Статистика ФИВТ ПМИ

Прикладной поток

Лекция 12

6. Критерии согласия

6.1. Общие критерии согласия

Задача. Пусть $X_1, ..., X_n$ — выборка из **нормального распределения** с параметрами a, σ. Найти оценку максимального правдоподобия.

Задача. Пусть $X_1, ..., X_n$ — выборка из **гамма-распределения** с параметрами α, β. Найти доверительный интервал для параметра α.

Задача. Пусть $X_1, ..., X_n$ — выборка из **пуассоновского распределения** с параметром θ. Построить асимптотический критерий проверки гипотез $\mathrm{H}_0 : \theta = \theta_0$ *vs*. $\mathrm{H}_1 : \theta \neq \theta_0$.

Откуда **это** известно?

Критерии согласия

$H_0: P = P_0$

т.е. $H_0: P \in \mathscr{P}_0$, где $\mathscr{P}_0 = \{P_0\}$

$H_0: P$ — нормальное распределение

т.е. $H_0: P \in \mathscr{P}_0$, где $\mathscr{P}_0 = \{P_{a,\sigma}$ — распределение $\mathcal{N}(a, \sigma^2)\}$

$H_0: P$ — гамма-распределение

т.е. $H_0: P \in \mathscr{P}_0$, где $\mathscr{P}_0 = \{P_{\alpha,\beta}$ — распределение $\Gamma(\alpha, \beta)\}$

Соответствующие критерии называются **критерии согласия**.

Если данные не противоречат проверяемым свойствам распределения (семейства распределений), то можно считать, что выборка **согласуется** с основной гипотезой.

Критерий Колмогорова

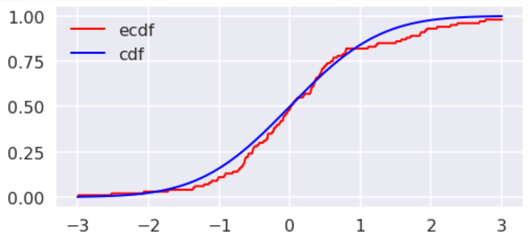
$X = (X_1, ..., X_n)$ — выборка из неизвестного распределения с ф.р. F.

$\mathrm{H}_0 : F = F_0$

$D_n = \sup\limits_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|$ — статистика критерия, где $\widehat{F}_n$ — ЭФР

Теорема Колмогорова:
Если F_0 непрерывна $\sqrt{n} D_n \xrightarrow{d_0} \xi \sim K$ — распр. Колмогорова.

Критерий: $\{\sqrt{n} D_n \geqslant c_\alpha\}$, где c_α — (1-α)-квантиль распр. K.



Критерий Колмогорова

$X = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения с ф.р. F.

$\mathrm{H}_0\colon F = F_0$

$D_n = \sup\limits_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|$ — статистика критерия, где $\widehat{F}_n$ — ЭФР

Критерий: $\{\sqrt{n} D_n \geqslant c_\alpha\}$, где c_α — $(1\text{-}\alpha)$-квантиль распр. K.

Свойства:

1. статистика D_n не зависит от конкретного F_0;
2. имеет низкую мощность;
3. не чувствителен к различиям на хвостах;
4. применим при $n \geqslant 20$;
5. выполнена состоятельность: $\beta(\mathrm{P}) \to 1$ при $n \to \infty$ и $\mathrm{P} \neq \mathrm{P}_0$.

Другие критерии согласия, основанные на различиях между F_0 и $\widehat{F}_n$.

1. Джини: $\int \left| \widehat{F}_n(x) - F_0(x) \right| dx$;
2. Крамера-фон Мизеса: $\int \left(\widehat{F}_n(x) - F_0(x) \right)^2 dx$;
3. Смирнова-Крамера-фон Мизеса: $\int \left(\widehat{F}_n(x) - F_0(x) \right)^2 dF_0(x)$;
4. Андерсона-Дарлинга: $\int \frac{\left(\widehat{F}_n(x) - F_0(x) \right)^2}{F_0(x)(1 - F_0(x))} dF_0(x)$;
5. Купера: $\sup\limits_{x\in\mathbb{R}} (\widehat{F}_n(x) - F_0(x)) + \sup\limits_{x\in\mathbb{R}} (F_0(x) - \widehat{F}_n(x))$;
6. Ватсона: $\int \left(\widehat{F}_n(x) - F_0(x) - \int \left(\widehat{F}_n(x) - F_0(x) \right) dF_0(x) \right) dF_0(x)$;
7. Фроцини: $\int \left| \widehat{F}_n(x) - F_0(x) \right| dF_0(x)$.

Пусть $\mathscr{P} = \{\mathrm{P}_{a,\sigma}\}$ — семейство распределений, где a — параметр сдвига, σ — параметр масштаба.

Для проверки $\mathrm{H}_0: \; \mathrm{P} \in \mathscr{P}$ существуют модификации этих критериев, которые используют $\widehat{a}$ и $\widehat{\sigma}$ — состоятельные оценки.

Зачем столько критериев?

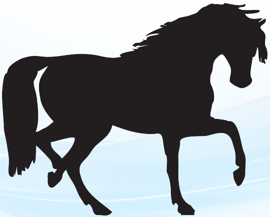
Пусть $\mathscr{P}$ — все парнокопытные звери.

$H_0: \ P$ — лошадь

S — критерий, проверящий некоторое свойство парнокопытного.

Если результат не свойственен лошади,

то гипотезу следует отклонить.



Животное 1

Животное 1

Животное 1

Животное 1

Животное 1

Животное 2

Животное 2

Животное 2

Животное 3

Животное 3

Животное 3

Животное 3

Животное 3

Животное 3

Животное 4

Животное 4

Выводы

- Разные критерии проверяют выполнение разных свойств распределения (или класса распределений).
  Если свойство не выполняется, критерий отвергает гипотезу.
  Лошади обычно не пятнистые.
- Иначе, возможно, критерий недостаточно мощный.
  Копыта есть у всех.
- Чем больше разных критериев, тем лучше.
- Комбинирование происходит с помощью МПГ.

QQ plot — графический способ

Пусть $\mathscr{P} = \{\mathsf{P}_{a,\sigma}\}$ — семейство распределений,

a — параметр сдвига,

σ — параметр масштаба.

$X = (X_1, ..., X_n)$ — выборка из неизв. распр. P.

Верно ли, что $\mathsf{P} \in \mathscr{P}$?

———————

Пусть $F_{a,\sigma}$ — функция распр. $\mathsf{P}_{a,\sigma}$.

По свойствам параметров: $F_{a,\sigma}(x) = F_{0,1}\left(\frac{x-a}{\sigma}\right)$.

Утверждение: $F_{a,\sigma}(X_i) \sim U[0,1]$

$$\Rightarrow \quad F_{a,\sigma}(X_{(1)}) = F_{0,1}\left(\frac{X_{(1)}-a}{\sigma}\right), ..., F_{a,\sigma}(X_{(n)}) = F_{0,1}\left(\frac{X_{(n)}-a}{\sigma}\right)$$

— вариационный ряд, построенный по выборке из $U[0,1]$.

QQ plot — графический способ

$$F_{a,\sigma}(X_{(1)}) = F_{0,1}\left(\frac{X_{(1)}-a}{\sigma}\right), \ldots, F_{a,\sigma}(X_{(n)}) = F_{0,1}\left(\frac{X_{(n)}-a}{\sigma}\right)$$

— вариационный ряд, построенный по выборке из $U[0,1]$.

$$\Rightarrow \quad X_{(i)} = a + \sigma \cdot F_{0,1}^{-1}\left(F_{a,\sigma}\left(X_{(i)}\right)\right)$$

Вместо выборочных квантилей $F_{a,\sigma}\left(X_{(i)}\right)$ подставим теоретические квантили $\mathsf{E}F_{a,\sigma}(X_{(i)}) = \frac{i}{n+1}$ (из ДЗ-1):

$$X_{(i)} \approx a + \sigma \cdot F_{0,1}^{-1}\left(\mathsf{E}F_{a,\sigma}\left(X_{(i)}\right)\right) = a + \sigma \cdot F_{0,1}^{-1}\left(\frac{i}{n+1}\right)$$

Вывод: точки $\left(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right)\right)$ примерно лежат на одной прямой.

Q-Q plot — график точек $\left(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right)\right)$.

QQ plot — графический способ

Q-Q plot — график точек $\left(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right)\right)$.

6. Критерии согласия

6.2. Критерии проверки нормальности

Критерий Жарка-Бера

$X = (X_1, \dots, X_n)$ — выборка из неизвестного распределения P.

$\mathsf{H}_0:$ P — нормальное распределение

$\varkappa = \frac{\mathsf{E}(\xi - a)^3}{\sigma^3}$ — коэффициент асимметрии (skewness)

$\gamma = \frac{\mathsf{E}(\xi - a)^4}{\sigma^4} - 3$ — коэффициент эксцесса (kurtosis)

Для нормального распределения $\varkappa = 0, \gamma = 0$.

Их оценки методом подстановки:

$$\widehat{\varkappa} = \frac{1}{nS^3} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^3, \quad \widehat{\gamma} = \frac{1}{nS^4} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^4 - 3.$$

Статистика критерия $T(X) = \frac{n}{6}\left(\widehat{\varkappa}^2 + \widehat{\gamma}^2/4\right) \sim \chi^2_2$.

Критерий: $\left\{T(x) > \chi^2_{2,1-\alpha}\right\}$.

p-value: $p(t) = 1 - F(t)$, где F — функция распр. χ^2_2.

Критерий Шапиро-Уилка

$X = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения P.

$\mathsf{H}_0\colon$ P — нормальное распределение

Статистика критерия

$$W(X) = \left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2 \Big/ nS^2$$

$$a = \frac{m^T V^{-1}}{(m^T V^{-2} m)^{1/2}}$$

m и V — вектор средних и матрица ковариаций вар. ряда $\mathcal{N}(0, 1)$.

При H_0 имеет табличное распределение.

На практике обычно он самый мощный для проверки нормальности.

О проверке нормальности

- Часто в предположении нормальности выборки применимы более мощные методы.
- Перед использованием таких методов нужно проверить нормальность.
- Методы могут быть устойчивыми к небольшим отклонениям от нормальности.
- Если метод устойчив к отклонениям, обычно достаточно посмотреть на Q-Q plot.
- Если метод неустойчив к отклонениям, желательно проверить нормальность критерием Шапиро-Уилка.

Критерии проверки нормальности [Кобзарь]

Сравнение критериев проверки нормальности [Кобзарь]

Наименование критерия (раздел)	Характер альтернативного распределения					Ранг
	асимметричное		симметричное		≈ нормальное	
	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 < 3$	$\alpha_4 > 3$	$\alpha_4 \approx 3$	
Критерий Шапиро–Уилка (3.2.2.1)	1	1	3	2	2	1
Критерий K^2 (3.2.2.16)	7	8	10	6	4	2
Критерий Дарбина (3.1.2.7)	11	7	7	15	1	3
Критерий Д'Агостино (3.2.2.14)	12	9	4	5	12	4
Критерий α_4 (3.2.2.16)	14	5	2	4	18	5
Критерий Васичека (3.2.2.2)	2	14	8	10	10	6
Критерий Дэвида–Хартли–Пирсона (3.2.2.10)	21	2	1	9	1	7
Критерий χ^2 (3.1.1.1)	9	20	9	8	3	8
Критерий Андерсона–Дарлинга (3.1.2.4)	18	3	5	18	7	9
Критерий Филлибена (3.2.2.5)	3	12	18	1	9	10
Критерий Колмогорова–Смирнова (3.1.2.1)	16	10	6	16	5	11
Критерий Мартинеса–Иглевича (3.2.2.14)	10	16	13	3	15	12
Критерий Лина–Мудхолкара (3.2.2.13)	4	15	12	12	16	13
Критерий α_3 (3.2.2.16)	8	6	21	7	19	14
Критерий Шпигельхальтера (3.2.2.11)	19	13	11	11	8	15
Критерий Саркади (3.2.2.12)	5	18	15	14	13	16
Критерий Смирнова–Крамера–фон Мизеса (3.1.2.2)	17	11	20	17	6	17
Критерий Локка–Спурье (3.2.2.7)	13	4	19	21	17	18
Критерий Оя (3.2.2.8)	20	17	14	13	14	19
Критерий Хегази–Грина (3.2.2.3)	6	19	16	19	21	20
Критерий Муроты–Такеучи (3.2.2.17)	15	21	17	20	20	21

Другие распределения [Кобзарь]

6. Критерии согласия

6.3. Критерий хи-квадрат

Критерий хи-квадрат

$X = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения P

со значениями в $\mathscr{X}$.

$\mathsf{H}_0\colon\ \mathsf{P} = \mathsf{P}_0$

Рассмотрим разбиение $\mathscr{X} = \bigsqcup\limits_{j=1}^{k} B_j$

Статистика критерия $\widehat{\chi} = \sum\limits_{j=1}^{k} \dfrac{\left(\mu_j - np_j^0\right)^2}{np_j^0}$,

где $\mu_j = \#\{i \mid X_i \in B_j\}$,

$p_j^0 = \mathsf{P}_0(X_1 \in B_j)$

Теорема Пирсона: $\widehat{\chi} \xrightarrow{d_0} \chi^2_{k-1}$.

Критерий: $\{\widehat{\chi} \geqslant \chi^2_{k-1,1-\alpha}\}$

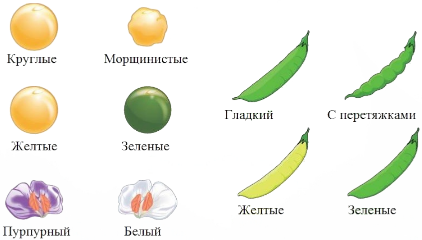
Свойства:

1. $n \geqslant 50$, $\forall i : np_j^0 \geqslant 5$ — необходимо для хорошего приближения;
2. обычно берут $k \sim \log_2 n$;
3. применим во многих задачах;
4. имеет маленькую мощность;
5. неоднозначное разбиение на интервалы;

Горошек Менделя

Мендель рассматривал

следующую классификацию гороха:



Рассмотрим непосредственно семена $(n = 556, k = 4)$

B_j — вид горошин	круглые желтые	морщин. желтые	круглые зеленые	морщин. зеленые
Вер-ти $p_j^0 = \mathrm{P}_0(X_1 \in B_j)$	$9/16$	$3/16$	$3/16$	$1/16$
Частота $\mu_j = \#\{X_i \in B_j\}$	315	101	108	32

$$\widehat{\chi} = \frac{(315-556\cdot 9/16)^2}{556\cdot 9/16} + \frac{(101-556\cdot 3/16)^2}{556\cdot 3/16} + \frac{(108-556\cdot 3/16)^2}{556\cdot 3/16} + \frac{(32-556\cdot 1/16)^2}{556\cdot 1/16} \approx 0.47$$

p-value $\approx 0.925 \Longrightarrow$ гипотеза не отвергается.

Обобщенный критерий хи-квадрат

$X = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения P со значениями в $\mathscr{X}$.

$\mathsf{H}_0\colon\ \mathsf{P} \in \mathscr{P}_0$, где $\mathscr{P}_0 = \{\mathsf{P}_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$

Рассмотрим разбиение $\mathscr{X} = \bigsqcup\limits_{j=1}^{k} B_j$

Статистика критерия

$$\widehat{\chi} = \sum_{j=1}^{k} \frac{\left(\mu_j - np_j^0(\widehat{\theta})\right)^2}{np_j^0(\widehat{\theta})},$$

где $\mu_j = \#\{i \mid X_i \in B_j\}$,

$p_j^0(\theta) = \mathsf{P}_\theta(X_1 \in B_j)$

ОМП $\widehat{\theta}$ получается из условия $\sum\limits_{j=1}^{k} \mu_j \ln p_j^0(\theta) \to \max\limits_{\theta}$

Теорема: в условиях регулярности $\widehat{\chi} \xrightarrow{d_0} \chi^2_{k-1-d}$.

Примеры на критерий хи-квадрат

Каждый человек имеет кровь, принадлежащую одной из четырех групп: 0, A, B, AB. Наследование управляется тремя генами: A, B, 0, при этом ген 0 подавляется генами A и B.

p, q, r — вероятности появления генов A, B, 0 у родителей.

Группа B_j	Гены родителей	Вероятность $p_j^0(\theta)$	Данные μ_j
0	00	r^2	121
A	AA, A0	$p^2 + 2pr$	120
B	BB, B0	$q^2 + 2qr$	79
AB	AB	$2pq$	33

H_0 : механизм наследования крови имеет место

Примеры на критерий хи-квадрат

B_j — группы 0, A, B, AB, т.е. $k = 4$.

$d = 2$ — количество параметров (используем $r = 1 - p - q$).

ОМП для параметра $\theta = (p, q)$ ищем из условия:

$$\sum_{j=1}^{k} \mu_j \ln p_j^0(\theta) = 121 \ln r^2 + 120 \ln(p^2 + 2pr) + 79 \ln(q^2 + 2qr) + 33 \ln 2pq \to \max_{p,q}$$

Получаем $\widehat{p} \approx 0.241, \widehat{q} \approx 0.167, \widehat{r} \approx 0.592$.

Оценки $p_j^0(\widehat{\theta})$: $\quad \widehat{p}_1 = 0.343, \widehat{p}_2 = 0.340, \widehat{p}_3 = 0.224, \widehat{p}_4 = 0.093$.

Отсюда $\widehat{\chi} = 0.001, pvalue = 0.97$.