Математическая статистика. ДЗ 3.

Прохоров Юрий, 776

Задача 1

Проводятся много экспериментов по селекции гороха. В таблицу записано, сколько раз появился каждый вид семян. Также указана теоретическая вероятность появления каждого вида семян.

№	вид семян	частота, ν_k	теоретическая вероятность, p_k
1	круглые и желтые	315	9/16
2	морщинистые и желтые	101	3/16
3	круглые и зеленые	108	3/16
4	морщинистые и зеленые	32	1/16

Проверить гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями на уровне значимости $\alpha = 0.1$.

Решение:

Так как теоретическая функция распределения разрывная, критерий Колмогорова применять нельзя, а критерий χ^2 Пирсона можно.

Данные уже сгруппированы хорошо ($\nu_k \geq 5$) и их достаточно много ($n \geq 50$).

Считаем статистику Пирсона ($n = 556,\ r = 4$):

$$T = \sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k} \approx 0.470$$

Согласно основной теореме, можно считать, что $T \sim \chi^2(r-1) = \chi^2(3)$.

Из таблицы для χ^2-распределения, находим соответствующий α-квантиль:

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

В нашем случае α-квантиль есть $t_\alpha = 6.251$.

$$T < t_\alpha \qquad \Longrightarrow \qquad \text{данные согласованы с гипотезой на уровне } \alpha = 0.1$$

По таблице можно оценить *p-значение* — наибольший уровень значимости, при котором гипотеза еще принимается. В нашем случае $0.9 < p < 0.95$.

Задача 2

Пусть $\mathbf{X}_n = (X_1, \dots, X_n)$ — простая выборка, $\nu_1, \dots, \nu_r$ — частоты попаданий элементов выборки в интервалы $\Delta_1, \dots, \Delta_r$ соответственно, $p_1, \dots, p_r$ — теоретические вероятности попаданий в эти интервалы.

Доказать, что для *статистики* χ^2 *Пирсона* выполнено

$$T_n = T(\mathbf{X}_n) = \sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k} \xrightarrow[n\to\infty]{d} \chi^2(r-1)$$

Решение:

1. Рассмотрим последовательность (по n) случайных векторов $\xi^{(n)}$ с компонентами

   $$\xi_k^{(n)} = \frac{\nu_k - np_k}{\sqrt{np_k}}, \qquad k = \overline{1,r}$$

   Тогда можно записать

   $$T_n = \sum_{k=1}^{n} \left(\xi_k^{(n)}\right)^2 = \left\|\xi^{(n)}\right\|_2^2$$

2. Представим $\xi^{(n)}$ в виде суммы независимых случайных векторов:

   $$\xi^{(n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i^{(n)}, \qquad \eta_i^{(n)} = \begin{bmatrix} \dots \\ \frac{\mathbb{I}\{X_i \in \Delta_k\} - p_k}{\sqrt{p_k}} \\ \dots \end{bmatrix}$$

   Все $\eta_i^{(n)}, i = \overline{1,n}$ независимы, потому что все X_i независимы.

3. Применим многомерную ЦПТ:

   ЦПТ для случайных векторов

   Пусть $\vec{X_1}, \dots, \vec{X_n}, \dots$ последовательность независимых и одинаково распределённых случайных векторов, каждый из которых имеет среднее $E\vec{X_1} = \vec{a}$ и невырожденную матрицу ковариаций Σ. Обозначим через $S_n = \vec{X_1} + \dots + \vec{X_n}$ вектор частичных сумм. Тогда при $n \to \infty$ имеет место слабая сходимость распределений векторов

   $\vec{\eta_n} = \frac{S_n - na}{\sqrt{n}} \xrightarrow{weak} \vec{\eta}$, где $\vec{\eta}$ имеет распределение $N(\vec{0}, \Sigma)$.

   Тут слабая сходимость — сходимость по распределению, а невырожденность ковариационой матрицы Σ можно опустить.

   Вычислим ковариационную матрицу $\Sigma = \mathbb{E}\big[(\eta_i - \mathbb{E}\eta_i)(\eta_i - \mathbb{E}\eta_i)^T\big]$. В нашем случае все $\mathbb{E}\eta_i = 0$.

   $$\Sigma_{kl} \overset{k\neq l}{=} \mathbb{E}\left[\frac{\mathbb{I}\{X_i \in \Delta_k\} - p_k}{\sqrt{p_k}} \cdot \frac{\mathbb{I}\{X_i \in \Delta_l\} - p_l}{\sqrt{p_l}}\right] = \frac{1}{\sqrt{p_k p_l}} \left[0 - p_k p_l - p_k p_l + p_k p_l\right] = -\sqrt{p_k p_l}$$

   $$\Sigma_{kk} = \mathbb{E}\left[\frac{\mathbb{I}\{X_i \in \Delta_k\} - p_k}{\sqrt{p_k}}\right]^2 = \frac{1}{p_k}\left[p_k - 2p_k^2 + p_k^2\right] = 1 - p_k$$

   Тогда по ЦПТ:

   $$\xi^{(n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i^{(n)} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma)$$

4. Из эквивалентного определения сходимости по распределению следует, что

   $$\xi_n \xrightarrow[n\to\infty]{d} \xi \qquad \Longrightarrow \qquad f(\xi_n) \xrightarrow[n\to\infty]{d} f(\xi) \text{ для любой непрерывной функции } f$$

   Тогда имеем, что

   $$T_n = \left\|\xi^{(n)}\right\|_2^2 \xrightarrow[n\to\infty]{d} \left\|\xi\right\|_2^2, \qquad \xi \sim \mathcal{N}(0, \Sigma)$$

5. Осталось найти распределение $\|\xi\|_2^2$. Сначала заметим, что матрица Σ является идемпотентной, т.е. $\Sigma^2 = \Sigma$.

   Представим Σ в виде

   $$\Sigma = I - bb^T, \qquad b = \begin{bmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_r} \end{bmatrix}$$

   Тогда имеем

   $$\Sigma^2 = I - 2bb^T + b \underbrace{b^T b}_{1} b^T = I - bb^T = \Sigma$$

   Тогда ранг матрицы

   $$\operatorname{rg}\Sigma = \operatorname{tr}\Sigma = \sum_{k=1}^{r}(1 - p_k) = r - 1$$

6. Матрица Σ идемпотентная и ранга $r-1$, значит собственные значения

   $$\lambda_1 = \ldots = \lambda_{n-1} = 1, \qquad \lambda_n = 0$$

   Из симметричности следует, что существует ортогональная матрица U, такая, что

   $$\Sigma = U^T \Lambda U, \qquad \Lambda = \operatorname{diag}(1, \ldots, 1, 0)$$

7. Рассмотрим случайную величину

   $$\eta = U\xi \qquad \Longrightarrow \qquad \eta \sim \mathcal{N}(0, U\Sigma U^T) = \mathcal{N}(0, \Lambda)$$

   Так как матрица U ортогональная, евклидова норма сохраняется и получается, что

   $$T_n \xrightarrow[n\to\infty]{d} \|\xi\|_2^2 = \|U^T\eta\|_2^2 = \|\eta\|_2^2 = \sum_{k=1}^{r-1} \eta_k^2,$$

   где $\eta_k \sim \mathcal{N}(0,1)$ — некоррелированные (а, значит, и независимые) компоненты случайного нормального вектора η.

8. Итак, имеем

   $$T_n \xrightarrow[n\to\infty]{d} \sum_{k=1}^{r-1} \eta_k^2 \sim \chi^2(r-1)$$