Урок 6

Непараметрические критерии

6.1. Как работают непараметрические критерии?

В прошлом уроке шла речь о параметрических критериях — критериях, которые предполагают, что поступающая на вход выборка взята из какого-то распределения, и проверяют гипотезы о значениях параметров этого распределения. В этом уроке все будет иначе. Непараметрические критерии применяются в задачах следующего типа. Есть выборка объема n из какого-то распределения $F(x)$:

$$X^n = (X_1, \ldots, X_n), X \sim F(x).$$

Проверяется гипотеза о равенстве нулю среднего значения случайной величины, из которой взята эта выборка.

Чтобы проверять любую гипотезу, нужна T-статистика, для которой должно быть известно нулевое распределение, то есть распределение при условии справедливости нулевой гипотезы. Если про исходное распределение $F(x)$ что-то известно и T-статистика выбрана удачно, то нулевое распределение статистики может быть выражено аналитически. Однако распределение F(x) может быть нестандартным, и о нём может быть ничего не известно.

Гипотезы про среднее значение можно проверять с использованием центральной предельной теоремы (фактически, с помощью Z-критерия). Но центральная предельная теорема не всегда применима: иногда распределения бывают слишком скошенные, иногда выборка недостаточно большая, чтобы распределение ее выборочного среднего можно было считать нормальным.

В таких ситуациях существует два варианта действий. Во-первых, имеющуюся выборку из неизвестного распределения можно преобразовать так, что о её распределении будет больше информации. Во-вторых, можно сделать какие-то предположения о функции распределения исходной выборки $F(x)$, и на основании этих предположений построить статистику, нулевое распределение которой можно оценить.

В методах, которые будут рассмотрены далее в этом уроке, в разных комбинациях используются эти два способа работы с выборками.

6.2. Критерии знаков

Критерии знаков — это одно из семейств непараметрических критериев. Эти критерии обладают невысокой мощностью, но они крайне универсальны и практически ничего не требуют от данных, поэтому они очень полезны на практике.

6.2.1. Критерий знаков для одной выборки

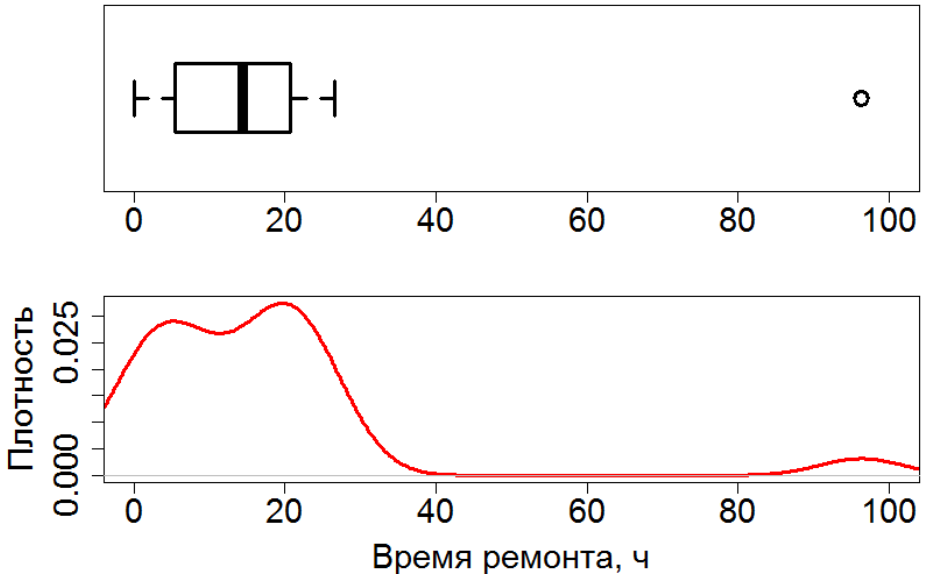

Рис. 6.1: Данные о времени ремонта интернет-оборудования клиентов провайдера Verizon

Для примера можно рассмотреть данные о времени ремонта интернет-оборудования клиентов провайдера Verizon. Выборка состоит из 23 наблюдений, и, как видно по графикам на рисунке 6.1, распределение признака не похоже на нормальное. Хочется понять, позволяют ли собранные данные утверждать, что среднее время ремонта составляет больше восьми часов. Использовать для этого параметрические критерии (например, критерий Стьюдента) не стоит, поскольку у распределения признака тяжелый правый хвост. Кроме того, объем выборки достаточно маленький, поэтому не получается воспользоваться центральной предельной теоремой.

Решить эту задачу можно с помощью критерия знаков.

выборка:	$X^n = (X_1, \ldots, X_n), X_i \neq m_0;$
нулевая гипотеза:	$H_0 \colon \operatorname{med} X = m_0;$
альтернатива:	$H_1 \colon \operatorname{med} X <\neq> m_0;$
статистика:	$T(X^n) = \sum_{i=1}^{n} [X_i > m_0];$
нулевое распределение:	$T(X^n) \sim Bin(n, \frac{1}{2}).$

Таблица 6.1: Описание одновыборочного критерия знаков

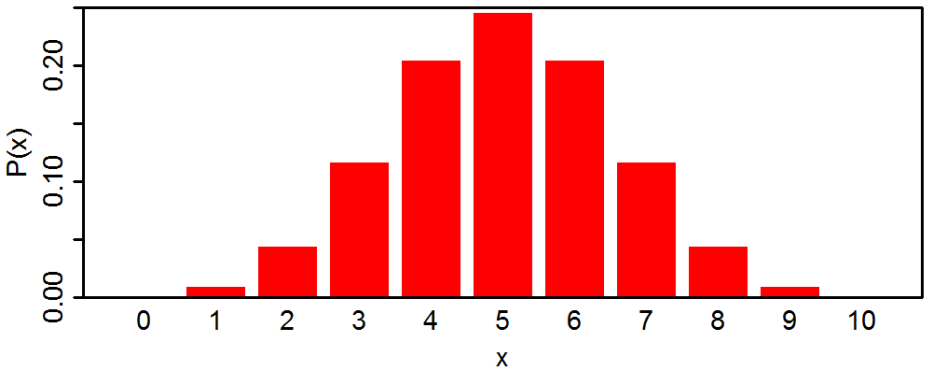

Рис. 6.2: Биномиальное распределение Bin(10, 0.5)

Единственное требование при применении этого критерий — в выборке не должно быть ни одного объекта,

значение признака которого в точности совпадает с m_0. Если нулевая гипотеза справедлива, то статистика этого критерия имеет биномиальное распределение (рисунок 6.2).

Итак, в задаче о времени ремонта интернет-оборудования проверяется нулевая гипотеза о том, что медиана времени ремонта составляет 8 часов:

$$H_0 : \operatorname{med} X = 8.$$

Односторонняя альтернатива утверждает, что ремонт в среднем длится дольше 8 часов:

$$H_1 : \operatorname{med} X > 8.$$

В имеющейся выборке ремонт занял больше 8 часов в 15 случаях из 23. Критерий знаков утверждает, что это недостаточно много. Его достигаемый уровень значимости (вероятность получить 15 из 23 в условиях справедливости нулевой гипотезы) $p = 0.105$. Нулевая гипотеза не отвергается, данные не позволяют утверждать, что ремонт в среднем длится дольше 8 часов.

6.2.2. Цензурированная выборка

Критерий знаков настолько нетребователен к данным, что его можно использовать даже на цензурированных выборках.

Пример. Наблюдаются пациенты с лимфоцитарной лимфомой, измеряемый признак — это время их жизни в неделях после того, как был поставлен диагноз. Исследование длится семь лет. В выборке есть один пациент, который после семи лет (362 недель наблюдений) остался жив. Поскольку исследование закончилось, неизвестно, сколько еще он прожил после этого. Такая выборка называется цензурированной сверху, поскольку на части объектов известна только нижняя граница значения признака.

Если требуется проверить гипотезу о том, что среднее время дожития составляет 200 недель, против односторонней альтернативы, что оно больше 200 недель, для этой выборки можно без проблем использовать критерий знаков. Его достигаемый уровень значимости $p = 0.9453$. То есть нулевую гипотезу нельзя отклонить против односторонней альтернативы.

6.2.3. Критерий знаков для связанных выборок

	$\mathrm{AUC}_{C4.5}$	$\mathrm{AUC}_{C4.5+m}$
adult (sample)	0.763	**0.768**
breast cancer	**0.599**	0.591
breast cancer wisconsin	0.954	**0.971**
cmc	0.628	**0.661**
ionosphere	0.882	**0.888**
iris	**0.936**	0.931
liver disorders	0.661	**0.668**
lung cancer	0.583	0.583
lymphography	0.775	**0.838**
mushroom	1.000	1.000
primary tumor	0.940	**0.962**
rheum	0.619	**0.666**
voting	0.972	**0.981**
wine	0.957	**0.978**

Таблица 6.2: Данные о качестве классификаторов

В этом примере рассматривается классификатор C4.5 (один из способов построения деревьев решений). Для этого классификатора на 14 стандартных наборах данных посчитали площадь под ROC-кривой на тестовой выборке. Эти данные находятся в первом столбце таблицы 6.2. Далее этот классификатор модифицировали — изменили один из его гиперпараметров, минимальное количество объектов в листе m. Для нового классификатора на тех же 14 наборах данных посчитали площадь под ROC-кривой на тестовой выборке, результат — второй столбец таблицы 6.2.

Используя полученные данные, нужно определить, какая из этих двух версий классификатора лучше. Во-первых, можно заметить, что из 14 датасетов на 10 площадь под ROC-кривой больше у второй версии

классификатора. На двух наборах данных показывает лучший результат первая версия классификатора, и еще на двух — ничья.

Чтобы по этим цифрам посчитать статистическую значимость, можно использовать критерий знаков для связанных выборок (таблица 6.3).

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n})$, $X_2^n = (X_{21}, \ldots, X_{2n})$, $X_{1i} \neq X_{2i}$, выборки связанные;
нулевая гипотеза:	$H_0 \colon \mathbf{P}(X_1 > X_2) = \frac{1}{2}$;
альтернатива:	$H_1 \colon \mathbf{P}(X_1 > X_2) <\neq> \frac{1}{2}$;
статистика:	$T\left(X_1^n, X_2^n\right) = \sum\limits_{i=1}^{n} \left[X_{1i} > X_{2i}\right]$;
нулевое распределение:	$T\left(X_1^n, X_2^n\right) \sim Bin(n, \frac{1}{2})$.

Таблица 6.3: Описание критерия знаков для связанных выборок

В этом уроке идёт речь о непараметрических критериях, которые проверяют гипотезы о средних, но под «средними» они часто понимают совершенно разные вещи. Так, одновыборочный критерий знаков под средним понимает медиану. Двухвыборочный критерий знаков гипотезу о средних формулирует в представленном выше экзотическом виде. Другие критерии могут использовать другие варианты нулевых гипотез, но, тем не менее, всё это — в каком-то виде утверждение о средних.

Статистика двухвыборочного критерия знаков — это сумма индикаторов того, что элемент первой выборки больше, чем соответствующий элемент второй выборки:

$$T\left(X_1^n, X_2^n\right) = \sum_{i=1}^{n} \left[X_{1i} > X_{2i}\right].$$

Если нулевая гипотеза справедлива, эта статистика, так же, как и в случае одновыборочного критерия, имеет биномиальное распределение (рисунок 6.2) с параметрами $n, \frac{1}{2}$:

$$T\left(X_1^n, X_2^n\right) \sim Bin(n, \frac{1}{2}).$$

В задаче о качестве классификаторов требуется проверить нулевую гипотезу о том, что их среднее качество одинаково:

$$H_0 \colon \mathbf{P}(\text{AUC}_{C4.5+m} > \text{AUC}_{C4.5}) = \frac{1}{2}.$$

Эта гипотеза проверяется против односторонней альтернативы о том, что качество модифицированного классификатора выше:

$$H_1 \colon \mathbf{P}(\text{AUC}_{C4.5+m} > \text{AUC}_{C4.5}) > \frac{1}{2}$$

Странно предполагать, что при настройке какого-то гиперпараметра получится в среднем падение качества классификатора, поэтому используется именно односторонняя альтернатива. Критерий знаков дает достигаемый уровень значимости $p = 0.019$. На уровне значимости 0.05 отвергается нулевая гипотеза о том, что у этих классификаторов качество одинаковое, против альтернативы о том, что второй классификатор лучше. Модифицированный алгоритм лучше на 83% датасетов. 95% нижний доверительный предел для доли датасетов, на которых модифицированный классификатор лучше, — 56.2%.

6.3. Ранговые критерии

Для проверки гипотез о средних критерии знаков выбрасывают большую часть информации, содержащуюся в выборке. Вместо исходных значений признака используется бинарный вектор. Ранговые критерии позволяют сохранить больше информации.

Выборку

$$X_1, \ldots, X_n$$

всегда можно превратить в вариационный ряд, то есть упорядочить её по неубыванию:

$$X_{(1)} \leq \ldots < \underbrace{X_{(k_1)} = \ldots = X_{(k_2)}}_{\text{связка размера } k_2 - k_1 + 1} < \ldots \leq X_{(n)}.$$

Если при этом есть какие-то части вариационного ряда, в которых элементы полностью совпадают, эти части называются «связками».

Рангом наблюдения X_i называется его позиция в вариационном ряду. Если X_i не попадает в связку, то

$$\operatorname{rank}(X_i) = r \colon X_i = X_{(r)},$$

а если X_i оказывается в связке $X_{(k_1)}, \ldots, X_{(k_2)}$, то

$$\operatorname{rank}(X_i) = \frac{k_1 + k_2}{2},$$

то есть в связке все объекты получают одинаковый средний ранг.

6.3.1. Критерий знаковых рангов Уилкоксона

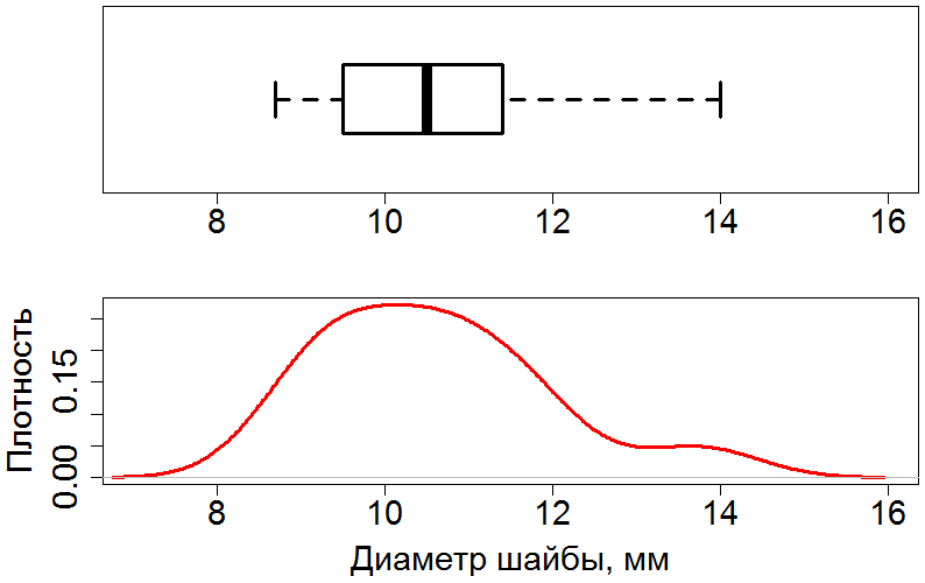

Рис. 6.3: Данные в задаче о размере шайбы

Использовать ранги можно для решения следующей задачи: имеются 24 шайбы, произведённые на одном и том же конвейере, для которых измерены диаметры. По этой выборке требуется понять, соответствует ли диаметр шайбы стандартному размеру в $m_0 = 10$ мм. Для этого будет использоваться критерий знаковых рангов, или, как его иногда называют, критерий знаковых рангов Уилкоксона (таблица 6.4).

выборка:	$X^n = (X_1, \ldots, X_n), X_i \neq m_0,$
	F_X симметрично относительно медианы;
нулевая гипотеза:	$H_0 \colon \operatorname{med} X = m_0;$
альтернатива:	$H_1 \colon \operatorname{med} X <\neq> m_0;$
статистика:	$W(X^n) = \sum_{i=1}^{n} \operatorname{rank}(\lvert X_i - m_0 \rvert) \cdot \operatorname{sign}(X_i - m_0);$
нулевое распределение:	табличное.

Таблица 6.4: Описание критерия знаковых рангов Уилкоксона

1	2	3	4	5	W
$-$	$-$	$-$	$-$	$-$	-15
$+$	$-$	$-$	$-$	$-$	-13
$-$	$+$	$-$	$-$	$-$	-11
$+$	$+$	$-$	$-$	$-$	-9
$-$	$-$	$+$	$-$	$-$	-9
...	...	...	...	...	...
$+$	$+$	$-$	$+$	$+$	9
$-$	$-$	$+$	$+$	$+$	9
$+$	$-$	$+$	$+$	$+$	11
$-$	$+$	$+$	$+$	$+$	13
$+$	$+$	$+$	$+$	$+$	15

Таблица 6.5: Возможные реализации знаков рангов и соответствующие им значения статистики

При справедливости нулевой гипотезы каждый из рангов в выборке мог с одинаковой вероятностью реализоваться с любым знаком ($sign(X_i - m_0)$): и с «+», и с «−». Таким образом, получается 2^n вариантов распределения знаков по рангам. Перебирая все эти варианты, для каждого из них можно вычислить значение статистики, пример перебора показан в таблице 6.5. Именно так строится нулевое распределение критерия знаковых рангов.

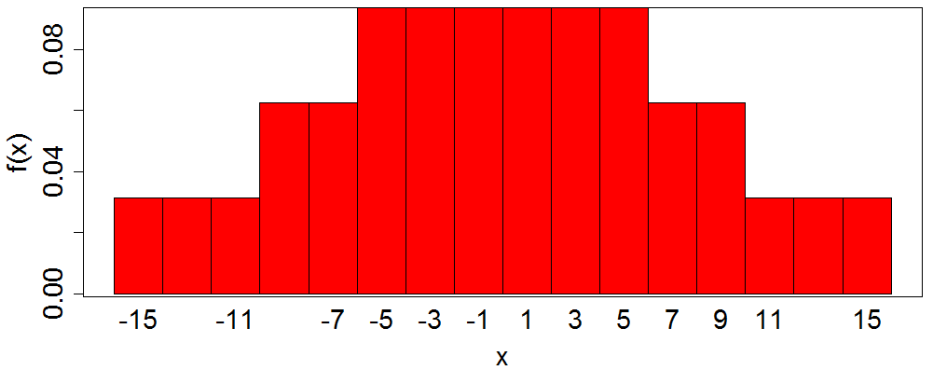

Рис. 6.4: Нулевое распределение при размере выборки $n = 5$

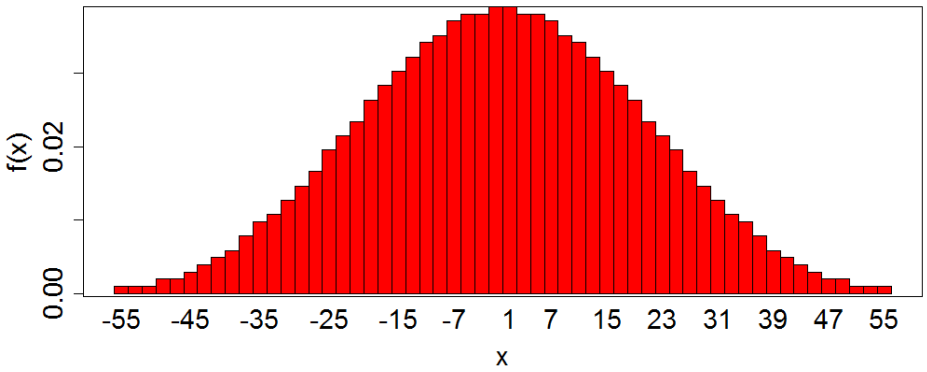

Рис. 6.5: Нулевое распределение при размере выборки $n = 10$

Нулевое распределение статистики при размерах выборки $n = 5, 10, 15$ показано на рисунках 6.4, 6.5, 6.6. Из них видно, что с ростом объёма выборки нулевое распределение становится похожим на нормальное. При размере выборки $n > 20$ можно использовать следующую нормальную аппроксимацию:

$$W \approx\sim N\left(0, \frac{n(n+1)(2n+1)}{6}\right).$$

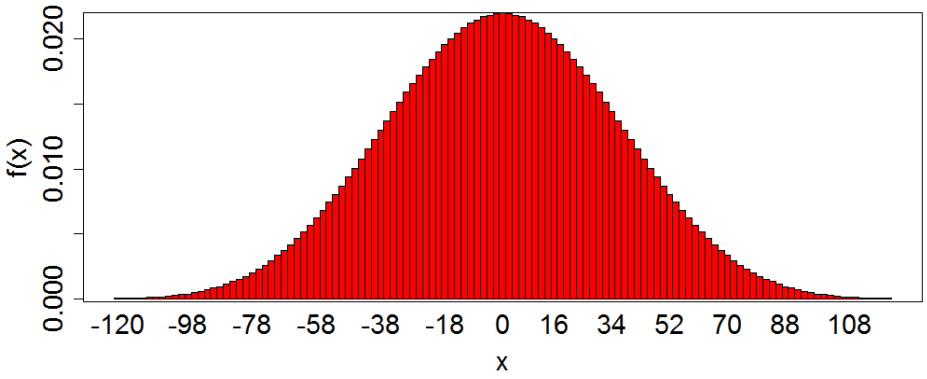

Рис. 6.6: Нулевое распределение при размере выборки $n = 15$

В задаче о диаметре шайбы проверяется нулевая гипотеза о том, что средний размер шайбы составляет 10 миллиметров:

$$H_0 \colon \operatorname{med} X = 10$$

против двусторонней альтернативы:

$$H_1 \colon \operatorname{med} X \neq 10$$

Критерий знаковых рангов даёт достигаемый уровень значимости $p = 0.0673$, нулевая гипотеза не отвергается. Выборочная медиана диаметра составляет 10.5 мм, 95% доверительный интервал: $[9.95, 11.15]$ мм. Доверительной интервал содержит целевое значение $m_0 = 10$. Так и должно быть, когда достигаемый уровень значимости выше порога.

6.3.2. Двухвыборочная задача со связанными выборками

Как и до этого в курсе, двухвыборочная задача со связанными выборками решается с использованием того же самого критерия, что и одновыборочная. Версия критерия знаков для двух связанных выборок показана в таблице 6.6.

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n})$, $X_2^n = (X_{21}, \ldots, X_{2n})$, $X_{1i} \neq X_{2i}$, выборки связанные;
нулевая гипотеза:	$H_0 \colon \operatorname{med}(X_1 - X_2) = 0$;
альтернатива:	$H_1 \colon \operatorname{med}(X_1 - X_2) \lessgtr 0$;
статистика:	$W(X_1^n, X_2^n) = \sum_{i=1}^{n} \operatorname{rank}(\lvert X_{1i} - X_{2i} \rvert) \cdot \operatorname{sign}(X_{1i} - X_{2i})$;
нулевое распределение:	табличное.

Таблица 6.6: Описание критерия знаковых рангов для связанных выборок

На рисунке 6.7 показан график, отражающий данные о депрессивности 9 пациентов, измеренной по шкале Гамильтона до и после первого приёма транквилизатора. Хочется понять, действует ли транквилизатор, то есть снижается ли у этих пациентов депрессивность. Формально проверяется нулевая гипотеза о равенстве нулю медианы попарных разностей депрессивности до и после приёма транквилизаторов:

$$H_0 \colon \operatorname{med}(X_2 - X_1) = 0.$$

Альтернативная односторонняя гипотеза — депрессивность снизилась:

$$H_1 \colon \operatorname{med}(X_2 - X_1) < 0$$

Критерий знаковых рангов даёт достигаемый уровень значимости $p = 0.019$, то есть нулевая гипотеза отвергается в пользу односторонней альтернативы. Медиана снижения составляет 0.49 пунктов. 95% нижний доверительный предел для снижения: 0.175 пунктов.

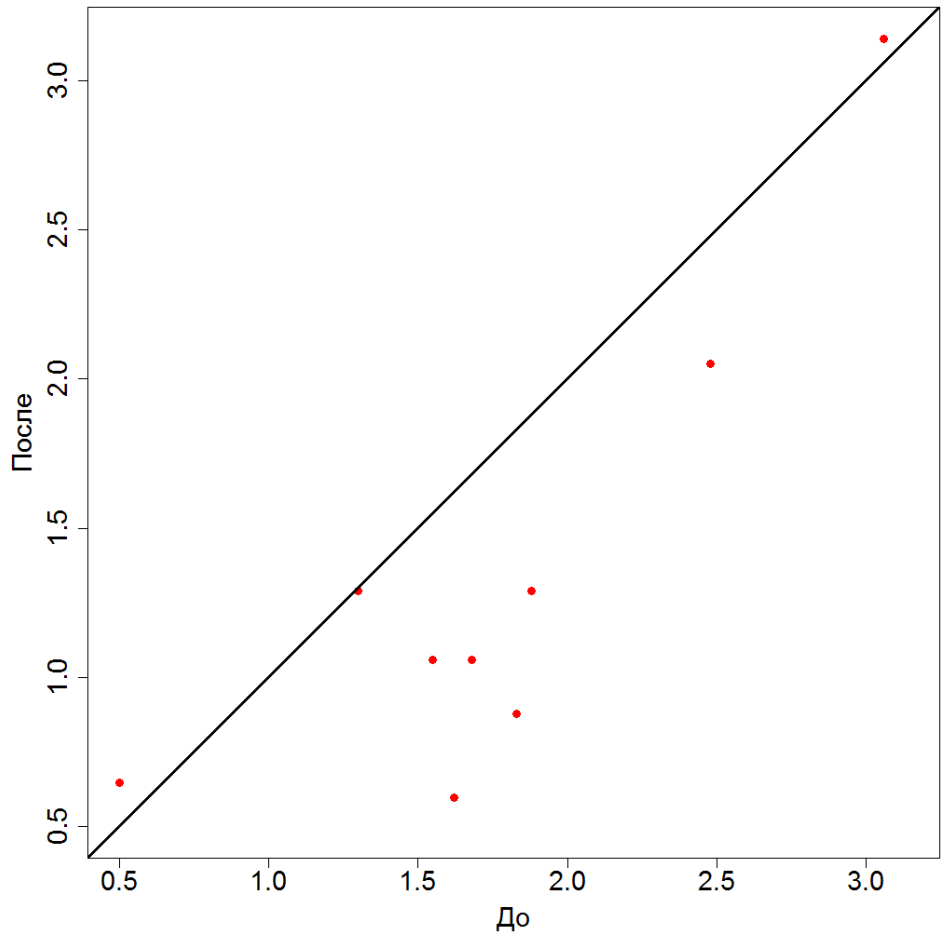

Рис. 6.7: Данные о депрессивности пациентов до и после приёма транквилизатора

6.3.3. Двухвыборочная задача с независимыми выборками

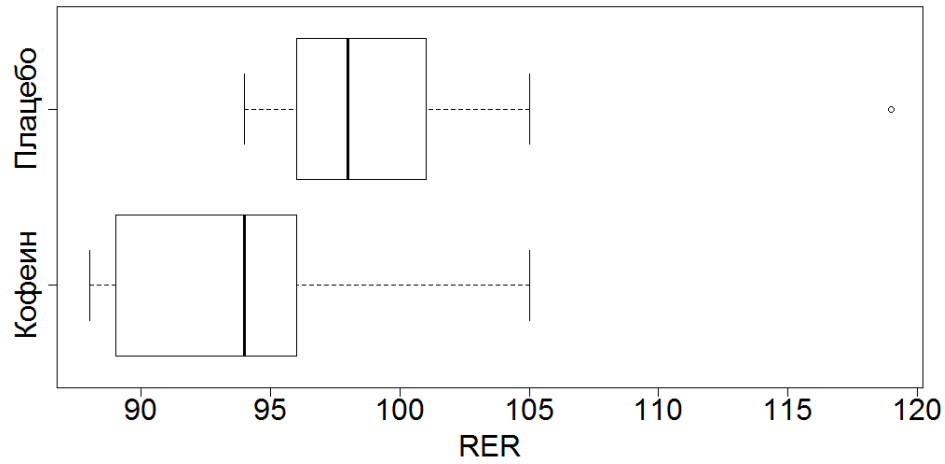

Рис. 6.8: Данные о респираторном обмене испытуемых после принятия кофеина или плацебо

В этой задаче измеряемый признак — это респираторный обмен, соотношение числа молекул углекислого газа и кислорода в выдыхаемом воздухе. Респираторный обмен является косвенным признаком того, из чего в данный момент мышцы вырабатывают энергию, из жиров или углеводов. В эксперименте измеряется респираторный обмен у 18 испытуемых в процессе физических упражнений (рисунок 6.8). За час до этого 9 из

них получили таблетку кофеина, а оставшиеся 9 — таблетку плацебо. Хочется понять, повлиял ли кофеин на среднее значение показателей респираторного обмена.

Эту задачу можно решить с помощью критерия Манна-Уитни, который иногда называют критерием Уилкоксона-Манна-Уитни (таблица 6.7).

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$, $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$,
нулевая гипотеза:	$H_0 \colon F_{X_1}(x) = F_{X_2}(x)$;
альтернатива:	$H_1 \colon F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$;
статистика:	$X_{(1)} \leq \ldots \leq X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$, $R_1(X_1^{n_1}, X_2^{n_2}) = \sum\limits_{i=1}^{n_1} \operatorname{rank}(X_{1i})$;
нулевое распределение:	табличное.

Таблица 6.7: Описание критерия Манна-Уитни

Относительно параметра Δ альтернатива в этом критерии может быть односторонней или двусторонней. Если справедлива альтернативная гипотеза и между распределениями действительно есть сдвиг, то средние значения признаков в выборках будут различаться. Поэтому это тоже в каком-то виде гипотеза о средних.

Для того чтобы построить статистику критерия Манни-Уитни, для объединенной выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$ строится вариационный ряд

$$X_{(1)} \leq \ldots \leq X_{(n_1+n_2)},$$

и подсчитываются ранги

$$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \operatorname{rank}(X_{1i}).$$

X_1	X_2	R_1
{1,2,3}	{4,5,6,7}	6
{1,2,4}	{3,5,6,7}	7
{1,2,5}	{3,4,6,7}	8
{1,2,6}	{3,4,5,7}	9
{1,2,7}	{3,4,5,6}	10
{1,3,4}	{2,5,6,7}	8
...	...	...
{3,5,7}	{1,2,4,6}	15
{3,6,7}	{1,2,4,5}	16
{4,5,6}	{1,2,3,7}	15
{4,5,7}	{1,2,3,6}	16
{4,6,7}	{1,2,3,5}	17
{5,6,7}	{1,2,3,4}	18

Таблица 6.8: Возможные распределения рангов между выборками

Статистикой будет сумма рангов элементов первой выборки в объединенном вариационном ряду. Нулевое распределение этой статистики, как и в предыдущем случае, табличное. Оно получается следующим образом. Если нулевая гипотеза справедлива, то каждый из рангов с одинаковой вероятностью мог реализоваться как в выборке X_1, так и в выборке X_2. Необходимо перебрать все возможные варианты того, как это могло произойти (таблица 6.8), всего таких вариантов $C_{n_1+n_2}^{n_1}$. На каждом из этих вариантов нужно вычислить значение статистики критерия Манни-Уитни, так и получается нулевое распределение.

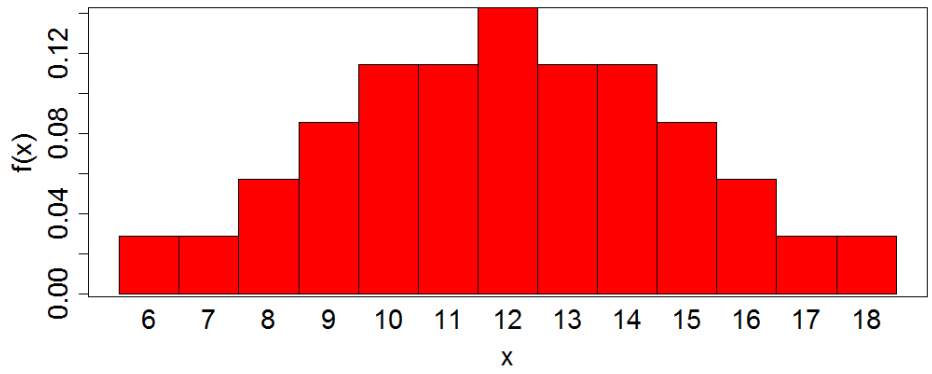

Рис. 6.9: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок $n_1 = 3,\ n_2 = 4$

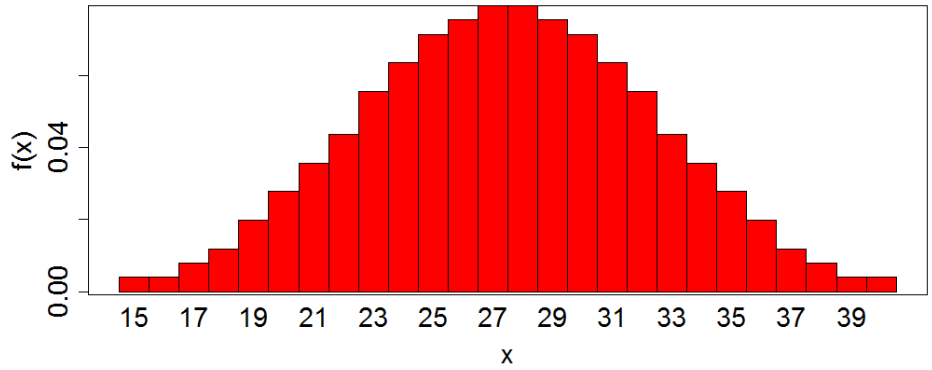

Рис. 6.10: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок $n_1 = n_2 = 5$

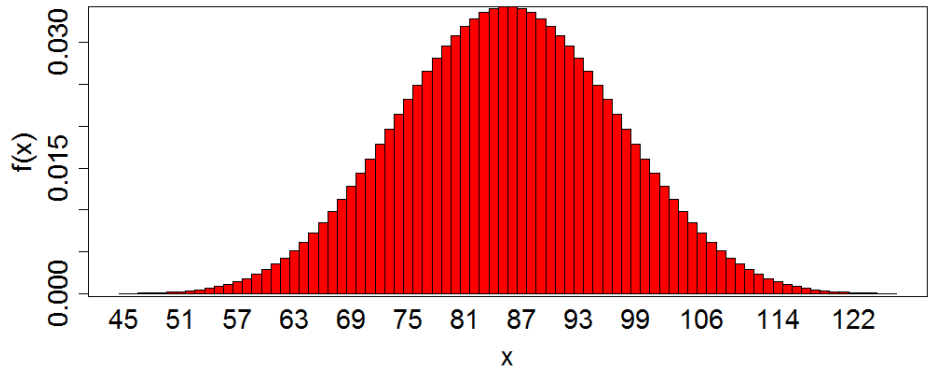

Рис. 6.11: Нулевое распределение статистики критерия Манна-Уитни при размерах выборок $n_1 = n_2 = 10$

Нулевые распределения статистики критерия Манна-Уитни для различных размеров выборок показаны на рисунках 6.9, 6.10, 6.11. Как и в предыдущем случае, при увеличении объёма выборок нулевое распределение стремится к нормальному. Для критерия Манни-Уитни также можно использовать нормальную аппроксимация нулевого распределения, если в каждой из выборок есть по меньшей мере десять объектов:

$$R_1 \sim N\left(\frac{n_1(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right).$$

В задаче о кофеине и респираторным обмене проверяется нулевая гипотеза H_0: среднее значение показателей в двух группах не отличаются, против двухсторонней альтернативы H_1: среднее значение показателя респираторного обмена отличается в двух группах. Критерий Манни-Уитни даёт достигаемый уровень значимости $p = 0.0521$, это совсем немного больше стандартного уровня значимости в 0.05. Сдвиг между

средними значениями показателей в двух выборках составляет 6 пунктов, 95% доверительный интервал — $[-0.00005, 12]$ пт, 0 в него всё-таки попадает, отвергнуть нулевую гипотезу нельзя.

6.4. Перестановочные критерии

При использовании ранговых критериев выборки превращают в ранги, затем делается какое-то дополнительное предположение, и на основании этого предположения получается, что разные конфигурации этих рангов при справедливости нулевой гипотезы могут реализоваться с равной вероятностью. Далее необходимо перебрать все конфигурации, и на каждой посчитать значение статистики — таким образом оценивается ее нулевое распределение.

Если в этом алгоритме пропустить первый пункт (не превращать наблюдения в ранги), а остальное делать точно так же, то получится алгоритм работы перестановочных критериев.

6.4.1. Одновыборочный перестановочный критерий

выборка:	$X_1^n = (X_1, \ldots, X_n)$, $F(X)$ симметрично относительно матожидания;
нулевая гипотеза:	$H_0 \colon \mathbb{E}X = m_0$;
альтернатива:	$H_1 \colon \mathbb{E}X <\neq> m_0$;
статистика:	$T(X^n) = \sum_{i=1}^{n} (X_i - m_0)$,
нулевое распределение:	порождается перебором 2^n знаков перед слагаемыми $X_i - m_0$.

Таблица 6.9: Описание одновыборочного перестановочного критерия

Имеется выборка размера n: и делается предположение, что функция распределения $F(x)$ симметрична относительно математического ожидания. Одновыборочный перестановочный критерий (таблица 6.9) проверяет нулевую гипотезу о значении математического ожидания случайной величины, из которой взята выборка.

Если нулевая гипотеза этого критерия справедлива, каждый из объектов выборки мог с одинаковой вероятностью реализоваться слева и справа от математического ожидания. Поэтому нужно перебрать все 2^n знаков, которые могут стоять в выражении для статистики перед разностью $x_i - m_0$. На основании этого перебора и будет восстановлено нулевое распределение статистики.

В качестве примера использования одновыборочного перестановочного можно вспомнить задачу анализа диаметра шайб (рисунок 6.3): по выборке из 24 элементов требуется понять, соответствует ли средний диаметр шайбы стандарту — 10 миллиметров:

$$H_0 \colon \mathbb{E}X = 10.$$

Эта нулевая гипотеза проверяется против двусторонней альтернативы о том, что средний диаметр шайбы не соответствует стандарту:

$$H_1 \colon \mathbb{E}X \neq 10.$$

Критерий знаковых рангов в этом случае давал достигаемый уровень значимости $p = 0.0673$, нулевое распределение показано на рисунке 6.12.

Нулевое распределение, полученное при использовании перестановочного критерия, показано на рисунке 6.13. Значение статистики, реализовавшейся в эксперименте: $T = 14.6$.

Чтобы вычислить достигаемый уровень значимости, нужно просуммировать высоты всех столбцов в распределении статистики, начиная от значения 14.6 и больше, а также от -14.6 и меньше (поскольку альтернатива двухсторонняя). В результате получается достигаемый уровень значимости $p = 0.1026$, то есть нулевая гипотеза не отвергается.

Фактически достигаемый уровень значимости перестановочного критерия — это доля перебираемых перестановок, на которых получается такое же или еще более экстремальное значение статистики.

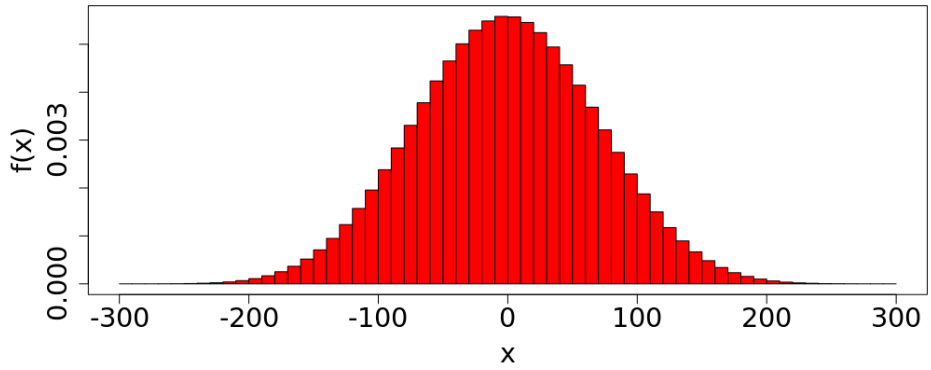

Рис. 6.12: Распределение тестовой статистики в решении задачи о размере шайб с использованием критерия знаковых рангов

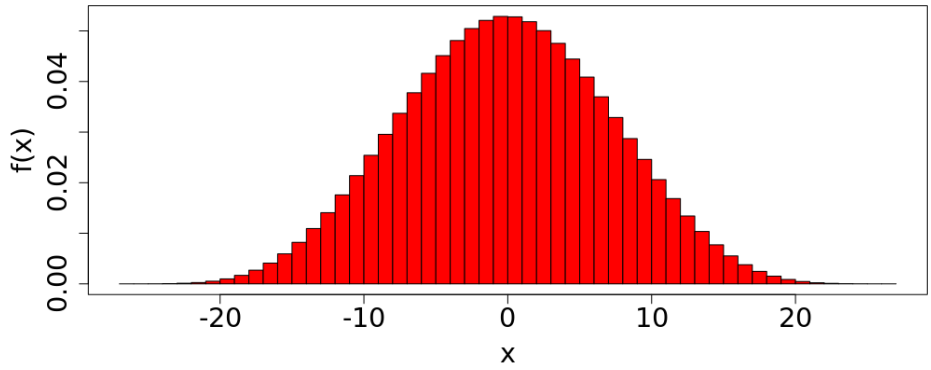

Рис. 6.13: Распределение тестовой статистики в решении задачи о размере шайб с использованием перестановочного критерия

6.4.2. Двувыборочный критерий для связанных выборок

Двухвыборочная задача со связанными выборками решается с использованием абсолютно такого же критерия: от двух связанных выборок происходит переход к одной выборке соответствующих попарных разностей (таблица 6.10).

выборки:	$X_1^n = (X_{11}, \ldots, X_{1n})$, $X_2^n = (X_{21}, \ldots, X_{2n})$, выборки связанные;
нулевая гипотеза:	$H_0 \colon \mathbb{E}(X_1 - X_2) = 0$;
альтернатива:	$H_1 \colon \mathbb{E}(X_1 - X_2) <\neq> 0$;
статистика:	$D^n = (X_{1i} - X_{2i})$, $T(X_1^n, X_2^n) = T(D^n) = \sum_{i=1}^{n} D_i$,
нулевое распределение:	порождается перебором 2^n знаков перед слагаемыми D_i.

Таблица 6.10: Описание двухвыборочного перестановочного критерия для связанных выборок

В качестве примера можно вспомнить задачу об эффективности транквилизатора. У девяти пациентов, до и после приема транквилизатора, была измерена депрессивность по шкале Гамильтона (рисунок 6.7). Требуется проверить нулевую гипотезу о том, что депрессивность не изменилась:

$$H_0 \colon \mathbb{E}(X_1 - X_2) = 0,$$

против односторонней альтернативы о том, что транквилизатор подействовал, то есть депрессивность снизи-

лась:

$$H_1 : \mathbb{E}\,(X_1 - X_2) > 0$$

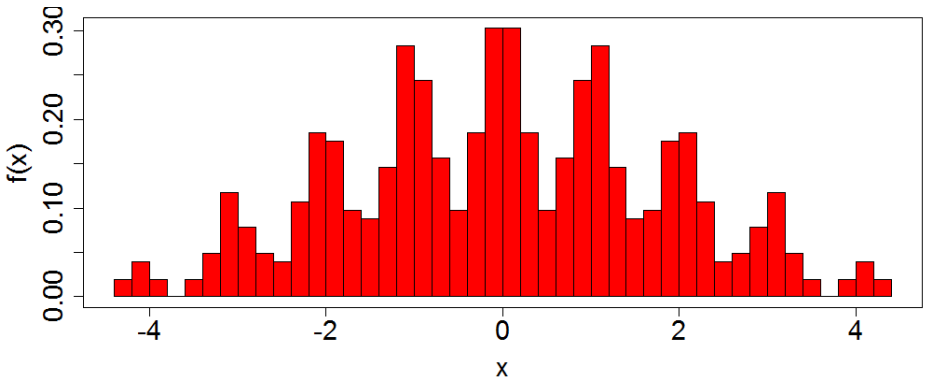

Рис. 6.14: Распределение порядковой статистики при использовании перестановочного критерия в задаче об эффективности транквилизатора

Критерий знаковых рангов давал достигаемый уровень значимости $p = 0.019$.

Нулевое распределение перестановочного критерия изображено на рисунке 6.14. Значение статистики, которое реализуется в эксперименте: $T = 3.887$. При суммировании высоты всех столбиков, начиная от 3.887 и направо, получается достигаемый уровень значимости $p = 0.0137$. Нулевая гипотеза отвергается в пользу односторонней альтернативы.

6.4.3. Перестановочный критерий для независимых выборок

Перестановочный критерий для независимых выборок выглядит абсолютно так же, как критерий Манна-Уитни за исключением того, что не производятся ранговые преобразования.

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$, $X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$,
нулевая гипотеза:	$H_0 : F_{X_1}(x) = F_{X_2}(x)$;
альтернатива:	$H_1 : F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$;
статистика:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$;
нулевое распределение:	порождается перебором $C_{n_1+n_2}^{n_1}$ размещений объединённой выборки.

Таблица 6.11: Описание двухвыборочного перестановочного критерия для несвязанных выборок

Нулевое распределение статистики этого критерия точно так же, как и для критерия Манна-Уитни, получается перебором всех $C_{n_1+n_2}^{n_1}$ размещений объединенной выборки по выборкам $X_1^{n_1}$ и $X_2^{n_2}$.

В задаче об анализе связи между кофеином и респираторным обменом (рисунок 6.8) проверялась нулевая гипотеза H_0 : среднее значение показателей респираторного обмена не отличается в двух группах пациентов (в одной пациенты принимали кофеин, в другой — плацебо) — против двусторонней альтернативы H_1 : что-то изменилось.

Критерий Манна-Уитни давал достигаемый уровень значимости $p = 0.0521$. На рисунке 6.15 показано нулевое распределение перестановочного критерия. Значение статистики, которое реализуется в эксперименте: $T = 6.33$, оно соответствует достигаемому уровню значимости $p = 0.0578$. Нулевая гипотеза все еще не отвергается.

6.4.4. Особенности перестановочных критериев

У перестановочных критериев есть некоторые особенности, о которых очень важно помнить.

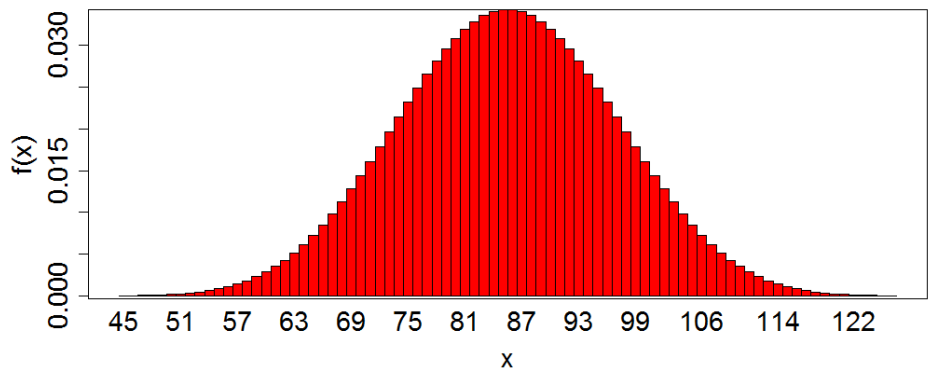

Рис. 6.15: Распределение нулевой статистики перестановочного критерия, полученной из данных эксперимента о связи между кофеином и респираторным обменом

Статистику для перестановочных критериев можно выбирать по-разному. В некоторых случаях это приводит к одному и тому же достигаемому уровню значимости, то есть ни на что не влияет. Например, в одновыборочной задаче, проверяя гипотезу о равенстве нулю математического ожидания

$$H_0 \colon \mathbb{E}X = 0, \; H_1 \colon \mathbb{E}X \neq 0,$$

в качестве статистики перестановочного критерия можно использовать как сумму элементов выборки, так и выборочное среднее:

$$T_1\left(X^n\right) = \sum_{i=1}^{n} X_i \quad \sim \quad T_2\left(X^n\right) = \bar{X}.$$

Нулевые распределения этих статистик будут отличаться только сдвигом и масштабом, поэтому достигаемый уровень значимости, посчитанный по ним, будет одним и тем же.

В других случаях, по-разному выбирая статистику для перестановочного критерия, можно получать разные достигаемые уровни значимости. Например, для статистик

$$T_2\left(X^n\right) = \bar{X} \quad \nsim \quad T_3\left(X^n\right) = \frac{\bar{X}}{S/\sqrt{n}}$$

нулевые распределения отличаются не только сдвигом и масштабом, поэтому достигаемый уровень значимости у критериев с этими двумя вариантами статистик тоже будет разный. Поэтому при выборе статистики для перестановочного критерия важно думать о том, какие из свойств исходной случайной величины для наиболее важны.

Перестановочные критерии придумал Рональд Фишер еще в начале XX века, однако их начали активно использовать только с появлением и широким распространением компьютеров, потому что для вычисления нулевых распределений этих критериев используются перестановки. В отличие от ранговых критериев, нормальных аппроксимаций для нулевого распределения в случае больших выборок не существует, поэтому единственный способ оценить нулевое распределение статистики — это перебрать много перестановок. Поэтому точно посчитать достигаемый уровень значимости перестановочного критерия на больших выборках достаточно сложно. Однако его можно посчитать приближённо. Для этого нужно взять какое-то случайное подмножество G' множества всех возможных перестановок G. При этом стандартное отклонение достигаемого уровня значимости будет примерно равно $\sqrt{\frac{p(1-p)}{|G'|}}$. На практике, чтобы получить хорошую аппроксимацию достигаемого уровня значимости, достаточно взять несколько тысяч перестановок.

6.5. Перестановки и бутстреп

Доверительные интервалы для параметров тесно связаны с проверкой точечных гипотез об их значениях. Например, z-критерии для средних связаны с нормальными доверительными интервалами, критерии Стьюдента соответствуют доверительным интервалам, построенным с использованием распределения Стьюдента. Для перестановочных критериев ближайшим аналогом в мире доверительных интервалов является метод бутстрепа, однако отношения между ними не такие взаимнооднозначные.

6.5.1. Проверка гипотез с помощью перестановочных критериев и метода бутстрепа

Перестановочные критерии принимают на вход выборку (или выборки), считают на них какую-то статистику. Далее делается дополнительное предположение о распределении, из которого эти выборки взяты. Это предположение порождает множество перестановок исходных данных, которые могли реализоваться с одинаковой вероятностью, если нулевая гипотеза справедлива. На этих перестановках вычисляется значение статистики и таким образом оценивается ее нулевое распределение.

Бутстреп-методы работают в каком-то смысле похоже. На вход они также принимают выборку или выборки и считают значение статистики, которая оценивает интересующий параметр. Далее на основании исходных данных генерируется множество бутстреп-псевдовыборок, и на этих псевдовыборках вычисляются значения интересующей статистики, то есть оценивается ее распределение.

Ключевых различий между этими методами несколько. Во-первых, перестановочный критерий использует дополнительное предположение, которое позволяет породить множество перестановок, которые используются для построения распределения. Во-вторых, перестановки, используемые в перестановочном критерии, — это выборки без возвращения, в то время как бутстреп-методы используют выборки с возвращением (бутстреп-псевдовыборки могут содержать по несколько копий элементов исходной выборки). Кроме того, распределения, которые получаются при использовании этих методов, абсолютно разные, потому что распределение статистики перестановочного критерия — это то распределение, которое статистика будет иметь при справедливости нулевой гипотезы, в то время как распределение бутстреп-статистики не подразумевает никакой нулевой гипотезы.

Чтобы лучше это понять, можно вспомнить пример с кофеином и респираторным обменом. В этой задаче проверялась гипотеза H_0: среднее значение показателя респираторного обмена не отличается в двух группах. Эта нулевая гипотеза проверялась против односторонней альтернативы H_1: под воздействием кофеина среднее значение показателя респираторного обмена снижается.

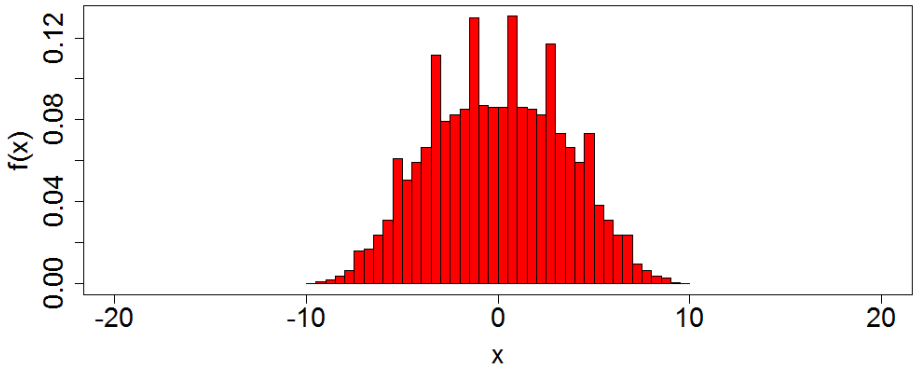

Рис. 6.16: Распределение нулевой статистики перестановочного критерия в задаче о респираторном обмене

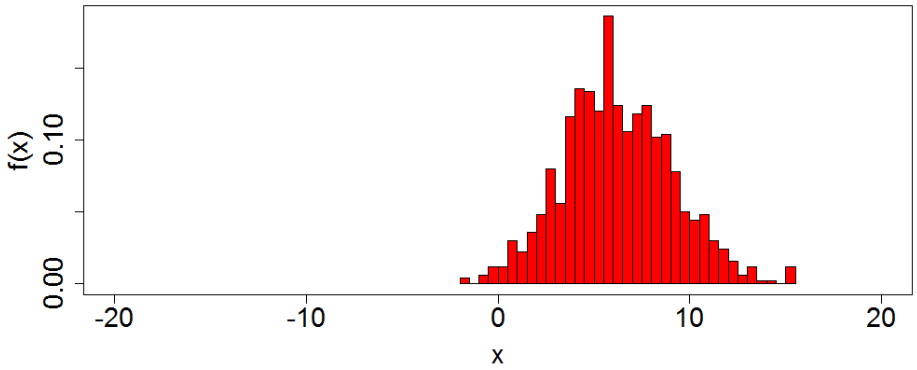

Рис. 6.17: Данные о респираторном обмене исследуемых после принятия кофеина или плацебо

На анализируемых данных разность выборочных средних значений показателей респираторного обмена у

пациентов, которые приняли плацебо и которые приняли кофеин, составляет $\bar{X}_{1n} - \bar{X}_{2n} = 6.33$.

На рисунке 6.16 показано нулевое распределение перестановочного критерия со статистикой $\bar{X}_{1n} - \bar{X}_{2n}$. На рисунке 6.17 — бутстреп-распределение той же самой статистики $\bar{X}_{1n} - \bar{X}_{2n}$. Ключевое различие между этими двумя распределениями в том, что они центрированы в разных местах. Перестановочное нулевое распределение центрировано в нуле — значении, соответствующем нулевой гипотезе. Бутстреп-распределение, в свою очередь, центрировано в выборочном среднем значений параметра. Параметр в данном случае — это разность средних, то есть центр бутстреп-распределения — это 6.33.

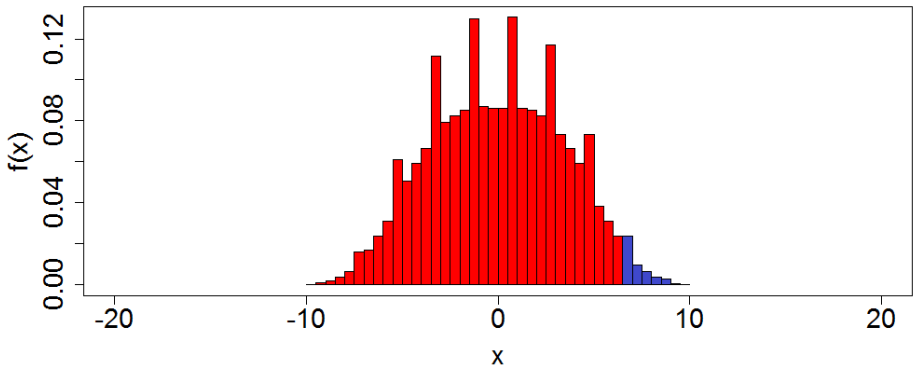

Рис. 6.18: Распределение статистики перестановочного критерия, синим показана доля перестановок, на которых среднее больше либо равно 6.33

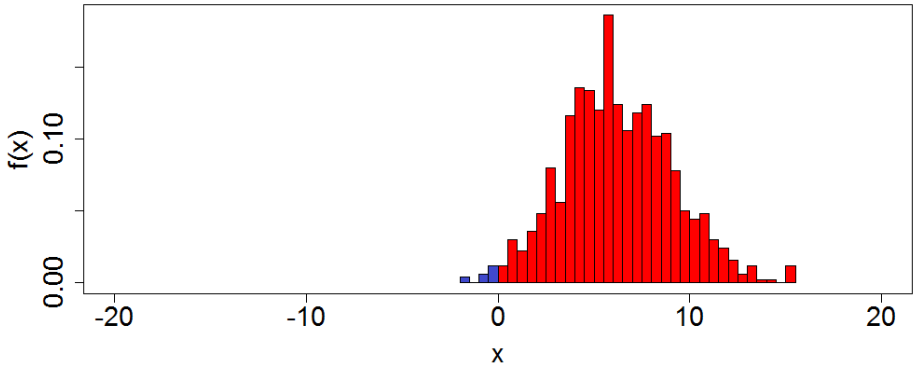

Рис. 6.19: Бутстреп-распределение статистики, синим показана доля псевдовыборок, на которых среднее меньше либо равно 0

Для перестановочного критерия доля перестановок, на которых среднее больше либо равно 6.33 (выборочное среднее, которое реализовано в данных) составляет примерно 0.03 (рисунок 6.18). Это и есть достигаемый уровень значимости перестановочного критерия, и он точный.

На бутстреп-распределении той же самой статистки доля псевдовыборок, на которых среднее меньше либо равно нулю, составляет 0.011 (рисунок 6.19). Эту величину можно считать приближенным достигаемым уровнем значимости бутстреп-критерия. То есть с помощью доверительных интервалов на основе бутстрепа тоже можно проверять гипотезы, однако нужно это делать немного иначе.

6.5.2. Различия перестановочного критерия и бутстрепа

Перестановочный критерий с помощью нулевого распределения статистики измеряет расстояние от 0 до $\bar{D}_n$ — значения параметра, реализовавшегося в эксперименте. Бутстреп-критерий измеряет, наоборот, расстояние от $\bar{D}_n$ до 0.

Перестановочный критерий является более точным, потому что в нем перебирается подмножество всех возможных перестановок данных, которые равновероятны при справедливости нулевой гипотезы. Бутстреп-критерий в описанном выше виде является только приближенным, поскольку в нем всегда перебирается

только конечное подмножество всех возможных бутстреп-псевдовыборок, потому что их заведомо слишком много.

Самое важное различие между этими двумя критериями заключается в том, что они проверяют разные гипотезы, поскольку перестановочный критерий использует дополнительные предположения. Перестановочный критерий проверяет гипотезу полного равенства распределений в двух выборках

$$H_0 \colon F_{X_1}(x) = F_{X_2}(x),$$

и проверяется она против альтернативы сдвига (в этом примере односторонней):

$$H_1 \colon F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta > 0$$

Предполагается, что никак иначе, кроме как сдвигом, эти распределения отличаться не могут.

Бутстреп-критерий проверяет всего лишь гипотезу о равенстве математических ожиданий:

$$H_0 \colon \mathbb{E}X_1 = \mathbb{E}X_2.$$

Гипотезу равенства он проверяет против односторонней альтернативы точно так же, как и перестановочный критерий:

$$H_1 \colon \mathbb{E}X_1 > \mathbb{E}X_2.$$

Однако эта гипотеза заведомо более общая: не используется предположение равенстве функций распределения в двух выборках.

6.5.3. Резюме

Проверка гипотез с помощью бутстрепа — это достаточно сложная задача. Показанный небольшой двухвыборочный пример позволяет понять плюсы и минусы бутстрепа и перестановочных критериев.

Бутстреп тоже позволяет проверять гипотезы, причём гораздо более широкого класса. Поскольку этот метод не использует дополнительных предположений, можно проверять только интересующий параметр (в примере это была разность математических ожиданий). Кроме того, с помощью бутстрепа можно проверять и другие крайне экзотические гипотезы, которые никакими другими методами проверить нельзя. Например, гипотеза о том, что распределение имеет ровно две моды.

Если предположения, лежащие в основе перестановочного критерия, выполняются, то перестановочный критерий, во-первых, точнее, поскольку его достигаемый уровень значимости точный, во-вторых, всегда мощнее, чем аналогичный критерий бутстрепа. Но бутстреп при этом гораздо более гибок, потому что его можно использовать в ситуациях, когда не выполняются предположения, использующиеся в перестановочном критерии.