Прикладная статистика

Слайды к лекции 4

7 февраля 2023

Сергей Александрович Спирин

sspirin@hse.ru

Чай с молоком

Как лучше?

Из книги:

R.A. Fisher. *The design of Experiments.*

Edinburgh: Oliver and Boyd, 1942

«Некая леди заявляет, что, попробовав чашку чая с молоком, она может определить, что было сначала налито в чашку — молоко или чай. Характерно, что леди не претендует на то, что она может безошибочно определить разницу во вкусе, но утверждает, что, пусть иногда ошибаясь, она чаще отвечает верно, чем неверно.»

Фишер предложил приготовить 8 чашек чая с молоком, из них в четыре сначала налить чай, потом молоко, а в другие четыре — наоборот, и предложить леди попробовать чай и указать, в каких четырёх чашках чай приготовлен правильно.

Вопросы:

(1) Возможно ли в принципе доказать на уровне достоверности 5%, что леди действительно может отличить один способ от другого по вкусу?

(2) Сколько раз из четырёх леди может ошибиться, чтобы вывод всё же можно было сделать?

https://en.wikipedia.org/wiki/Lady_tasting_tea

Проверка гипотез
(*напоминание*)

Имеются наблюдения:

– выборка из генеральной совокупности *или*

– результаты нескольких экспериментов

Имеется нулевая гипотеза (H_0).

Как правило, она предполагает отсутствие эффекта, новизны, вины подсудимого и т.п.

Но возможна ситуация, когда она состоит, наоборот, в наличии болезни, дефекта и т.п.

В любом случае нулевая гипотеза — та, которую опаснее ошибочно отвергнуть.

Имеется альтернативная гипотеза (H_1 или H_A).

Она может быть просто отрицанием нулевой гипотезы или быть более детализованной.

Выбирается критерий (он же тест): алгоритм, позволяющий на основании наблюдений сделать выбор из двух гипотез.

Проверка гипотез
(*продолжение*)

Ошибка первого рода (**Type I error**): принятие альтернативной гипотезы, в то время как верна нулевая.

Ошибка второго рода (**Type II error**): принятие нулевой гипотезы, в то время как верна альтернативная.

Всегда стараются минимизировать вероятность ошибки первого рода.

Проверка гипотез
(*продолжение*)

Условная вероятность отклонения критерием нулевой гипотезы при условии её истинности, называется уровнем значимости (**significance level**) критерия.

Уровень значимости (или просто значимость, significance) обычно обозначается через α. Чем меньше α, тем лучше. Приличным считается уровень значимости не более 1/20.

Условная вероятность отклонения нулевой гипотезы при условии истинности альтернативной называется мощностью (**power**) критерия. При прочих равных чем мощнее критерий, тем лучше.
На практике оценить мощность бывает трудно или даже невозможно.

Проверка гипотез
(*продолжение*)

Практически все критерии устроены так:

- выбирается **статистика**, то есть числовая функция от наблюдений;
- выбирается **критическое множество** (для значений выбранной статистики);
- нулевая гипотеза отвергается, если статистика попадает в критическое множество.

Тем самым значимость равна вероятности для статистики попасть в критическое множество при условии H_0, а мощность — вероятности того же события при условии H_1.

Проверка гипотез
(*продолжение*)

Основой используемых критериев являются результаты математической статистики (раздел математики).

Наблюдения рассматриваются как реализации одинаково распределённых независимых случайных величин.

При таком подходе H_0 и H_1 превращаются в утверждения о законе распределения этих случайных величин.

Примеры H_0 и H_1:

- H_0: величины распределены нормально со средним 0 и дисперсией 1,
H_1 : это не так;
- H_0: величины распределены нормально (но параметры распределения неизвестны),
H_1: это не так;
- H_0: математическое ожидание величин равно 0,
H_1: матожидание отлично от 0 (двусторонняя альтернатива);
- H_0: математическое ожидание величин равно 0,
H_1: матожидание больше 0 (односторонняя альтернатива)

Проверка гипотез
(*продолжение*)

Очень часто статистика (в смысле функции от наблюдений) пересчитывается в так называемое p-значение (**P-value**).

Оно равно вероятности для статистики принять такое же или «более необычное с точки зрения H_0» значение.

(*Что такое «более необычное», зависит от H_1*).

Поскольку p-значение — это функция от статистики, а статистика — функция от наблюдений, само p — тоже функция от наблюдений и тем самым тоже статистика!

Основное свойство p-значения: при справедливости H_0 p распределено равномерно на отрезке [0,1]

Проверка гипотез
(*окончание*)

Основное свойство p-значения: при справедливости H_0 p распределено равномерно на отрезке [0,1]

Если статистика пересчитана в p, то критическим множеством, соответствующим значимости α, является отрезок [0,α]

Пример 1

Рыболовы Артур и Борис поспорили, кто из них круче. В течение недели они ловили рыбу каждый день одинаковое время. Артур поймал 43 рыбёшки, а Борис 28. Можно ли уверенно заявить, что метод Артура лучше?

Пример 1

Рыболовы Артур и Борис поспорили, кто из них круче. В течение недели они ловили рыбу каждый день одинаковое время. Артур поймал 43 рыбёшки, а Борис 28. Можно ли уверенно заявить, что метод Артура лучше?

Решение.
Сначала нужно определиться с уровнем уверенности. Пусть это будет стандартная 1/20.

Нулевая гипотеза: в среднем оба рыболова ловят одинаково.

Альтернатива: один из них ловит лучше (двусторонняя! Мы заранее не знали, кто победит).

Статистика: разность числа пойманных рыб, она равна 15.

Как она распределена при нулевой гипотезе?

Таблица стандартного нормального распределения

	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
-5	$2{,}9\cdot10^{-7}$	$4{,}8\cdot10^{-7}$	$7{,}9\cdot10^{-7}$	$1{,}3\cdot10^{-6}$	$2{,}1\cdot10^{-6}$	$3{,}4\cdot10^{-6}$	$5{,}4\cdot10^{-6}$	$8{,}5\cdot10^{-6}$	$1{,}3\cdot10^{-5}$	$2{,}1\cdot10^{-5}$
-4	$3{,}2\cdot10^{-5}$	$4{,}8\cdot10^{-5}$	$7{,}2\cdot10^{-5}$	$1{,}1\cdot10^{-4}$	$1{,}6\cdot10^{-4}$	$2{,}3\cdot10^{-4}$	$3{,}4\cdot10^{-4}$	$4{,}8\cdot10^{-4}$	$6{,}9\cdot10^{-4}$	$9{,}7\cdot10^{-4}$
-3	0,0013	0,0019	0,0026	0,0035	0,0047	0,0062	0,0082	0,011	0,014	0,018
-2	0,023	0,029	0,036	0,045	0,055	0,067	0,081	0,097	0,12	0,14
-1	0,16	0,18	0,21	0,24	0,27	0,31	0,34	0,38	0,42	0,46

$F(-1{,}96) \approx 0{,}025$
$F(-1{,}65) \approx 0{,}05$

Пример 1

Рыболовы Артур и Борис поспорили, кто из них круче. В течение недели они ловили рыбу каждый день одинаковое время. Артур поймал 43 рыбёшки, а Борис 28. Можно ли уверенно заявить, что метод Артура лучше?

Решение.

Если среднее число пойманных за неделю рыб равно μ, то реальное число распределено по Пуассону со средним μ.

Можно считать, что $\mu = 35{,}5$ (среднее между двумя наблюдениями).

При таком среднем распределение Пуассона практически не отличается от нормального распределения со средним 35,5 и дисперсией тоже 35,5
(у распределения Пуассона матожидание всегда равно дисперсии).

Разность двух независимых и так распределённых величин распределена нормально со средним 0 и дисперсией 71 (то есть с $\sigma = 8{,}43$).

Пересчитываем нашу статистику в Z-score = $15/8{,}43 = 1{,}78 < 1{,}96$
(по таблице это соответствует вероятности чуть больше 0,07)

Нулевая гипотеза **не отклоняется**
(уверенности в преимуществе Артура нет).

Почему в примере 1 альтернатива двусторонняя?

Примите нулевую гипотезу и посмотрите на распределение p-значения.

Если всё правильно, оно должно быть равномерным на отрезке [0,1]

Если бы мы рассматривали одностороннюю альтернативу, p не могло бы принять значения, большие ½, значит это было бы неправильно.

Пример 1

Рыболовы Артур и Борис поспорили, кто из них круче. В течение недели они ловили рыбу каждый день одинаковое время. Артур поймал 43 рыбёшки, а Борис 28. Можно ли уверенно заявить, что метод Артура лучше?

Другое решение.
Рассмотрим всех пойманных рыб (их 71 штука) и примем за нулевую гипотезу, что каждая из них с равной вероятностью попадалась либо Артуру, либо Борису.

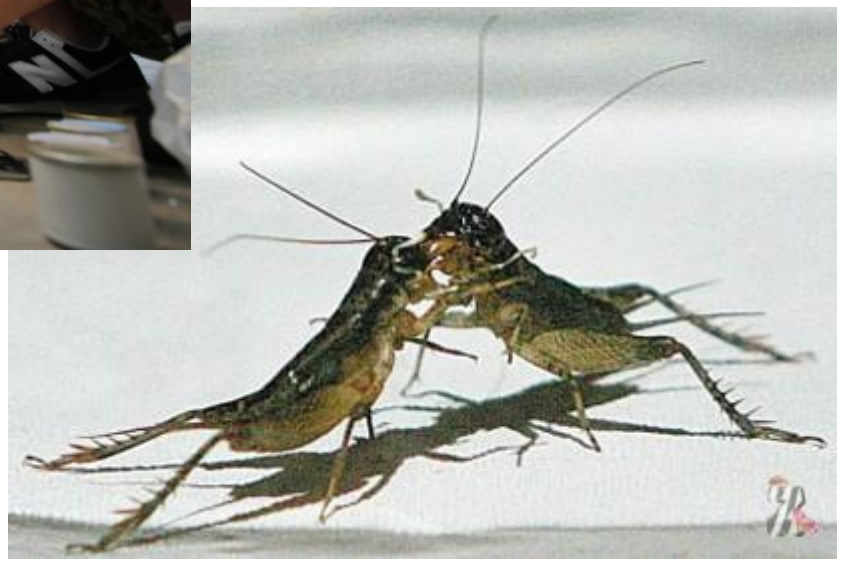
Тогда наше p-значение равно удвоенной (почему?) вероятности иметь не более 28 успехов при 71 испытании с вероятностью успеха ½.

Такая вероятность даётся биномиальным распределением. В данном случае матожидание числа успехов равно 35,5, а дисперсия равна
$n \cdot p \cdot (1-p) = 71 \cdot ½ \cdot ½ = 17{,}75$.

Биномиальное распределение при таких значениях хорошо приближается нормальным. Среднее равно 35,5 и сигма равна корню из 17,75 = 4,21 , откуда $Z = (28-35{,}5)/4{,}21 = 1{,}78$

Тот же результат! ☺

Бои сверчков



Пример 2

Некоторые любители боёв сверчков добавляют в рацион боевых сверчков женьшень. Для проверки эффективности этого приёма устроили 12 боёв, в каждом из которых одного из участников кормили женьшенем, а другого — нет. Сверчки, которых кормили женьшенем, победили в шести боях и потерпели поражение в трёх, а в остальных случаях победителя выявить не удалось. Можно ли утверждать, что женьшень способствует победам?

Пример 2

Некоторые любители боёв сверчков добавляют в рацион боевых сверчков женьшень. Для проверки эффективности этого приёма устроили 12 боёв, в каждом из которых одного из участников кормили женьшенем, а другого — нет. Сверчки, которых кормили женьшенем, победили в шести боях и потерпели поражение в трёх, а в остальных случаях победителя выявить не удалось. Можно ли утверждать, что женьшень способствует победам?

Решение.

Нулевая гипотеза: вероятность победы равна вероятности поражения (независимо от женьшеня).

P-значение равно вероятности получить шесть или более успехов при девяти (почему?) испытаниях с вероятностью успеха ½.

Такая вероятность равна ¼.

Поэтому на разумном уровне значимости отклонить нулевую гипотезу нельзя.

Примеры из прошлой лекции

1. При посеве на питательную среду проб воды из водоёма А выросло 49 колоний бактерий, а при посеве (такого же количества и объёма) проб из водоёма Б — 32 колонии. На основании этих данных было сделано утверждение, что в первом водоёме загрязнённость бактериями выше. Как посчитать P-value этого утверждения?

2. Возникло предположение, что очистные сооружения некоторого небольшого города не полностью очищают воду от бактерий. При посеве пробы воды, взятой выше города, выросло 23 колонии, а пробы, взятой ниже города — 41 колония. Как посчитать уровень значимости, на котором данные результаты подтверждают предположение?

Пример 3

Двадцати испытуемым дали препарат, призванный улучшить реакцию.
У каждого испытуемого измеряли время реакции на звуковой сигнал до приёма препарата и через 3 ч. после него. Выяснилось, что в среднем время реакции уменьшилось на 0,07 сек. Достаточно ли этих данных, чтобы сделать заключение об эффективности препарата?

Пример 3

Двадцати испытуемым дали препарат, призванный улучшить реакцию.
У каждого испытуемого измеряли время реакции на звуковой сигнал до приёма препарата и через 3 ч. после него. Выяснилось, что в среднем время реакции уменьшилось на 0,07 сек. Достаточно ли этих данных, чтобы сделать заключение об эффективности препарата?

Ответ: этих данных недостаточно.

Пример 3

Двадцати испытуемым дали препарат, призванный улучшить реакцию.
У каждого испытуемого измеряли время реакции на звуковой сигнал до приёма препарата и через 3 ч. после него. Выяснилось, что в среднем время реакции уменьшилось на 0,07 сек. Достаточно ли этих данных, чтобы сделать заключение об эффективности препарата?

Ответ: этих данных недостаточно.

Для ответа на вопрос нужно выяснить некоторые детали. Например:

- У скольких испытуемых время реакции увеличилось, а у скольких — уменьшилось?
- Какова дисперсия изменения времени реакции?
- Если расположить **абсолютные** значения изменений по возрастанию, то какие места в этом ряду займут положительные изменения?

Парные критерии

Имеется n парных наблюдений: $x_1, \ldots, x_n, y_1, \ldots, y_n$
Обычно речь идёт об n независимых объектах/экспериментах, а x_i и y_i — измерения одной и той же характеристики i-го объекта в разных условиях (например, до и после какого-то воздействия).

Нужно проверить, влияет ли изменение условий на данную характеристику.

Парные критерии

Имеется n парных наблюдений: $x_1, \ldots, x_n, y_1, \ldots, y_n$
Обычно речь идёт об n независимых объектах/экспериментах, а x_i и y_i — измерения одной и той же характеристики i-го объекта в разных условиях (например, до и после какого-то воздействия).

Нулевая гипотеза: нет систематической разницы между x и y.
Альтернативная гипотеза: есть систематическая разница.

Обычное уточнение: заранее предполагается, что каждое значение y_i отличается от значения x_i сдвигом на случайную величину r.

Тогда нулевая гипотеза может быть: Er = 0 или медиана r равна 0 или даже что r распределена симметрично относительно 0.
Альтернативная гипотеза: всё то же, с заменой 0 на неизвестное ненулевое число (двусторонняя) или же на положительное число (односторонняя).

Парные критерии

Имеется n парных наблюдений: $x_1, \ldots, x_n, y_1, \ldots, y_n$

Можно рассмотреть разности $r_i = x_i - y_i$ и поставить вопрос о равенстве среднего или медианы разностей нулю, или даже о симметричности распределения разностей.

Это эквивалентная формулировка.

Критерий знаков (sign test)

Есть **пары** наблюдений (X_i, Y_i). Считаем, что у нас есть генеральная совокупность пар чисел.

Нулевая гипотеза — медиана разностей $X_i - Y_i$ равна 0 (эквивалентная формулировка: вероятность того, что $X > Y$, равна вероятности того, что $X < Y$).

Альтернативная гипотеза: медиана $X_i - Y_i$ меньше нуля (односторонняя) или медиана не равна 0 (двусторонняя).

Статистика: число случаев, когда $X_i < Y_i$.

P-value может быть посчитано, исходя из биномиального распределения.

Сводим анализ пары измерений к двум числам «успехов»

Парный критерий Уилкоксона

Пары наблюдений (X_i, Y_i).

Нулевая гипотеза — разности $X_i — Y_i$ распределены симметрично относительно нуля.

Альтернативная гипотеза содержательно состоит в том, что на Y действует некоторый фактор, приводящий к увеличению Y по сравнению с соответствующим X.
Точная формулировка H_1 : значения разностей $Y_i — X_i$ распределены симметрично относительно некоторого $M > 0$.
Двусторонний вариант H_1 : разности распределены симметрично относительно некоторого $M \neq 0$.

Упорядочим абсолютные значения разностей $|X_i — Y_i|$:

$$\{|X_1 - Y_1|, \ldots, |X_l - Y_l|\} = \{Z_1, \ldots, Z_l\}, \text{ причём } Z_1 < Z_2 < \ldots < Z_l.$$

Статистика W равна сумме **рангов** тех пар, для которых $X < Y$.

При больших l и нулевой гипотезе величина W распределена нормально со средним $l(l+1)/4$ и дисперсией $l(l+1)(2l+1)/24$.

Z-test

Пары наблюдений ($X_i,\ Y_i$).

Нулевая гипотеза — **математическое ожидание** (= среднее по ген. совокупности) разностей $\mathrm{E}R_i = \mathrm{E}X_i - \mathrm{E}Y_i$ равно нулю.

Альтернативная гипотеза: среднее меньше 0 (односторонняя) или среднее не равно 0 (двусторонняя).

Посчитаем среднее разностей по выборке: $\bar{R} = \bar{X} - \bar{Y} = \sum (X_i - Y_i)/n$

Если наблюдений много, то это среднее при нулевой гипотезе распределено нормально со средним 0

Оценим дисперсию разностей s^2 как среднее величин $(R_i - \bar{R})^2$

(лучше в качестве оценки брать не $\sum_i (R_i - \bar{R})^2/n$, а $\sum_i (R_i - \bar{R})^2/(n-1)$, хотя при больших n это не очень существенно)

Тогда дисперсия среднего при нулевой гипотезе равна s^2/n (почему?).

Квадратный корень из s^2/n называется **стандартной ошибкой** *SE*

Статистика: $Z = \bar{R}/SE$, при нулевой гипотезе имеет распределение N(0,1)

Можно применять только при n > 100 !!!

Критерий Стьюдента (t-test)

Применяется, если *n* невелико **и мы твёрдо уверены**, что разности *Y* – *X* распределены нормально.
Если такой уверенности нет, применять нельзя — используйте критерий Уилкоксона или знаков.

Статистика выглядит так же, как для Z-теста : $t = (\overline{X} - \overline{Y})/(s^2/n)^{1/2}$
Здесь уже существенно при подсчёте s^2 делить на (n – 1), а не на n.

Распределение этой статистики при малом *n* другое, это «распределение Стьюдента с *n* – 1 степенью свободы»

Чем больше *n* , тем сильнее распределение *t* похоже на распределение *Z* (то есть стандартное нормальное, N(0,1)).

Для малых *n* оно существенно отличается от нормального: чем больше *t* (и меньше *n*), тем сильнее вероятность получить такое или большее значение будет отличаться (в бóльшую сторону) от аналогичной вероятности для Z.

"Плотность t убывает медленнее, чем плотность Z"

"У распределения t тяжёлые хвосты"

Распределение Стьюдента

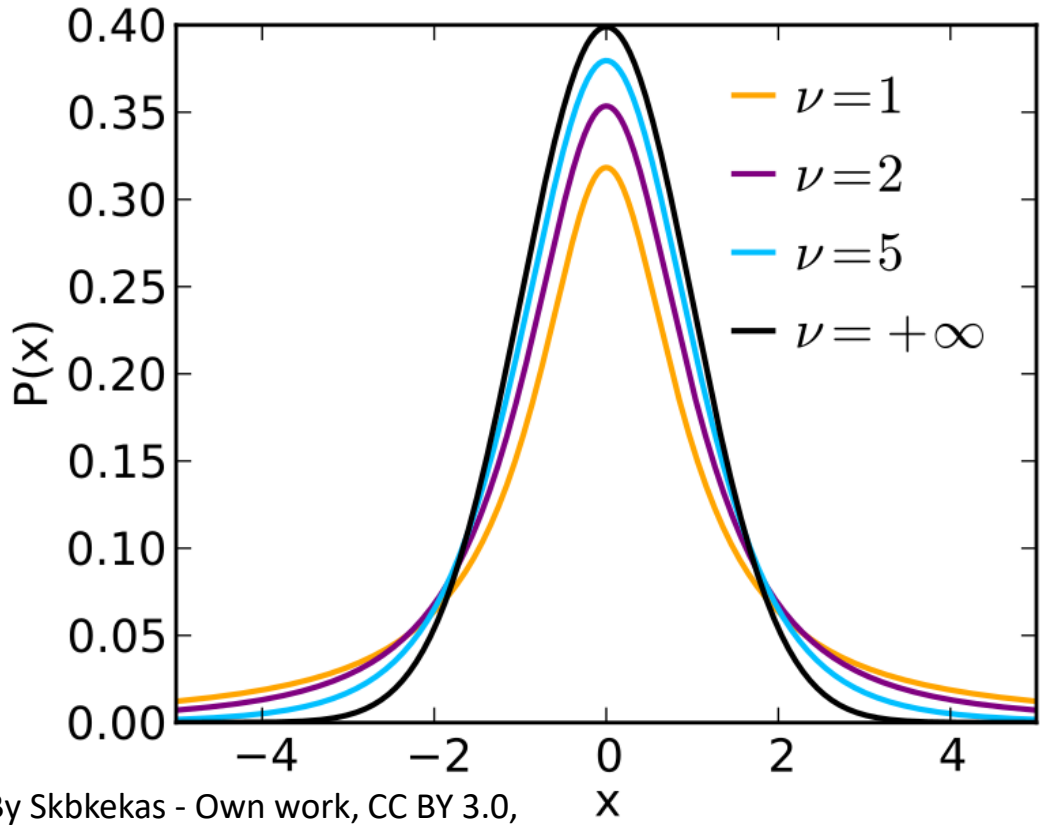
односторонний	**75%**	**80%**	**85%**	**90%**	**95%**	**97,5%**	**99%**	**99,5%**	**99,75%**	**99,9%**	**99,95%**
двусторонний	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99,5%**	**99,8%**	**99,9%**
1	1,000	1,376	1,963	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,080	1,386	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
…											
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Последняя строчка соответствует нормальному распределению

https://ru.wikipedia.org/wiki/Распределение_Стьюдента

Распределение Стьюдента

Плотность вероятности: $$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



Пример 4

Фирма испытывает новый крем для обуви,
Для этих испытаний разработан индекс изношенности верха обуви (положительное число).

Двоим испытателям поручили каждый день намазывать один ботинок новым кремом, а другой — старым. Через месяц значения индекса изношенности были такие:

Новый крем: 10,8 у первого, 9,5 у второго.
Старый крем: 15,4 у первого, 13,8 у второго.
Выводы?

Пример 4

Фирма испытывает новый крем для обуви,
Для этих испытаний разработан индекс изношенности верха обуви (положительное число).

Двоим испытателям поручили каждый день намазывать один ботинок новым кремом, а другой — старым, Через месяц значения индекса изношенности были такие:

Новый крем: 10,8 у первого, 9,5 у второго,
Старый крем: 15,4 у первого, 13,8 у второго.
Выводы?

$n = 2,\ n^{1/2} \approx 1{,}42$
$X_1 = 15{,}4 - 10{,}8 = 4{,}6$
$X_2 = 13{,}8 - 9{,}5 = 4{,}3$
$\bar{X} = 4{,}45$
$s^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = 2 \cdot 0{,}15^2 = 0{,}045 \Rightarrow s \approx 0{,}21$
$t = 4{,}45/(0{,}21/1{,}42) = 30{,}1 > 6{,}314$
Вывод: можно отклонить нулевую гипотезу. Новый крем действительно лучше.

Парные критерии

- Критерий знаков (sign test)
- Критерий Уилкоксона (Wilcoxon test)
- Z-критерий (Z-test) (только для большого числа наблюдений!)
- Критерий Стьюдента (t-test) (Осторожно! При малых *n* можно применять только при уверенности, что *y* – *x* распределено хотя бы приблизительно нормально!)

Критерий знаков можно применять всегда.

Для критерия Уилкоксона желательно, чтобы все значения разностей были различны.

Бывают ситуации, когда можно выбирать между критерием знаков, критерием Уилкоксона и Z-критерием, а бывают, когда между первыми двумя и критерием Стьюдента.

При этом эти критерии содержательно разные и дадут разные ответы.

При больших *n* (> 100) критерий Стьюдента превращается в Z-критерий

Критерий Стьюдента (t-test)

Выборка: n чисел $X_1, \ldots, X_n$.

Условие: генеральная совокупность имеет нормальное распределение!

Нулевая гипотеза: среднее значение генеральной совокупности равно некоторому заранее заданному числу μ.

(например, в случае разностей для парного теста $\mu = 0$).

Альтернативная гипотеза: среднее > μ (или среднее $\neq \mu$).

Дисперсия неизвестна!

Вычисляем среднее по выборке: $\overline{X} = \sum_i x_i / n$

Оцениваем дисперсию по формуле $s^2 = \sum_i (X_i - \overline{X})^2/(n-1)$

Статистика

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

При нулевой гипотезе (независимо от дисперсии) t подчиняется **распределению Стьюдента с $n-1$ степенью свободы**.

Пример 5

В статье «Использование марихуаны и гнев» (Journal of Psychology, 1988, т. 122, стр. 33) Сью Стоунер сообщил, что в выборке из 17 потребителей марихуаны среднее и стандартное отклонение по шкале выражения гнева – 42,72 и 6,05 соответственно. Проверьте, значимо ли отличие от неупотребляющих (среднее 41,6). Какие предположения необходимы, чтобы для указанных условий испытание считалось действительным?