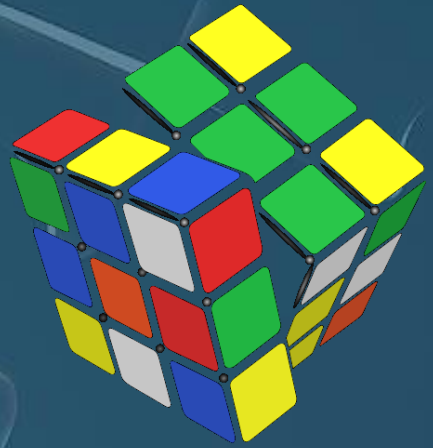
СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ



{ статистическая гипотеза - критерии принятия гипотез - критерий согласия Пирсона - критерий проверки – пример - критерии согласия Колмогорова и Смирнова }

Статистическая гипотеза

В математической статистике считается, что данные, получаемые в результате наблюдений, подчинены некоторому неизвестному вероятностному распределению, и задача состоит в том, чтобы извлечь из эмпирических данных правдоподобную информацию об этом неизвестном распределении. Один из подходов к этой общей задаче, состоит в проверке гипотез.

- *Статистической гипотезой* называют предположение о распределении вероятностей, которое необходимо проверить по имеющимся данным.
- Пусть $X(x_1, x_2, \ldots, x_n)$ - независимая выборка, соответствующая неизвестной функции распределения $F_\xi(t)$. *Простой гипотезой* называют предположение, состоящее в том, что неизвестная функция $F_\xi(t)$ отвечает некоторому совершенно конкретному вероятностному распределению.

  Пример простой гипотезы: H - данные являются выборкой из равномерного распределения на отрезке $[-1,1]$.

- *Сложной гипотезой* называют предположение о том, что неизвестная функция $F_\xi(t)$ принадлежит некоторому множеству распределений.

Критерии принятия гипотез

- Проверить статистическую гипотезу H - это значит на основе имеющихся данных $X(x_1, x_2, \ldots, x_n)$ *принять* или *отвергнуть* сделанное предположение.

Для этого используется подход, основанный на выборе так называемого *критического множества* S. Если данные наблюдений $X(x_1, x_2, \ldots, x_n)$ попадают в критическое множество, то гипотеза H отвергается, если они находятся вне этого множества, то гипотеза H – принимается.

Это правило называется *критерием*, основанным на критическом множестве S.

Если $X(x_1, x_2, \ldots, x_n) \in S \Rightarrow \neg H$ гипотеза H отвергается.

Если $X(x_1, x_2, \ldots, x_n) \notin S \Rightarrow H$ гипотеза H принимается.

Критерии принятия гипотез

В силу случайной природы наблюдаемых данных возможна первая ситуация, в то время, когда гипотеза H справедлива. В силу нашего правила мы отвергнем гипотезу H и, тем самым, допустим ошибку. В случае простой гипотезы вероятность такой ошибки равна

- $P_H(X(x_1, x_2, \ldots, x_n) \in S)$

Эту вероятность называют также *уровнем значимости* статистического критерия.

На практике уровень значимости критерия задается изначально, исходя из реальных приложений и последствий возможных ошибок.

Критерий согласия Пирсона

Рассмотрим независимую выборку $X(x_1, x_2, \ldots, x_n)$. Предположим неизвестную функцию распределения $F(t)$. Нас интересует вопрос о том, согласуются ли данные наблюдений $(x_1, x_2, \ldots, x_n)$ с простой гипотезой

$$H_o: F(t) = F_o(t),$$

где $F_o(t)$ – некоторая конкретная функция распределения.

Разобьем множество R на конечное множество непересекающихся подмножеств $D_1, \ldots, D_r$. P_o – вероятность, соответствующую функции распределения $F_o(t)$, обозначим $p^o_i = P_o(D_i)$, $i = 1, \ldots, r$.

Их сумма равна единице (правило нормировки): $\sum_{i=1}^{r} p_i^o = 1$

Критерий согласия Пирсона

Группируем выборочные данные по разрядам D_i

$$m_i = \left| \{ j : x_j \in D_i \} \right|, \quad i = 1,2,\ldots, r$$

и определяем эмпирические частоты m_i / n.
В силу случайных колебаний они будут отличаться от теоретических вероятностей $p^o{}_i$.

Чтобы контролировать это различие, следует подобрать хорошую меру расхождения между экспериментальными данными и гипотетическим теоретическим распределением.

По аналогии с идеей метода наименьших квадратов в качестве такой меры расхождения можно взять, например

$$\sum_{i=1}^{r} c_i \left(\frac{m_i}{n} - p_i^o \right)^2$$

где c_i – достаточно произвольные числа .

Критерий согласия Пирсона

$$\sum_{i=1}^{r} c_i \left(\frac{m_i}{n} - p_i^0 \right)^2$$

К. Пирсон показал, что если выбрать $c_i = n / p^o{}_i$, то полученная величина будет обладать рядом замечательных свойств.

Она называется *статистикой Пирсона* χ^2 .

$$X^2 = \sum_{i=1}^{r} \frac{n}{p_i^0} \left(\frac{m_i}{n} - p_i^0 \right)^2 = \sum_{i=1}^{r} \frac{(m_i - np_i^0)^2}{np_i^0}$$

Критерий согласия Пирсона

- Теорема К. Пирсона

Предположим, что гипотеза H_0 верна. Тогда при неограниченном росте объема выборки распределение величины X^2 сходится к распределению хи-квадрат с $(r-1)$ степенями свободы, где r – число параметров теоретического закона выборки, то есть

$$\lim_{n\to\infty} P\{X^2 \le t\} = P\{\chi^2_{r-1} \le t\} \quad \forall t \in R$$

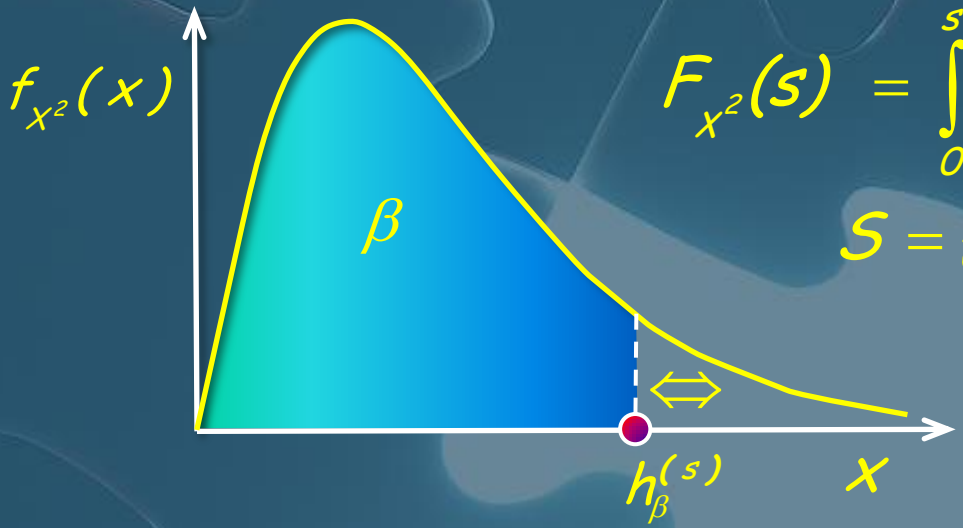
Смысл теоремы: при большом объеме выборки распределение X^2 можно считать распределением хи-квадрат с $(r-1)$ степенью свободы.

Если гипотеза H_0 неверна, то величина X^2 стремится в бесконечность.

Критерий согласия Пирсона

То обстоятельство, что поведение распределения X^2 различно в зависимости от того верна или не верна гипотеза H_0, дает возможность построить критерий для её проверки. Зададимся *уровнем значимости* (*допустимой вероятностью ошибки*) $\varepsilon > 0$ и возьмем квантиль распределения X^2, определяемый формулой ниже

$$X^2 \Rightarrow h_\beta^{(r-1)}, 0 < \varepsilon < 1 \Leftrightarrow F_{x^2}(h_\beta^{(1-r)}) = \beta \Leftrightarrow$$



$$F_{x^2}(s) = \int_0^s f_{x^2}(x)dx$$

Определим критическое множество

$$S = \{(x_1, \ldots, x_2) : X^2 > h_{1-\varepsilon}^{(r-1)}\}$$

$$P\{\chi^2_{r-1} > h_{1-\varepsilon}^{(r-1)}\} = \varepsilon$$

Критерий согласия Пирсона

Действия: определим X^2 и сравниваем её с квантилем $h_{1-\varepsilon}^{(r-1)}$.

Если неравенство справедливо $X^2 > h_{1-\varepsilon}^{(r-1)}$

гипотеза H_0 *отвергается*

(выборка обнаруживает значимое отклонение от гипотезы),

если нет $X^2 \le h_{1-\varepsilon}^{(r-1)}$ *то гипотеза* H_0 *принимается*

(выборка совместима с гипотезой H_0*).*

При таком решающем правиле мы может допустить ошибку, отвергнув верную гипотезу H_0 .

Из теоремы Пирсона вытекает, что при больших n величина вероятности этой ошибки близка к ε .

Границы применимости критерия согласия Пирсона

Утверждения теоремы Пирсона относятся к выборкам с пределу при $n \to \infty$. На практике мы имеем дело лишь с выборками ограниченного объема . Поэтому, применяя вышеописанный критерий, необходимо проявлять осторожность .

Согласно рекомендациям, применение критерия дает хорошие результаты, когда все ожидаемые частоты $np_i^0 \geq 10$. Если какие-то из этих чисел малы, то рекомендуется, укрупняя некоторые группы, перегруппировать данные таким образом, чтобы ожидаемые частоты всех групп были не меньше десяти .

Если число r достаточно велико, то порог для ожидаемых частот может быть понижен до 5 или даже до 3 , если r имеет порядок нескольких десятков .

Практически считается достаточным, чтобы $n > 50 - 60$ и $m_i > 5 - 8$

Пример

@ *Радиоактивное вещество наблюдалось в течение 2680 равных интервалов времени (по 7.5 секунд каждый). В каждом из интервалов регистрировалось число частиц, попавших в счетчик. В таблице приведены числа m_i интервалов времени, в течении которых в счетчик попадало ровно i частиц.*

i	m_i	i	m_i
0	57	6	273
1	203	7	139
2	383	8	45
3	525	9	27
4	532	>10	16
5	408	Итого :	$n = m_1 + .. + m_{>10} = 2680$

@ *Проверить, используя критерий хи - квадрат, гипотезу о согласии наблюденных данных с законом распределения Пуассона. Уровень значимости* ε *принять равным 5 %*

$$P(i,\lambda) = \frac{e^{-\lambda}\lambda^i}{i!}$$

Вычислим оценку параметра распределения λ

$$P(i,\tilde{\lambda}) = P(i,3.87) = \frac{e^{-3.87}3.87^i}{i!}$$

$$\tilde{\lambda} = \frac{\sum_{i=0}^{10} im_i}{n} = \frac{0\cdot 57 + 1\cdot 203 + 2\cdot 383 + 3\cdot 525 + 4\cdot 532}{2680} +$$

$$+ \frac{5\cdot 408 + 6\cdot 273 + 7\cdot 139 + 8\cdot 45 + 9\cdot 27 + 10\cdot 16}{2680} = 3.87$$

Вычисляем теоретические вероятности P_i попадания в счетчик i частиц при наличии закона Пуассона

Пример

i	p_i	np_i	m_i-np_i	$(m_i-np_i)^2$	$(m_i-np_i)^2/np_i)$
0	0.021	54.8	2.2	4.84	0.088
1	0.081	211.2	-8.2	67.24	0.318
2	0.156	406.8	-23.8	566.44	1.392
3	0.201	524.2	0.8	0.64	0.001
4	0.195	508.6	23.4	547.56	1.007
5	0.151	393.8	14.2	201.64	0.512
6	0.097	253.0	20.0	400.00	1.581
7	0.054	140.8	-1.8	3.24	0.023
8	0.026	67.8	-22.8	519.84	7.667
9	0.011	28.7	-1.7	2.89	0.101
>10	0.007	18.3	-2.3	5.29	0.289
	1.000				$\chi^2 = 13.049$

$$X_k^2 = \sum_{i=0}^{10} \frac{(m_i - np_i^0)^2}{np_i^0}$$

Пример

$$X_k^2 = \sum_{i=0}^{10} \frac{(m_i - np_i)^2}{np_i} = 13.05$$

Число степеней свободы: $k = l - r - 1 = 11 - 1 - 1 = 9$

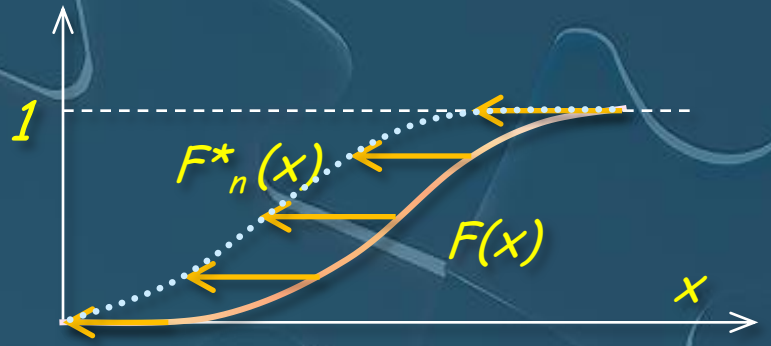
В таблице для статистики Пирсона для $k = 9$ **и** $\chi^2 = 13.05$ находим вероятность того, что величина χ^2 превзойдет значение, полученное по выборке

$$\varepsilon_k = P(\chi^2 \geq X_n^2) = 0.166$$

Так как $\varepsilon_k > \varepsilon = 0.05$, то отклонения от закона Пуассона незначимы.

Критерии согласия Колмогорова - Смирнова

Критерий согласия Колмогорова применим в том случае, когда параметры теоретического закона распределения определяются не по данным исследуемой выборки. За меру расхождения принимается наибольшее значение статистики D : абсолютной величины разности статистической и теоретической функций :



- $D_n = \sup_{x \in R} |F_n^*(x) - F(x)|$

При неограниченном росте объема выборки величина $\lambda = \sqrt{n}D$ независимо от вида закона распределения СВ X стремится к закону распределения Колмогорова.

$$\varepsilon_n = P(D \geq D_n) = P(\lambda) = 1 - K(\lambda)$$

- Статистика Смирнова: $\omega_n^2 = \int_{-\infty}^{\infty} (F_n^*(t) - F(t))^2 dF(t)$