Методы машинного обучения

Лекция 3

Основы статистического анализа

Статистические гипотезы

Статистическая гипотеза — гипотеза о виде распределения и свойствах случайной величины, которые можно подтвердить или опровергнуть применением статистических методов к данным выборки.

Нулевая гипотеза — принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами. Часто в качестве нулевой гипотезы выступают предположения об отсутствии взаимосвязи или корреляции между исследуемыми переменными, об отсутствии различий (однородности) в распределениях (параметрах распределений) в двух и/или более выборках. Для обозначения нулевой гипотезы часто используют символ H_0.

Проверка гипотез

Нулевая гипотеза H_0 считается верной, пока нельзя доказать обратное. Проверка фальсифицируемости нулевой гипотезы — общепринятый способ обеспечения строгости исследования. Если прямого способа проверки у нас нет, приходится прибегать к проверкам косвенным. **Статистика как наука даёт чёткие условия, при наступлении которых нулевая гипотеза может быть отвергнута**.

- Если некоторое явление логически неизбежно следует из гипотезы, но в природе не наблюдается, то это значит, что гипотеза неверна.
- Если происходит то, что при гипотезе происходить не должно, это тоже означает ложность гипотезы.

Заметим, что строго говоря, **косвенным образом доказать гипотезу нельзя, хотя опровергнуть — можно.**

При статистическом выводе исследователь пытается показать несостоятельность нулевой гипотезы, несогласованность её с имеющимися опытными данными, то есть отвергнуть гипотезу. При этом подразумевается, что должна быть принята другая, альтернативная (конкурирующая) гипотеза, исключающая нулевую гипотезу. Если же данные, наоборот, подтверждают нулевую гипотезу, то она не отвергается.

Проверка статистической значимости

Смысл статистической значимости заключается в том, чтобы определить, имеет ли под собой какое-то основание разница между двумя показателями, или же она случайна.

Выберем уровень вероятности $\varepsilon > 0$. Условимся считать событие практически невозможным, если его вероятность меньше ε. Когда речь идет о проверке гипотез, число ε называют уровнем значимости.

Уровень значимости — это вероятность того, что мы сочли различия существенными, в то время как они на самом деле случайны. Уровень значимости показывает степень достоверности выявленных различий между выборками, т.е. показывает, насколько мы можем доверять тому, что различия действительно есть.

Уровни значимости: $p \leq 0{,}05$, $p \leq 0{,}01$, $p \leq 0{,}001$.

Определение. *Событие* A *называется критическим для гипотезы* H*, или* **критерием** *для* H*. Если* $P(A|H) \leq \varepsilon$*, то* ε *называют* **гарантированным уровнем значимости** *критерия* A *для* H*.*

Критерий для проверки гипотезы – это решающее правило, отвергающее или принимающее нулевую гипотезу на основе выборочных наблюдений.

Ошибки первого и второго рода

Возможны ошибки двух родов: первого рода (α) и второго рода (β).

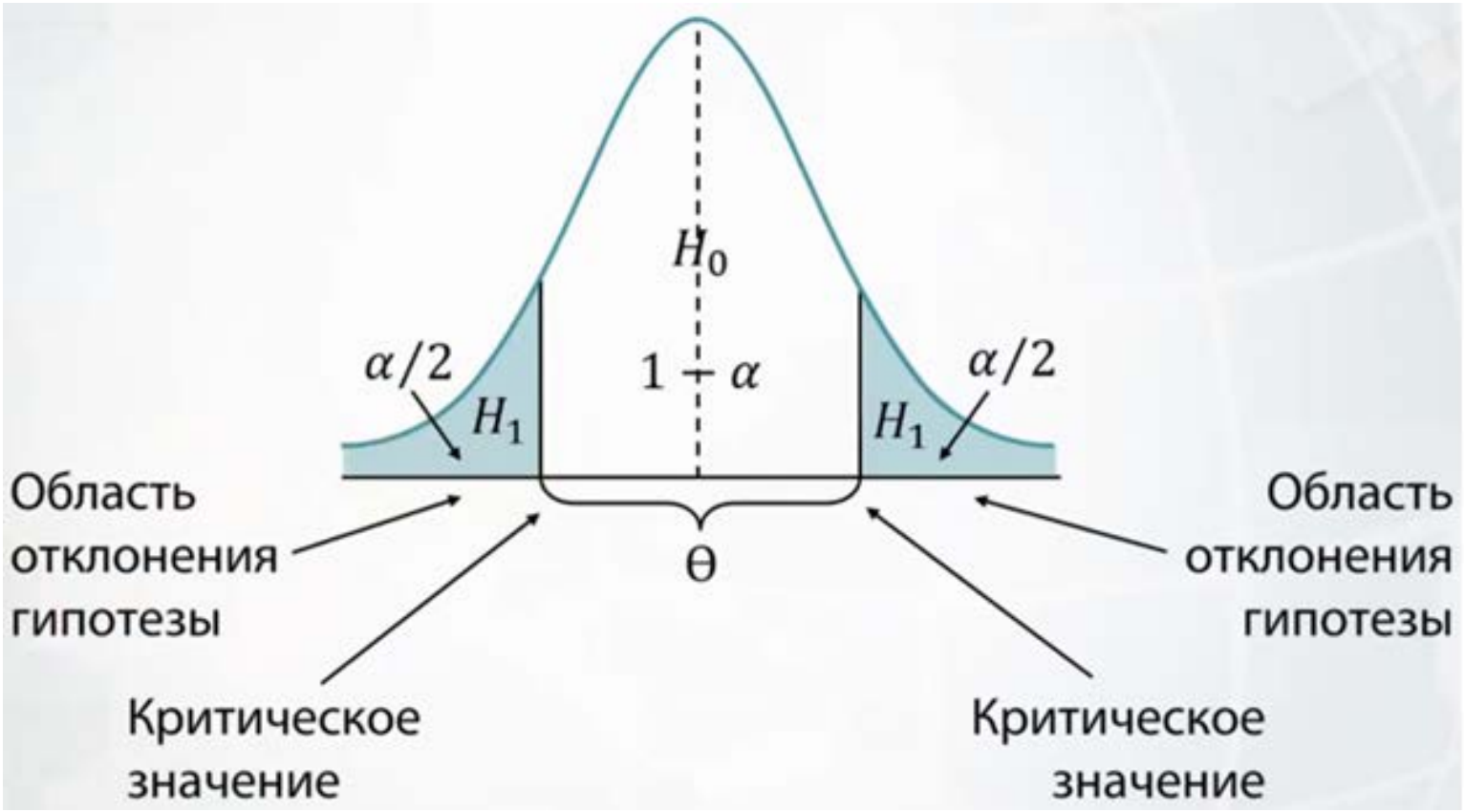
Ошибка первого рода состоит в том, что гипотеза H_0 будет отвергнута, хотя на самом деле она правильная. Вероятность допустить такую ошибку называют **уровнем значимости** и обозначают буквой **α** («альфа»).

Ошибка второго рода состоит в том, что гипотеза H_0 будет принята, но на самом деле она неправильная. Вероятность совершить эту ошибку обозначают буквой **β** («бета»). Значение **(1-β)** называют **мощностью критерия** – это вероятность отклонения ложной гипотезы, т.е. способность выявлять даже мелкие различия. Чем мощнее критерий, тем лучше он отвергает нулевую гипотезу.

В практических задачах, как правило, задают ***уровень значимости***, и наиболее часто выбирают значения: α=0.1, **α=0.05,** α=0.01.

Нулевая гипотеза H_0	**Ложная**	**Истинная**
Отклоняется	Ошибки нет (1-β)	Ошибка I рода (ложно-положительный вывод), ложная тревога (α)
Принимается ***(не отклоняется)***	Ошибка II рода (ложно-отрицательный вывод) , пропуск цели (β)	Ошибки нет (1-α)

Области принятия и отклонения гипотезы для двусторонней альтернативной гипотезы



Методика проверки статистических гипотез

Пусть задана случайная выборка $X^m = (x_1, \dots, x_m)$ — последовательность *m* объектов из множества *X*. Предполагается, что на множестве *X* существует некоторая неизвестная вероятностная мера P.

1. Формулируются *нулевая* $\boldsymbol{H_0}$ и альтернативная $\boldsymbol{H_1}$ гипотезы о распределении вероятностей на множестве *X*.
2. Задаётся некоторая статистика (функция выборки – *на след. слайде*) $T: X^m \to \mathbb{R}$, для которой в условиях справедливости гипотезы $\boldsymbol{H_0}$ выводится функция распределения $F(T)$ и/или плотность распределения $p(T)$.
3. Фиксируется *уровень значимости* — допустимая для данной задачи вероятность *ошибки первого рода*, то есть того, что гипотеза на самом деле верна, но будет отвергнута процедурой проверки. Это должно быть достаточно малое число $a \in [0,1]$. На практике часто полагают a=0.05.
4. На множестве допустимых значений статистики T выделяется *критическое множество* Ω_a наименее вероятных значений статистики T , такое, что $P\{T \in \Omega_a | H_0\} = \alpha$.
5. Собственно *статистический тест* (*статистический критерий*) заключается в проверке условия:
   - Если $T(X^m) \in \Omega_\alpha$, то делается вывод «данные противоречат нулевой гипотезе при уровне значимости α». Гипотеза отвергается.
   - Если $T(X^m) \notin \Omega_a$, то делается вывод «данные не противоречат нулевой гипотезе при уровне значимости α». Гипотеза принимается.

Статистика (функция выборки)

Статистикой называется произвольная измеримая функция выборки $\mathrm{T}: \mathrm{X}^{\mathrm{m}} \to \mathbb{R}$, которая не зависит от неизвестных параметров распределения.

Условие измеримости статистики означает, что эта функция является случайной величиной. Независимость этой функции от неизвестных параметров заключается в том, что исследователь может по имеющимся в его распоряжении данным найти значение этой функции, а следовательно — основывать на этом значении оценки и прочие статистические выводы.

Пример. Статистики, используемые для оценки моментов (выборочные моменты)

Выборочное среднее:	$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
Выборочная дисперсия:	$s^2 = s_m^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$
Несмещённая оценка дисперсии:	$s^2 = s_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$
Выборочный момент *k*-го порядка (выборочное среднее—момент первого порядка):	$M_k = \frac{1}{m}\sum_{i=1}^{m} x_i^k$
Выборочный центральный момент k-го порядка (выборочная дисперсия — центральный момент второго порядка):	$\widetilde{M_k} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^k$

Замечание о проверке статистических гипотез

Важно! Если данные не противоречат нулевой гипотезе, это ещё не значит, что гипотеза верна.

Тому есть две причины:

- По мере увеличения длины выборки нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия данных гипотезе, и она будет отвергнута. То есть многое зависит от объёма данных, если данных не хватает, можно принять даже самую неправдоподобную гипотезу.
- Выбранная статистика T может отражать не всю информацию, содержащуюся в гипотезе H_0. В таком случае увеличивается вероятность ошибки второго рода — нулевая гипотеза может быть принята, хотя на самом деле она не верна.

Критерий согласия Пирсона Хи-квадрат (Chi-square test)

Критерий согласия проверяет, согласуется ли заданная выборка с заданным фиксированным распределением (семейством распределений) или с другой выборкой, т.е. критерий согласия для проверки гипотезы о законе распределения.

Критерий хи-квадрат — любая статистическая проверка гипотезы, в которой выборочное распределение критерия имеет распределение хи-квадрат при условии верности нулевой гипотезы.

Пусть X — исследуемая случайная величина. Требуется проверить гипотезу H_0 о том, что данная случайная величина подчиняется закону распределения *F(x)*. Дана выборка $X^n = (x_1, \dots, x_n)$, $x_i \in [a, b]$, $\forall i = 1 \dots n$. Необходимо построить эмпирический закон распределения *F'(x)* случайной величины *X*.

- Разделим $[a, b]$ на *k* непересекающихся интервалов $[a_i, b_i]$,*i=1…k*.
- Пусть n_j — количество наблюдений в j-м интервале;
- $p_j = F(b_j) - F(a_j)$ — вероятность попадания наблюдения в j-й интервал при выполнении гипотезы H_0';
- $E_j = np_j$ — ожидаемое число попаданий в j-й интервал;
- тогда распределение Хи-квадрат с числом степеней свободы k−1 будет иметь следующую статистику (критерий Хи-квадрат **Пирсона**):

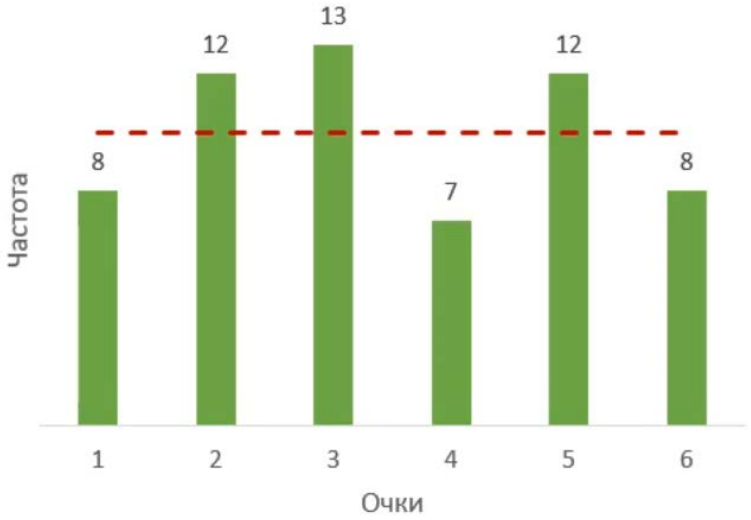
$$\boldsymbol{\chi}^2 = \sum_{j=1}^{k} \frac{(n_j - E_j)^2}{E_j} \sim \boldsymbol{\chi}^2_{k-1}$$

В зависимости от значения критерия χ2 гипотеза H_0 может приниматься либо отвергаться:

$\boldsymbol{\chi}^2_1 < \boldsymbol{\chi}^2 < \boldsymbol{\chi}^2_2$ — гипотеза H_0 выполняется.

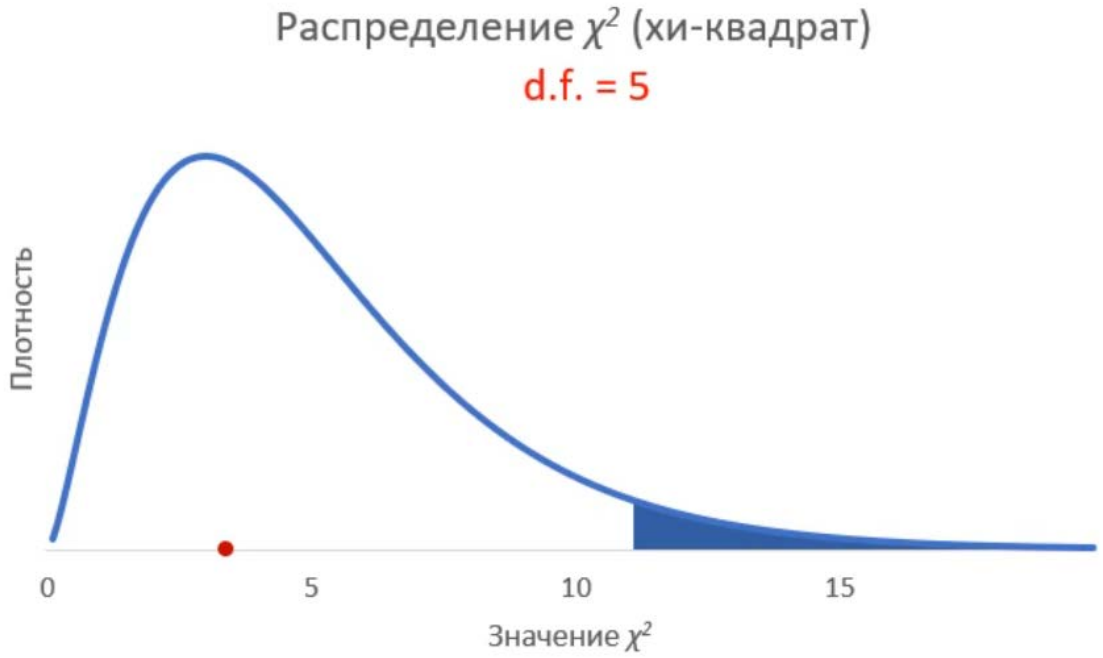
$\boldsymbol{\chi}^2 \geq \boldsymbol{\chi}^2_2$ — попадает в правый «хвост» распределения, гипотеза H_0 отвергается.

Пример: Критерий согласия Пирсона Хи-квадрат

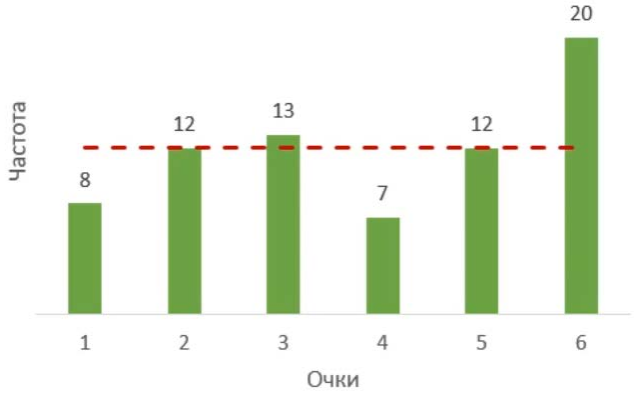


Очки	Частота		
	Наблюдаемая O	Ожидаемая E	
1	8	10	0,4
2	12	10	0,4
3	13	10	0,9
4	7	10	0,9
5	12	10	0,4
6	8	10	0,4
Итого	**60**		**3,4**

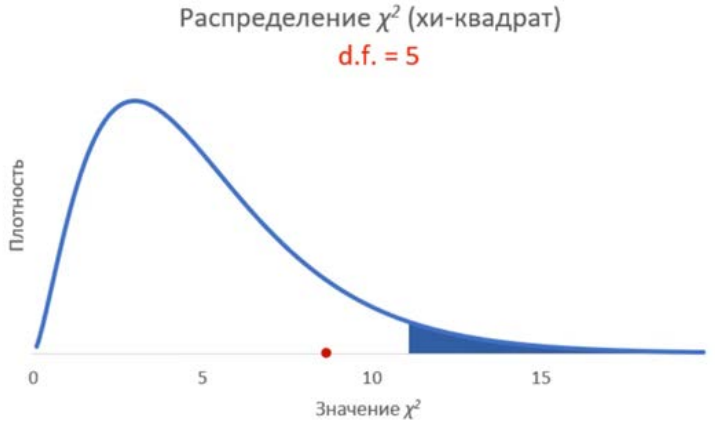
d.f.	5
χ^2	3,4
$\chi^2_{0,05;\,5}$	11,0705
p-value	0,63857
Тест χ^2	0,63857

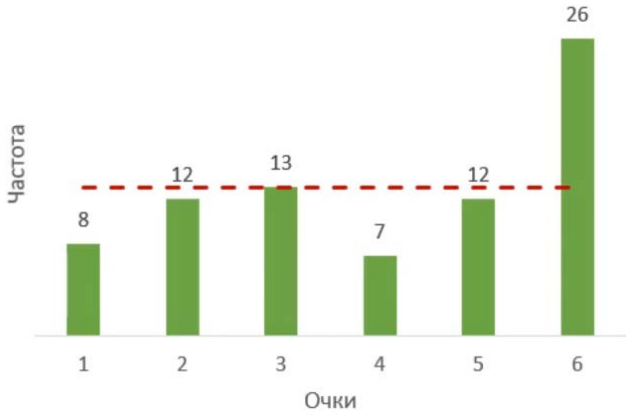


Пример: Критерий согласия Пирсона Хи-квадрат

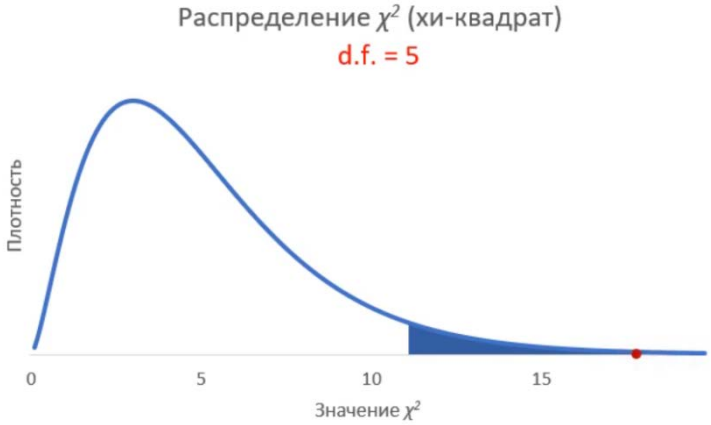


d.f.	5
χ^2	8,83333
$\chi^2_{0,05;\,5}$	11,0705
p-value	0,1159
Тест χ^2	0,1159





d.f.	5
χ^2	17,8462
$\chi^2_{0,05;\,5}$	11,0705
p-value	0,00315
Тест χ^2	0,00315



Проверка гипотезы о математическом ожидании

Может быть два варианта, когда дисперсия известна и когда неизвестна. Рассмотрим случай, когда она неизвестна.

Сравнение выборочного среднего с заданным значением

Задана выборка $x^m = (x_1, \dots, x_m)$, $x_i \in \mathbb{R}$, $x_i \sim N(\mu, \sigma^2)$.

Дополнительное предположение: выборка простая и нормальная.

Нулевая гипотеза $H_0: \bar{x} = \mu_0$ (выборочное среднее равно заданному числу μ_0), альтернативная $H_1: \bar{x} \neq \mu_0$

Статистика критерия:

$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{m}}}$$

имеет t-распределение Стьюдента (числитель-нормальное/знаменатель хи-квадрат) с *m-1* степенями свободы, где:

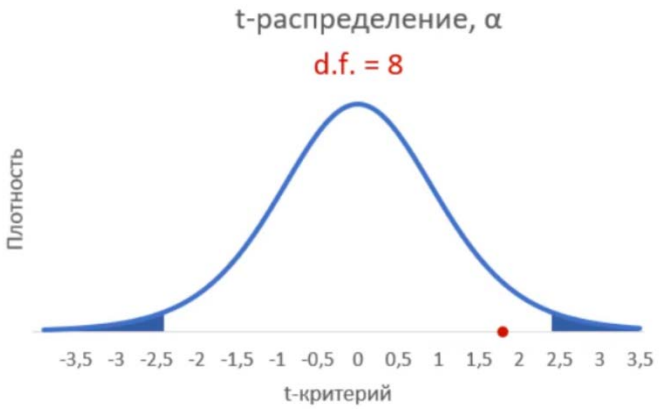
- $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ — выборочное среднее,
- $s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$ — выборочная дисперсия.

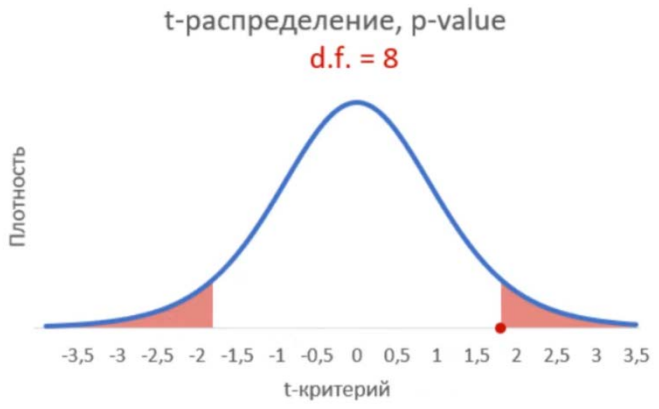
Критерий (при уровне значимости):

Гипотеза H_0 принимается, если $|t| < t_{1-\alpha/2}$, в противном случае отвергается.

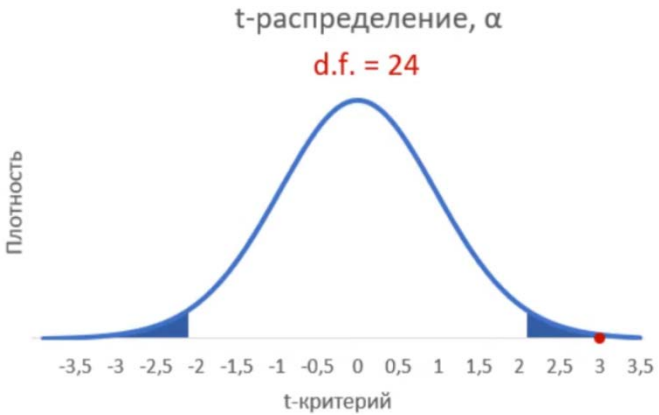
Пример: Проверка гипотезы о математическом ожидании

Показатель	Значение
μ	50
$X_{ср}$	50,3
n	9
s	0,5
tфакт	1,8
α	0,05
d.f.	8
tкритич	2,306
p-value	0,10955





Показатель	Значение
μ	50
$X_{ср}$	50,3
n	25
s	0,5
tфакт	3
α	0,05
d.f.	24
tкритич	2,064
p-value	0,00621



Проверка гипотезы о математическом ожидании - Две выборки

Специальный случай двухвыборочных критериев согласия. Проверяется гипотеза сдвига, согласно которой распределения двух выборок имеют одинаковую форму и отличаются только сдвигом на константу.

Критерий Стьюдента. Рассмотрим теперь задачу сравнения средних значений двух нормальных выборок.

Пусть $x_1, \dots, x_n; y_1, \dots, y_m$ —нормальные независимые выборки из законов распределения с параметрами (a_1, σ_1^2) и (a_2, σ_2^2) соответственно.

Рассмотрим проверку гипотезы:

$$H_0: a_1 = a_2 \text{ против альтернативы } a_1 \neq a_2$$

Относительно параметров σ_1^2 и σ_2^2 выделим следующие четыре варианта предположений:

а) обе дисперсии известны и равны между собой;

б) обе дисперсии известны, но не равны между собой;

в) обе дисперсии неизвестны, но предполагается, что они равны между собой;

г) обе дисперсии неизвестны, их равенство не предполагается.

Проверка гипотезы о математическом ожидании - Две выборки

Для построения критерия проверки гипотезы H_0 проведем следующие рассуждения.

От выборок $x_1, \dots, x_n$ и $y_1, \dots, y_m$ перейдем к выборочным средним $\bar{x}$ и $\bar{y}$. Согласно свойствам нормального распределения и выдвинутой гипотезе, величины $\bar{x}$ и $\bar{y}$ имеют нормальные распределения с одними тем же средним и дисперсиями σ_1^2/n и σ_2^2/m.

Далее перейдем к статистике, основанной на выборочных средних $\bar{x}$ и $\bar{y}$ и дисперсиях σ_1^2 и σ_2^2 (если они известны) или их оценках s_1^2 и s_2^2 (если дисперсии неизвестны).

$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ — выборочное среднее,

$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$ — выборочная дисперсия.

Проверка гипотезы о математическом ожидании - Дисперсии известны

а) Обе дисперсии известны и равны между собой;

$$\frac{(\bar{x}-\bar{y})}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

Статистика имеет стандартное нормальное распределение, так как является линейной комбинацией независимых нормальных величин. Гипотеза H принимается на уровне значимости α, если

$$\left|\frac{(\bar{x}-\bar{y})}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| < z_{1-\alpha/2}$$

в противном случае гипотеза отвергается в пользу альтернативы $a_1 \neq a_2$

б) Обе дисперсии известны, но не равны между собой;

$$\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$$

Статистика имеет также стандартное нормальное распределение. Правило принятия гипотезы аналогично правилу пункта а).

Проверка гипотезы о математическом ожидании - Дисперсии неизвестны

в) Обе дисперсии неизвестны, но предполагается, что они равны между собой;

В случае, когда обе дисперсии неизвестны, но предполагаются равными между собой, мы имеем две оценки s_1^2 и s_2^2 одной и той же величины дисперсии $\sigma_1^2 = \sigma_2^2$. В связи с этим разумно перейти к объединенной оценке: $s^2 = \frac{s_1^2(n-1)+s_2^2(m-1)}{(n-1)+(m-1)}$

Критерий для проверки гипотезы H_0: $a_1 = a_2$ опирается на статистику $\frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n}+\frac{1}{m}}}$,

которая имеет распределение Стьюдента с n+m−2 степенями свободы.

г) Обе дисперсии неизвестны, их равенство не предполагается.

В случае неизвестных дисперсий, равенство которых не предполагается, используется аналог статистики пункта б) с заменой неизвестных дисперсий их оценками: $\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}}$

В этой ситуации указать точное распределение введенной статистики затруднительно. Известно, однако, что это распределение близко к распределению Стьюдента с числом степеней свободы, равным: $\frac{(s_1^2/n+s_2^2/m)^2}{\frac{(s_1^2/n)^2}{n-1}+\frac{(s_2^2/m)^2}{m-1}}$

Критерий проверки гипотезы устроен так же, как и в пункте в).

P-value или p-значение

P-value или p-значение – одна из ключевых величин, используемых в статистике при тестировании гипотез. Она показывает вероятность получения наблюдаемых результатов при условии, что нулевая гипотеза верна, или вероятность ошибки в случае отклонения нулевой гипотезы.

Доверительный интервал (Confidence interval)

В математической статистике — интервал, в пределах которого с заданной вероятностью лежат выборочные оценки статистических характеристик генеральной совокупности.

Если оценку среднего требуется связать с **определённой вероятностью**, то интересующий параметр генеральной совокупности нужно оценивать не одним числом, а интервалом. Доверительным интервалом называют интервал, в котором с определённой вероятностью *P* находится значение оцениваемого показателя генеральной совокупности. Доверительный интервал, в котором с вероятностью $P = 1 - \alpha$ (α - квантиль стандартного нормального распределения) находится случайная величина $\bar{X}$, рассчитывается следующим образом:

$$\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

где $z_{1-\frac{\alpha}{2}}$ - критическое значение стандартного нормального распределения для уровня значимости $\alpha = 1 - P$, которое можно найти в приложении к практически любой книге по статистике.

Доверительный интервал (Confidence interval)

На практике среднее значение генеральной совокупности μ и дисперсия σ^2 не известны, поэтому дисперсия генеральной совокупности заменяется дисперсией выборки S^2 , а среднее генеральной совокупности - средним значением выборки $\bar{X}$. Таким образом, доверительный интервал в большинстве случаев рассчитывается так:

$$\bar{X} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}$$

где S -несмещённое выборочное стандартное отклонение, имеет распределение Стьюдента с n-1 степенями свободы t(n-1). Пусть $t_{\alpha,n-1}$—α - квантили распределения Стьюдента.

Формулу доверительного интервала можно использовать для оценки среднего генеральной совокупности, если

- известно стандартное отклонение генеральной совокупности;
- или стандартное отклонение генеральной совокупности не известно, но объём выборки - больше 30.

Пример: Доверительный интервал

Пример 1. Собрана информация из 100 случайно выбранных кафе в некотором городе о том, что среднее число работников в них составляет 10,5 со стандартным отклонением 4,6. Определить доверительный интервал 95% числа работников кафе.

Решение:

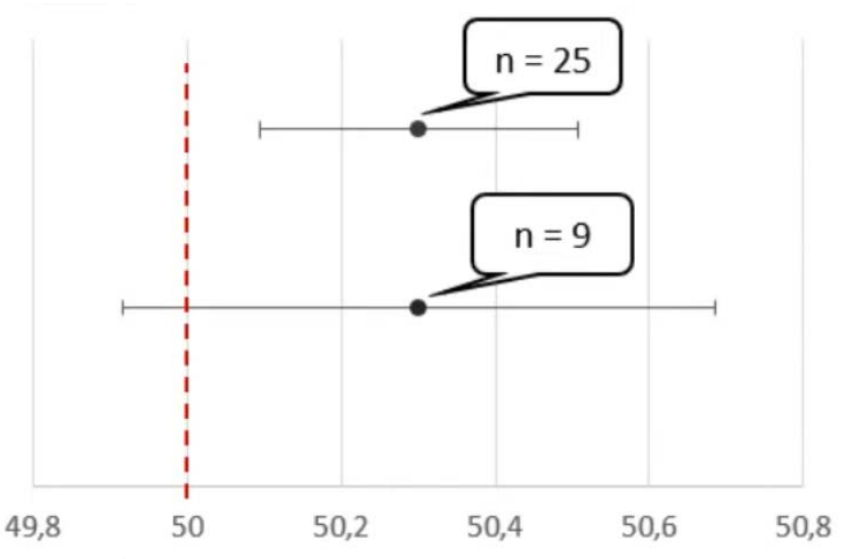
$$10{,}5 - 1{,}96 \cdot \frac{4{,}6}{\sqrt{100}} \leq \mu \leq 10{,}5 + 1{,}96 \cdot \frac{4{,}6}{\sqrt{100}}$$

где $z_{0{,}05} = 1{,}96$ - критическое значение стандартного нормального распределения для уровня значимости α = 0,05.

Таким образом, доверительный интервал 95% среднего числа работников кафе составил от 9,6 до 11,4.

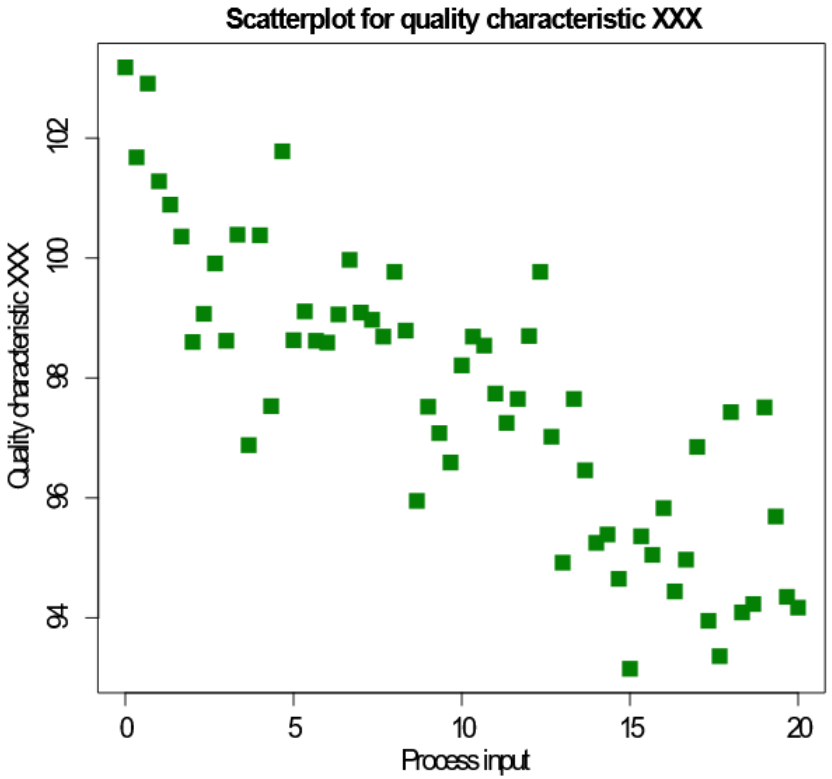
Пример 2. - со слайда 14.

Показатель	n = 9	n = 25
μ	50	50
$X_{ср}$	50,3	50,3
n	9	25
s	0,5	0,5
α	0,05	0,05
Δ_X	0,384	0,206
μ_1	**49,916**	**50,094**
μ_2	**50,684**	**50,506**



Корреляция - определение

Корреляция (от лат. correlatio «соотношение, взаимосвязь»), или корреляционная зависимость — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.



	1	2	3	4	5	6	7	8	9	10	11	12
1		0,27	-0,28	0,38	-0,45	0,45	-0,08	-0,02	-0,16	-0,25	0,05	-0,18
2	0,27		-0,43	0,45	-0,29	0,26	0,41	-0,30	0,32	0,27	-0,25	0,12
3	-0,28	-0,43		-0,91	0,40	-0,37	-0,47	0,49	-0,47	-0,28	0,50	-0,29
4	0,38	0,45	-0,91		-0,48	0,45	0,45	-0,49	0,44	0,11	-0,45	0,13
5	-0,45	-0,29	0,40	-0,48		-0,94	-0,12	0,25	-0,18	0,20	0,15	0,27
6	0,45	0,26	-0,37	0,45	-0,94		0,01	-0,10	0,07	-0,23	0,00	-0,31
7	-0,08	0,41	-0,47	0,45	-0,12	0,01		-0,71	0,82	0,44	-0,65	0,28
8	-0,02	-0,30	0,49	-0,49	0,25	-0,10	-0,71		-0,75	-0,52	0,81	-0,42
9	-0,16	0,32	-0,47	0,44	-0,18	0,07	0,82	-0,75		0,55	-0,58	0,28
10	-0,25	0,27	-0,28	0,11	0,20	-0,23	0,44	-0,52	0,55		-0,47	0,62
11	0,05	-0,25	0,50	-0,45	0,15	0,00	-0,65	0,81	-0,58	-0,47		-0,65
12	-0,18	0,12	-0,29	0,13	0,27	-0,31	0,28	-0,42	0,28	0,62	-0,65	

Корреляция и взаимосвязь величин

Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой статистической связи в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно-следственный характер.

- Рассматривая пожары в конкретном городе, можно выявить весьма высокую корреляцию между ущербом, который нанёс пожар, и количеством пожарных, участвовавших в ликвидации пожара, причём эта корреляция будет положительной.
- Из этого не следует вывод «увеличение количества пожарных приводит к увеличению причинённого ущерба», и тем более не будет успешной попытка минимизировать ущерб от пожаров путём ликвидации пожарных бригад.

Отсутствие корреляции между двумя величинами ещё не значит, что между ними нет никакой связи. Например, зависимость может иметь сложный нелинейный характер, который корреляция не выявляет.

Положительная и отрицательная корреляция

- Некоторые виды коэффициентов корреляции могут быть положительными или отрицательными.
  - В первом случае предполагается, что мы можем определить только наличие или отсутствие связи, а во втором — также и её направление.
- Если предполагается, что на значениях переменных задано отношение строгого порядка, то в этом случае:
  - **Отрицательная** корреляция — корреляция, при которой увеличение одной переменной связано с уменьшением другой.
  - **Положительная** корреляция в таких условиях — это такая связь, при которой увеличение одной переменной связано с увеличением другой переменной.
- Возможна также ситуация отсутствия статистической взаимосвязи — например, для независимых случайных величин.

Показатели корреляции

Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные. Так, для измерения переменных с интервальной и количественной шкалами необходимо использовать коэффициент **корреляции Пирсона** (корреляция моментов произведений).

Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, необходимо использовать ранговую корреляцию Спирмена или τ (тау) Кендалла.

Ковариация

Важной характеристикой совместного распределения двух случайных величин является ковариация (или корреляционный момент).

- Ковариация является совместным центральным моментом второго порядка.
- Ковариация определяется как математическое ожидание произведения отклонений случайных величин:

$$cov_{XY} = \boldsymbol{M}\big[\big(X - \boldsymbol{M}(X)\big)\big(Y - \boldsymbol{M}(Y)\big)\big] = M(XY) - M(X)M(Y)$$

где M – математическое ожидание.

Свойства ковариации:

- Ковариация двух независимых случайных величин *X* и *Y* равна нулю
- Абсолютная величина ковариации двух случайных величин X и Y не превышает среднего геометрического их дисперсий:$|cov_{XY}| \leq \sqrt{D_X D_Y}$.
- Ковариация имеет размерность, равную произведению размерности случайных величин, то есть величина ковариации зависит от единиц измерения независимых величин.
  - Данная особенность ковариации затрудняет её использование в целях корреляционного анализа.

Линейный коэффициент корреляции

Для устранения недостатка ковариации был введён линейный коэффициент корреляции (или коэффициент корреляции Пирсона).

- Коэффициент корреляции рассчитывается по формуле:

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

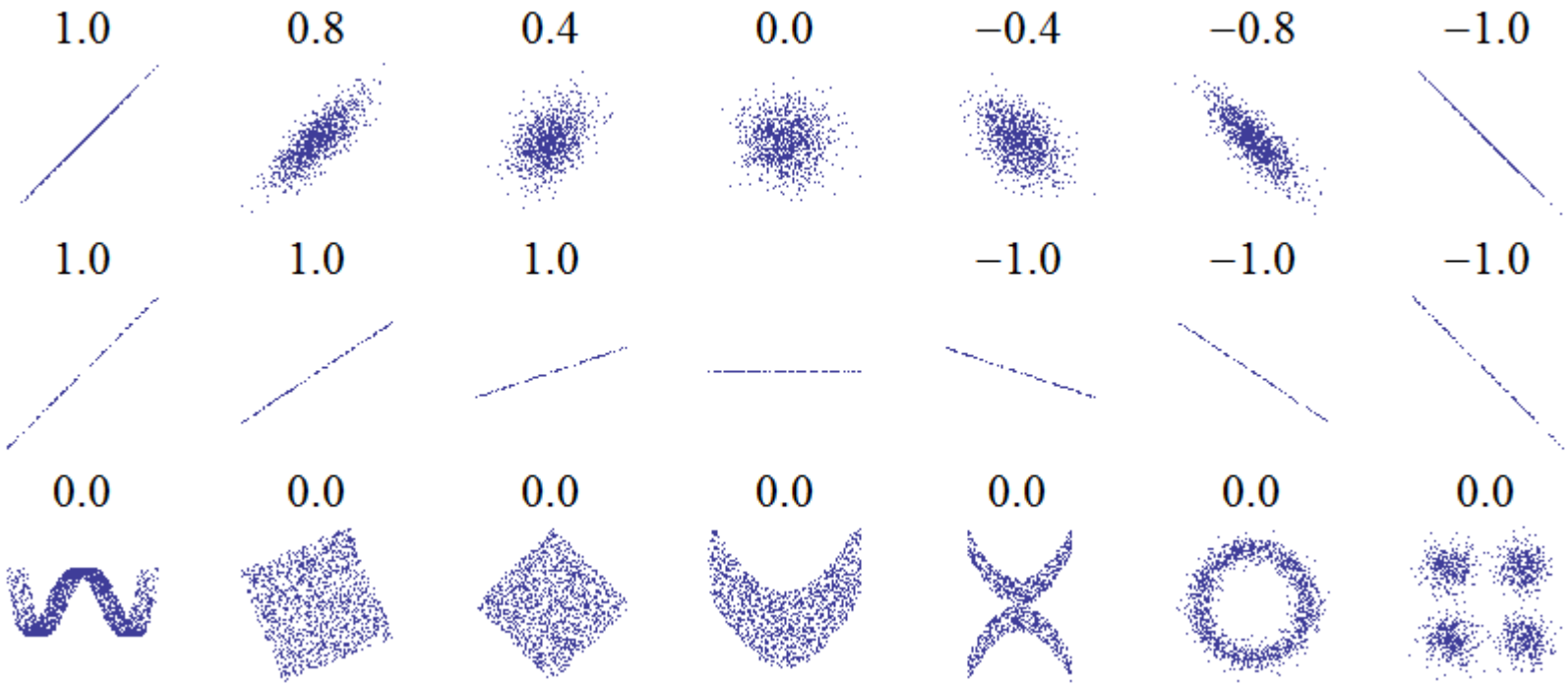
где $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$, $\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ — среднее значение выборок.

- Коэффициент корреляции изменяется в пределах

от -1 до +1.

Ограничения корреляционного анализа

- Применение возможно при наличии достаточного количества наблюдений для изучения.
- Коэффициент корреляции отражает «зашумлённость» линейной зависимости (верхняя строка)
- Коэффициент корреляции не описывает наклон линейной зависимости (средняя строка)
- Коэффициент корреляции совсем не подходит для описания сложных, нелинейных зависимостей (нижняя строка).
- Для распределения, показанного в центре рисунка, коэффициент корреляции не определен, так как дисперсия *y* равна нулю.



Отбор признаков на основе корреляции

Отбор признаков на основе меры корреляции (англ. *Correlation Feature Selection*, CFS) оценивает подмножества признаков на базе следующей гипотезы:

«*Хорошие поднаборы признаков содержат признаки, сильно коррелирующие с классификацией, но не коррелирующие друг с другом*».

Следующее равенство даёт оценку поднабора признаков *S*, состоящего из *k* признаков:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

В формулу входят среднее значение всех корреляций признак-классификация $\overline{r_{cf}}$, и среднее значение всех корреляций признак-признак $\overline{r_{ff}}$.

Критерий CFS (Correlation Feature Selection)

Критерий CFS определяется следующим образом:

$$CFS = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_k f_1})}} \right]$$

Переменные r_{cf_i} и $r_{f_i f_j}$ являются корреляциями, но не обязательно коэффициентами корреляции Пирсона.

Пусть x_i будет индикаторной функцией вхождения в множество для признака f_i. Тогда формула выше может быть переписана как задача оптимизации:

$$CFS = \max_{x \in \{0,1\}^n} \left[\frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} x_i + \sum_{i \neq j} 2 b_{ij} x_i x_j} \right]$$

Комбинаторные задачи выше являются, фактически, смешанными 0-1 задачами линейного программирования, которые могут быть решены с помощью алгоритма ветвей и границ.

Спасибо за внимание