** и **female**) с помощью функции mosaicplot().

Пример использования функции mosaicplot().

>mosaicplot(x, col=c("royalblue","purple","sienna","mediumblue"))