Семинар 3

Суслова Ирина

19 февраля 2024

1 Критерий согласия хи-квадрат Пирсона

Условия.
Дана дискретная случайная величина ξ, принимающая значения $1, 2, \dots, N$ с некоторыми неизвестными вероятностями $p_1, \dots, p_N$, которые образуют вектор $p = (p_1, \dots, p_N)$. Имеется выборка $X = (X_1, X_2, \dots X_n)$ и вектор вероятностей $p^0 = (p_1^0, \dots, p_N^0)$, причем $0 < p_j < 1$ для всех $j = 1, \dots, N$. Выдвинута простая гипотеза:

$$\mathrm{H}_1 : \ p = p^0$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.
Алгоритм

1. Вычислить частоты исходов

$$\nu_j = \sum_{i=1}^{n} I(X_i = j), \ j = 1, \dots, N$$

2. Вычислить статистику критерий

$$T_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^0)^2}{np_j^0}$$

3. Выбрать в качестве критческой области $\Omega_2 = \{x \in \Omega : \ T_{\chi^2} > t_\alpha\}$, и найти t_α из условия на уровень значимости:

$$\mathbb{P}_1(T_{\chi^2} > t_\alpha) = \alpha$$

.

4. Принять решение по следующей схеме:

$$\mathrm{H}_1 \text{ отвергается} \Longleftrightarrow T_{\chi^2} > t_\alpha$$

Известно, что при истинности гипотезы H_1 статистика $T_{\chi^2} \xrightarrow[n\to\infty]{d} \chi^2(N-1)$, поэтому границу t_α можно вычислять как $(1-\alpha)$-квантиль распределения $\chi^2(N-1)$.

Замечание 1. Для того, чтобы воспользоваться фактом сходимости распределения статистики критерия к распределению хи-квадрат, критерий рекомендуется применять при $n \geq 50$ и $\nu_j \geq 5$ для всех $j = 1, \dots, N$

Замечание 2. .Вектор $\nu = (\nu_1, \dots, \nu_N)$ имеет полиномиальное распределение $M(n, p_1, \dots, p_N)$ с функцией вероятности

$$\mathbb{P}(\nu_1 = k_1, \dots, \nu_N = k_N) = \frac{N!}{k_1!k_2!\dots k_N!} p_1^{k_1} \dots p_N^{k_N}, \; k_1 + \dots + k_N = n.$$

Про это распределение известно, что

$$\nu_j \in \mathrm{Bi}(n, p_j), \; \mathbb{E}\nu_j = np_j, \; \mathbb{D}\nu_j = np_j(1-p_j),$$
$$\forall i, j : i \neq j \; \mathrm{cov}(\nu_i.\nu_j) = -np_ip_j$$

.

Замечание 3. Если случайная величина ξ имеет непрерывное распределение, то, чтобы воспользоваться критерием хи-квадрат, можно применить метод группировки наблюдений: разбить пространство значений на N непересекающщхся интервалов, задать гипотетические вероятности попасть в эти интервалы и применить критерий.

Замечание 4. Статистика T_{χ^2} представляет собой меру хи-квадрат отклонения эмпирических данных от гипотетических. Чтобы лучше себе представить содержимое этого выражения, надо вспомнить, что согласно закону больших чисел, $\nu_j/n \xrightarrow{\mathbb{P}}$ с ростом объема выборки. Поэтому при достаточно больших n и при условии истинности гипотезы H_1 мы ожидаем, что разница $(\nu_j - np_j^0)^2$ будет небольшой. Весовой коэффициент $1/np_j^0$ позволяет получить сходимость статистики к распределению хи-квадрат, и таким образом позволяет воспользоваться таблицами распределений хи-квадрат

Задача 1.1 (Задача 1)**.**
Монету подбросили 4040 рз. Решка выпала 2048 раз, орел выпал 1992 раза На уровне значимости $\alpha = 0.05$ проверить гипотезу о симметричности монеты

Решение:
По условию задачи нам даны частоты исходов: $\nu_1 = 2048 > 5, \; \nu_2 = 1992 > 5, \; N = 2, \; n = 4040 > 50$
Гипотеза о симметричности монеты:

$$\mathrm{H}_1 : p = p^0 = [1/2, 1/2]$$

Вычислим реализацию статистики критерия:

$$T_{\mathrm{ch}^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^0)^2}{np_j^0} = 0.3881 + 0.3881 = 0.7762$$

Теперь заглянем в таблицу квантилей и найдем квантиль $t_\alpha = \chi^2_{1-\alpha}(N-1) = \chi^2_{0.95}(1) = 3.841$. Видим, что $T_{\chi^2} = 0.7762 < 3.841$, поэтому ответ: "Данные гипотезе не противоречат"

2 Критерий хи-квадрат для сложной гипотезы

Условия.
Дана дискретная случайная величина ξ, принимающая значения $1, 2, \dots, N$ с некоторыми неизвестными вероятностями $p_1, \dots, p_N$, которые образуют вектор $p = (p_1, \dots, p_N)$. Имеется выборка $X = (X_1, X_2, \dots X_n)$ и гладкая вектор-функция $p^0(\theta) = (p_1^0(\theta), \dots, p_N^0(\theta))$, $\theta = (\theta_1, \dots, \theta_r) \in \Theta$, где $r < N-1$. Выдвинута сложная гипотеза:

$$\mathrm{H}_1: \; p = p^0(\theta), \text{ для некоторого } \theta \in \Theta$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.
Алгоритм

1. Вычислить частоты исходов

$$\nu_j = \sum_{i=1}^{n} I(X_i = j), \; j = 1, \dots, N$$

2. Оценить неизвестный параметр θ. Для этого используют оценку максимального правдоподобия:

$$\hat{\theta} = \arg\max_{\theta} \prod_{j=1}^{N} (p_j^0(\theta))^{\nu_j},$$

которая находится из системы r уравнний

$$\sum_{j=1}^{N} \frac{\nu_j}{p_j^0(\theta)} \cdot \frac{\partial p_j^0(\theta)}{\partial \theta_k} = 0$$

Решение обозначим $\hat{\theta}$

3. Вычислить статистику критерий

$$T_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^0(\hat{\theta}))^2}{np_j^0(\hat{\theta})}$$

4. Выбрать в качестве критческой области $\Omega_2 = \{x \in \Omega : \; T_{\chi^2} > t_\alpha\}$, и найти t_α из условия на уровень значимости:

$$\mathbb{P}_1(T_{\chi^2} > t_\alpha) = \alpha$$

.

5. Принять решение по следующей схеме:

$$\mathrm{H}_1 \text{ отвергается} \Longleftrightarrow T_{\chi^2} > t_\alpha$$

Известно, что при истинности гипотезы H_1 и выполнении условий

1. $\sum\limits_{j=1}^{N} p_j^\theta = 1, \forall\theta \in \Theta$
2. $p_j^0 \geq c > 0,\ \forall j$ и функции p_j^0 дважды непрерывно дифференцируемы
3. матрица $\|\partial p_j^0(\theta)/\partial\theta_k\|$ имеет ранг r для всех $\theta \in \Theta$

статистика $T_{\chi^2} \xrightarrow[n\to\infty]{d} \chi^2(N-1-r)$, поэтому границу t_α можно вычислять как $(1-\alpha)$-квантиль распределения $\chi^2(N-1-r)$.
Для критерия хи-квадрат для сложной гипотезы справедливы такое же замечание на условия применимости и замечание про случайную величину из непрерывного распределения, как и для простой гипотезы
Замечание. Оценка максимального правдоподобия - такое значение неизвестного параметра, при котором вероятность получить имеющиеся данные максимальна.

Задача 2.1 (Задача 1)**.**
Среди 2020 семей, имеющих двух детей, 527 детей с двумя мальчиками, 476 с двумя девочками, а остальные 1017 семей имеют и девочку, и мальчика. Можно ли с уровнем значимости 0.05 считать, что количество мальчиков в семье с двумя детьми - биномиальная величина?

Решение:
По условию задачи пусть ξ-кол-во мальчиков в семье, принимает 3 значения: $0, 1, 2$. По условию нам даны частоты исходов: $\nu_0 = 476 > 5,\ \nu_1 = 1017 > 5,\ \nu_2 = 527,\ N = 3,\ n = 2020 > 50$
Гипотеза о симметричности монеты:

$$\mathrm{H}_1 : \xi \in \mathrm{Bi}(2, \theta)$$

Сначала найдем гипотетические вероятности:

$$p_0^0(\theta) = \mathbb{P}_0(\xi = 0) = (1-\theta)^2,$$

$$p_1^0(\theta) = \mathbb{P}_0(\xi = 1) = 2\theta(1-\theta),$$

$$p_2^0(\theta) = \mathbb{P}_0(\xi = 2) = \theta^2$$

Оценим неизвестный параметр, который у нас один, т.е $r = 1$, решив одно уравнение

$$\sum_{j=1}^{2} \frac{\nu_j}{p_j^0(\theta)} \frac{\partial p_j^0(\theta)}{\partial\theta} = 0$$

Подставив выражения для $p_j^0(\theta)$ и решив полученное уравнение относительно параметра θ, получим

$$\hat{\theta} = 0.5126$$

и оценку гипотетических вероятностей

$$p_0^0(\hat{\theta}) = 0.2375,\ p_1^0(\hat{\theta}) = 0.4997,\ p_2^0(\hat{\theta}) = 0.2628$$

Вычислим реализацию статистики критерия:

$$T_{\mathrm{ch}^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^0(\hat{\theta})^2}{np_j^0(\hat{\theta})} = 0.1158$$

Теперь заглянем в таблицу квантилей и найдем квантиль $t_\alpha = \chi^2_{1-\alpha}(N - 1 - r) = \chi^2_{0.95}(1) = 3.841$. Видим, что $T_{\chi^2} = 0.1158 < 3.841$, поэтому ответ: "Данные гипотезе не противоречат"