Прикладная статистика

Слайды к лекции 5

16 февраля 2023

Сергей Александрович Спирин

sspirin@hse.ru

Парадокс двух выборок

Молодые пациенты

	Лечили	Не лечили
Осложнения	6	20
Нет осложнений	14	30

Итого без лечения 20/50 = 40% осложнений, а с лечением 6/20 = 30%
Помогает!

Старые пациенты

	Лечили	Не лечили
Осложнения	36	12
Нет осложнений	24	4

Итого без лечения 12/16 = 75% осложнений, а с лечением 36/60 = 60%
Старым тоже помогает!

А если не различать молодых и старых и объединить данные?

Парадокс двух выборок

Молодые пациенты

	Лечили	Не лечили
Осложнения	6	20
Нет осложнений	14	30

Старые пациенты

	Лечили	Не лечили
Осложнения	36	12
Нет осложнений	24	4

Все пациенты

	Лечили	Не лечили
Осложнения	42	32
Нет осложнений	38	34

Без лечения 32/66 < 50% осложнений, а с лечением 42/80 > 50%

???

Таблица сопряжённости

Испытывается новая методика лечения овец, заражённых некоторым заболеванием. Из 50 овец, которых лечили старым методом, умерли 25, а из 60, леченных новым методом — 20. Как посчитать P-value утверждения, что новый метод действительно лучше?

Составим таблицу

	Старый	Новый
Умерло	25	20
Выжило	25	40

Это называется «таблица сопряжённости 2×2»

Таблица сопряжённости

	Старый	Новый
Умерло	25	20
Выжило	25	40

Общий случай

n_{11}	n_{12}
n_{21}	n_{22}

Нулевая гипотеза — строки и столбцы независимы

Для проверки составляется «таблица ожидаемых значений».

Ожидаемое значение на пересечении строки и столбца равно доле строки, умноженной на долю столбца и умноженной на общее количество примеров.

В нашем случае общее количество N =110.

Доля первой строки $p_1 = (n_{11} + n_{12})/N = 45/110$

Доля второй строки $p_2 = (n_{21} + n_{22})/N = 65/110$

Доля первого столбца $q_1 = (n_{11} + n_{21})/N = 50/110$

Доля второго столбца $q_2 = (n_{12} + n_{22})/N = 60/110$

Ожидаемое количество в верхней левой ячейке

$$E_{11} = N\, p_1 q_1 = N \cdot (n_{11} + n_{12})/N \cdot (n_{11} + n_{21})/N = (n_{11} + n_{12}) \cdot (n_{11} + n_{21})/N$$

...

Таблица сопряжённости

Наблюдаемое

n_{11}	n_{12}
n_{21}	n_{22}

Ожидаемое

E_{11}	E_{12}
E_{21}	E_{22}

	Старый	Новый
Умерло	25	20
Выжило	25	40

	Старый	Новый
Умерло	20,45	24,55
Выжило	29,55	35,45

$E_{11} = (n_{11} + n_{12})\cdot(n_{11} + n_{21})/N$, E_{12} = (аналогично), ...

Статистика $\chi = \sum_{ij} (n_{ij} - E_{ij})^2/E_{ij}$

Такая статистика при нулевой гипотезе (= независимость строк от столбцов) распределена по закону «хи-квадрат с одной степенью свободы»

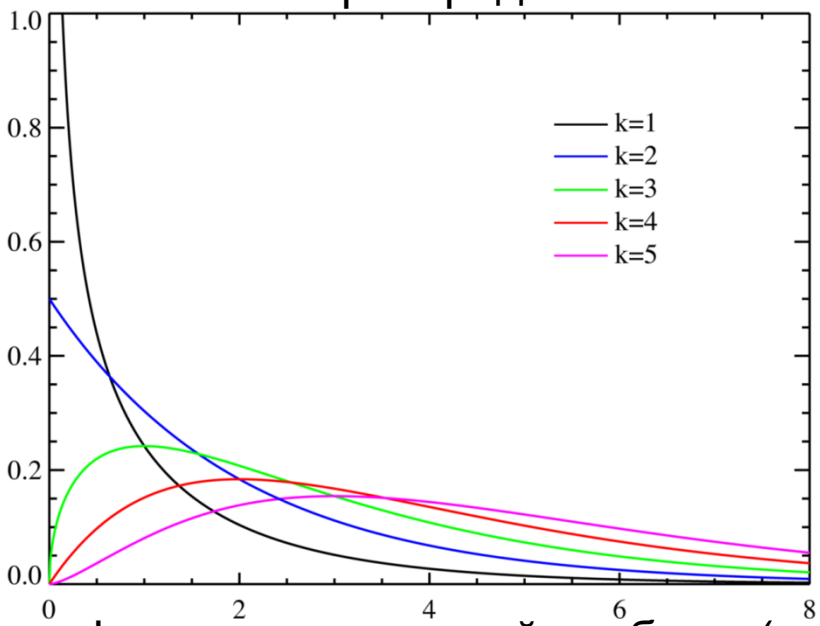
В нашем случае получается $\chi = 3{,}13$

Распределение хи-квадрат

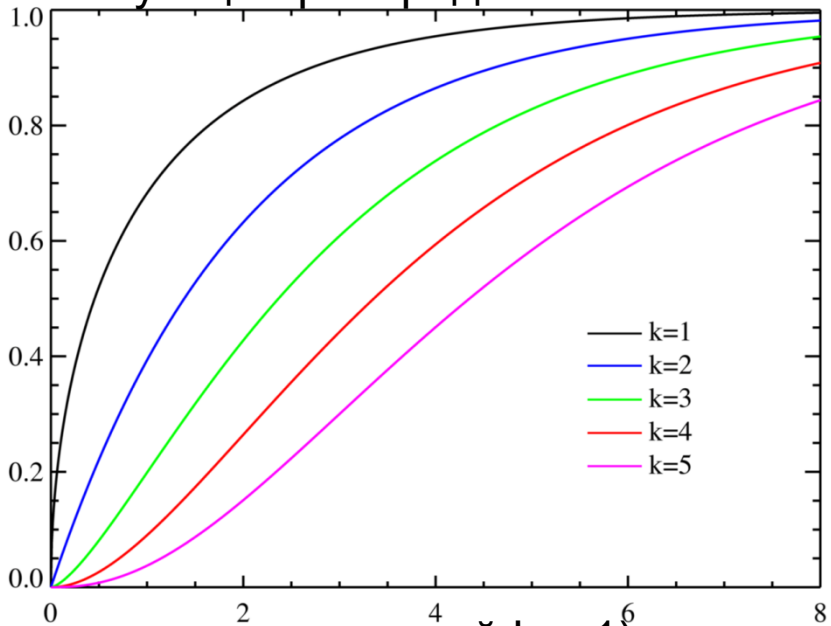
Если величина ξ распределена по стандартному нормальному закону, то величина ξ^2 распределена по закону «хи-квадрат с одной степенью свободы»

Это значит, что $P(\xi^2 > x) = 2F(-x)$, где F — функция распределения стандартного нормального закона N(0,1)

Плотность распределения



Функция распределения



k — число степеней свободы (нас пока интересует случай k = 1)

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

Распределение хи-квадрат

$$\chi_n^2 = \xi_1^2 + \ldots + \xi_n^2$$

где $\xi_1, \ldots, \xi_n$ – независимые нормально распределённые с.в. со средним 0 и дисперсией 1

Table C χ^2 **critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42

При χ = 3,13 из таблицы получаем, что p между 0,1 и 0,05

Можно посчитать, что p = 0,077

Распределение хи-квадрат

$$\chi_n^2 = \xi_1^2 + \ldots + \xi_n^2$$

где $\xi_1, \ldots, \xi_n$ – независимые нормально распределённые с.в. со средним 0 и дисперсией 1

Table C χ^2 **critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42

При χ = 3,13 из таблицы получаем, что p между 0,1 и 0,05

Можно посчитать, что p = 0,077

Критерий сопряжённости хи-квадрат всегда двусторонний!

В нашем случае можно поделить p пополам и получить p для одностороннего варианта, равное 0,038 — такое же, что получилось другим методом

А если число испытаний (и/или успехов) мало?

Выборку подростков можно разделить по полу, с одной стороны, и на тех, которые и в настоящее время сидят или нет на диете, с другой. Мы предполагаем, что доля лиц на диете выше среди девушек, чем среди юношей, и хотим проверить, значительна ли разница долей, которые мы наблюдаем.

	Boys	Girls	total
dieting	1	9	10
not dieting	11	3	14
total	12	12	24

(оценка вероятности при нулевой гипотезе будет слишком неточной)

Точный тест Фишера

	men	women	total
dieting	1	9	10
not dieting	11	3	14
totals	12	12	24

	men	women	total
dieting	*a*	*b*	*a* + *b*
not dieting	*c*	*d*	*c* + *d*
totals	*a* + *c*	*b* + *d*	*n*

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Гипергеометрическое распределение

Эта формула даёт вероятность данной таблицы
при условии данных сумм по строкам и столбцам

Точный тест Фишера

	men	women	total
dieting	1	9	10
not dieting	11	3	14
totals	12	12	24

	men	women	total
dieting	*a*	*b*	*a* + *b*
not dieting	*c*	*d*	*c* + *d*
totals	*a* + *c*	*b* + *d*	*n*

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Чтобы получить p-значение, надо найти наименьшее значение в таблице.

Затем посчитать по приведённой формуле вероятность такой таблицы и всех таблиц, получаемых из данной уменьшением этого минимального значения при сохранении сумм по строкам и столбцам. В данном случае величине *a* надо придать значения 1 (как в исходной таблице) и 0 (в этом случае *b*=10, *c* = 12, *d* =2).

Наконец, надо сложить эти вероятности, получится p-значение.

Сравнение долей при малых числах

Сравнивают результаты двух методов лечения А и В при редком заболевании. Из пяти больных, леченных методом А, выздоровел один, а вот все четверо леченных методом В выздоровели. Что можно сказать об относительной эффективности методов лечения?

Критерии независимости

Для каждого объекта (в каждом эксперименте) измеряется по два свойства

- H_0: Свойства независимы
- H_1: Свойства зависимы

Пример

Владелец ресторана провел исследования случайной выборки из 385 клиентов, чтобы определить есть ли связь удовлетворенности клиентов с полом и возрастом.

	Молодой муж.	Молодая жен.	Пожилой муж.	Пожилая жен.
Удовл.	25	30	135	112
Не удовл.	8	16	22	37

Предположение о независимости

	Молодой муж.	Молодая жен.	Пожилой муж.	Пожилая жен.	**TOTAL**
Удовл.	25	30	135	112	302
Не удовл.	8	16	22	37	83
TOTAL	33	46	157	149	385

Если пол / возраст не влияют на удовлетворенность, то
P (удовл. и молодой муж.) =P (удовл.) * P (молодой муж.)

P(удовл.) = 302/385
P(молодой муж.) = 33/385
P(удовл. и молодой муж.) = $302*33/385^2$
Ожидаемое количество = 302*33/385

Наблюдаемое и ожидаемое

Наблюдаемое

	Молодой муж.	Молодая жен.	Пожилой муж.	Пожилая жен.	**TOTAL**
Удовл.	25	30	135	112	302
Не удовл.	8	16	22	37	83
TOTAL	33	46	157	149	385

Ожидаемое

	Молодой муж.	Молодая жен.	Пожилой муж.	Пожилая жен.	**TOTAL**
Удовл.	**25,9**	**36,1**	**123,1**	**116,9**	302
Не удовл.	**7,1**	**9,9**	**33,9**	**32,1**	83
TOTAL	33	46	157	149	385

Тест независимости χ-квадрат

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(25-25,9)^2}{25,9} + \frac{(30-36,1)^2}{36,1} + \ldots = 11,1$$

d.f. = (n–1)×(m–1) = 3

Хи-квадрат: предупреждение

- Тест независимости хи-квадрат применим, только если ожидаемое значение в каждой ячейке больше 5.

- Если это условие не выполняется, для таблиц 2×2 надо использовать точный критерий Фишера
(существует его обобщение на таблицы большего размера)

Точный критерий Фишера

(обобщение на таблицы m×n)

P-value по точному критерию Фишера равно дроби.

В числителе — число возможных расстановок объектов таких, чтобы получались таблицы, удовлетворяющие условиям :

(а) сумма чисел по каждому столбцу та же, что у данной таблицы

(б) сумма чисел по каждой строке та же, что у данной таблицы

(в) отличие от «ожидаемой» таблицы не меньше, чем у данной таблицы.

В знаменателе — число возможных расстановок объектов таких, чтобы получались таблицы, удовлетворяющие условиям (а) и (б).

Для таблицы 2×2 условие (в) легко проверяемо, можно взять элемент таблицы, наиболее отклоняющийся вниз от ожидаемого, и перебрать расстановки по всем таблицам, в которых этот элемент принимает значения от 0 до данного включительно (остальные элементы в случае 2×2 однозначно задаются условиями (а) и (б)).

В общем случае нужно использовать какую-то меру отличия, например ту же статистику хи-квадрат.

Проверка независимости

Исходные данные — выборка объектов, для каждого имеются два свойства. Нужно проверить, независимы ли свойства.

Свойства бывают категориальные и численные.

Если оба свойства **категориальные**, то проверка независимости — это исследование таблицы сопряжённости (хи-квадрат или точный кр. Фишера)

Если одно свойство **категориальное**, а другое **численное**, то проверка независимости — это дисперсионный анализ (ANOVA) или его обобщения.
(Если категории две, то это то же, что сравнение двух выборок).

Если оба свойства **численные**, то исследуется корреляция между свойствами.
(А ещё численные свойства можно превратить в категориальные, пожертвовав частью информации).

Критерий согласия Пирсона

Менеджер продуктового магазина хочет определить, будет ли определенный продукт будет продаваться одинаково хорошо в любом из пяти мест в магазине. В каждом из мест собраны данные о продажах

Место	1	2	3	4	5
Кол-во продаж	43	29	52	34	48

Достаточно ли доказательств, чтобы утверждать наличие разницы?

Критерий согласия Пирсона (χ^2)

Наблюдаемое и ожидаемое

- Ожидаемое: (43+29+52+34+48)/5 = 41,2
- Разность наблюдаемого (observed, *O*) и ожидаемого (expected, *E*) имеет в предположении H_0 нормальное распределение со средним 0 и дисперсией *E*
- $(O - E)/E^{1/2} \sim N(0; 1)$
- Разные *O* зависимы, но зависимость линейна: $\sum O = nE$
- Величина

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

распределена как χ^2 с $n - 1$ степенью свободы
(в нашем случае $n = 5$, поэтому d.f. = 4).

Критерий согласия Пирсона (χ^2)

Место	1	2	3	4	5
Кол-во продаж	43	29	52	34	48

Всего = 43+29+…+48=206

Ожидаемое в каждой позиции при независимости: 206/5=41,2

H_0: Распределение равномерное
H_A: Распределение не равномерное

d.f. = 4

Пример

Генетик утверждает, что у плодовых мушек некоторая пара признаков появляется в соотношении 1:3:3:9. Предположим, что выборка из 4000 плодовых мух содержит 226, 764, 733 и 2277 мух каждого типа соответственно. На 10% уровне значимости, достаточно ли доказательств, чтобы отвергнуть гипотезу генетиков?

χ^2 : предупреждение

Критерий согласия хи-квадрат применим, только если ожидаемое значение в **каждой** ячейке больше 5, а общее количество объектов не менее 20.

Критерии согласия

Нередко бывает нужно установить отличие распределения наблюдений от некоторого заранее заданного распределения.

Например, H_0 может состоять в том, что распределение равномерное, а H_1 — что любое другое.

В случае дискретных распределений (и достаточно больших чисел для каждого исхода) следует использовать критерий хи-квадрат Пирсона.

Как быть в случае непрерывных распределений?

Самый простой выход: разбить область значений на несколько подобластей, для каждой посчитать теоретическую вероятность попадания в неё, потом посчитать реальные числа попаданий и применить хи-квадрат.

Недостаток такого подхода: зависимость от разбиения.

Критерий согласия Колмогорова

Обозначим через Φ функцию распределения, отвечающую нулевой гипотезе
(то есть $\Phi(x) = P(\xi<x)$ при $\xi \sim H_0$).

Обозначим через F эмпирическую функцию распределения нашей выборки:

$F(x) = \#(x_i < x)/n$

Статистика: $D = \max|\Phi(x) - F(x)|$

При нулевой гипотезе (независимо от исходного распределения!) и при больших n величина $n^{½}D$ распределена приблизительно по Колмогорову

(см. https://ru.wikipedia.org/wiki/Распределение_Колмогорова),
для малых n существуют таблицы
(напр., http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/),
но в наше время проще оценить p-значение вычислительным экспериментом.

Критерий согласия Колмогорова

Важное дополнение:

Нельзя (по крайней мере без специальных корректировок) использовать критерий Колмогорова, если параметры того распределения, с которым вы сравниваете выборку, подбирались, исходя из той же выборки

Проверка нормальности

Предположим, что мы хотим убедиться в том, что наблюдения $x_1, \ldots, x_n$ происходят из нормального распределения.

Прежде всего, надо оценить по выборке среднее $m = (x_1 + \ldots + x_n)/n$ и дисперсию $s^2 = \sum_i (x_i - m)^2/(n-1)$.

Далее можно использовать статистику Колмогорова:

$$K = n^{1/2} \max_i \max\left(|F(z_i) - i/n|, |F(z_i) - (i-1)/n|\right)$$

где $\{z_i\} = \{(x_i - m)/s\}$ и $z_1 \le z_2 \le \ldots \le z_n$ (нормализованные и упорядоченные наблюдения), а F — функция распределения для N(0;1).

К сожалению, ситуация усложняется тем, что мы вывели *m* и *s* из наблюдений, а значит таблицы для распределения Колмогорова впрямую использовать нельзя. Нужны либо специальные таблицы для такого варианта (распределение Лилиефорса, Lilliefors distribution), либо вычислительный эксперимент (вычисление той же статистики для данных, симулированных исходя из распределения N(0;1)).

Проверка нормальности

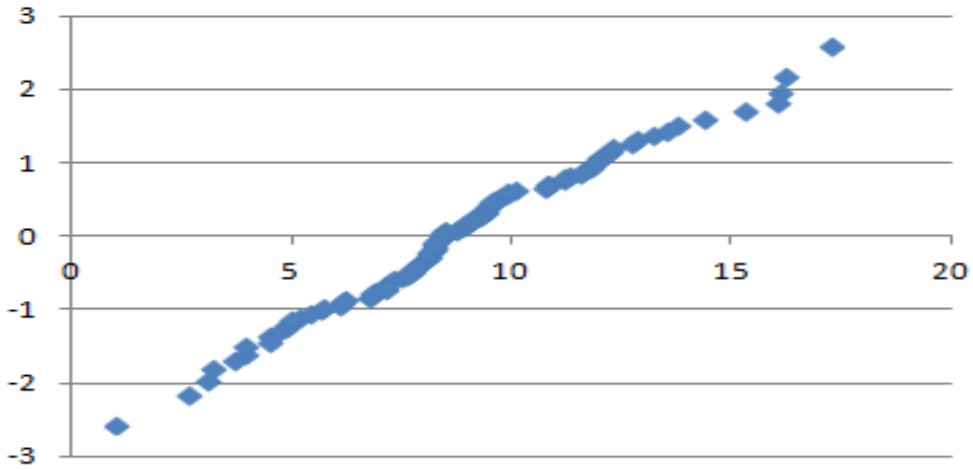
F и $z_1, z_2, \ldots, z_n$ — те же, что на предыдущем слайде.

Возьмём числа $F(z_1), F(z_2), \ldots, F(z_n)$ и нарисуем на плоскости точки с абсциссами $F(z_i)$ и ординатами $(i - ½)/n$ (середины скачков эмпирической функции распределения). Если исходное распределение было нормальным, то эти точки (**P-P plot**) будут хорошо ложиться на прямую.

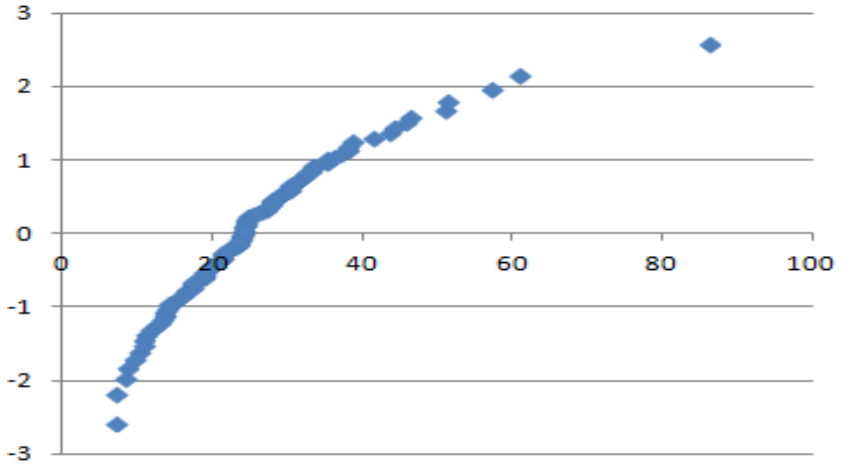
Ещё нагляднее будет, если рисовать **Q-Q plot**:
точки с абсциссами $z_1, z_2, \ldots, z_n$ и ординатами $F^{-1}((i - ½)/n)$.

Наконец, в последней картинке можно заменить $z_1, z_2, \ldots, z_n$ на исходные $x_1, x_2, \ldots, x_n$ (ведь они связаны линейным преобразованием), у точки с абсциссой x_i должна быть ордината $F^{-1}((\mathrm{rank}(x_i) - ½)/n)$.

Проверка нормальности (Q-Q plots)



Вероятно, нормально



Не нормально

По горизонтальной оси — значения из проверяемой выборки.
По вертикальной оси — числа $F^{-1}((\mathrm{rank}(x_i) - ½)/n)$, где n — размер выборки, $\mathrm{rank}(x_i)$ — ранг соответствующего значения (порядковый номер в **упорядоченной** выборке), F — функция распределения стандартного нормального распределения N(0;1)

Проверка нормальности — метод моментов

Обозначим $m = (x_1 + \ldots + x_n)/n$ — оценка среднего, $s^2 = \sum_i (x_i - m)^2/(n-1)$ — оценка дисперсии.

Величина $\sum|x_i - m|/ns$ называется первым абсолютным центральным моментом.
Для (любого) нормального распределения он стремится к 0,8 с ростом n.

Величина $S = \sum(x_i - m)^3/ns^3$ называется третьим центральным моментом или **асимметрией** (skewness) выборки.
Для нормального распределения она стремится к 0 с ростом n.

Величина $K = \sum(x_i - m)^4/ns^4$ называется четвёртым центральным моментом или **эксцессом** (kurtosis) выборки.
Для нормального распределения эксцесс стремится к 3 с ростом n.

Сравнивая эти три величины с их «нормальными» значениями, можно оценить, насколько распределение близко к нормальному.

Критерий Харке – Бера (Jarque–Bera test) использует статистику $$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

которая при больших n распределена как хи-квадрат с двумя степенями свободы, если исходная выборка происходила из нормального распределения.
Критерий Шапиро – Уилка (Shapiro–Wilk test) использует довольно сложную статистику, основанную на сопоставлении чисел выборки с их рангами в этой выборке.

Задача сравнения двух выборок

Пусть $X_1, \ldots, X_k$ и $Y_1, \ldots, Y_l$ — две выборки.

Нулевая гипотеза H_0 — эти выборки из одной генеральной совокупности.

Альтернативная гипотеза H_1 — эти выборки из разных генеральных совокупностей

Критерий Смирнова

Универсальный критерий для проверки **несовпадения** распределений, представленных двумя выборками размеров k и l

(без предположения о характере различий: положительном сдвиге и т.п.)

Сравниваем две эмпирические функции распределения:

$$F_X(t) = \#(X < t)/k \quad \text{и} \quad F_Y(t) = \#(Y < t)/l$$

Статистика $D_{kl} = \max_t | F_X(t) - F_Y(t) |$

Величина

$$K = (kl/(k+l))^{½} D_{kl}$$

при больших k и l (и справедливости H_0, то есть совпадении распределений), распределена по Колмогорову, см. https://ru.wikipedia.org/wiki/Распределение_Колмогорова

(критические значения: K = 1,949 для α = 0,001 ; K = 1,358 для α = 0,05).

Для малых k и l существуют таблицы, но проще оценить P вычислительным экспериментом.

Задача сравнения двух выборок
(распространённое уточнение)

Пусть $X_1, \ldots, X_k$ и $Y_1, \ldots, Y_l$ — две выборки.

Нулевая гипотеза H_0 — эти выборки из одной генеральной совокупности.

Альтернативная гипотеза H_1 — эти выборки из разных генеральных совокупностей, и вторая отличается от первой каким-то фактором, способствующим увеличению значений.

Последнему условию можно придавать немного разный точный смысл. Математики обычно рассматривают вариант, когда распределение для *Y* отличается от распределения для *X* сдвигом на положительную величину: $Y \sim X + c$, где c — константа. Альтернативная гипотеза тогда состоит в том, что $c > 0$.

Бывают и другие уточнения H_1, например, что *Y* отличается от *X* масштабом: $Y \sim c\,X$, $H: c > 1$

Сведение к таблице сопряжённости

Возьмём какой-то порог S и получим таблицу сопряжённости:

Число X < S	Число X ≥ S
Число Y < S	Число Y ≥ S

При нулевой гипотезе зависимости между строками нет, так что если мы её обнаружим (например, с помощью точного критерия Фишера), можно будет отклонить H_0.

Недостаток: произвол в выборе порога

(*у рецензента может возникнуть подозрение, что вы его подбирали, стараясь минимизировать P-value*)

Можно взять в качестве S, например, медиану объединённой выборки.

Двувыборочный критерий Уилкоксона (Wilcoxon rank-sum test)

Объединим и упорядочим выборки:

$$\{X_1, \ldots, X_k\} \cup \{Y_1, \ldots, Y_l\} = \{Z_1, \ldots, Z_{k+l}\}, \text{ причём } Z_1 < Z_2 < \ldots < Z_{k+l}.$$

Таким образом, номер i значения Z_i — это его ранг в объединённой выборке, а каждое Z_i равно либо какому-нибудь X_j, либо какому-нибудь Y_j.

Обозначим ранг каждого Z через $r(Z)$:

$$r(Z_i) = i$$

Статистика Уилкоксона (Wilcoxon) R — это сумма рангов X в объединённой выборке:

$$R = \sum_{Zi = Xj} i = \sum r(X_j)$$

Нулевая гипотеза H_0 отклоняется, если $R < C$, где C — критическое значение.

(разумеется, есть и двусторонний вариант этого теста)

Двувыборочный критерий Уилкоксона

При больши́х k и l величина R при условии H_0 имеет нормальное распределение со средним $k(k+1)/2 + kl/2$ и дисперсией $kl(k + l +1)/12$

При малых k или l нужно смотреть в специальные таблицы для распределения Уилкоксона (или провести вычислительный эксперимент)

При совпадении части значений нужно усреднить ранги этих значений

При совпадении большого числа значений критерий неприменим

Пример

Любители устроили соревнование в скорости улиток двух пород:
8 крапчатых и 12 поперечнополосатых.
К финишу улитки пришли в таком порядке:
ПККПКПККККПКППКПППППП
По результатам возникло предположение, что крапчатые в среднем ползают быстрее.
Как оценить P-value сделанного утверждения?

Критерий Манна – Уитни (*U* test)

Исходные данные те же, что у критерия Уилкоксона.
Статистика Манна – Уитни (Mann-Whitney) — число случаев (из общего числа kl пар наблюдений), когда какой-нибудь X_i больше какого-нибудь Y_j :

$$U = \#\{i, j \mid X_i > Y_j\}$$

Нетрудный факт (упражнение — доказать): для любых исходных данных
(при условии, что все X и Y различны):

$$U = R - k(k+1)/2$$

где R — статистика Уилкоксона.

Поэтому U (при больших k, l) распределено нормально со средним $kl/2$ и дисперсией $kl(k + l + 1)/12$.

Содержательно критерии Уилкоксона и Манна – Уитни — одно и то же.

(см. https://en.wikipedia.org/wiki/Mann–Whitney_U_test)

Двувыборочный Z-тест

Оцениваем средние каждой выборки:

$\bar{X} = (X_1 + \ldots + X_k)/k$

$\bar{Y} = (Y_1 + \ldots + Y_l)/l$

Оцениваем дисперсии:

$s_X^2 = \sum_i (X_i - \bar{X})^2 / (k-1)$, $s_Y^2 = \sum_j (Y_j - \bar{Y})^2 / (l-1)$

Если выборки достаточно большие, величина

$Z = (\bar{X} - \bar{Y}) / (s_X^2/k + s_Y^2/l)^{1/2}$

будет при H_0 (то есть равенстве средних) иметь стандартное нормальное распределение

Z-тест годится для любых «приличных» распределений (достаточно, чтобы существовала дисперсия), но требует большого объёма данных (не менее 100 в каждой выборке)

Двувыборочный критерий Стьюдента

Если выборки небольшие (меньше 100 элементов хотя бы в одной из них), то можно применить критерий Стьюдента (t-тест), но должны выполняться условия:

(1) обе генеральные совокупности нормально распределены

(2) дисперсии обеих г.с. равны
(это условия не только на H_0, но и на H_1!):

Оцениваем средние каждой выборки:

$$\bar{X} = (X_1 + \ldots + X_k)/k$$

$$\bar{Y} = (Y_1 + \ldots + Y_l)/l$$

Оцениваем (общую для двух выборок) дисперсию:

$$s^2 = (\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2) \,/\, (k + l - 2)$$

Теперь при H_0 (т.е., равенстве средних) величина $t = (\bar{X} - \bar{Y}) \,/\, (s^2/k + s^2/l)^{1/2}$
распределена по Стьюденту с $k + l - 2$ степенями свободы

При больших объёмах данных t-тест эквивалентен Z-тесту

Пример

Отделение физкультуры в больнице зафиксировало пульс пяти бегунов в состоянии покоя 60, 58, 59, 61 и 67, соответственно, в то время как для семи нетренированных людей – 83, 60, 75, 71, 91, 82, и 84. Можно ли на уровне значимости 1% считать, что есть различие между пульсом бегунов и нетренированных?
Предположите нормальное распределение и равные дисперсии.

(Средние равны 61 и 78, станд. отклонения – 3,54 и 10,23, соответственно)

Пример

Исследователь полагает, что новая диета должна увеличить вес лабораторных мышей. Если десять контрольных мышей на старой диете весили 4 унции со стандартным отклонением в 0,3 унции, в то время как вес десяти мышей на новой диете был 4,8 унции со стандартным отклонением в 0,2 унции, чему равно P-значение для утверждения о приросте веса?