Лекция 5
Непараметрические критерии

Юрий Котов
Senior Data Engineer
T-bank

План лекции

- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

Непараметрические критерии

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.

План лекции

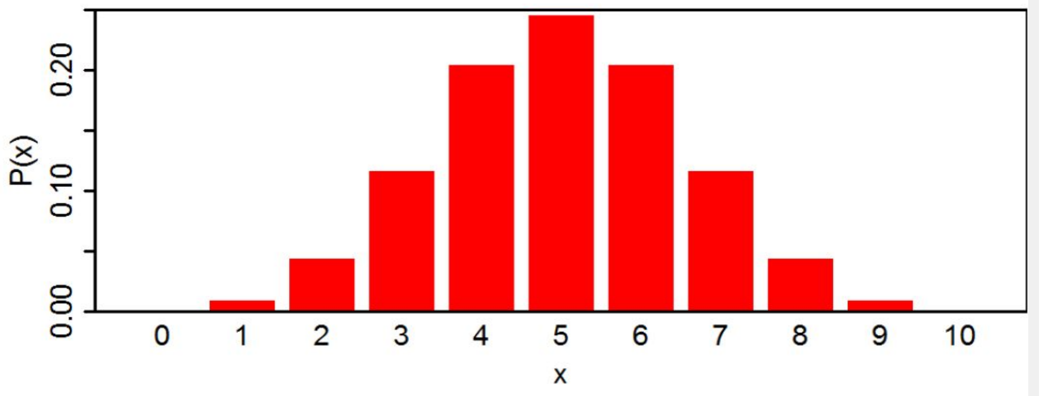
- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

Критерий знаков для одной выборки

Условия применения

- В выборке не должно быть ни одного объекта значение признака которого в точности совпадает с m_0

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



Критерий знаков для одной выборки

Предположим, мы хотим проверить гипотезу о том, что медиана зарплат в компании составляет **50 000 рублей**. Для этого мы взяли выборку из 10 сотрудников и их зарплаты, чтобы сравнить их с этой медианой.

Гипотезы:

- Нулевая гипотеза (H_0): медиана зарплат в компании равна 50 000 рублей.
- Альтернативная гипотеза (H_1): медиана зарплат не равна 50 000 рублей.

Сотрудник	Зарплата (тыс. рублей)
1	48
2	52
3	49
4	50
5	51
6	47
7	53
8	46
9	55
10	50

Критерий знаков для одной выборки

Сотрудник	Зарплата (тыс. рублей)	Знак
1	48	-
2	52	+
3	49	-
4	50	исключаем
5	51	+
6	47	-
7	53	+
8	46	-
9	55	+
10	50	исключаем

Подсчёт знаков:

- Количество положительных знаков (+) = 4
- Количество отрицательных знаков (-) = 4
- Два наблюдения (50 000 рублей) были исключены, так как они равны медиане

Критерий знаков для одной выборки

Тестирование гипотезы:

При n = 8 и при равенстве знаков вероятность получить такое распределение знаков равна 0.5 для каждого знака.

- Нулевая гипотеза предполагает, что положительных и отрицательных знаков должно быть примерно поровну.
- Так как количество знаков почти поровну (4 плюса и 4 минуса), это подтверждает нашу нулевую гипотезу.
- Мы не можем отвергнуть гипотезу о том, что медиана зарплат действительно равна 50 000 рублей, так как распределение знаков не свидетельствует о значительных отклонениях.

Критерий знаков для связанных выборок

Условия применения

- Связанные выборки (парные наблюдения)
- $X_{1i} \neq X_{2i}$

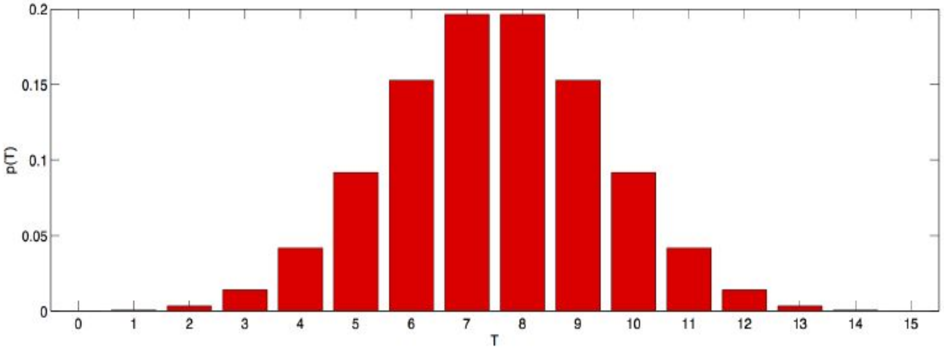
выборки: $X_1^n = (X_{11}, \ldots, X_{1n})$
$X_2^n = (X_{21}, \ldots, X_{2n}), X_{1i} \neq X_{2i}$
выборки связанные

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

нулевое распределение: $Bin(n, \frac{1}{2})$



Критерий знаков для связанных выборок

Мы хотим проверить, влияет ли лекарство на артериальное давление. Мы измеряем давление каждого участника **до** и **после** применения лекарства и применяем критерий знаков для анализа изменений.

Гипотезы

- Нулевая гипотеза (H_0): Лекарство не влияет на артериальное давление, медиана разностей давления до и после приема лекарства равна нулю.
- Альтернативная гипотеза (H_1): Лекарство влияет на давление, медиана разностей не равна нулю (есть систематическое снижение или повышение).

Пациент	Давление до	Давление после
1	150	140
2	145	142
3	160	155
4	155	152
5	148	150
6	165	160
7	158	154
8	152	150
9	149	147
10	153	153

Критерий знаков для связанных выборок

Пациент	Давление до	Давление после	Разница
1	150	140	+10
2	145	142	+3
3	160	155	+5
4	155	152	+3
5	148	150	-2
6	165	160	+5
7	158	154	+4
8	152	150	+2
9	149	147	+2
10	153	153	0

- Количество положительных знаков (+) = 8
- Количество отрицательных знаков (-) = 1
- Одно наблюдение (разница = 0) исключаются из анализа

Критерий знаков для связанных выборок

Частота «успеха» $H = K/n$ $H = 8/9 \sim 0.88$

Найдем выборочное значение статистики критерия

$$Z = \frac{H - 1/2}{\sqrt{1/4n}} = 2\sqrt{n}(H - 1/2) \qquad 2 * \sqrt{9}(0.88 - 1/2) = 2.28$$

Посчитаем для α = 5% уровня значимости. Аппроксимируем закон распределения статистики критерия при условии истинности основной гипотезы стандартизованным нормальным распределением *N*(0, 1).

По таблицам математической статистики находим квантиль на уровне 1–α/2: $z_{1-\alpha/2} = z_{0.975} = 1.96$
Таким образом область допустимых значения (-1.96; 1.96)
2.28 не входит в нашу область допустимых значений, значит можем отклонить нулевую гипотезу в пользу альтернативной

Причины использовать критерии знаков

- Если у вас небольшая выборка и данные сильно асимметричны или содержат выбросы
- Если исследование направлено на выявление тенденций увеличения или уменьшения

План лекции

- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

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:

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

Ранг наблюдения X_i:

если X_i не в связке, то $\operatorname{rank}(X_i) = r \colon X_i = X_{(r)}$,

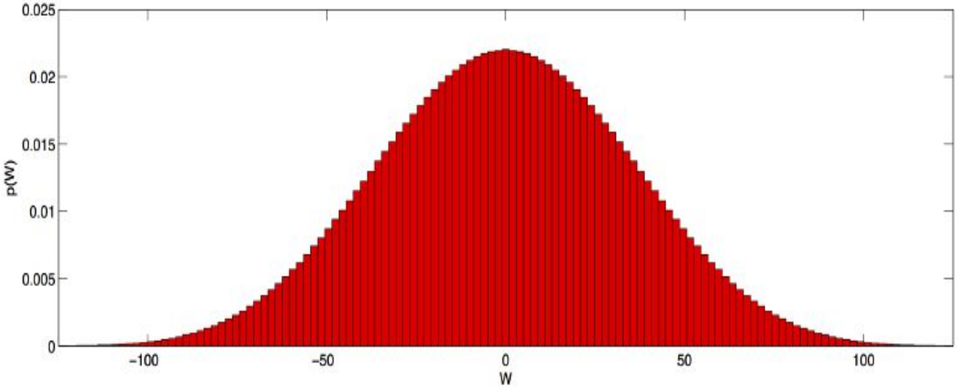
если X_i в связке $X_{(k_1)}, \ldots, X_{(k_2)}$, то $\operatorname{rank}(X_i) = \frac{k_1 + k_2}{2}$.

Одновыборочный критерий знаковых рангов Уилкоксона

Условия применения

- В выборке не должно быть ни одного объекта значение признака которого в точности совпадает с m_0
- Функция распределения симметрично относительно медианы

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



Одновыборочный критерий знаковых рангов Уилкоксона

Хотим проверить, отличается ли медиана уровня холестерина у группы пациентов от значения 200 мг/дл.

Уровень холестерина: 195, 210, 198, 205, 192, 220, 215, 205, 200, 190

Мы хотим проверить гипотезу о том, что медиана выборки равна 200 мг/дл.

- Нулевая гипотеза H_0 : Медиана уровня холестерина равна 200 мг/дл.
- Альтернативная гипотеза H_1: Медиана уровня холестерина не равна 200 мг/дл.

Одновыборочный критерий знаковых рангов Уилкоксона

Считаем разность с предполагаемой медианой

- $195 - 200 = -5$
- $210 - 200 = 10$
- $198 - 200 = -2$
- $205 - 200 = 5$
- $192 - 200 = -8$
- $220 - 200 = 20$
- $215 - 200 = 15$
- $205 - 200 = 5$
- $200 - 200 = 0$
- $190 - 200 = -10$

Разность с учетом знаков:

$-5, 10, -2, 5, -8, 20, 15, 5, -10$

Теперь мы игнорируем знаки и ранжируем разности по их абсолютным величинам и проставляем ранги:

- $|2|$ — ранг 1
- $|5|$ — ранг 3 (так как три значения 5 совпадают, их ранг будет средним $(2+3+4)/3=3$)
- $|8|$ — ранг 5
- $|10|$ — ранг 6.5 (так как два значения совпадают, их ранг $(6+7)/2=6.5$)
- $|15|$ — ранг 8
- $|20|$ — ранг 9

Одновыборочный критерий знаковых рангов Уилкоксона

Ранжированные абсолютные значения:

- 2 (ранг 1)
- 5 (ранг 3)
- 5 (ранг 3)
- 5 (ранг 3)
- 8 (ранг 5)
- 10 (ранг 6.5)
- 10 (ранг 6.5)
- 15 (ранг 8)
- 20 (ранг 9)

Возвращаем знаки рангам

- −5 (ранг 3)
- +10 (ранг 6.5)
- −2 (ранг 1)
- +5 (ранг 3)
- −8 (ранг 5)
- +20 (ранг 9)
- +15 (ранг 8)
- +5 (ранг 3)
- −10 (ранг 6.5)

Одновыборочный критерий знаковых рангов Уилкоксона

- Сумма положительных рангов: T_+ = 6.5+3+9+8+3=29.5
- Сумма отрицательных рангов: T_- = 3+1+5+6.5=15.5

$$W = min(T_+, T_-) = min(29.5, 15.5) = 15.5$$

Одновыборочный критерий знаковых рангов Уилкоксона

Теперь нужно сравнить полученное значение W=15.5 с критическим значением. Для одновыборочного критерия Уилкоксона критическое значение зависит от числа ненулевых наблюдений и уровня значимости α. У нас есть 9 ненулевых наблюдений, и при уровне значимости α=0.05 критическое значение из таблицы равно 8.

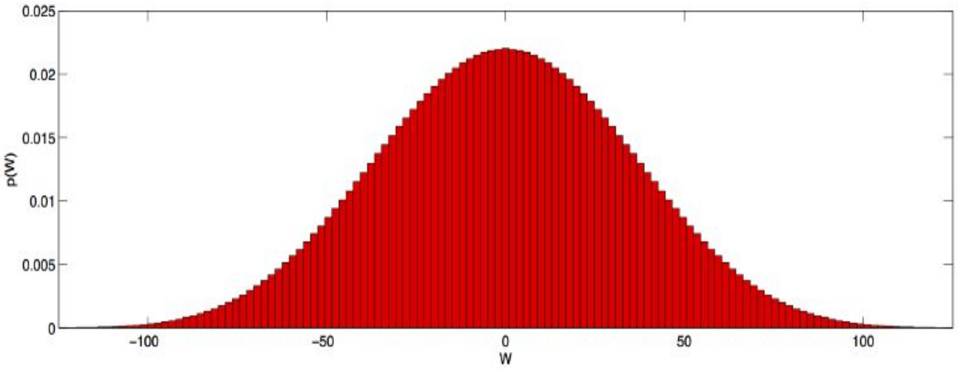
Поскольку W=15.5 больше критического значения 8, у нас **нет оснований отвергнуть нулевую гипотезу**. Это означает, что нет достаточных доказательств того, что медиана уровня холестерина в выборке отличается от 200 мг/дл.

Критерий знаковых рангов Уилкоксона для связанных выборок

Условия применения

- Связанные выборки (парные наблюдения)
- $X_{1i} \neq X_{2i}$

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



Критерий знаковых рангов Уилкоксона для связанных выборок

Предположим, мы проводим исследование на 8 пациентах, оценивая их артериальное давление до и после применения нового лекарства. Нам нужно проверить, оказывает ли лекарство значительное влияние на артериальное давление.

Данные:

- Измерения **до лечения**: 120, 135, 130, 110, 140, 125, 115, 135
- Измерения **после лечения**: 110, 130, 127, 105, 135, 121, 107, 130

Критерий знаковых рангов Уилкоксона для связанных выборок

Предположим, мы проводим исследование на 8 пациентах, оценивая их артериальное давление до и после применения нового лекарства. Нам нужно проверить, оказывает ли лекарство значительное влияние на артериальное давление.

Данные:

- Измерения **до лечения**: 120, 135, 130, 110, 140, 125, 115, 135
- Измерения **после лечения**: 110, 130, 127, 105, 135, 121, 107, 130

Нулевая гипотеза: Влияния лекарства нет $med(X_1 - X_2) = 0$

Альтернативная гипотеза: Есть влияние лекарства $med(X_1 - X_2) > 0$

Критерий знаковых рангов Уилкоксона для связанных выборок

Для каждой пары измерений (до и после лечения) находим разность:

- 120 - 110 = 10
- 135 - 130 = 5
- 130 - 127 = 3
- 110 - 105 = 5
- 140 - 135 = 5
- 125 - 121 = 4
- 115 - 107 = 8
- 135 - 130 = 5

Критерий знаковых рангов Уилкоксона для связанных выборок

- Игнорируем нулевые разности
- Определение знака разностей
- Ранги
  - 3 < 4 < 5 = 5 = 5 = 5 < 8 < 10
  - (3 + 4 + 5 + 6) / 4 = 4.5

Сумма положительных рангов: Поскольку все разности положительные, сумма положительных рангов будет равна 36.

Сумма отрицательных рангов: Поскольку отрицательных разностей нет, сумма отрицательных рангов равна 0.

W = 36

Критерий знаковых рангов Уилкоксона для связанных выборок

Теперь мы сравниваем полученное значение $W = 36$ с критическим значением для критерия Уилкоксона, которое зависит от уровня значимости α и числа наблюдений $n = 8$. Для небольших выборок критические значения можно найти в таблице критических значений критерия Уилкоксона.

При $n = 8$ и уровне значимости $\alpha = 0.05$, критическое значение равно 5. Поскольку $W = 36$ значительно больше критического значения 5, мы отвергаем нулевую гипотезу.

Критерий Манна-Уитни

Условия применимости

- Группы однородны по независимым переменным
- Наблюдения независимы (для зависимых – критерий Уилкоксона.
- Группы могут быть небольшими, минимум 8 наблюдений, но чем больше наблюдений, тем точнее работает критерий
- Повторяющихся значений практически нет
- Форма распределения в группах должна быть схожей

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$

выборки независимые

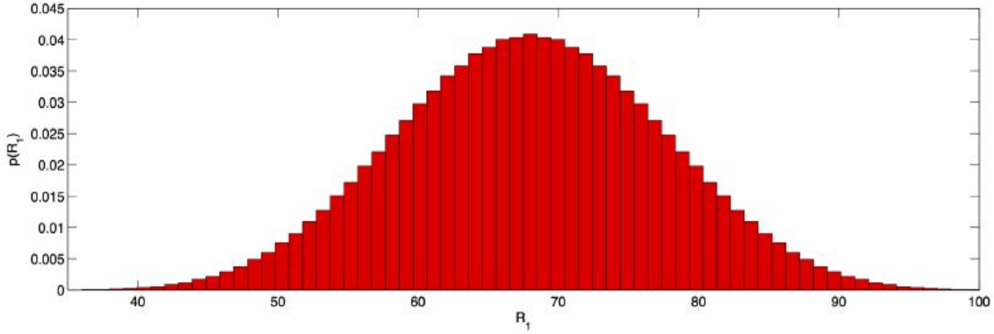
нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$

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

статистика: $X_{(1)} \le \dots \le X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$

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

нулевое распределение: табличное



Критерий Манна-Уитни

Хотим сравнить уровень тревожности у двух групп студентов: одна группа занималась йогой, а другая нет. Уровень тревожности измерен по 10-балльной шкале. Получены следующие данные:

- Группа 1 (занимались йогой): 5, 8, 7, 6, 7
- Группа 2 (не занимались йогой): 9, 10, 8, 9, 9

Критерий Манна-Уитни

Хотим сравнить уровень тревожности у двух групп студентов: одна группа занималась йогой, а другая нет. Уровень тревожности измерен по 10-балльной шкале. Получены следующие данные:

- Группа 1 (занимались йогой): 5, 8, 7, 6, 7
- Группа 2 (не занимались йогой): 9, 10, 8, 9, 9

Нулевая гипотеза: Распределение уровня тревожности в двух группах одинаковое, то есть между ними нет различий.

Альтернативная гипотеза: Распределение уровня тревожности в одной группе отличается от другой.

Критерий Манна-Уитни

- Группа 1 (занимались йогой): 5, 8, 7, 6, 7
- Группа 2 (не занимались йогой): 9, 10, 8, 9, 9

5 — ранг 1

6 — ранг 2

7 — средний ранг 3.5

8 — средний ранг 5.5

9 — средний ранг 8

10 — ранг 10

5 (1), 6 (2), 7 (3.5), 7 (3.5), 8 (5.5), 8 (5.5), 9 (8), 9 (8), 9 (8), 10 (10)

Критерий Манна-Уитни

Группа 1 (йога)
- 5 — ранг 1
- 9 — ранг 5.5
- 7 — ранг 3.5
- 6 — ранг 2
- 7 — ранг 3.5

Группа 2 (без йоги)
- 9 — ранг 8
- 10 — ранг 10
- 8 — ранг 5.5
- 9 — ранг 8
- 9 — ранг 8

Сумма рангов

T_1= 1 + 5.5 + 3.5 + 2 + 3.5 = 15.5

T_2= 8 + 10 + 5.5 + 8 + 8 = 39.5

Критерий Манна-Уитни

Формула для статистики критерия Манна-Уитни:

$$U_1 = n_1 \cdot n_2 + \frac{n_1(n_1+1)}{2} - T_1$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

- n_1— количество наблюдений в группе 1 (йога),
- n_2— количество наблюдений в группе 2 (без йоги),
- T_1 и T_2 — суммы рангов для групп.

$$U_1 = 24.5$$

$$U_2 = 0.5$$

$$U = \min(U_1, U_2) = \min(24.5, 0.5) = 0.5$$

Критерий Манна-Уитни

Для выборок $n_1 = 5$ и $n_2 = 5$и уровня значимости $\alpha = 0.05$ критическое значение U можно найти в таблице критических значений Манна-Уитни. Оно равно 2.

Полученное значение U = 0.5 меньше критического значения 2, следовательно, мы отвергаем нулевую гипотезу.

Еще пару примеров применения критерия: [1] и [2]

Critical Values for the Mann-Whitney U-Test

Level of significance: 5% (P = 0.05)

Size of the largest sample (n_2)

Size of the smallest sample (n_1)	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	13	13
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	17	18	19	20	21	22	23
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23	24	25	27	28	29	30	32	33
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30	32	33	35	37	38	40	42	43
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54
8				13	15	17	19	22	24	26	29	31	34	36	38	41	43	45	48	50	53	55	57	60	62	65
9					17	20	23	26	28	31	34	37	39	42	45	48	50	53	56	59	62	64	67	70	73	76
10						23	26	29	33	36	39	42	45	48	52	55	58	61	64	67	71	74	77	80	83	87
11							30	33	37	40	44	47	51	55	58	62	65	69	73	76	80	83	87	90	94	98
12								37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109
13									45	50	54	59	63	67	72	76	80	85	89	94	98	102	107	111	116	120
14										55	59	64	67	74	78	83	88	93	98	102	107	112	118	122	127	131
15											64	70	75	80	85	90	96	101	106	111	117	122	125	132	138	143
16												75	81	86	92	98	103	109	115	120	126	132	138	143	149	154
17													87	93	99	105	111	117	123	129	135	141	147	154	160	166
18														99	106	112	119	125	132	138	145	151	158	164	171	177
19															113	119	126	133	140	147	154	161	168	175	182	189
20																127	134	141	149	156	163	171	178	186	193	200
21																	142	150	157	165	173	181	188	196	204	212
22																		158	166	174	182	191	199	207	215	223
23																			175	183	192	200	209	218	226	235
24																				192	201	210	219	228	238	247
25																					211	220	230	239	249	258
26																						230	240	250	260	270
27																							250	261	271	282
28																								272	282	293
29																									294	305
30																										317

Причины использовать ранговые критерии

- Если у вас небольшая выборка и данные сильно асимметричны или содержат выбросы
- Если исследование направлено на выявление тенденций увеличения или уменьшения
- В качестве выборки у нас только ранговые(порядковые) значения

План лекции

- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

Как работают перестановочные критерии

Основная идея заключается в том, чтобы оценить, насколько необычно наблюдаемое значение статистики теста, предполагая, что нулевая гипотеза верна. Для этого генерируется распределение статистики теста путем многократного перемешивания (перестановки) данных между группами, что позволяет построить эмпирическое распределение.

Шаги проведения перестановочного теста

1. Формулирование гипотез
2. Вычисление статистики теста на исходных данных
3. Перестановки и вычисление статистики для каждой перестановки
4. Построение эмпирического распределения
5. Оценка значения p-value

Одновыборочный перестановочный критерий

выборка: $X_1^n = (X_1, \ldots, X_n)$

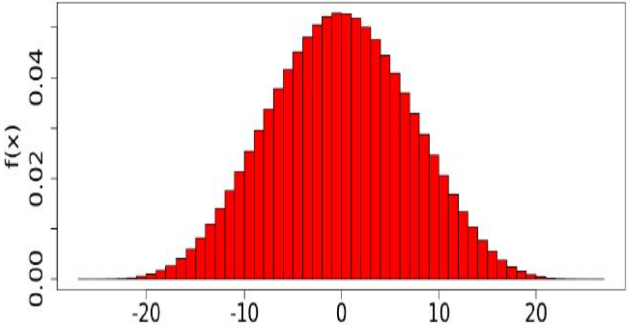
$F(X)$ симметрично относительно матожидания

нулевая гипотеза: $H_0 \colon \mathbb{E}X = m_0$

альтернатива: $H_1 \colon \mathbb{E}X <\neq> m_0$

статистика: $T(X^n) = \sum\limits_{i=1}^{n} (X_i - m_0)$

нулевое распределение: порождается перебором 2^n знаков перед слагаемыми $X_i - m_0$



Достигаемый уровень значимости — доля перестановок знаков, на которых получилось такое же или ещё более экстремальное значение статистики.

Одновыборочный перестановочный критерий

Компания хочет проверить, насколько удовлетворены сотрудники текущими условиями труда. Сотрудники оценили удовлетворенность по шкале от 1 до 10. Компания предполагает, что средний уровень удовлетворенности должен быть не ниже 7.

Уровни удовлетворенности сотрудников: 6, 7, 8, 5, 9, 6, 7, 6, 8, 5

Гипотеза: Проверяем, отличается ли средняя оценка от 7

Уровень значимости: $\alpha = 0.05$

Одновыборочный перестановочный критерий

Вычисляем среднее для выборки

$$\text{Среднее выборки} = \frac{6+7+8+5+9+6+7+6+8+5}{10} = 6.7$$

Разница между средней выборки и ожидаемого значения

$$D_{\text{наблюдаемая}} = |6.7 - 7| = 0.3$$

Одновыборочный перестановочный критерий

1. Вычислим отклонения каждого значения от 7:
   Получим отклонения: −1, 0, 1, −2, 2, −1, 0, −1, 1, −2
2. Перестановка будет заключаться в случайной смене знаков этих отклонений, то есть на каждом этапе мы случайно заменяем отклонение на противоположное (например, −1 на 1, 2 на −2, и так далее).
3. После каждой перестановки пересчитываем "перестановочное среднее" и разницу между ним и ожидаемым значением (в нашем случае 7). Для каждой перестановки получим значение статистики D_i
4. Повторяем шаг 3 минимум 1000 раз
5. Вычисляем p-value: доля перестановок, для которых разница D_i оказалась больше или равна 0.3

Двухвыборочный перестановочный критерий для связанных выборок

выборки: $X_1^n = (X_{11}, \dots, X_{1n})$

$X_2^n = (X_{21}, \dots, X_{2n})$

выборки связанные

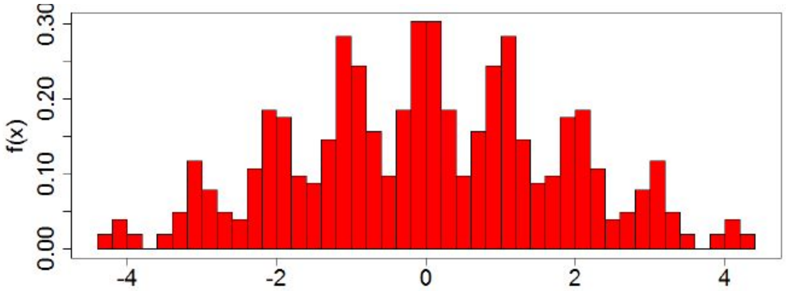
нулевая гипотеза: $H_0 \colon \mathbb{E}\,(X_1 - X_2) = 0$

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



Двухвыборочный перестановочный критерий для связанных выборок

У нас есть группа студентов, которые прошли курс подготовки, и мы измерили их результаты до и после курса. Мы хотим выяснить, повысились ли их результаты после курса.

Результаты **до курса**: 65, 70, 78, 60, 68, 72

Результаты **после курса**: 70, 75, 80, 62, 71, 74

Нулевая гипотеза H_0: Нет значимого изменения в средних результатах до и после курса.

Альтернативная гипотеза H_1: Есть значимое изменение в средних результатах после курса.

ITMO

Двухвыборочный перестановочный критерий для связанных выборок

Найдем разности между результатами "после" и "до" для каждой пары наблюдений

Разности: 5, 5, 2, 2, 3, 2

Найдем среднее этих разностей

$$\text{Средняя разность} = \frac{5+5+2+2+3+2}{6} = 3.17$$

Эта **наблюдаемая средняя разность** будет нашим значением статистики, которое мы будем проверять с помощью перестановок.

Двухвыборочный перестановочный критерий для связанных выборок

1. В каждой перестановке случайным образом меняем знак у каждой разности (то есть, делаем её положительной или отрицательной), чтобы смоделировать возможное распределение разностей в случае, если бы не было значимого изменения.
2. Для каждой перестановки пересчитываем среднюю разность. Это будет наше значение статистики D_i для данной перестановки.
3. Повторяем шаги перестановки минимум 1000 раз, чтобы создать распределение возможных значений средней разности D_i при условии, что изменения были случайными.

Двухвыборочный перестановочный критерий для связанных выборок

Вычисление p-value

Теперь, когда у нас есть распределение значений средней разности D_i из перестановок, можно определить, насколько часто перестановочные значения средней разности были больше или равны (по абсолютному значению) наблюдаемой средней разности $D_{\text{наблюдаемая}} = 3.17.$

p-value — это доля перестановок, в которых абсолютное значение средней разности было больше или равно 3.17.

Двухвыборочный перестановочный критерий для независимых выборок

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$

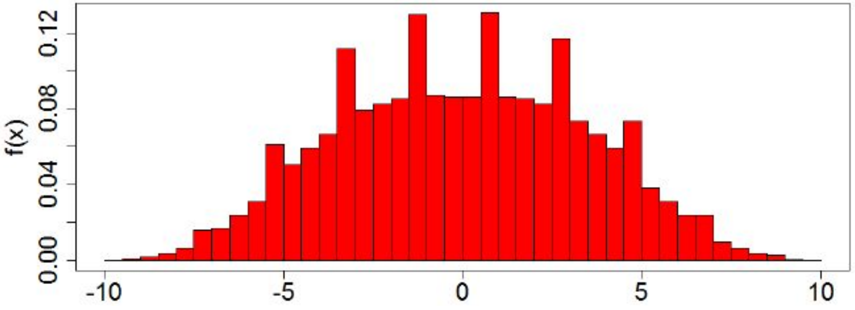
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$

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: $T\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$

нулевое распределение: порождается перебором $C_{n_1+n_2}^{n_1}$ размещений объединённой выборки



Двухвыборочный перестановочный критерий для независимых выборок

Две группы студентов прошли разные курсы подготовки. Мы хотим выяснить, отличается ли уровень знаний между студентами, прошедшими разные курсы, на основе их итоговых тестовых баллов.

Группа A: 85, 88, 90, 82, 87, 91

Группа B: 78, 81, 79, 80, 77, 83

Нулевая гипотеза H_0: Средние баллы групп А и В не отличаются.

Альтернативная гипотеза H_1: Средние баллы групп А и В отличаются.

Двухвыборочный перестановочный критерий для независимых выборок

Среднее для группы A:

$$\text{Среднее A} = \frac{85 + 88 + 90 + 82 + 87 + 91}{6} = 87.17$$

Среднее для группы B:

$$\text{Среднее B} = \frac{78 + 81 + 79 + 80 + 77 + 83}{6} = 79.67$$

Рассчитаем разницу между средними:

$$D_{\text{наблюдаемая}} = 87.17 - 79.67 = 7.5$$

Эта **наблюдаемая разница** (7.5) будет нашим значением статистики.

Двухвыборочный перестановочный критерий для независимых выборок

Генерация перестановок:

1. Объединим все значения из групп A и B в одну выборку:

   85,88,90,82,87,91,78,81,79,80,77,83

2. Перемешиваем значения случайным образом и делим их на две новые группы той же размерности, что и исходные группы (по 6 значений в каждой). Это дает нам пару случайных "групп" без учета их принадлежности к исходным группам.
3. Рассчитываем среднее для каждой новой группы, находим разницу между ними и записываем это значение
4. Повторяем этот процесс большое больше 1000 раз, чтобы получить распределение значений разностей

Двухвыборочный перестановочный критерий для независимых выборок

Теперь, когда у нас есть распределение значений D_i для перестановок, мы можем определить, насколько часто значение разности было больше или равно наблюдаемой разнице $D_{\text{наблюдаемая}} = 7.5$ (в абсолютном значении).

1. p-value — это доля перестановок, в которых значение разности $|D_i| \geq 7.5$.

Предположим, что уровень значимости выбран $\alpha = 0.05$:

- Если p-value < 0.05, то мы отвергаем нулевую гипотезу и считаем, что средние значимо отличаются.
- Если p-value > 0.05, у нас нет оснований отклонить нулевую гипотезу, и мы можем предположить, что средние не отличаются.

План лекции

- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

Бутстреп

Суть метода: путем многократного повторного извлечения наблюдений из исходной выборки мы можем построить эмпирическое распределение, которое поможет апроксимировать исходное неизвестное распределение необходимой нам статистики

- Имеет в основе идею ресэмплинга (многократную генерацию повторных выборок на основе одних и тех же данных)
- Можно оценивать средние/дисперсию/квантили, строить доверительные интервалы, тестировать гипотезы
- Выборка может быть небольшой, но отражающей генеральную совокупность
- Случайные величины в выборке независимы

Бутстреп и доверительные интервалы

- Для каждой подвыборки, сгенерированной во время проведения бутстрепа мы строим точечную оценку интересующего нас параметра
- Собрав достаточно точечных оценок, мы получаем их распределение
- Для этого распределения можно построить доверительный интервал с интересующими нас границами

Бутстреп

У нас есть данные о стоимости товара:

50, 60, 70, 65, 55, 75, 62, 58, 73, 69

Мы хотим:

1. Оценить среднее значение стоимости.
2. Найти 95%-ный доверительный интервал для среднего значения с помощью бутстрепа.

Бутстреп

Из этой выборки будем случайно извлекать элементы **с возвращением**, создавая множество бутстреп-выборок.

1. Выбираем случайные значения из данных с возвращением, чтобы создать первую бутстреп-выборку, равную по размеру исходной выборке (10 элементов).
2. Повторяем процесс многократно (минимум 1000 раз)

Для каждой из 1000 бутстреп-выборок рассчитываем среднее значение. Это даст нам 1000 значений средних, которые будут распределены вокруг среднего исходных данных. Эти средние значения формируют **бутстреп-распределение** средней стоимости.

Теперь, когда у нас есть 1000 средних значений, полученных из бутстреп-выборок, мы можем построить доверительный интервал. Для 95%-го доверительного интервала возьмем 2.5-й и 97.5-й процентили бутстреп-распределения.

Бутстреп

1. Допустим, после генерации бутстреп-выборок и вычисления средних мы получили распределение, по которому 2.5-й и 97.5-й процентили составляют:

   Нижняя граница: 58.2 Верхняя граница: 71.8

2. Таким образом, наш 95%-ный бутстреп-доверительный интервал для среднего будет равен: [58.2,71.8]

Вывод

Средняя стоимость товара по данным составляет примерно 63.7, и с 95%-й уверенностью можно утверждать, что истинное среднее значение находится в диапазоне от 58.2 до 71.8.

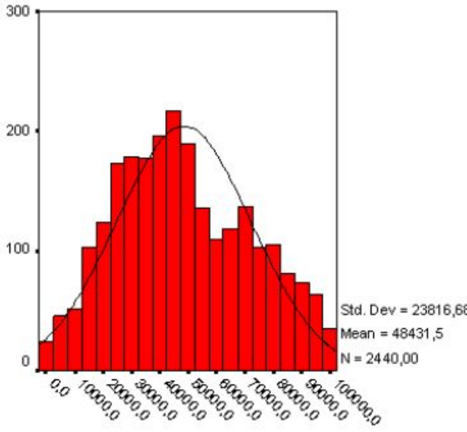
Отличие бутстрепа от перестановочного критерия

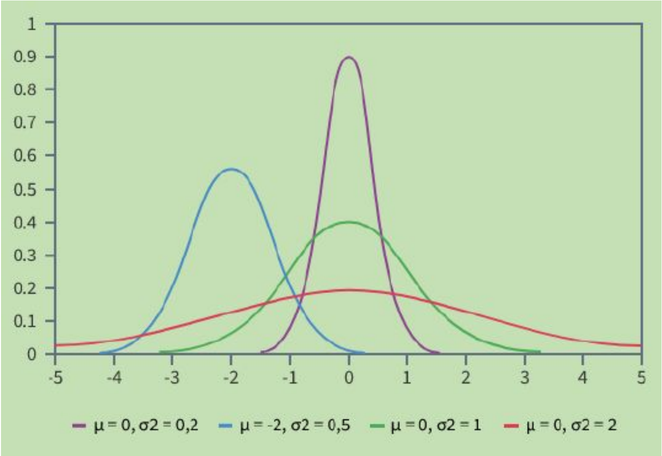
Характеристика	**Бутстреп**	**Перестановочные критерии**
Цель	Оценка распределения статистики, таких как среднее, медиана или стандартное отклонение	Проверка значимости различий между группами
Метод выборки	Выборка **с возвращением** из исходной выборки	Перемешивание данных между группами **без возвращения**
Когда применяется	Когда требуется оценить доверительный интервал или стандартную ошибку статистики, особенно для небольших выборок	Для проверки гипотез о различиях между группами, обычно двух выборок
Предположения	Не требует предположений о распределении данных, требует независимости наблюдений	Требует условия независимости групп (для несвязанных выборок)

План лекции

- Критерии знаков
- Ранговые критерии
- Перестановочные критерии
- Бутстреп
- Критерии для проверки нормальности выборки

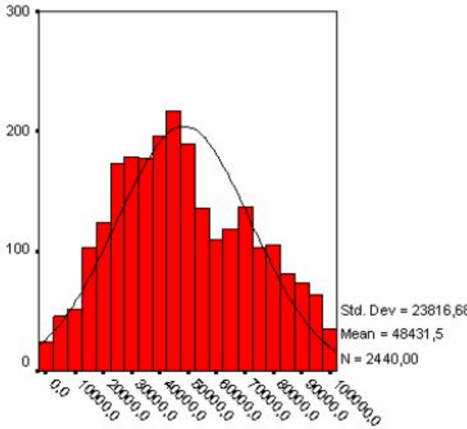
Как проверить нормальность распределения?

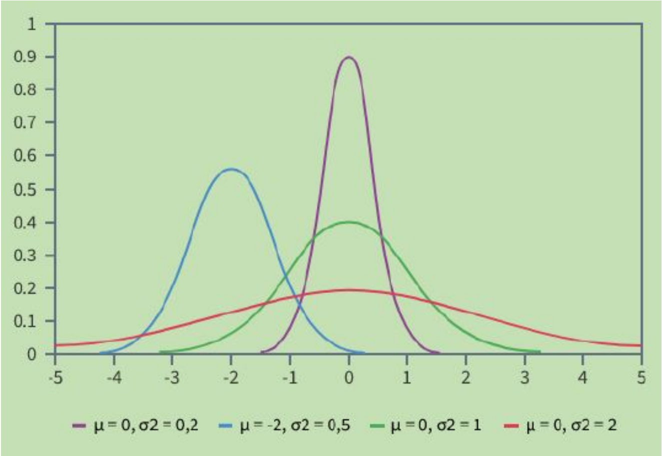




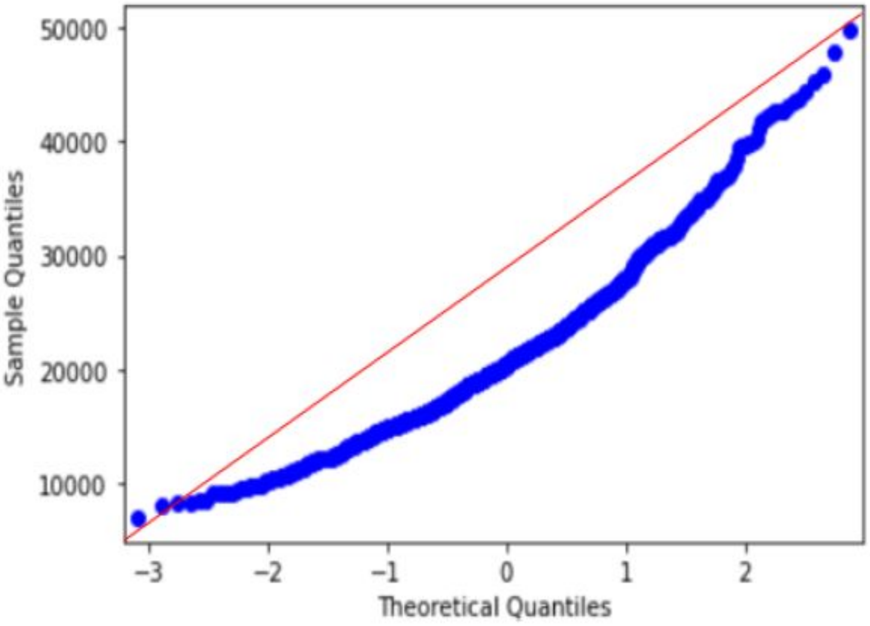
Как проверить нормальность распределения?

- Q-Q plot
- Тест Колмогорова Смирнова
- Тест Шапиро Уилка





Q-Q plot(quantile-quantile plot)



Как построить:

1. Упорядочиваем нашу выборку
2. Считаем экспериментальные квантили
3. Считаем теоретические квантили
4. Строим график с координатами: (теоретическое значение, экспериментальное)

Q-Q plot(quantile-quantile plot)

Как трактовать

- Точки совпадают с прямой => идеальный вариант нормального распределения
- Точки кривой выше прямой => значения в выборке выше, относительно нормального распределения
- Точки кривой ниже прямой => значения в выборке ниже, чем должны быть при нормальном распределении
- Справа от нуля - область от середины до хвоста с максимальными значениями, слева - от середины до минимальных

Критерий Колмогорова-Смирнова

Критерий согласия Колмогорова используется для проверки простых гипотез о принадлежности анализируемой выборки некоторому полностью известному закону распределения

Критерий однородности Смирнова используется для проверки гипотезы о принадлежности двух независимых выборок одному закону распределения, то есть о том что для эмпирических распределения соответствуют одному и тому же закону

В основе критерия - **статистика Колмогорова-Смирнова**, которая является оценкой расстояния между эмпирической выборочной функцией распределения и кумулятивной функцией теоретического распределения, либо между эмпирическими функциями двух выборок

В случае двух выборок распределение, рассматриваемое в рамках нулевой гипотезы, должно быть непрерывным

Критерий Колмогорова-Смирнова

$$X_1^{n1} = (X_{11}, \dots, X_{1n1})$$

$$X_2^{n2} = (X_{21}, \dots, X_{2n2})$$

$$\mathrm{H}_0: F_{X1}(x) = F_{X2}(x)$$

$$\mathrm{H}_A: F_{X1}(x) \neq F_{X2}(x)$$

$F_{n1X1}(x)$, $F_{n2X2}(x)$ - эмпирические функции распределения, построенные по выборкам X_1^{n1} и X_2^{n2}

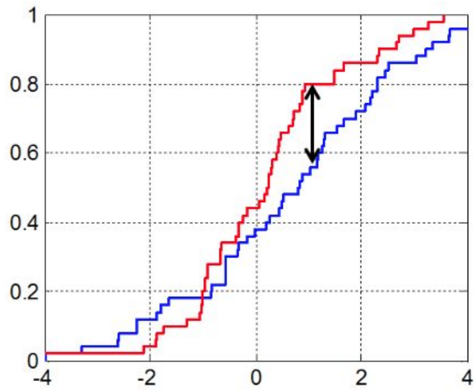
$$\lambda_{\text{наблюдаемое}} = D(X_1^{n1}, X_2^{n2}) * \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}$$

Статистика Колмогорова:

$$D(X_1^{n_1}, X_2^{n2}) = \sup_{-\infty < x < +\infty} |F_{n1X1}(x) - F_{n2X2}(x)|$$

sup – супремум – наибольшее расстояние между функциями

При справедливости H_0 распределение статистики D_n будет одинаковым для любых непрерывных распределений



[Пример использования критерия]()

Критерий Колмогорова-Смирнова

Критические числа Колмогорова–Смирнова

Степень свободы	Проверка единичной выборки *			Проверка двух выборок **	
N	$D_{0,10}$	$D_{0,05}$	$D_{0,01}$	$D_{0,05}$	$D_{0,01}$
1	0,950	0,975	0,995	–	–
2	0,776	0,842	0,929	–	–
3	0,642	0,708	0,828	–	–
4	0,564	0,624	0,733	1,000	1,000
5	0,510	0,565	0,669	1,000	1,000
6	0,470	0,521	0,618	0,833	1,000
7	0,438	0,486	0,577	0,857	0,857
8	0,411	0,457	0,543	0,750	0,875
9	0,388	0,432	0,514	0,668	0,778
10	0,368	0,410	0,490	0,700	0,800
11	0,352	0,391	0,468	0,636	0,727
12	0,338	0,375	0,450	0,583	0,667
13	0,325	0,361	0,433	0,538	0,692
14	0,314	0,349	0,418	0,571	0,643

Наблюдаемое значение сравнивается с критическими из таблицы:

- Если наблюдаемое значение БОЛЬШЕ критического, то различия считаются ЗНАЧИМЫМИ
- Если наблюдаемое значение МЕНЬШЕ критического, то нет оснований отвергать нулевую гипотезу => оба эмпирических распределения соответствуют одному закону распределения

Условия применимости

- Не требует равенства дисперсий, ни одинакового размера групп
- Хорошо подходит для проверки двух групп на однородность по статистическим признакам
- Не подходит для оценки эффекта в A/B тестах

Критерий Шапиро-Уилка

- Наиболее эффективный критерий проверки гипотезы о принадлежности выборки нормальному закону распределения
- Подходит и для больших и малых выборок (но не менее 3х наблюдений)
- W-статистика принимает значения на интервале [0;1]
- Коэффициенты a_t есть в таблицах, как и критические значения

$X^n = (X_1, \dots, X_n)$

$\mathrm{H}_0: X \sim N(\mu, \sigma^2)$

$\mathrm{H}_A: X \sim N(\mu, \sigma^2)$

Если $W_{obs} < W_{cr}$, то нулевая гипотеза нормальности распределения отклоняется на уровне значимости

Таблица 82

Критические значения критерия экспоненциальности W_{E_n} Шапиро–Уилка [13, 298]

n	α = 0,90		α = 0,95		n	α = 0,90		α = 0,95	
	$W_1(\alpha)$	$W_2(\alpha)$	$W_1(\alpha)$	$W_2(\alpha)$		$W_1(\alpha)$	$W_2(\alpha)$	$W_1(\alpha)$	$W_2(\alpha)$
7	0,033	0,225	0,025	0,260	22	0,022	0,069	0,020	0,080
8	0,032	0,200	0,025	0,230	23	0,021	0,065	0,019	0,075
9	0,031	0,177	0,025	0,205	24	0,021	0,062	0,019	0,069
10	0,030	0,159	0,025	0,184	25	0,020	0,058	0,018	0,065
11	0,030	0,145	0,025	0,166	26	0,020	0,056	0,018	0,062
12	0,029	0,134	0,025	0,153	27	0,020	0,054	0,017	0,058
13	0,028	0,124	0,025	0,140	28	0,019	0,052	0,017	0,056
14	0,027	0,115	0,024	0,128	29	0,019	0,050	0,016	0,054
15	0,026	0,106	0,024	0,119	30	0,019	0,048	0,016	0,053
16	0,025	0,098	0,023	0,113	31	0,018	0,047	0,016	0,051
17	0,024	0,093	0,023	0,107	32	0,018	0,045	0,015	0,050
18	0,024	0,087	0,022	0,101	33	0,018	0,044	0,015	0,048
19	0,023	0,083	0,022	0,096	34	0,017	0,043	0,014	0,046
20	0,023	0,077	0,021	0,090	35	0,017	0,041	0,014	0,045
21	0,022	0,074	0,020	0,085					

Статистика: W = $\frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - X_m)}$, $x_{(i)}$ – вариационный ряд, X_m – выборочное среднее

$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)}, m = (m_1, \dots, m_n)^T$ - матожидания порядковых статистик стандартного нормального распределения, V – ковариационная матрица.

Плюсы и минусы непараметрических критериев

Плюсы

- Отсутствие требований к распределению данных
- Устойчивость к выбросам
- Меньшие требования к размеру выборки
- Простота в применении и интерпретации (но не для всех критериев)

Минусы

- Меньшая мощность теста
- Менее информативные результаты
- Проблемы с большими выборками
- Игнорирование величины различий

Отличия параметрических от непараметрических критериев

Параметрические критерии	**Непараметрические критерии**
Предполагают нормальное распределение данных, гомогенность дисперсий	Не требуют предположений о распределении, подходят для любых типов распределений
Числовые данные, интервальная или шкала отношений	Порядковые или количественные данные, допускающие ранжирование
Чувствительны к выбросам и нарушениям нормальности	Более устойчивы к выбросам и подходят для сильно смещённых данных
Мощность выше, если данные соответствуют предположениям (нормальность и т.п.)	Меньше мощность, но подходят для разнообразных распределений
Прямое понимание разницы в значениях (например, средних)	Оценка различий в рангах или медианах, а не в средних

Спасибо за внимание!