Лекция 7.

Курс «Продуктовая аналитика»

Коваленко Ангелина

Повторение

Повторение. Проверка гипотез

- В проверке гипотез делается предположение о распределении данных, и наша задача состоит в том, чтобы определить, содержит ли выборка достаточно информации, чтобы отвергнуть это предположение или нет

- Чтобы иметь возможность отвергнуть предположение, нам необходимо зафиксировать альтернативу - другое предположение о распределении данных, относительно которого мы будем решать, отвергать основную гипотезу или нет

- Альтернативная гипотеза шире нулевой

Проверка гипотез

Правило, позволяющее принять или отвергнуть нулевую гипотезу на основании выборки называется **статистическим критерием**

Обычно статистический критерий задается с помощью функции от выборки $T(x_1, \ldots, x_n)$ - статистики критерия

Проверка гипотез

Статистика любого критерия $T(x_1, \ldots, x_n)$ должна обладать двумя важными свойствами

- При верной H_0 статистика T должна принимать умеренные значения, а при неверной H_0 - другие, экстремальные

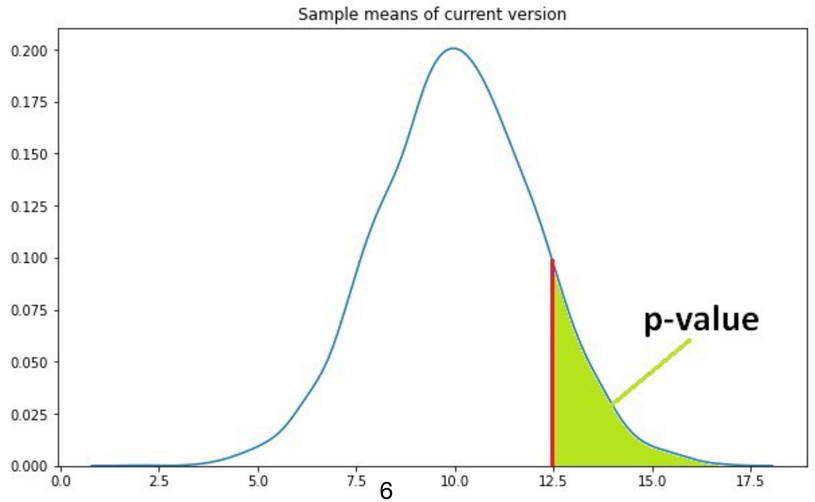
- При верной H_0 статистика T должна иметь известное распределение G_0, а при неверной H_0 - распределение отличное от G_0 (возможно неизвестное)

Проверка гипотез

Фактический уровень значимости или p-value - вероятность для статистики T при верной H_0 принять значение $t = T(x)$, которое получилось на выборке $x = (x_1, . . , x_n)$, или еще более экстремальное

Если для статистики T экстремальными значения являются большие значения, то это можно записать так:

$$p(x) = \mathbb{P}(T(X) \geqslant t \mid H_0)$$



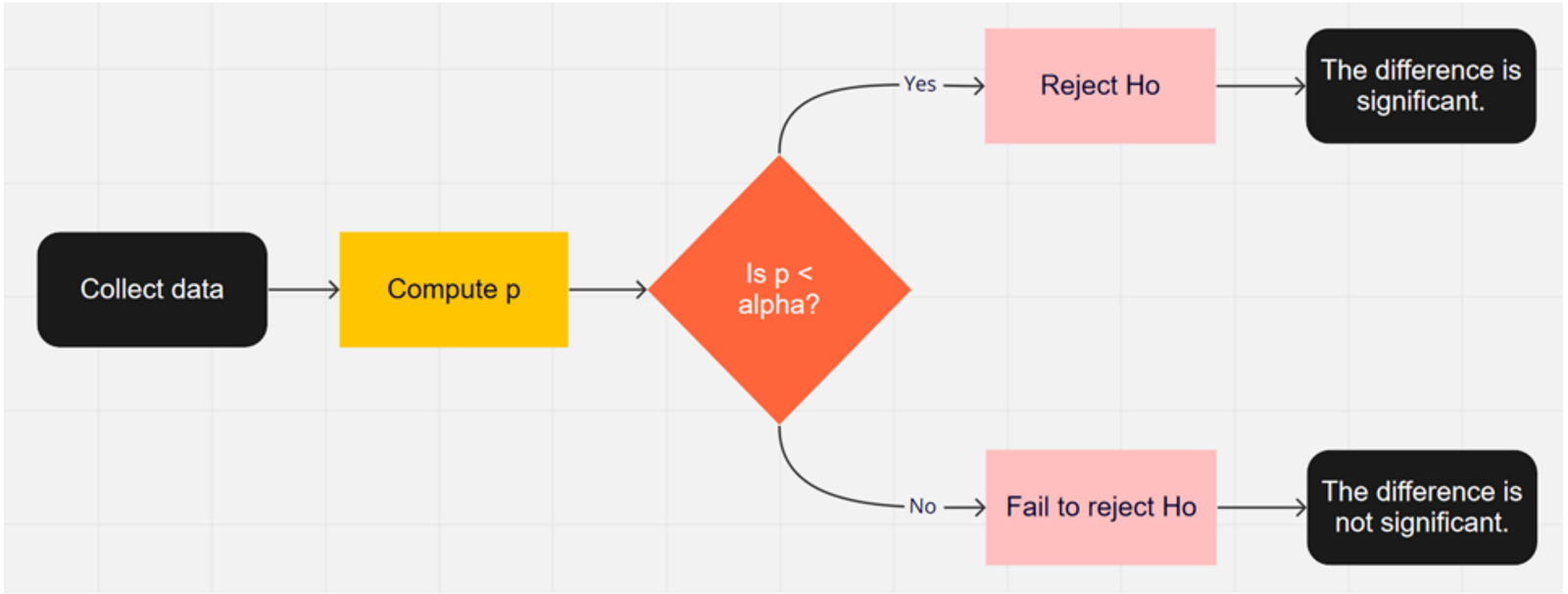
Проверка гипотез

Если p-value мало, то данные свидетельствуют против нулевой гипотезы H_0 в пользу альтернативы H_1.

Если значения p-value недостаточно мало, то данные не свидетельствуют против H_0 в пользу альтернативы H_1

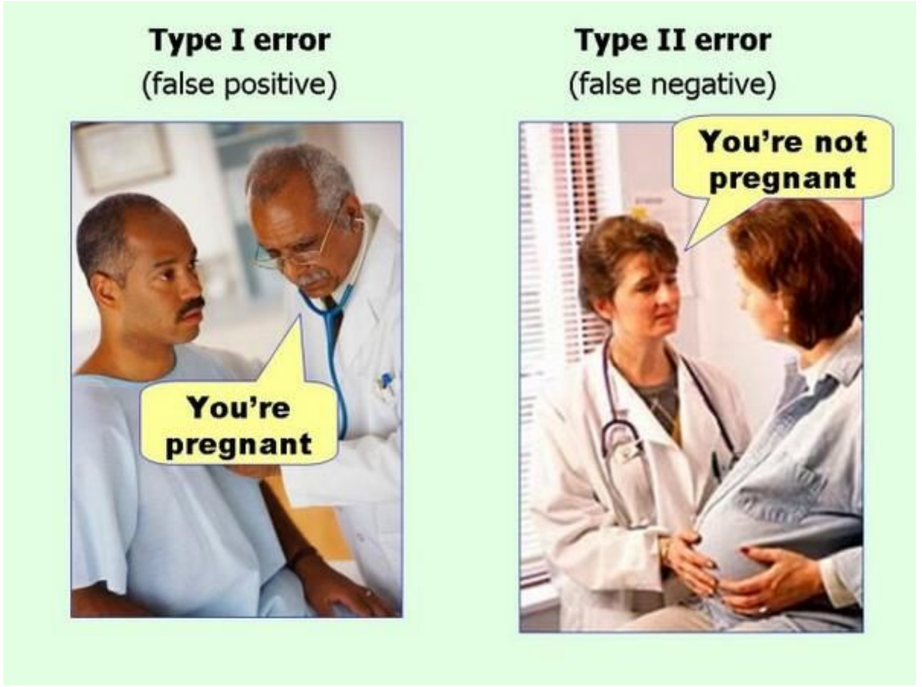
Проверка гипотез

Нулевая гипотеза H_0 отвергается при $p(x) \leqslant \alpha, \alpha$ - уровень значимости, который **мы** задаем



Проверка гипотез

Hypothesis Testing Errors	Null Hypothesis not rejected	Null hypothesis rejected
Null hypothesis is TRUE	Correct Conclusion	ERROR - Type A
Null Hypothesis is FALSE	ERROR - Type B	Correct Conclusion



Качество критерия определяется вероятностями двух этих ошибок. Эти вероятности обозначают α_I и α_{II} соответственно. Величину 1 – α_{II} называют мощностью.

Когда применяется проверка гипотез в жизни?

- Правда ли что маркетинговая кампания дала какой-то эффект на пользователей которые ее увидели?
- Правда ли что маркетинговая кампания А работает лучше чем маркетинговая кампания В?
- Правда ли что одна группа пользователей отличается от другой?
- Правда ли что изменение дало положительный эффект и пользователи стали больше времени проводить на сайте?
- А/Б тесты - правда ли что изменение, которое мы тестируем, даст положительный эффект?

Что мы можем оценивать?

- Доли (пропорции/конверсии) vs Среднее (среднее/медиана/мода)
- Зависимые vs Независимые выборки
- Параметрические vs непараметрические распределения
- Можем оценивать изменение в целом, а можем в конкретную сторону (стало лучше/хуже)

Критерии для проверки гипотез

Критерии согласия

Проверка распределений

Критерии однородности

Сравнение выборок

Параметрические

Непараметрические

Критерии согласия

Пусть нам дана выборка $X_1, \ldots, X_n \sim F$, где F - некоторое неизвестное нам распределение. В критериях согласия в качестве H_0 рассматривается гипотеза о принадлежности F какому-то параметрическому семейству. Альтернативой считаем принадлежность всем остальным распределениям.

В критериях согласия рассматриваются:

- $H_0 : F \in \mathbf{F}_\theta$ (нулевая гипотеза)
- $H_1 : F \notin \mathbf{F}_\theta$ (альтернативная гипотеза)

Где $\mathbf{F}_\theta$ - некоторое параметрическое семейство распределений

Проверка нормальности

Под гипотезой нормальности понимается сложная гипотеза

$$H_0 : F \in \left\{ F_{\mu,\sigma \in \mathbb{R}, \sigma > 0} \right\}$$

где класс $\left\{ F_{\mu,\sigma} \right\}_{\mu \in \mathbb{R}, \sigma > 0}$ образуют функции распределение нормального закона.

Напомним, что по определению $Y \sim N\left(\mu, \sigma^2\right)$, если $Y = \mu + \sigma X$, где $X \sim N(0,1)$. Поэтому можно записать, что

$$F_{\mu,\sigma}(x) = \Phi(\frac{x - \mu}{\sigma}),$$

где Φ - функция распределения $N(0,1)$.

Проверка нормальности

(a) Критерий Шапиро-Уилка (Shapiro-Wilk)

Критерий Шапиро-Уилка базируется на статистике, которая является отношением квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии:

$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$, где a_i- некоторые константы.

При верной H_0 распределение SW_n является табличным. На этом факте и основан критерий Шапиро-Уилка.

Проверка нормальности

(б) Критерий Харке-Бера (Jarque-Bera)

Этот критерий основан на статистике, которая использует выборочные коэффициенты асимметрии и эксцесса:

$$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3\mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2},$$

где μ_k - центрированный выборочный момент порядка k

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k.$$

Данная статистика сходится к распределению χ_2^2. На этом факте и основан критерий Харке-Бера.

Критерии однородности

Параметрические и непараметрические критерии

Мы изучим параметрические и непараметрические критерии.

В параметрических критериях делается предположение о том, что выборки взяты из некоторого параметрического семейства распределений, а в непараметрических — нет.

Параметрические и непараметрические критерии

В целом, непараметрические критерии менее чувствительные (потому что более общие), но зато они не требуют идеальных условий, например, нормальности данных.

Часто бывает так, что при совсем небольших отклонениях от идеальных условий непараметрические критерии работают значительно лучше параметрических

Параметрические критерии

Параметрические критерии для долей

Одновыборочный Z-критерий для доли

Выборка: $\boldsymbol{X} = (X_1, \ldots, X_n),$ где $X_i \sim Ber(p)$

Нулевая гипотеза: $H_0 : p = p_0$

Альтернатива: $H_1 : p \neq p_0$ или $p > p_0$ или $p < p_0$

Статистика: $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Нулевое распределение: $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \to N(0,1)$$

Одновыборочный Z-критерий для доли

Выборка: $X = (X_1, \ldots, X_n),$ где $X_i \sim Ber(p)$

Нулевая гипотеза: $H_0 : p = p_0$

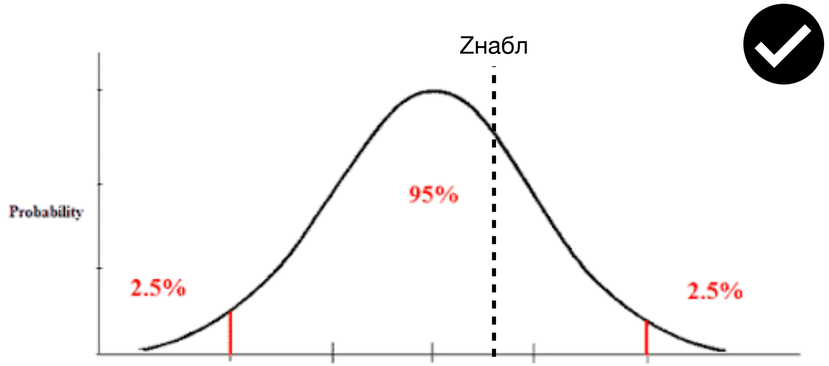
Альтернатива: $H_1 : p \neq p_0$ или $p > p_0$ или $p < p_0$

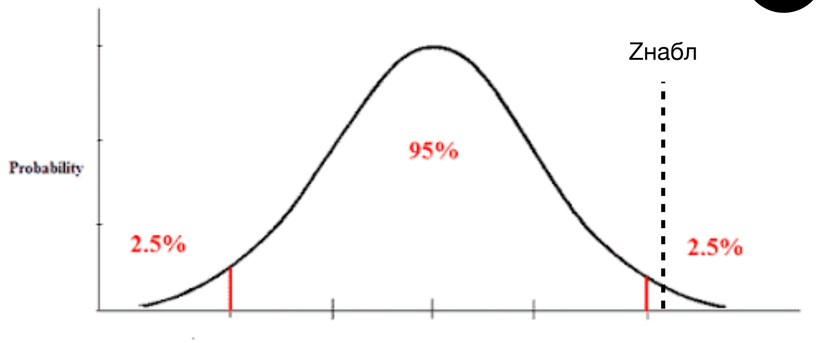
Статистика:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Нулевое распределение:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \to N(0,1)$$





Одновыборочный Z-критерий (для доли)

Пример:

Вася работает лидом тестировщиков. Он знает, что 20% приходящих в тестирование задач имеют баги. Он порекомендовал изменить процесс код ревью перед отправкой в тестирование и хочет проверить, изменилась ли частота ошибок. За следующий месяц в тестирование ушло 400 задач, Вася обнаружил, что 60 из них имеют ошибку.

$H_0 : p = 0.2$

$H_1 : p \neq 0.2$

Одновыборочный Z-критерий (для доли)

Пример:

Вася работает лидом тестировщиков. Он знает, что 20% приходящих в тестирование задач имеют баги. Он порекомендовал изменить процесс код ревью перед отправкой в тестирование и хочет проверить, изменилась ли частота ошибок. За следующий месяц в тестирование ушло 400 задач, Вася обнаружил, что 60 из них имеют ошибку.

$$H_0 : p = 0.2$$

$$H_1 : p \neq 0.2$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Одновыборочный Z-критерий (для доли)

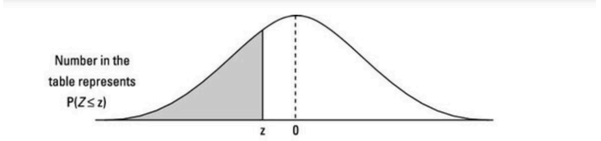
Пример:

Вася работает лидом тестировщиков. Он знает, что 20% приходящих в тестирование задач имеют баги. Он порекомендовал изменить процесс код ревью перед отправкой в тестирование и хочет проверить, изменилась ли частота ошибок. За следующий месяц в тестирование ушло 400 задач, Вася обнаружил, что 60 из них имеют ошибку.

$$H_0 : p = 0.2$$

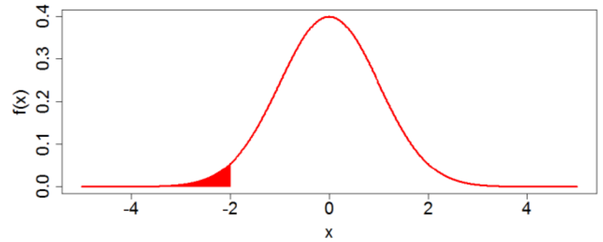
$$H_1 : p \neq 0.2$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

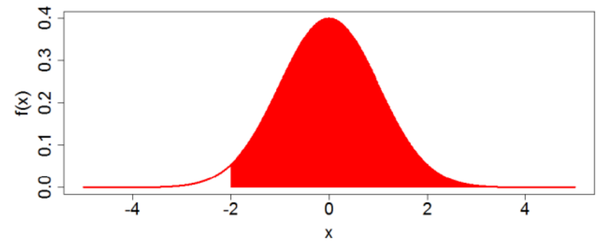


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

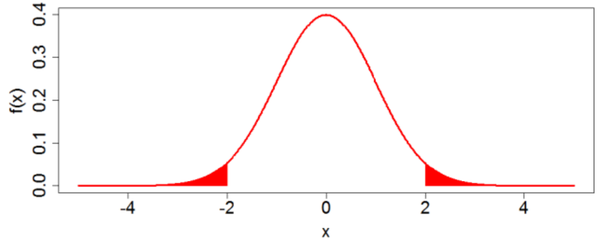
Одновыборочный Z-критерий (для доли)



$$H_1 : \ \mu < \mu_0$$



$$H_1 : \ \mu > \mu_0$$



$$H_1 : \ \mu \neq \mu_0$$

Одновыборочный Z-критерий (для доли)

Пример:

Вася работает лидом тестировщиков. Он знает, что 20% приходящих в тестирование задач имеют баги. Он порекомендовал изменить процесс код ревью перед отправкой в тестирование и хочет проверить, изменилась ли частота ошибок **в лучшую сторону**. За следующий месяц в тестирование ушло 400 задач, Вася обнаружил, что 60 из них имеют ошибку.

$$H_0 : p = 0.2$$

$$H_1 : p < 0.2$$

Z-критерий для разности независимых долей

Выборки: $\boldsymbol{X} = (X_1, \ldots, X_n)$, где $X_i \sim Ber(p)$

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, где $Y_i \sim Ber(p)$, выборки независимы,

Нулевая гипотеза: $H_0 : p_x = p_y$

Альтернатива: $H_1 : p_x \neq p_y$

Статистика: $z = \dfrac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$, где $P = \dfrac{m_x + m_y}{n_x + n_y}$, m_i - число 1 в выборке

Нулевое распределение: $z = \dfrac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}} \to N(0,1)$

Z-критерий для разности независимых долей

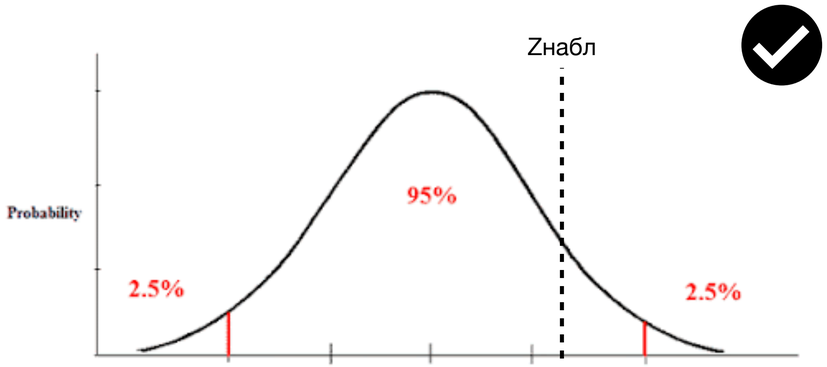
Выборки: $\boldsymbol{X} = (X_1, \ldots, X_n)$, где $X_i \sim Ber(p)$

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, где $Y_i \sim Ber(p)$, выборки независимы

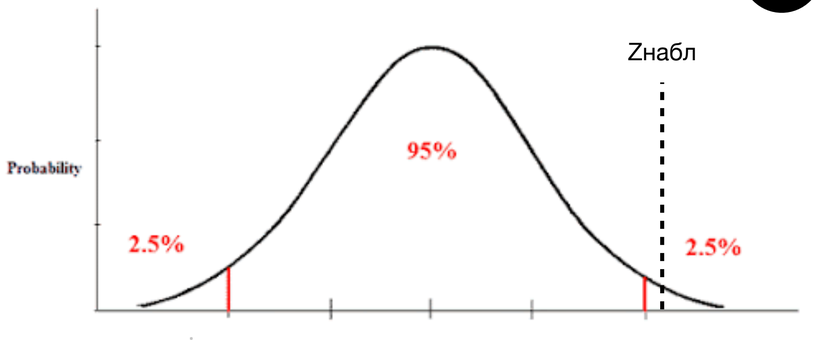
Нулевая гипотеза: $H_0 : p_x = p_y$

Альтернатива: $H_1 : p_x \neq p_y$

Статистика: $$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}, \text{ где } P = \frac{m_x + m_y}{n_x + n_y},$$ m_i - число 1 в выбоке



Нулевое распределение: $$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}} \to N(0,1)$$



Пример

Результаты опроса пользователей о том нравится ли им дизайн приложения

Нравится дизайн	Школьники	Взрослые	Всего
Да	58	52	110
Нет	42	48	90
Всего	100	100	200
Доля	0.58	0.52	0.55

Правда ли что школьникам сайт нравится больше?

Пример

Результаты опроса пользователей о том нравится ли им дизайн приложения

Нравится дизайн	Школьники	Взрослые	Всего
Да	58	52	110
Нет	42	48	90
Всего	100	100	200
Доля	0.58	0.52	0.55

Правда ли что школьникам сайт нравится больше?

$$H_0 : p_s = p_a$$

$$H_1 : p_a < p_s$$

Пример

Результаты опроса пользователей о том нравится ли им дизайн приложения

Нравится дизайн	Школьники	Взрослые	Всего
Да	58	52	110
Нет	42	48	90
Всего	100	100	200
Доля	0.58	0.52	0.55

Правда ли что школьникам сайт нравится больше?

$$H_0 : p_s = p_a$$

$$H_1 : p_a < p_s$$

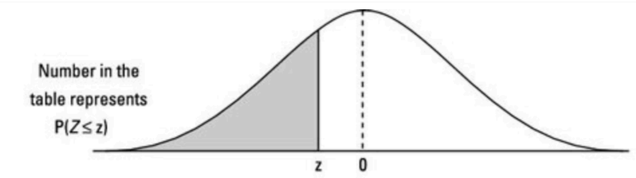
$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$, где $P = \frac{m_x + m_y}{n_x + n_y}$, m_i - число 1 в выборке

Пример

Результаты опроса пользователей о том нравится ли им дизайн приложения

Нравится дизайн	Школьники	Взрослые	Всего
Да	58	52	110
Нет	42	48	90
Всего	100	100	200
Доля	0.58	0.52	0.55

Есть ли разница между двумя группами?



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867

Параметрические критерии для средних

Одновыборочный Z-критерий для среднего

Выборка: $\boldsymbol{X} = (X_1, \ldots, X_n),$ где $X_i \sim N(\mu, \sigma^2)$, σ - известна

Нулевая гипотеза: $H_0 : \mu = \mu_0$

Альтернатива: $H_1 : \mu \neq \mu_0$ или $\mu > \mu_0$ или $\mu < \mu_0$

Статистика: $$Z_n = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Нулевое распределение: $Z_n \to N(0,1)$

Одновыборочный Z-критерий для среднего

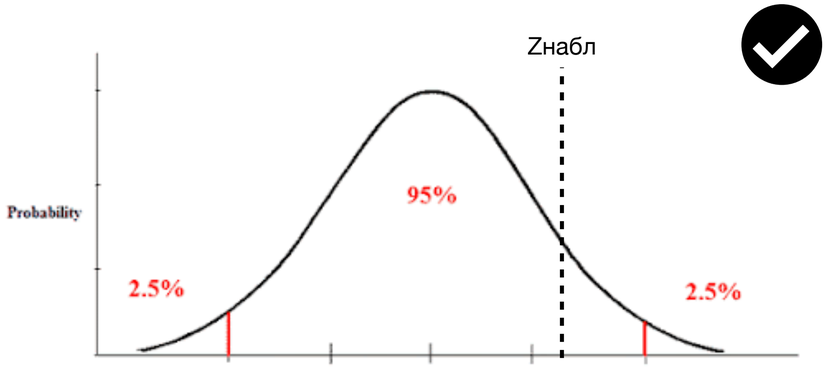
Выборка: $\boldsymbol{X} = (X_1, \ldots, X_n)$, где $X_i \sim N(\mu, \sigma^2)$, σ - известна

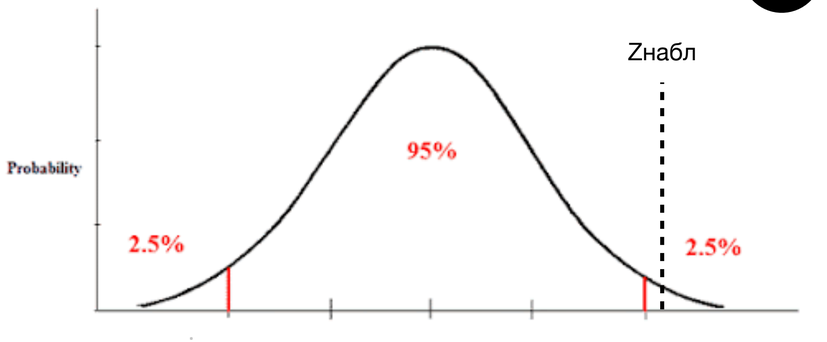
Нулевая гипотеза: $H_0 : \mu = \mu_0$

Альтернатива: $H_1 : \mu \neq \mu_0$ или $\mu > \mu_0$ или $\mu < \mu_0$

Статистика: $Z_n = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Нулевое распределение: $Z_n \to N(0,1)$





Одновыборочный t-критерий для среднего

Выборка: $X = (X_1, \ldots, X_n),$ где $X_i \sim N(\mu, \sigma^2)$, σ - **неизвестна**

Нулевая гипотеза: $H_0 : \mu = \mu_0$

Альтернатива: $H_1 : \mu \neq \mu_0$ или $\mu > \mu_0$ или $\mu < \mu_0$

Статистика: $Z_n = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$, где $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Нулевое распределение: $T_n \to t_{n-1}$

Двухвыборочный t-критерий (Уэлча) для разности независимых выборок

Выборки:

$$\boldsymbol{X} = (X_1, .., X_n), \text{ где } X_i \sim N(\mu_1, \sigma_1^2)$$

$$\boldsymbol{Y} = (Y_1, .., Y_n), \text{ где } Y_i \sim N(\mu_2, \sigma_2^2),$$

X, Y независимы, σ_1, σ_2 неизвестны

Нулевая гипотеза:

$$H_0 : \mu_1 = \mu_2$$

Альтернатива:

$$H_1 : \mu_1 \neq \mu_2 \text{ или } \mu_1 > \mu_2 \text{ или } \mu_1 < \mu_2$$

Статистика:

$$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \text{ где } S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Нулевое распределение:

$$T_n \approx t_k$$

Двухвыборочный t-критерий (Уэлча) для разности независимых выборок

Вы запустили сайт для просмотра сериалов и монетизируете его через показ рекламы. Маркетолог предположил, что ему стоит больше таргетироваться на женщин, потому что прочитал, что они дольше времени проводят за просмотром сериалов, а значит увидят больше рекламы. Проверьте гипотезу маркетолога.

	Male	Female
Mean	1.3h	1.6h
Standard deviation	0.5h	0.3h
Number of users	22	24

Двухвыборочный t-критерий (Уэлча) для разности независимых выборок

Задача сравнения средних двух нормальных выборок при неизвестных и неравных дисперсиях известна как проблема Беренса-Фишера. Точного решения этой задачи нет. Однако рассмотренная аппроксимация (критерий Уэлча) достаточно точна в двух ситуациях:

1. Если выборки одинакового объема $n1 = n2$.

2. Если знак неравенства между $n1$ и $n2$ такой же, как между $\sigma1$ и $\sigma2$, то есть выборка с большей дисперсией имеет больший объем.

z-test vs t-test в теории

Z-test

- Распределение нормальное и дисперсия известна
- Распределение биномиальное, $np>10$, $nq>10$

T-test

- Распределение нормальное, дисперсия неизвестна, $n < 30$

z-test vs t-test на практике

Z-test

- Распределение нормальное и дисперсия известна - считай, никогда
- Распределение биномиальное, np>10, nq>10

T-test

- Распределение нормальное, дисперсия неизвестна, ~~n < 30~~

z-test vs t-test на практике

Z-test

- Распределение нормальное и дисперсия известна - считай, никогда
- Распределение биномиальное, np>10, nq>10

T-test

- Распределение нормальное, дисперсия неизвестна, ~~n < 30~~

Для нормального распределения всегда используйте t-тест, если заранее не знаете дисперсию генеральной совокупности (читай - всегда). t-распределение сходится к нормальному, поэтому это правильное распределение для любого N.

Не нужно беспокоиться о том, когда перейти на z-тест, потому что t-распределение «переключается» за вас.

Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Критерии Согласия Проверяем принадлежность классу распределений (согласованность выборки с распределением)	Общие	Непрерывные распределения	$H_0 : F \in \mathbf{F}_\theta$	Критерий Колмогорова	$D_n = sup \lvert \widehat{F_n}(u) - F_0(u) \rvert$	Максимальное расстояние между функциями распределений	`scipy.stats.kstest`
		Дискретные распределения		Критерий Пирсона (хи-квадрат)	$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$	Отклонения количества каждых значений от ожидаемых количеств этих значений	`scipy.stats.chisquare`
	Специальные	Нормальное распределение	$H_0 : F \sim N(\mu, \sigma)$	Критерий Шапиро-Уилка	$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$	Отношение квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии	`scipy.stats.shapiro`
				Критерий Харке-Бера	$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3 \mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2}$	Стандартизировнные коэффициенты эксцесса и асимметрии	`scipy.stats.jarque_bera`
Параметрические Проверяем параметры (среднее/ пропорции), предполагая, что выборки пришли из конкретного распределения	Одновыборочные	$X_i \sim Ber(p)$	$H_0 : p = p_0$	Z-критерий	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.proportion.proportions_ztest`
		$X_i \sim N(\mu, \sigma^2)$ σ неизвестна	$H_0 : \mu = \mu_0$	t-критерий Стьюдента	$T_n = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_1samp`
	Двухвыборочные	$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки независимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$X_i \sim N(\mu_x, \sigma_x^2), Y_i \sim N(\mu_y, \sigma_y^2)$ σ_x, σ_y неизвестны (и могут быть не равны) выборки независимы	$H_0 : \mu_1 = \mu_2$	Двухвыборочный t-критерий Уэлча	$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_ind`
Непараметрические							
Критерии корреляции							

Двухвыборочные критерии для зависимых выборок

Двухвыборочные критерии. Зависимые выборки.

Рассмотрим теперь случаи связанных выборок, когда элементы X_i и Y_i соответствуют одному и тому же объекту, но измерения сделаны в разные моменты

Размеры выборок в этом случае должны совпадать

$$n_1 = n_2 = n_3$$

Двухвыборочный t-критерий для зависимых выборок.

Рассмотрим выборку, образованную разностями X_i и Y_i:

$$Z_i = Y_i - X_i, i = 1,..,n$$

Сравнение среднего в зависимых выборках ничем не отличается от сравнения среднего разности Z_i с нулем

Двухвыборочный t-критерий для зависимых выборок.

Выборки: $\boldsymbol{X} = (X_1, .., X_n),\ \boldsymbol{Y} = (Y_1, .., Y_n),\ Z_i = Y_i - X_i$ и $Z_i \sim N(\mu, \sigma^2)$

X, Y зависимые, σ неизвестна

Нулевая гипотеза: $H_0 : \mu = 0$

Альтернатива: $H_1 : \mu \neq 0$ или $\mu > 0$ или $\mu < 0$

Статистика: $T_n = \dfrac{\overline{Z}}{S\sqrt{n}}$, где $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$

Нулевое распределение: $T_n \sim t_{n-1}$

Z-критерий для разности зависимых долей

Выборки: $\boldsymbol{X} = (X_1, \ldots, X_n),$ где $X_i \sim Ber(p)$

$\boldsymbol{Y} = (Y_1, \ldots, Y_n),$ где $Y_i \sim Ber(p)$, выборки зависимы

Нулевая гипотеза: $H_0 : p_x = p_y$

Альтернатива: $H_1 : p_x \neq p_y$

Статистика:

$$z = \frac{c - b}{\sqrt{c + b - \frac{(c-b)^2}{n}}}$$

Y \ X	0	1
0	a	b
1	c	d

Пример

- В январе маркетолог отпросил сотню пользователей нравится ли им дизайн приложения
- Через два месяца у этих же ста человек спросил тот же вопрос
- Правда ли, что число тех, кому нравится дизайн изменилось

		Февраль	
		0	1
Апрель	0	20	10
	1	20	50

Пример

Апрель \ Февраль	0	1
0	20	10
1	20	50

$$\boldsymbol{X} = (X_1, \ldots, X_n), \text{ где } X_i \sim Ber(p)$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n), \text{ где } Y_i \sim Ber(p)$$

Выборки зависимы

$$H_0 : p_x = p_y$$

$$H_1 : p_x \neq p_y$$

Пример

Февраль

Апрель		0	1
	0	20	10
	1	20	50

$$z_{obs} = \frac{20 - 10}{\sqrt{20 + 10 - \frac{\sqrt{(20-10)^2}}{100}}}$$

$$\boldsymbol{X} = (X_1, \ldots, X_n), \text{ где } X_i \sim Ber(p)$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n), \text{ где } Y_i \sim Ber(p)$$

Выборки зависимы

$$H_0 : p_x = p_y$$

$$H_1 : p_x \neq p_y$$

Пример

Февраль

Апрель		0	1
	0	20	10
	1	20	50

$$z_{obs} = \frac{20 - 10}{\sqrt{20 + 10 - \frac{\sqrt{(20-10)^2}}{100}}}$$

$$z_{0.975} = 1.96$$

$$X = (X_1, \ldots, X_n), \text{ где } X_i \sim Ber(p)$$

$$Y = (Y_1, \ldots, Y_n), \text{ где } Y_i \sim Ber(p)$$

Выборки зависимы

$$H_0 : p_x = p_y$$

$$H_1 : p_x \neq p_y$$

Пример

Апрель \ Февраль	0	1
0	20	10
1	20	50

$$z_{obs} = \frac{20 - 10}{\sqrt{20 + 10 - \frac{\sqrt{(20-10)^2}}{100}}}$$

$$z_{0.975} = 1.96$$

$$X = (X_1, \ldots, X_n), \text{ где } X_i \sim Ber(p)$$

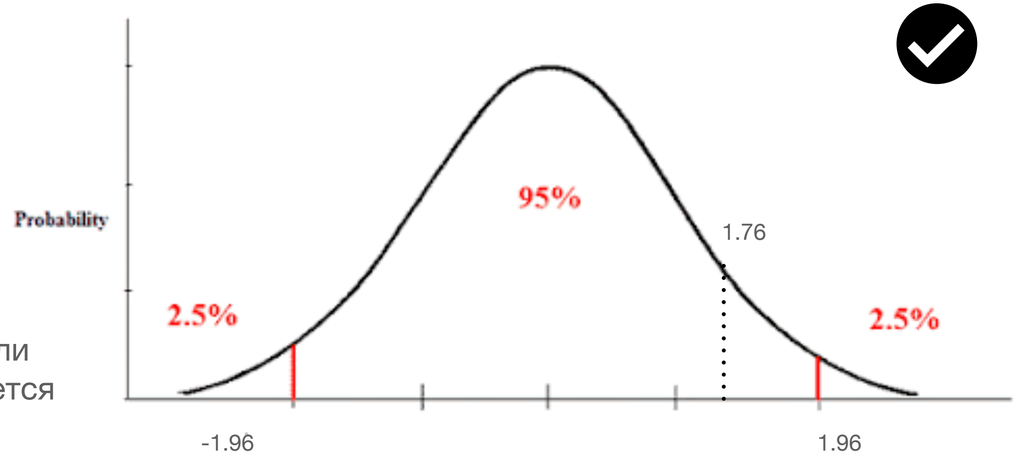
$$Y = (Y_1, \ldots, Y_n), \text{ где } Y_i \sim Ber(p)$$

Выборки зависимы

$$H_0 : p_x = p_y$$

$$H_1 : p_x \neq p_y$$

Гипотеза о том что пользователи не поменяли мнения не отвергается



Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Критерии Согласия Проверяем принадлежность классу распределений (согласованность выборки с распределением)	Общие	Непрерывные распределения	$H_0 : F \in \mathbf{F}_\theta$	Критерий Колмогорова	$D_n = sup \lvert \widehat{F_n}(u) - F_0(u) \rvert$	Максимальное расстояние между функциями распределений	`scipy.stats.kstest`
		Дискретные распределения		Критерий Пирсона (хи-квадрат)	$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$	Отклонения количества каждых значений от ожидаемых количеств этих значений	`scipy.stats.chisquare`
	Специальные	Нормальное распределение	$H_0 : F \sim N(\mu, \sigma)$	Критерий Шапиро-Уилка	$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$	Отношение квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии	`scipy.stats.shapiro`
				Критерий Харке-Бера	$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3 \mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2}$	Стандартизировнные коэффициенты эксцесса и асимметрии	`scipy.stats.jarque_bera`
Параметрические Проверяем параметры (среднее/ пропорции), предполагая, что выборки пришли из конкретного распределения	Одновыборочные	$X_i \sim Ber(p)$	$H_0 : p = p_0$	Z-критерий	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.proportion.proportions_ztest`
		$X_i \sim N(\mu, \sigma^2)$ σ неизвестна	$H_0 : \mu = \mu_0$	t-критерий Стьюдента	$T_n = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_1samp`
	Двухвыборочные	$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки независимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$X_i \sim N(\mu_x, \sigma_x^2), Y_i \sim N(\mu_y, \sigma_y^2)$ σ_x, σ_y неизвестны (и могут быть не равны) выборки независимы	$H_0 : \mu_1 = \mu_2$	Двухвыборочный t-критерий Уэлча	$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_ind`
Непараметрические							
Критерии корреляции							

Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Критерии Согласия Проверяем принадлежность классу распределений (согласованность выборки с распределением)	Общие	Непрерывные распределения	$H_0 : F \in \mathbf{F}_\theta$	Критерий Колмогорова	$D_n = sup \lvert \widehat{F_n}(u) - F_0(u) \rvert$	Максимальное расстояние между функциями распределений	`scipy.stats.kstest`
		Дискретные распределения		Критерий Пирсона (хи-квадрат)	$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$	Отклонения количества каждых значений от ожидаемых количеств этих значений	`scipy.stats.chisquare`
	Специальные	Нормальное распределение	$H_0 : F \sim N(\mu, \sigma)$	Критерий Шапиро-Уилка	$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$	Отношение квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии	`scipy.stats.shapiro`
				Критерий Харке-Бера	$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3 \mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2}$	Стандартизировнные коэффициенты эксцесса и асимметрии	`scipy.stats.jarque_bera`
Параметрические Проверяем параметры (среднее/ пропорции), предполагая, что выборки пришли из конкретного распределения	Одновыборочные	$X_i \sim Ber(p)$	$H_0 : p = p_0$	Z-критерий	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.proportion.proportions_ztest`
		$X_i \sim N(\mu, \sigma^2)$ σ неизвестна	$H_0 : \mu = \mu_0$	t-критерий Стьюдента	$T_n = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_1samp`
	Двухвыборочные	$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки независимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$X_i \sim N(\mu_x, \sigma_x^2), Y_i \sim N(\mu_y, \sigma_y^2)$ σ_x, σ_y неизвестны (и могут быть не равны) выборки независимы	$H_0 : \mu_1 = \mu_2$	Двухвыборочный t-критерий Уэлча	$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_ind`
		$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки зависимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{c - b}{\sqrt{c + b - \frac{(c-b)^2}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
Непараметрические							
Критерии корреляции							

Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Критерии Согласия Проверяем принадлежность классу распределений (согласованность выборки с распределением)	Общие	Непрерывные распределения	$H_0 : F \in \mathbf{F}_\theta$	Критерий Колмогорова	$D_n = sup \mid \widehat{F_n}(u) - F_0(u) \mid$	Максимальное расстояние между функциями распределений	`scipy.stats.kstest`
		Дискретные распределения		Критерий Пирсона (хи-квадрат)	$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$	Отклонения количества каждых значений от ожидаемых количеств этих значений	`scipy.stats.chisquare`
	Специальные	Нормальное распределение	$H_0 : F \sim N(\mu, \sigma)$	Критерий Шапиро-Уилка	$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$	Отношение квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии	`scipy.stats.shapiro`
				Критерий Харке-Бера	$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3 \mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2}$	Стандартизировнные коэффициенты эксцесса и асимметрии	`scipy.stats.jarque_bera`
Параметрические Проверяем параметры (среднее/ пропорции), предполагая, что выборки пришли из конкретного распределения	Одновыборочные	$X_i \sim Ber(p)$	$H_0 : p = p_0$	Z-критерий	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.proportion.proportions_ztest`
		$X_i \sim N(\mu, \sigma^2)$ σ неизвестна	$H_0 : \mu = \mu_0$	t-критерий Стьюдента	$T_n = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_1samp`
	Двухвыборочные	$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки независимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$X_i \sim N(\mu_x, \sigma_x^2), Y_i \sim N(\mu_y, \sigma_y^2)$ σ_x, σ_y неизвестны (и могут быть не равны) выборки независимы	$H_0 : \mu_1 = \mu_2$	Двухвыборочный t-критерий Уэлча	$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_ind`
		$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки зависимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{c - b}{\sqrt{c + b - \frac{(c-b)^2}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$Z_i = Y_i - X_i, Z_i \sim N(\mu, \sigma^2)$ выборки зависимы	$H_0 : \mu = 0$	Двухвыборочный t-критерий	$T_n = \frac{\overline{Z}}{S\sqrt{n}}$	Аналог t-критерия стьюдента для одной выборки для статистики Y-X	`scipy.stats.ttest_rel`
Непараметрические							
Критерии корреляции							

Непараметрические критерии

Перейдем к **непараметрическим** критериям.

Мы не будем предполагать, что данные имеют нормальное распределение.

Достоинством непараметрических критерием является то, что они не требуют нормальности данных, пригодны для выборок малого размера и слабо реагируют на присутствие «выбросов» в данных. Начнем с модификации известных нам критериев.

Критерии знаков

Критерий знаков

- Превратим выборку в нули и единицы

Критерий знаков

- Превратим выборку в нули и единицы
- Часть информации о выборке будет потеряна

Критерий знаков

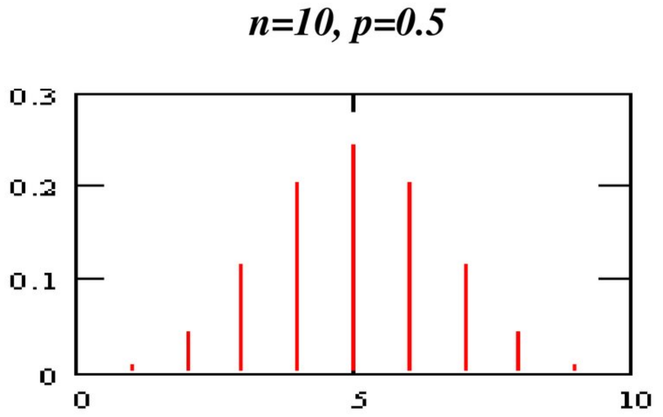
- Превратим выборку в нули и единицы
- Часть информации о выборке будет потеряна
- Но зато сможем воспользоваться биномиальным распредением

Критерий знаков

- Превратим выборку в нули и единицы:
  - Проверим что центр (медиана) распределения приходится на проверяемое число m_0 -
  - Присвоим 1 всем элементам большим m_0, 0 всем элементам меньше либо равным m_0

- Введем статистику $T = \sum_{i=1}^{n} [X_i > m_0]$
- $T = \sum_{i=1}^{n} [X_i > m_0] \sim Ber(0.5, n)$

Биномиальное распределение *Bin (n, p)*



$m = 10 \cdot 0{,}5 = 5 \quad D = 10 \cdot 0{,}5 \cdot 0{,}5 = 2{,}5 \quad k^* = 5$

Критерий знаков - одновыборочный

Выборка: $\mathrm{X} = (X_1, \ldots, X_n), X_i \sim F_X$

Нулевая гипотеза: $H_0 : Med(X) = m_0$

Альтернатива: $H_1 : Med(X) \neq m_0$ или $Med(X) > m_0$ или $Med(X) < m_0$

Статистика: $$T = \sum_{i-1}^{n} \left[X_i > m_0\right]$$

Нулевое распределение: $$T = \sum_{i-1}^{n} \left[X_i > m_0\right] \sim Ber(0.5,n)$$

Пример

Менеджер локального филиала банка S говорит, что среднее количество сберегательных счетов в день составляет 64. Консультант из этого же филиала считает, что это число больше 64. Консультант собрал статистику по открытиям счетов в день за 10 случайных дней. Можем ли мы опровергнуть утверждение менеджера с уровнем значимости 0.05?

День	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CC	60	66	65	70	68	72	46	63	77	75

Пример

Менеджер локального филиала банка S говорит, что среднее количество сберегательных счетов в день составляет 64. Консультант из этого же филиала считает, что это число больше 64. Консультант собрал статистику по открытиям счетов в день за 10 случайных дней. Можем ли мы опровергнуть утверждение менеджера с уровнем значимости 0.05?

День	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CC	60	66	65	70	68	72	46	76	77	75

$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

Пример

День	1	2	3	4	5	6	7	8	9	10
СС	60	66	65	70	68	72	46	76	77	75

$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
СС	0	1	1	1	1	1	0	0	1	1

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75

$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

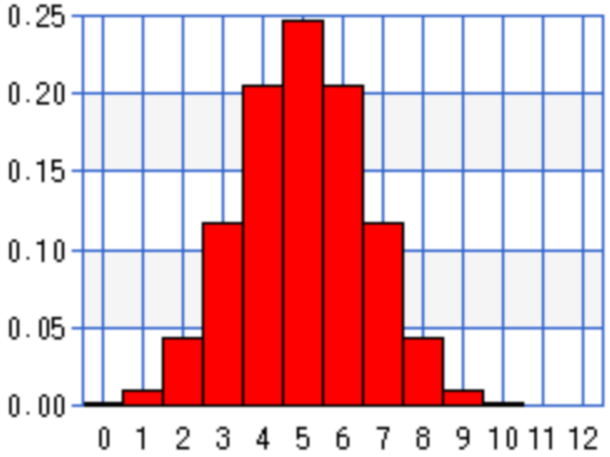
Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75



$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

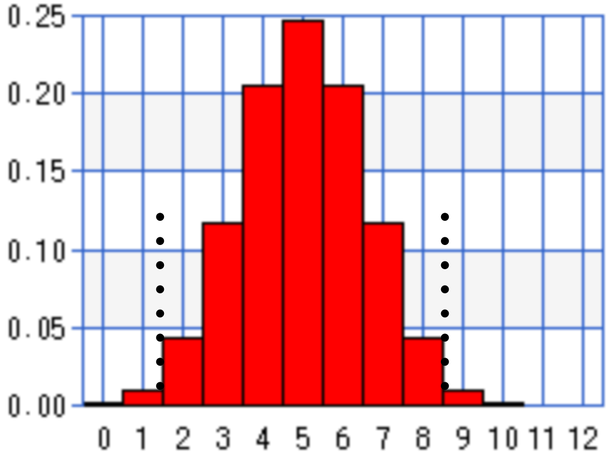
Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75



$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

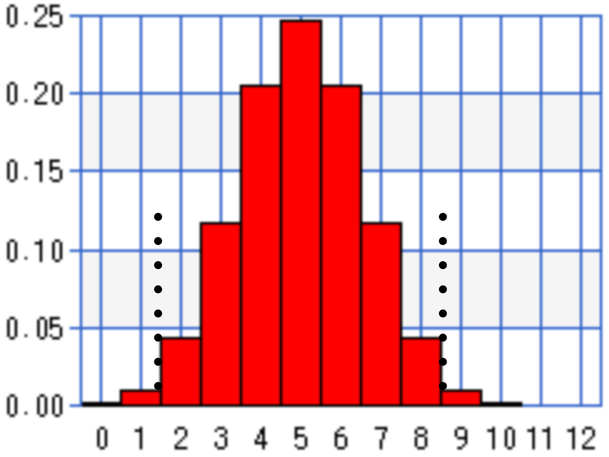
Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75



$$T_{crit} = 8$$

$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

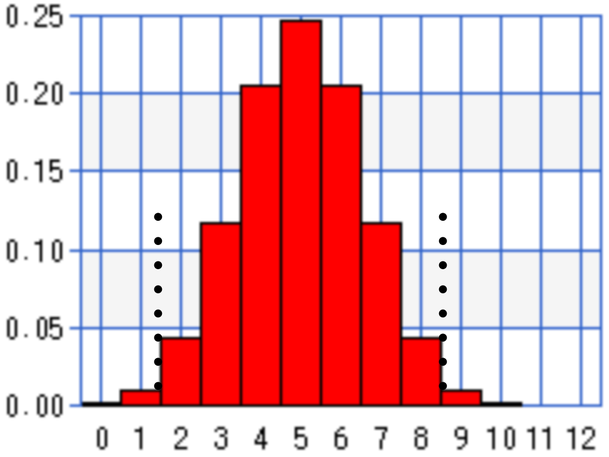
Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75



$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

$$T_{crit} = 8$$

$$T_{obs} < T_{crit}$$

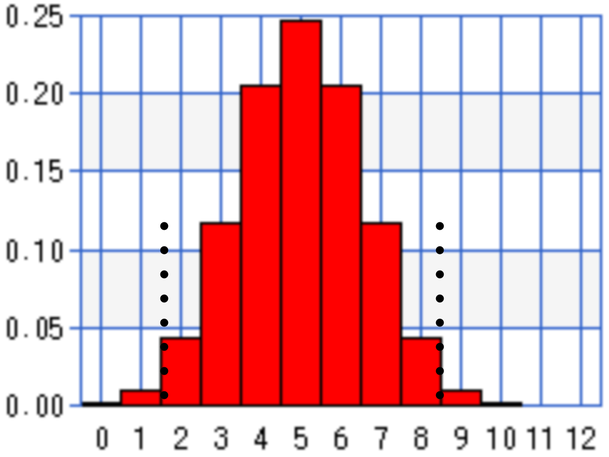
Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Пример

День	1	2	3	4	5	6	7	8	9	10
CC	60	66	65	70	68	72	46	76	77	75



$$H_0 : m_0 = 64$$

$$H_1 : m_0 > 64$$

$$T_{crit} = 8$$

$$T_{obs} < T_{crit}$$

H_0 не отвергается

Всем наблюдениям большим 64 присвоим 1, меньшим - 0

День	1	2	3	4	5	6	7	8	9	10
CC	0	1	1	1	1	1	0	1	1	1

$$T_{obs} = 7$$

Критерий знаков - двухвыборочный

Выборки: $\mathrm{X} = (X_1, \ldots, X_n), X_i \sim F_X$

$\mathrm{Y} = (Y_1, \ldots, Y_n), Y_i \sim F_Y$

X, Y зависимые

Нулевая гипотеза: $H_0 : \mathbb{P}(X > Y) = 0.5$

Альтернатива: $H_0 : \mathbb{P}(X > Y) \neq 0.5$

Статистика: $T = \sum_{i-1}^{n} \left[X_i > Y_i\right]$

Нулевое распределение: $T = \sum_{i-1}^{n} \left[X_i > Y_i\right] \sim Ber(0.5,n)$

Критерий знаков - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60

$$H_0 : P(Y > X) = 0.5$$

$$H_1 : P(Y > X) > 0.5$$

Критерий знаков - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
Sign	0	0	0	1	1	1	0	1	1	0

$$H_0 : P(Y > X) = 0.5$$

$$H_1 : P(Y > X) > 0.5$$

$$T_{obs} = 5$$

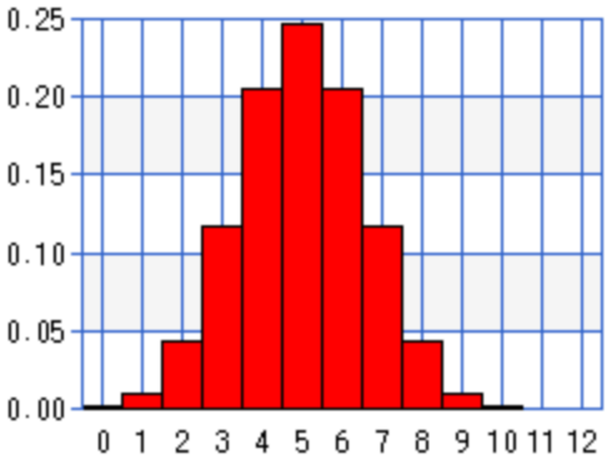
Критерий знаков - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
Sign	0	0	0	1	1	1	0	1	1	0

$$H_0 : P(Y > X) = 0.5$$

$$H_1 : P(Y > X) > 0.5$$



$$T_{obs} = 5$$

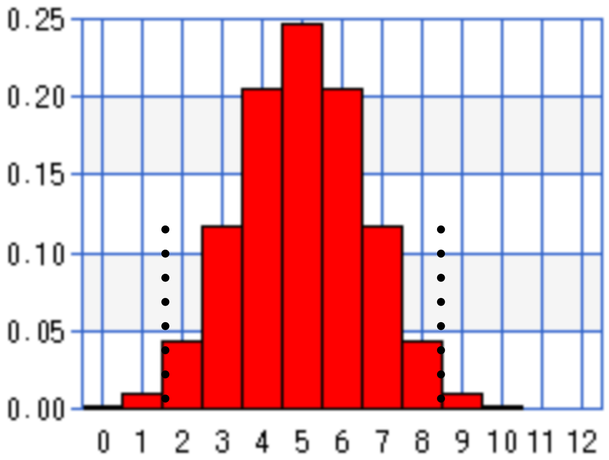
Критерий знаков - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
Sign	0	0	0	1	1	1	0	1	1	0

$$H_0 : P(Y > X) = 0.5$$

$$H_1 : P(Y > X) > 0.5$$



$$T_{obs} = 5$$

$$T_{crit} = 8$$

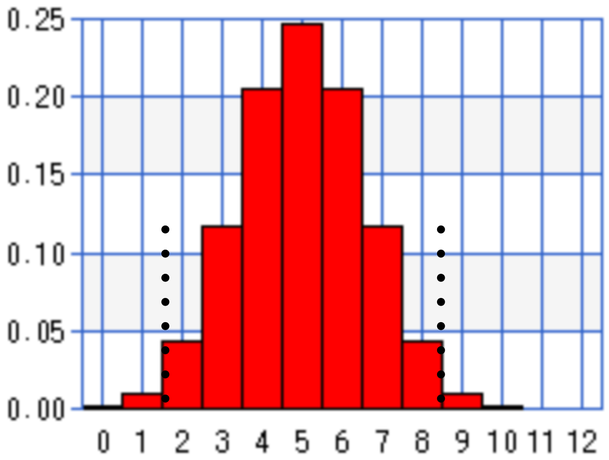
Критерий знаков - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
Sign	0	0	0	1	1	1	0	1	1	0

$$H_0 : P(Y > X) = 0.5$$

$$H_1 : P(Y > X) > 0.5$$



$$T_{obs} = 5$$

$$T_{crit} = 8$$

H_0 не отвергается

Критерии рангов

Критерий рангов

В критерии знаков, мы превращали наблюдения в 0 и 1, определяя в какую сторону от среднего отклонено наблюдение.

Тем самым мы теряли часть информации о наблюдении - насколько велико это отклонение.

Чтобы сохранять больше информации о выборке, можно превращать знаки в ранги.

Критерий рангов

Выборка $x_1, x_2, \ldots, x_n$

Упорядочим по возрастанию $x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$

Правило проставления ранга

- Порядковый номер наблюдения - ранг
- Если встречаются несколько одинаковых значений, им присваивается одинаковое значение ранга, равное среднему арифметическому их порядковых номеров

Критерий Уилкоксона (одновыборочный)

Выборки: $\mathrm{X} = (X_1, \ldots, X_n), X_i \sim F_X$

F_X симметрична относительно медианы

Нулевая гипотеза: $H_0 : Med(X) = m_0$

Альтернатива: $H_0 : Med(X) \neq m_0$

Статистика: $$W = \sum_{i-1}^{n} rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

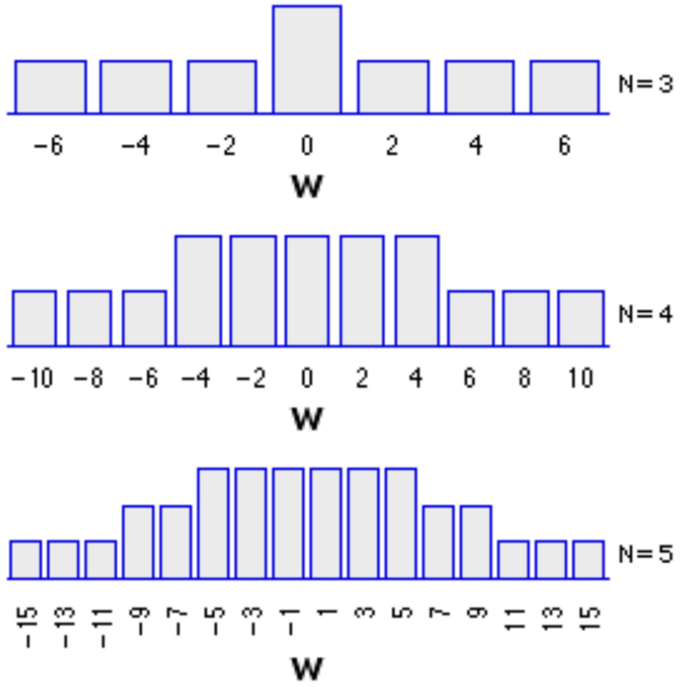
Нулевое распределение: У статистики табличное распределение

Статистика распределения

$$W = \sum_{i-1}^{n} rank(|X_i - m_0|) \cdot sign(X_i - m_0)$$

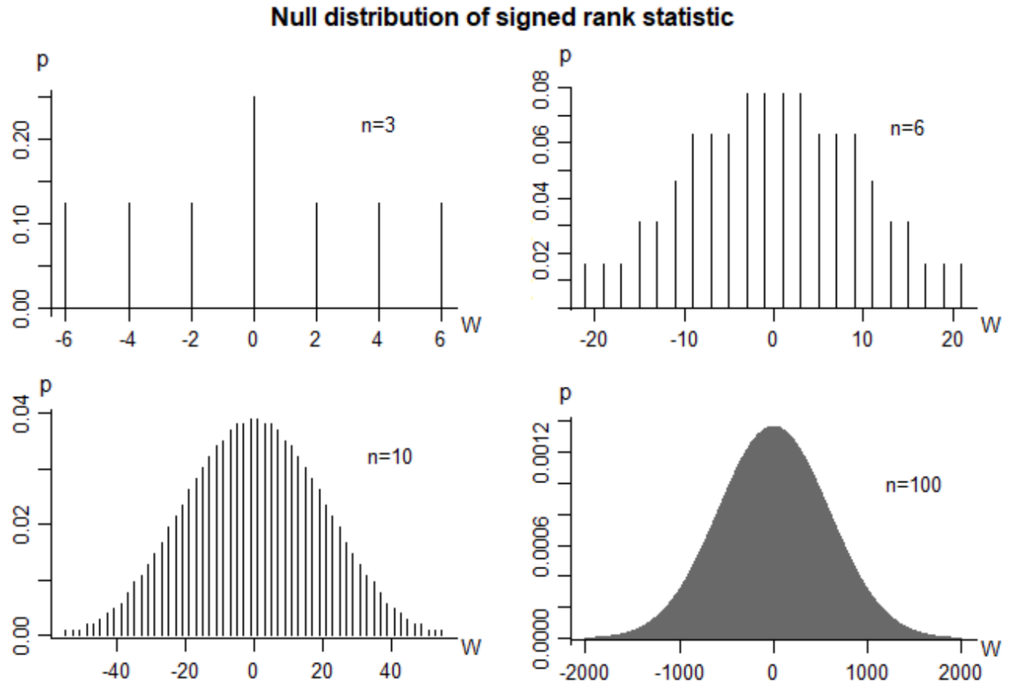
Пусть в выборке пять наблюдений

1	2	3	4	5	
-	-	-	-	-	-15
+	-	-	-	-	-13
-	+	-	-	-	-11
…	…	…	…	…	
-	+	+	+	+	13
+	+	+	+	+	15



Статистика распределения

$$W = \sum_{i-1}^{n} rank(|X_i - m_0|) \cdot sign(X_i - m_0) \sim N(0, \frac{n \cdot (n+1) \cdot (2n+1)}{6})$$



Критерий Уилкоксона (двухвыборочный)

Выборки: $\mathsf{X} = (X_1, \ldots, X_n), X_i \sim F_X$

$\mathsf{Y} = (Y_1, \ldots, Y_n), Y_i \sim F_Y$

X, Y зависимые

Нулевая гипотеза: $H_0 : Med(X - Y) = 0$

Альтернатива: $H_1 : Med(X - Y) \neq 0$

Статистика: $W = \sum_{i-1}^{n} rank(|X_i - Y_i|) \cdot sign(X_i - Y_i)$

Нулевое распределение: У статистики табличное распределение

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	**48**	**54**	**67**	**56**	**55**	**55**	**90**	**71**	**72**	**69**
После	**47**	**52**	**60**	**60**	**58**	**60**	**70**	**81**	**87**	**60**

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
|X-Y|	1	2	7	4	3	5	20	10	15	9

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
|X-Y|	1	2	7	4	3	5	20	10	15	9
rank|X-Y|	1	2	6	4	3	5	10	8	9	7

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
|X-Y|	1	2	7	4	3	5	20	10	15	9
rank|X-Y|	1	2	6	4	3	5	10	8	9	7
sign(X-Y)	+	+	+	-	-	-	+	-	-	+

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
|X-Y|	1	2	7	4	3	5	20	10	15	9
rank|X-Y|	1	2	6	4	3	5	10	8	9	7
sign(X-Y)	+	+	+	-	-	-	+	-	-	+

$$W_{obs} = -3$$

Критерий рангов - пример

Даны баллы студентов до и после апелляции. Правда ли, что в среднем апелляция не повышает балл за контрольную?

До	48	54	67	56	55	55	90	71	72	69
После	47	52	60	60	58	60	70	81	87	60
|X-Y|	1	2	7	4	3	5	20	10	15	9
rank|X-Y|	1	2	6	4	3	5	10	8	9	7
sign(X-Y)	+	+	+	-	-	-	+	-	-	+

$$W_{obs} = -3$$

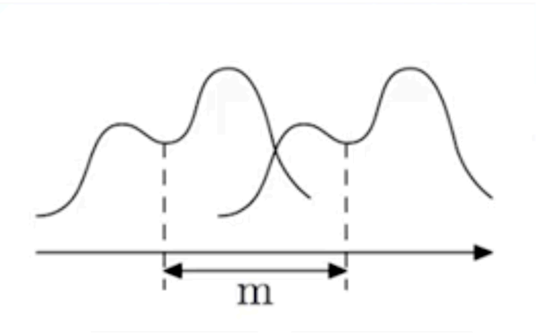
$$W_{crit} = W_{1-\alpha} = 33$$

H_0 не отвергается

Критерий Манна-Уитни

В отличие от критериев знаков, в ранговый критерий можно сформулировать для независимых выборок.

Предположим, что распределения одинаковые по форме, но отличаются сдвигом.



Объединим обе выборки в одну общую и посчитаемся всех чисел ранги.

Посчитаем сумму рангов только для первой выборки - X.

Если сумма оказалась большой, это значит, что X - где-то правее Y, если маленькой - то левее

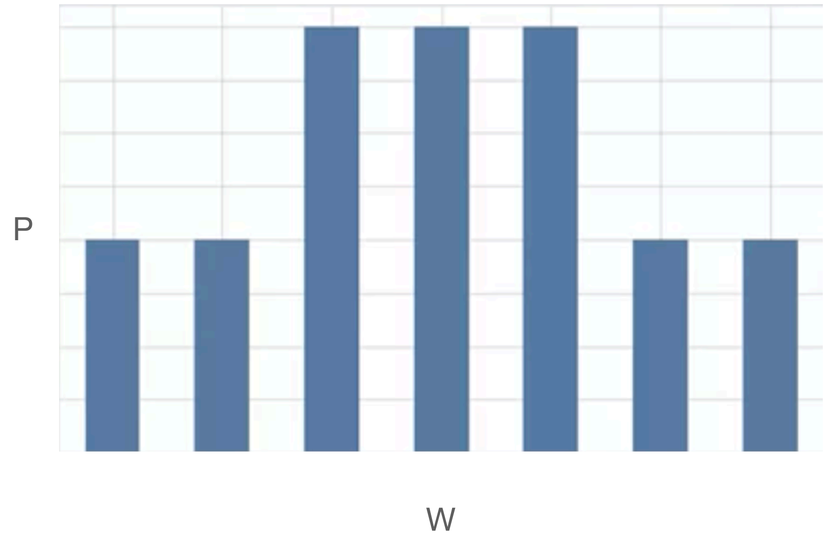
Критерий рангов - распределение

rank(X)	rank(Y)
{1,2}	{3,4,5}
{1,3}	{2,4,5}
{1,4}	{2,3,5}
{1,5}	{2,3,4}
{2,3}	{1,4,5}
…	

Критерий рангов - распределение

rank(X)	rank(Y)
{1,2}	{3,4,5}
{1,3}	{2,4,5}
{1,4}	{2,3,5}
{1,5}	{2,3,4}
{2,3}	{1,4,5}

…



Всего $C_{n_x+n_y}^{n_x}$ вариантов

Распределение статистики снова табличное

Для больших объемов выборки можно использовать нормальное распределение

Критерий Манна-Уитни / U-test (двухвыборочный)

Выборки: $X = (X_1, \ldots, X_{n_1}), X_i \sim F_X$

$Y = (Y_1, \ldots, Y_{n_2}), Y_i \sim F_Y$

X, Y независимые

Нулевая гипотеза: $H_0 : F_X = F_y$

Альтернатива: $H_0 : F_X \neq F_Y$

Статистика: $$U = \sum_{i=1}^{n_x} rank(X_i) - \frac{n(n+1)}{2}$$

Нулевое распределение: У статистики табличные значения

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3	9
4	7
2	5
6	10
2	6
5	8

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3	9
4	7
2	5
6	10
2	6
5	8

Rank	Value
1.5	2
1.5	2
3	3
4	4
5.5	5
5.5	5
7.5	6
7.5	6
9	7
10	8
11	9
12	10

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3 - 3	9 -7
4 - 4	7 - 9
2 - 1.5	5 - 5.5
6- 7.5	10 - 12
2 - 1.5	6 - 7.5
5 - 5.5	8 - 10

S = 23 S = 55

Rank	Value
1.5	2
1.5	2
3	3
4	4
5.5	5
5.5	5
7.5	6
7.5	6
9	7
10	8
11	9
12	10

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3 - 3	9 -7
4 - 4	7 - 9
2 - 1.5	5 - 5.5
6- 7.5	10 - 12
2 - 1.5	6 - 7.5
5 - 5.5	8 - 10

S = 23 S = 55

$U_a = 2 \quad U_b = 34$

Rank	Value
1.5	2
1.5	2
3	3
4	4
5.5	5
5.5	5
7.5	6
7.5	6
9	7
10	8
11	9
12	10

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3 - 3	9 -7
4 - 4	7 - 9
2 - 1.5	5 - 5.5
6- 7.5	10 - 12
2 - 1.5	6 - 7.5
5 - 5.5	8 - 10

S = 23 S = 55

$U_a = 2 \quad U_b = 34$

Rank	Value
1.5	2
1.5	2
3	3
4	4
5.5	5
5.5	5
7.5	6
7.5	6
9	7
10	8
11	9
12	10

$$min(U_a, U_b) = U_a = 2 < U_{crit} = 5$$

Критерий Манна-Уитни - пример

Сделали две имейл рассылки с рекламой черной пятницы на две разные рандомные группы - A и B и замерили количество товаров, купленных пользователями каждой группы.

A	B
3 - 3	9 -7
4 - 4	7 - 9
2 - 1.5	5 - 5.5
6- 7.5	10 - 12
2 - 1.5	6 - 7.5
5 - 5.5	8 - 10

S = 23 S = 55

$U_a = 2 \quad U_b = 34$

Rank	Value
1.5	2
1.5	2
3	3
4	4
5.5	5
5.5	5
7.5	6
7.5	6
9	7
10	8
11	9
12	10

$$min(U_a, U_b) = U_a = 2 < U_{crit} = 5$$

H_0 **отвергается!**

Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Критерии Согласия Проверяем принадлежность классу распределений (согласованность выборки с распределением)	Общие	Непрерывные распределения	$H_0 : F \in \mathbf{F}_\theta$	Критерий Колмогорова	$D_n = sup \lvert \widehat{F_n}(u) - F_0(u) \rvert$	Максимальное расстояние между функциями распределений	`scipy.stats.kstest`
		Дискретные распределения		Критерий Пирсона (хи-квадрат)	$T_n = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}$	Отклонения количества каждых значений от ожидаемых количеств этих значений	`scipy.stats.chisquare`
	Специальные	Нормальное распределение	$H_0 : F \sim N(\mu, \sigma)$	Критерий Шапиро-Уилка	$SW_n = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$	Отношение квадрата линейной оценки стандартного отклонения к смещенной оценке дисперсии	`scipy.stats.shapiro`
				Критерий Харке-Бера	$JB_n = n(\frac{S^2}{6} + \frac{(K-3)^2}{24}, S = \mu_3 \mu_2^{3/2}, K = \frac{\mu_4}{\mu_2^2}$	Стандартизировнные коэффициенты эксцесса и асимметрии	`scipy.stats.jarque_bera`
Параметрические Проверяем параметры (среднее/ пропорции), предполагая, что выборки пришли из конкретного распределения	Одновыборочные	$X_i \sim Ber(p)$	$H_0 : p = p_0$	Z-критерий	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.proportion.proportions_ztest`
		$X_i \sim N(\mu, \sigma^2)$ σ неизвестна	$H_0 : \mu = \mu_0$	t-критерий Стьюдента	$T_n = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_1samp`
	Двухвыборочные	$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки независимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{P(1-P)(\frac{1}{n_x} + \frac{1}{n_y})}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$X_i \sim N(\mu_x, \sigma_x^2), Y_i \sim N(\mu_y, \sigma_y^2)$ σ_x, σ_y неизвестны (и могут быть не равны) выборки независимы	$H_0 : \mu_1 = \mu_2$	Двухвыборочный t-критерий Уэлча	$T_n = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	Отношение разницы средних и разницы отклонений	`scipy.stats.ttest_ind`
		$X_i \sim Ber(p), Y_i \sim Ber(p)$ выборки зависимы	$H_0 : p_x = p_y$	Двухвыборочный Z-критерий	$z = \frac{c - b}{\sqrt{c + b - \frac{(c-b)^2}{n}}}$	Отношение разницы долей и разницы отклонений	`statsmodels.stats.weightstats.ztest`
		$Z_i = Y_i - X_i, Z_i \sim N(\mu, \sigma^2)$ выборки зависимы	$H_0 : \mu = 0$	Двухвыборочный t-критерий	$T_n = \frac{\overline{Z}}{S\sqrt{n}}$	Аналог t-критерия стьюдента для одной выборки для статистики Y-X	`scipy.stats.ttest_rel`
Непараметрические							
Критерии корреляции							

Класс критериев	Тип критериев	Условия	Гипотеза	Критерий	Статистика	Смысл статистики	Питон
Непараметрические	Одновыборочные	$X = (X_1, \dots, X_n), X_i \sim F_X$	$H_0 : Med(X) = m_0$	Критерий Знаков	$T = \sum_{i-1}^{n} [X_i > m_0]$	Количество значений больше медианы	`statsmodels.stats.descriptivestats.sign_test`
		$X = (X_1, \dots, X_n), X_i \sim F_X$ F_X симметрична относительно медианы	$H_0 : Med(X) = m_0$	Критерий Уилкоксона	$W = \sum_{i-1}^{n} rank(\lvert X_i - m_0 \rvert) \cdot sign(X_i - m_0)$	Сумма рангов разностей значений и медианы	scipy.stats.wilcoxon
	Двухвыборочные	$X = (X_1, \dots, X_n), X_i \sim F_X$ $Y = (Y_1, \dots, Y_n), Y_i \sim F_Y$ X, Y зависимые	$H_0 : Med(X - Y) = 0$	Критерий Уилкоксона (Wilcoxon signed-rank)	$W = \sum_{i-1}^{n} rank(\lvert X_i - Y_i \rvert) \cdot sign(X_i - Y_i)$	Сумма рангов разности значений до/после	scipy.stats.wilcoxon
		$X = (X_1, \dots, X_{n_1}), X_i \sim F_X$ $Y = (Y_1, \dots, Y_{n_2}), Y_i \sim F_Y$ X, Y независимые	$H_0 : F_X = F_y$	Критерий Манна-Уитни (Mann-Whitney/ Wilcoxon rank-sum)	$U = \sum_{i=1}^{n_x} rank(X_i) - \frac{n(n+1)}{2}$	Сумма рангов, проставленных по объединению двух выборок, для одной выборки	**`scipy.stats.mannwhitneyu`**