

Анализ данных с R

Методические указания к спецкурсу

«Вычислительные методы и пакеты в статистическом исследовании»

Антон Коробейников

asl@math.spbu.ru

Санкт-Петербург – 2010

1 Проверка гипотез однородности и смежные вопросы

1.1 Введение

В математической статистике совершенно особое место отводится проверке так называемых «гипотез однородности»¹. Этот факт связан с тем, для практических приложений, как правило, характерны задачи сравнения двух (и более) групп наблюдений. Естественно, что сравнивать выборки можно совершенно разными способами: исследователя может интересовать различие в средних, медианах, дисперсиях при разнообразных предположениях относительно самих наблюдений. В следующих разделах будет предпринята попытка рассказать о некоторых классических способах сравнения различных характеристик выборок.

1.2 Параметрические критерии проверки однородности выборок

Здесь мы остановимся на задачах проверки равенства средних и дисперсий для выборок из нормальной совокупности. Величины, имеющие нормальное распределение естественным образом возникают в большинстве экспериментов, связанных с измерением какой-либо характеристики при наличии ошибок измерения. Если предполагать, что ошибка прибора имеет нормальное распределение, то среднее отвечает за систематическую ошибку, а дисперсия — за величину случайной ошибки.

Еще раз подчеркнем: критерии представленные в данном разделе предполагают, что выборка имеет нормальное распределение. Если это заранее не известно, то этот факт нужно проверять отдельно (вставить ссылку на соотв. раздел). В противном случае все выводы, полученные на основе этих критериев, будут ошибочными.

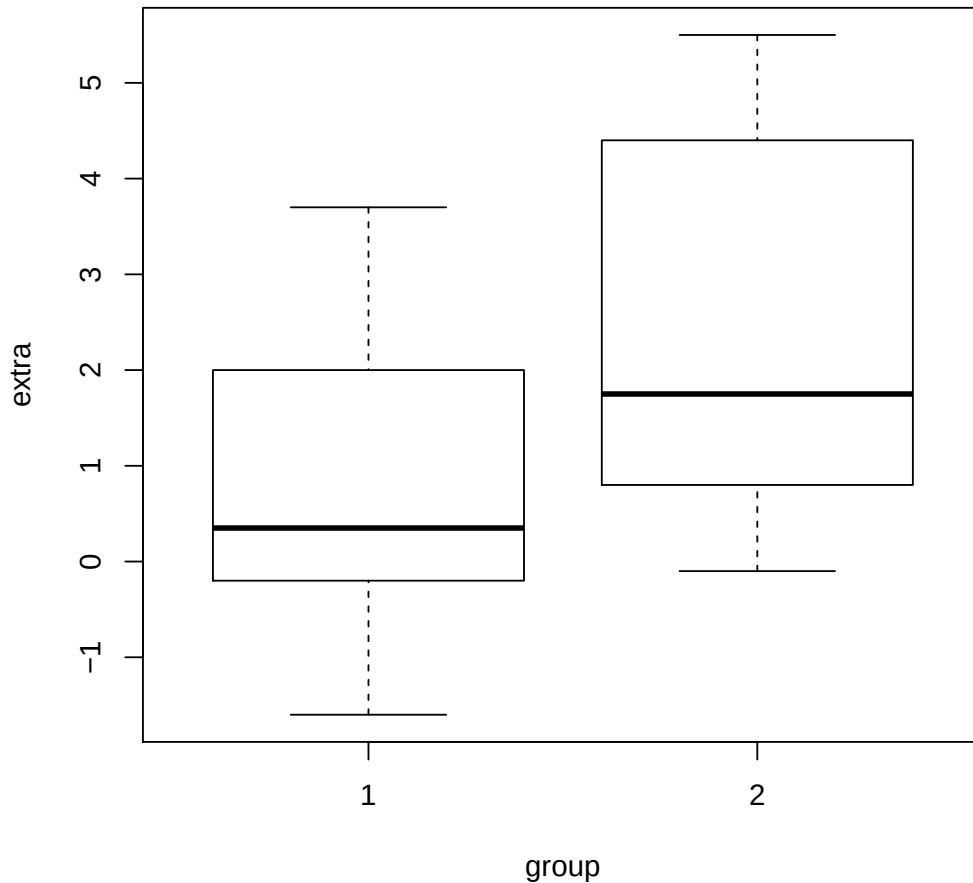
1.2.1 Двухвыборочный критерий Стьюдента равенства средних

Двухвыборочный t -критерий используется для проверки гипотезе о равенстве средних в двух независимых выборках, имеющих нормальное распределение. В классической своей постановке критерий проводится в предположении равенства дисперсий в двух выборках. В R реализована модификация критерия, позволяющая избавиться от этого предположения.

Возьмем классический набор данных, использовавшийся еще в оригинальной работе Стьюдента: производилось сравнение влияния двух различных снотворных на увеличение продолжительности сна. В R этот массив данных доступен под названием `sleep` в пакете `stats`. В столбце `extra` содержится среднее приращение продолжительности сна после начала приема лекарства (по отношению к контрольной группе), а в столбце `group` — код лекарства (первое или второе).

¹Выборки называются однородными, если они одинаково распределены.

```
> plot(extra ~ group, data = sleep)
```



Естественно, что влияние лекарства на каждого человека индивидуально, но среднее увеличение продолжительности сна можно считать логичным показателем «силы» лекарства. Попробуем проверить при помощи t-критерия, значимо ли различие в средних для этих двух выборок (соответствующих двум разным лекарствам):

```
> with(sleep, t.test(extra[group == 1], extra[group == 2], var.equal = FALSE))
```

Welch Two Sample t-test

data: extra[group == 1] and extra[group == 2]

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 of x mean of y

0.75 2.33

Параметр `var.equal` позволяет выбрать желаемый вариант критерия: оригинальный t-критерий Стьюдента в предположении одинаковых дисперсий или же t-критерий в модификации Уэлча (Welch), свободный от этого предположения.

Хотя формально гипотеза о равенстве средних не отвергается на стандартных уровнях значимости, мы видим, что возвращаемое p-значение достаточно маленькое, поэтому к данному результату стоит относиться достаточно критично. Возможно, стоит попробовать другие методы для проверки гипотезы, увеличить количество наблюдений, еще раз убедиться в нормальности распределений и т.д.

Можно ли использовать t-критерий, если необходимо сравнить среднее в зависимых выборках (например, при сравнении какого-либо жизненного показателя у пациента до и после лечения)? Ответ на этот вопрос: «да, можно». Но не обычный, а модифицированный специальным образом под такую ситуацию. В литературе такая модификация обычно называется парным t-критерием. Ничего специального для использования парного t-критерия в R делать не надо, достаточно выставить опцию `paired` в `TRUE`:

```
> with(sleep, t.test(extra[group == 1], extra[group == 2], paired = TRUE))
```

Paired t-test

data: `extra[group == 1]` and `extra[group == 2]`

t = -4.0621, df = 9, p-value = 0.002833

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.459858 -0.7001142

sample estimates:

mean of the differences

-1.58

Здесь мы видим, что парный t-критерий отвергает гипотезу о равенстве средних с достаточно большой надежностью².

²Заметим, что использованные здесь выборки не были «парными», поэтому, строго говоря, применять парный t-критерий мы права не имели, и все выводы носят сугубо иллюстративный характер.

1.2.2 Двухвыборочный критерий Фишера равенства дисперсий

Естественной характеристикой «размаха» распределения при нормальной модели является дисперсия. Предположим, что нам потребовалось проверить гипотезу об отсутствии различий в дисперсиях двух выборок. При этом мы не хотим делать абсолютно никаких допущений относительно значений средних в этих выборках. Такая задача может возникнуть естественным образом, например, при сравнении точности двух приборов. Напомним, что при наличии ошибок измерений мерой систематической ошибки прибора является среднее, а случайной — дисперсия. Систематическую ошибку, как правило, можно убрать за счет точной калибровки прибора. Случайную же ошибку убрать почти никогда не представляется возможным. В связи с этим задача проверки равенства дисперсий (например при сравнении эталонного прибора и проверяемого) становится достаточно актуальной.

Решением такой задачи служит F-критерий Фишера. В R он реализован в функции `var.test()`:

```
> x <- rnorm(50, mean = 0, sd = 2)
> y <- rnorm(30, mean = 1, sd = 1)
> var.test(x, y)
```

F test to compare two variances

data: x and y

F = 3.621, num df = 49, denom df = 29, p-value = 0.0004068

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

1.819250 6.812518

sample estimates:

ratio of variances

3.620952

В этом примере участвуют две выборки с разным количеством наблюдений, разными средними и дисперсиями. Гипотеза о равенстве дисперсий безусловно отвергается (p-значение мало). Кроме самого значения тестовой статистики, p-значения и величин степеней свободы функция выводит оценку отношения дисперсий и доверительный интервал от него. Значение оценки 3.55 не очень сильно отличается от истинного значения отношения дисперсий 4.

На самом деле, критерий проверяет несколько более общую гипотезу о том, что отношение дисперсий двух выборок имеет какое-то наперед заданное значение (понятно, что проверка гипотезы

о равенстве дисперсий является частным случаем такой гипотезы). Предполагаемое значение этого отношения можно задать посредством опции `ratio`:

```
> x <- rnorm(50, mean = 0, sd = 2)
> y <- rnorm(30, mean = 1, sd = 1)
> var.test(x, y, ratio = 4)
```

F test to compare two variances

data: x and y

F = 0.9389, num df = 49, denom df = 29, p-value = 0.8274

alternative hypothesis: true ratio of variances is not equal to 4

95 percent confidence interval:

1.886866 7.065719

sample estimates:

ratio of variances

3.755531

Тут мы видим, что при задании истинного значения отношения дисперсий гипотеза не отвергается.

1.3 Непараметрические критерии проверки однородности выборок

Критерии, приведенные в предыдущем разделе работают только в предположении нормальности распределения данных. Что делать, если мы заранее знаем, что выборки имеют другое распределение, или по каким-либо причинам проверить нормальность не можем? На помощь нам в таких случаях приходят так называемые непараметрические критерии, т.е. критерии свободные от предположения какой-либо параметрической модели данных. Естественно, ввиду того, что эти критерии оперируют гораздо меньшим «объемом информации», то они не смогут заметить те тонкие различия, которые были бы обнаружены параметрическими критериями.

1.3.1 Критерий Вилкоксона

Критерий Вилкоксона (двухвыборочный вариант которого еще известен под именем критерия Манна-Уитни) является непараметрическим аналогом t-критерия.

Стандартный набор данных `airquality` содержит информацию о величине озона в воздухе в Нью-Йорке с мая по сентябрь 1973 года. Проверим, например, гипотезу о том, что распределения уровня озона в мае и в августе было одинаковым:

```
> wilcox.test(Ozone ~ Month, data = airquality, subset = Month %in% c(5, 8))
```

Wilcoxon rank sum test with continuity correction

data: Ozone by Month

W = 127.5, p-value = 0.0001208

alternative hypothesis: true location shift is not equal to 0

Критерий отвергает гипотезу о согласии распределений уровня озона в мае и в августе с достаточно большой надежностью. В принципе, это достаточно легко вытекает из «общих соображений», так как уровень озона в воздухе сильно зависит от солнечной активности, температуры и ветра. И если с солнечной активностью все в порядке:

```
> wilcox.test(Solar.R ~ Month, data = airquality, subset = Month %in% c(5, 8))
```

Wilcoxon rank sum test with continuity correction

data: Solar.R by Month

W = 422.5, p-value = 0.4588

alternative hypothesis: true location shift is not equal to 0

то распределения ветра и температуры, напротив, сильно различаются:

```
> wilcox.test(Temp ~ Month, data = airquality, subset = Month %in% c(5, 8))
```

Wilcoxon rank sum test with continuity correction

data: Temp by Month

W = 27, p-value = 1.747e-10

alternative hypothesis: true location shift is not equal to 0

```
> wilcox.test(Wind ~ Month, data = airquality, subset = Month %in% c(5, 8))
```

Wilcoxon rank sum test with continuity correction

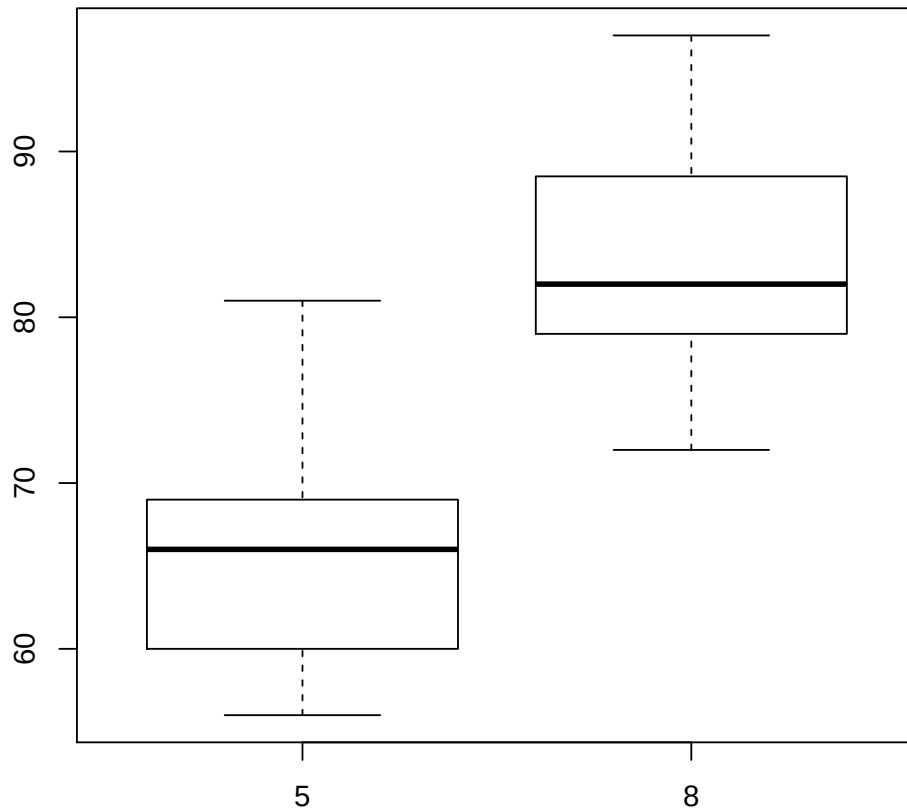
data: Wind by Month

W = 687.5, p-value = 0.003574

alternative hypothesis: true location shift is not equal to 0

Эти различия легко видны, скажем, на изображении ящиков с усами:

```
> boxplot(Temp ~ Month, data = airquality, subset = Month %in% c(5, 8))
```



но в сложных случаях только использование критериев позволяет получать объективные результаты.

По-умолчанию критерий проверяет гипотезу о том, что распределения двух выборок отличаются лишь постоянным и известным сдвигом (который, в свою очередь, по умолчанию равен нулю). Задать его можно при помощи параметра `mu`. Например:

```
> x <- rnorm(50, mean = 0)
> y <- rnorm(50, mean = 2)
> wilcox.test(x, y)
```

Wilcoxon rank sum test with continuity correction

data: x and y

W = 111, p-value = 4.207e-15

alternative hypothesis: true location shift is not equal to 0

```
> wilcox.test(x, y, mu = -2)
```

Wilcoxon rank sum test with continuity correction

data: x and y

W = 1003, p-value = 0.08926

alternative hypothesis: true location shift is not equal to -2

Критерий Манна-Уитни так же, как и t-критерий, бывает парным. Для использования парной модификации необходимо выставить опцию `paired` в значение `TRUE`. Задание опции `conf.int` позволяет получить оценку различия в сдвиге между двумя выборками³ и доверительный интервал для него:

```
> x <- rnorm(50, mean = 0)
```

```
> y <- rnorm(50, mean = 2)
```

```
> wilcox.test(x, y, conf.int = TRUE)
```

Wilcoxon rank sum test with continuity correction

data: x and y

W = 291, p-value = 3.903e-11

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-2.257208 -1.504091

sample estimates:

difference in location

-1.903832

```
> wilcox.test(x, y, mu = -2)
```

Wilcoxon rank sum test with continuity correction

data: x and y

W = 1322, p-value = 0.6221

alternative hypothesis: true location shift is not equal to -2

³Естественно, при условии, что кроме сдвига распределения двух выборок ничем не отличаются

1.3.2 Непараметрические критерии сравнения масштаба

Дисперсия является естественным параметром масштаба для выборки из нормальной совокупности. Этот факт и позволяет в случае нормального распределения заменить гипотезу о совпадении масштабов на гипотезу о совпадении дисперсий. При отказе от нормальной модели дисперсия уже не является характеристикой масштаба и поэтому надо честно проверять гипотезу именно о совпадении масштабов распределений двух выборок.

Формально, мы предполагаем, что одна из выборок имеет распределение с плотность $f(x - a)$, другая — $sf(s(x - a))$. Здесь функция плотности f и параметр сдвига a считаются неизвестными. Мы заинтересованы в проверке совпадения масштабов у двух выборок, то есть проверке того, что $s = 1$.

В стандартном пакете `stats` реализованы два классических непараметрических критерия, позволяющих проверить равенство масштабов: критерий Ансари-Брэдли (Ansari-Bradley) и критерий Муда (Mood). Начнем с первого.

```
> ansari.test(runif(50), runif(50, max = 2))
```

Ansari-Bradley test

```
data: runif(50) and runif(50, max = 2)
```

```
AB = 1483, p-value = 0.004127
```

```
alternative hypothesis: true ratio of scales is not equal to 1
```

Как несложно понять, распределение выборок в данном случае отличается только масштабом. Критерий отвергает гипотезу о совпадении дисперсий на стандартных уровнях значимости как это и должно быть. Все аналогично для критерия Муда:

```
> mood.test(runif(50), runif(50, max = 2))
```

Mood two-sample test of scale

```
data: runif(50) and runif(50, max = 2)
```

```
Z = -3.2719, p-value = 0.001068
```

```
alternative hypothesis: two.sided
```

2 Проверка гипотез нормальности распределения

2.1 Введение

Большая часть инструментов статистического вывода работает в предположении о том, что выборка получена из нормальной совокупности. За примерами далеко ходить не надо: t-критерий, критерий Фишера, построение доверительных интервалов для линейной регрессии и проверка гипотезы о линейной независимости двух выборок.

Вообще, как уже отмечалось раньше, нормальное распределение естественным образом возникает практически везде, где речь идет об измерении с ошибками. Более того, в силу центральной предельной теоремы, распределение многих выборочных величин (например, выборочного среднего) при достаточно больших объемах выборки хорошо аппроксимируется нормальным распределением вне зависимости от того, какое распределение было у выборки исходно.

В связи с этим становится понятным, почему проверке распределения на нормальность стоит уделить внимание. В дальнейшем речь пойдет о так называемых критериях согласия (goodness-of-fit tests): проверяться будет не просто факт согласия с нормальным распределением с определенными фиксированными значениями параметров, а несколько более общий факт принадлежности распределения к семейству нормальных распределений со всевозможными значениями параметров.

Основные классические критерии проверки на нормальность собраны в пакете `nortest`. Напомним, что установить пакет с CRAN можно при помощи вызова функции `install.packages()`:

```
install.packages(pkgs=c("nortest"))
```

а подключить установленный пакет при помощи функции `library()`:

```
library(nortest)
```

Естественно, может возникнуть вопрос: «А зачем столько много разных критериев для проверки одного факта? Нельзя ли выбрать наилучший и всегда его использовать?». Ответ на этот вопрос: «В общем случае, к сожалению, нельзя». Попробуем это объяснить.

Напомним, что ошибкой первого рода статистического критерия называется факт принятия альтернативной гипотезы, в ситуации когда верна гипотеза по-умолчанию. Например, пусть статистический критерий используется для разграничения доступа к какому-нибудь ресурсу. Тогда отказ в доступе для авторизованного пользователя и будет ошибкой первого рода для такого критерия. Ясно, что возможна «симметричная» ошибочная ситуация, заключающаяся в предоставлении доступа к ресурсу неавторизованному пользователю. Такая ошибка называется ошибкой второго рода: принятие критерием гипотезы по-умолчанию в ситуации, когда она не верна (то есть имеет место альтернативная гипотеза).

Как правило, чувствительность критерия к ошибкам первого рода мы настраиваем самостоятельно (как раз выбором тех самых «стандартных уровней значимости», с которым и сравниваем p -значение, выданное критерием). С ошибкой второго рода все хуже: ее вероятность сильно зависит от выбранной альтернативной гипотезы и является неотъемлемой характеристикой самого критерия. В редких случаях удастся построить критерий, который является наилучшим (это так называемые равномерно наиболее мощные критерии), и, к сожалению, это невозможно для нашей задачи. В нашем случае нулевая гипотеза формулируется просто: выборка имеет нормальное распределение с некоторыми неизвестными параметрами. А альтернативная гипотеза — полное ее отрицание. Альтернативная гипотеза гораздо «богаче» нулевой, туда входят все распределения, отличные от нормального. Для того и понадобилась целая батарея критериев: какие-то работают лучше против одного семейства альтернатив, другие — против другого. Использование всего набора позволяет быть хоть как-то уверенным в том, что распределение, не являющееся нормальным будет «распознано» хотя бы одним из критериев.

2.2 Собственно, сами критерии

2.2.1 Критерий Лиллифорса

Критерий Лиллифорса является вариантом известного классического критерия Колмогорова-Смирнова, специально модифицированного для проверки нормальности⁴. Критерий реализован в функции `lillie.test()`:

```
lillie.test(rnorm(100, mean = 6, sd = 4));
```

```
lillie.test(runif(100, min = 2, max = 4));
```

2.2.2 Критерии Крамера-фон Мизеса и Андерсона-Дарлинга

Первый критерий известен в русскоязычной литературе под именем критерия ω^2 или критерия Смирнова. Они менее известны, но обычно работают гораздо лучше, нежели критерий Лиллифорса. Реализованы в функциях `cvm.test()` и `ad.test()` соответственно:

```
cvm.test(rnorm(50, mean = 6, sd = 4));
```

```
ad.test(runif(50, min = 2, max = 4));
```

2.2.3 Критерий Шапиро-Франсиса

Этот критерий работает достаточно хорошо в большинстве не очень «сложных» случаев. Получить p -значение можно посредством функции `sf.test()`:

⁴Эта модификация существенна, для проверки гипотезы нормальности нельзя использовать классический непараметрический критерий Колмогорова-Смирнова, реализованный в функции `ks.test()`

```
sf.test(rexp(50, rate = 2));
```

2.2.4 Критерий хи-квадрат Пирсона

В отличие от задач проверки пропорций, критерий хи-квадрат обычно очень плохо работает в задачах проверки распределения на нормальность. Вероятность ошибки второго рода очень велика для достаточно широкого класса альтернативных распределений. В связи с этим, использовать его не рекомендуется.

Тем не менее: реализация его предоставлена функцией `pearson.test()`. У этой функции есть булевская опция `adjusted`, которая позволяет внести поправки в p -значение из-за наличия двух неизвестных параметров. Рекомендуемая последовательность действий такая: получить два p -значения, одно, соответствующее `adjusted=TRUE`, второе — `adjusted=FALSE`. Истинное p -значение обычно находится между. Кроме того, полезно поварьировать объем выборки и посмотреть, насколько сильно меняется p -значение. Если влияние объема выборки сильное, то от использования критерия стоит отказаться во избежание ошибок.

```
pearson.test(rnnorm(50, mean = 6, sd = 4));
```

```
pearson.test(runif(50, min = -1, max = 1));
```