Глава 15

Проверка на непараметрични хипотези

Основните задачи за проверка на непараметрични хипотези са:

- Проверка на хипотеза за вида на разпределението на случайна величина;
- Проверка на хипотезата за независимост на две случайни величини.

15.1 Проверка на хипотеза за вида на разпределението. χ^2 критерий на Пирсън

Нека са извършени n случайни наблюдения над статистическата променлива X и е получена извадката с обем n:

$$x_1, x_2, \ldots, x_n. \tag{15.1}$$

След като сме подредили числата от тази извадка и сме получили вариационния ред

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)} \quad (x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}), \tag{15.2}$$

можем да намерим статистическата функция на разпределение и да построим нейната графика. На следващия етап от обработката на статистическите данни може да получим таблицата на честотите и да построим

полигона на честотите (ако X е дискретна) или да получим интервалната таблица на честотите и да построим хистограмата на относителните честоти и кумулативната крива (ако X е непрекъсната). От вида на тези графики (както и от други съображения) може да се направи предположение за вида на разпределението на с.в. X. Допълнително това предположение може да се затвърди след като оценим параметрите на разпределението. След този анализ на статистическия материал може да се направи непараметричната основна хипотеза

$H_0 : X$ има функция на разпределение $F_0(x)$ (плътност на разпределение $f_0(x)$).

при алтернативна хипотеза

$H_1 : X$ има функция на разпределение $F_1(x) \neq F_0(x)$ (плътност на разпределение $f_1(x) \neq f_0(x)$).

Един от най-използваните начини за проверка на тази хипотеза е посредством **критерия χ^2 („хи квадрат")** на Пирсън. Проверката на хипотезата H_0 се базира на таблицата на честотите (ако X е дискретна)

y_k	y_1	y_2	$\ldots$	$\ldots$	y_m
n_k	n_1	n_2	$\ldots$	$\ldots$	n_m
w_k	w_1	w_2	$\ldots$	$\ldots$	w_m

или на интервалната таблица на честотите (ако X е непрекъсната)

I_k	$[c_0, c_1)$	$[c_1, c_2)$	$\ldots$	$\ldots$	$[c_{m-1}, c_m]$
n_k	n_1	n_2	$\ldots$	$\ldots$	n_m
w_k	w_1	w_2	$\ldots$	$\ldots$	w_m

Процедурата за прилагана на критерия χ^2 за проверка на хипотезата H_0 е следната:

1. Намират се оценките за неизвестните параметри на предполагаемия закон за разпределение $F_0(x)$;

2. При условие, че X има функция на разпределение $F_0(x)$ (плътност на разпределение $f_0(x)$) се определят теоретичните вероятности p_k:

   2.1. Ако X е дискретна случайна величина, се пресмятат теоретичните вероятности

$$p_k = P(X = y_k), \quad k = 1, 2, \ldots, m.$$

2.2. Ако X е непрекъсната случайна величина, се пресмятат теоретичните вероятности

$$p_k = P(X \in I_k), \quad k = 1, 2, \ldots, m.$$

Забележка 15.1. *Критерият χ^2 е основан на факта, че случайните величини*

$$\frac{n_k - np_k}{\sqrt{np_k}}, \quad k = 1, 2, \ldots, m$$

имат разпределение, което е близко до стандартното нормално разпределение $N(0,1)$. За да бъде това твърдение достатъчно точно, е необходимо за всяко $k = 1, 2, \ldots, m$ да е изпълнено условието $np_k \geq 5$. Ако $np_k < 5$ за някое k, то трябва да обединим $k-$тата колона в съответната таблица с някоя от съседните колони. Броят на колоните в тази таблица, които остават след такова обединение, ще означаваме отново с m.

3. Пресмята се наблюдаваната стойност на критерия

$$\chi_0^2 = \sum_{k=1}^{m} \frac{(n_k - np_k)^2}{np_k} \,. \tag{15.3}$$

Ще отбележим, че вместо (15.3) може да се използва формулата

$$\chi_0^2 = \sum_{k=1}^{m} \frac{n_k^2}{np_k} - n \,. \tag{15.4}$$

4. При избрано ниво на значимост α от Таблица 5 се определя квантилът

$$\chi^2_{1-\alpha}(f),$$

където $f = m - r - 1$ е степента на свобода на разпределението χ^2;

m е окончателният брой на честотите в съответната таблица;

r е броят на параметрите на теоретичното разпределение, които са определени с помощта на извадката.

5. Взема се решение:
Ако $\chi_0^2 < \chi^2_{1-\alpha}(f)$, то основната хипотеза H_0 се приема с доверителна вероятност $p = 1 - \alpha$;
Ако $\chi_0^2 \geq \chi^2_{1-\alpha}(f)$, то основната хипотеза H_0 се отхвърля.

Забележка 15.2. *Сумата на емпиричните вероятности $w_k = n_k/n$ винаги е равна на единица. Аналогично, сумата на теоретичните вероятности също трябва да е равна на единица. Поради това, при пресмятането на крайните вероятности p_1 и p_m от таблиците трябва да се спази следното правило: Ако според хипотезата H_0 стойностите на случайната величина X са неограничени отгоре, то $p_m = P(X \geq y_m)$(ако X е дискретна) или $p_m = P(X \geq c_{m-1})$ (ако X е непрекъсната). Ако стойностите на X са неограничени отдолу, то $p_1 = P(X \leq y_1)$ (ако X е дискретна) или $p_1 = P(X < c_1)$ (ако X е непрекъсната).*

Пример 15.1. *В телефонна централа в течение на 60 минути е наблюдавана случайната величина $X-$ брой на свързванията в минута и е получена следната таблица на честотите:*

y_k	0	1	2	3	4	5	6	7
n_k	8	17	16	10	6	2	0	1

Тук $n = 60$, а оценките за математическото очакване и дисперсията са $EX \approx \overline{x} = 2$, $DX \approx \overline{s}^2 = 2.1356$. Понеже X е дискретна безразмерна случайна величина, приема неотрицателни цели стойности и $EX \approx DX \approx 2$, то издигаме основната хипотеза

$H_0 : X$ е разпределена по закона на Поасон с параметър $\lambda = 2$
с алтернативна хипотеза

$H_1 : X$ не е разпределена по закона на Поасон с параметър $\lambda = 2$.

Теоретичните вероятности при $k = 0, 1, 2, 3, 4, 5, 6$ са равни на

$$p_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{2^k}{k!} e^{-2},$$

а теоретичната вероятност p_7 е равна на

$$p_7 = P(X \geq 7) = 1 - (p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6).$$

Ако случайната величина X е разпределена по закона на Поасон, то нейните стойности са неограничени отгоре. Затова, съгласно забележка 15.2 теоретична вероятността p_7 не е равна на $P(X = 7)$, а на $P(X \geq 7)$.

От Таблица 1 (с $\lambda = 2$) определяме p_k и пресмятаме величините $np_k = 60 \cdot p_k$. Получените стойности са дадени в долната таблица.

y_k	0	1	2	3	4	5	6	≥ 7
p_k	0.1353	0.2707	0.2707	0.1804	0.0402	0.0361	0.0546	0.0120
np_k	8.106	16.242	16.242	10.824	2.412	2.166	0.720	3.276

Забелязваме, че в последните четири колони $np_k < 5$. *Обединяваме резултатите от последните три колони и получаваме следната обединена таблица.*

y_k	0	1	2	3	4	≥ 5
n_k	8	17	16	10	6	3
p_k	0.1353	0.2707	0.2707	0.1804	0.0402	0.1027
np_k	8.106	16.242	16.242	10.824	2.412	6.162

Пресмятаме наблюдаваната стойност на критерия

$$\chi_0^2 = \frac{(8-8.106)^2}{8.106} + \frac{(17-16.242)^2}{16.242} + \frac{(16-16.242)^2}{16.242} + \frac{(10-10.824)^2}{10.824} + \frac{(6-2.412)^2}{2.412} + \frac{(3-6.162)^2}{6.162} \approx 7.075\,.$$

Понеже $m = 6$ *и* $r = 1$ (*от извадката е оценен един параметър* $\lambda = 2$), *то степента на свобода е равна на* $f = 6 - 1 - 1 = 4$. *Нека е избрано ниво на значимост* $\alpha = 0.05$. *Тогава от Таблица* 5 *определяме квантила* $\chi^2_{0.95}(4) = 9.49$. *Понеже* $\chi_0^2 < \chi^2_{0.95}(4)$, *то хипотезата* H_0 *се приема с доверителна вероятност* $p = 0.95$.

Пример 15.2. *Направени са* $n = 100$ *наблюдения на непрекъснатата случайна величина* X *и е получена следната интервална таблица на честотите.*

k	1	2	3	4	5	6	7	8
I_k	$[0,2)$	$[2,4)$	$[4,6)$	$[6,8)$	$[8,10)$	$[10,12)$	$[12,14)$	$[14,16]$
n_k	45	16	15	9	7	3	3	2

Оценките за математическото очакване, дисперсията и стандартното отклонение са $EX \approx \overline{x} = 3.96$, $DX \approx \overline{s}^2 = 13.224$, $\sigma_X \approx \overline{s} = 3.64$. *Понеже с.в.* X *е неотрицателна непрекъсната и* $EX \approx \sigma_X \approx 4$, *то правим хипотезата*

$H_0 : X$ *има показателно разпределение с параметър* $\mu = \dfrac{1}{4} = 0.25$ *с алтернативна хипотеза*

$H_1 : X$ няма показателно разпределение с параметър $\mu = 0.25$.

Да проверим тази хипотеза с ниво на значимост $\alpha = 0.10$.

В условията на хипотезата H_0 случайната величина X има функция на разпределение

$$F_0(x) = 1 - e^{-\mu x}, \quad \text{при } x \geq 0\,.$$

Понеже $\mu c_k = 0.25 \cdot 2k = 0.5k$, то теоретичните вероятности p_k при $k = 1, 2, 3, 4, 5, 6, 7$ са равни на

$$\begin{aligned} p_k &= P(X \in I_k) = F_0(c_k) - F_0(c_{k-1}) \\ &= (1 - e^{-\mu c_k}) - (1 - e^{-\mu c_{k-1}}) = e^{-0.5(k-1)} - e^{-0.5k}. \end{aligned} \tag{15.5}$$

Понеже X е неограничена отгоре, то съгласно забележка 15.2 теоретичната вероятност p_8 е равна на

$p_8 = P(X \geq 14) = 1 - P(X < 14) = 1 - (1 - e^{-0.25 \cdot 14}) = e^{-3.5} = 0.0302.$

След пресмятането на p_k по формула (15.5) и определянето на $np_k = 100p_k$ получаваме следната таблица

k	1	2	3	4	5	6	7	8
p_k	0.3935	0.2386	0.1447	0.0878	0.0533	0.0323	0.0196	0.0302
np_k	39.35	23.86	14.47	8.78	5.33	3.23	1.96	3.02

Забелязваме, че в последните три колони $np_k < 5$. Обединяваме резултатите от тези три колони и получаваме следната обединена таблица

I_k	$[0, 2)$	$[2, 4)$	$[4.6)$	$[6, 8)$	$[8, 10)$	$[10, +\infty)$
n_k	45	16	15	9	7	8
p_k	0.3935	0.2386	0.1447	0.0878	0.0533	0.0821
np_k	39.35	23.86	14.47	8.78	5.33	8.21

Пресмятаме наблюдаваната стойност на критерия

$$\begin{aligned} \chi_0^2 = &\frac{(45 - 39.35)^2}{39.35} + \frac{(16 - 23.86)^2}{23.86} + \frac{(15 - 14.47)^2}{14.47} \\ &+ \frac{(9 - 8.78)^2}{8.78} + \frac{(7 - 5.33)^2}{5.33} + \frac{(8 - 8.21)^2}{8.21} \approx 3.95\,. \end{aligned}$$

Понеже $m = 6$ и $r = 1$ (от извадката е оценен един параметър $\mu = 0.25$), то степента на свобода е равна на $f = 6 - 1 - 1 = 4$. От Таблица 5 определяме квантила $\chi^2_{0.90}(4) = 7.78$. Понеже $\chi_0^2 < \chi^2_{0.90}(4)$, то хипотезата H_0 се приема с доверителна вероятност $p = 0.90$.

15.2 Проверка на хипотезата за нормалност на генералното разпределение

Нормално разпределени случайни величини се срещат много в практиката и затова често се налага да се проверява хипотезата, че дадена статистическа променлива има нормално разпределение. Тъй като нормалното разпределение е непрекъснато, проверка на хипотезата за нормалност следва да се извършва единствено за непрекъснати статистически променливи. Проверката на хипотезата за нормалност на генералното разпределение протича през следните етапи:

- Анализ на статистическите данни и формиране на хипотезата;
- Проверка на хипотезата чрез критерий за проверка;
- Допълнителни проверки.

Ще разгледаме тези етапи по-подробно.

15.2.1 Анализ на статистическите данни и формиране на хипотезата

След като сме получили интервалната таблица на честотите на изследваната непрекъсната случайна величина X, можем да построим хистограмата на честотите и кумулативната крива (полигона на кумулативните относителни честоти). Ако хистограмата има изразена камбановидна форма, то можем да предположим, че X има нормално разпределение. След като оценим математическото очакване и стандартното отклонение на X чрез средното $\overline{x}$ и статистическото стандартното отклонение $\overline{s}$, можем да построим графиките на плътността и функцията на разпределение на случайна величина, която има нормално разпределение $N(\overline{x}, \overline{s})$. Тези графики се сравняват визуално със съответните хистограма и кумулативна крива. Наличието на „близост" между тези графики е факт в полза на предположението за нормалност на разпределението, а липсата на такава близост – срещу това предположение. Построяване на горните графики и тяхното сравняване е програмно осигурено и лесно може да се направи с компютър. Допълнителна графична проверка за нормалност на разпределението може да се извърши с т.нар. DETRENDED

PROBABILITY PLOT, който също е реализиран програмно и се осъществява по следния начин: Наблюденията се представят като точки върху екрана, като абсцисата на всяка точка е поредната варианта на извадката, а ординатата е съответния квантил на нормалното разпределение. Ако извадката наистина има нормално разпределение, то така построените точки лежат близо до права линия. На екрана се показва най-добрата такава права. Всяко закономерно отклонение от нея говори за разпределение, различно от нормалното. След като всички посочени по-горе проверки засилват нашето първоначално предположение, можем да направим основната (H_0) и алтернативната (H_1) хипотези:

$H_0 : X$ има нормално разпределение;

$H_1 : X$ има разпределение, различно от нормалното.

15.2.2 Проверка на хипотезата чрез критерий за проверка

Съществуват няколко метода за проверка на хипотезата за нормалност на дадено генерално разпределение, които са основани на $P-$стойността на критерия за проверка, използван в тях. Едни от най-често използваните критерии (тестове) са: χ^2 (хи квадрат) на К. Пирсън, на Колмогоров – Смирнов и на Д'Агостино – Пирсън. Характерно за всички тези методи е това, че според издигнатите основна (H_0) и алтернативна (H_1) хипотеза се изчислява $P-$ стойността на критерия за проверка. Това пресмятане е програмно осигурено в съществуващите статистически програмни продукти.

За изхода от проверката на всеки един от тези методи са възможни два различни случая:

1. За $P-$стойността на критерия за проверка е в сила: $P \leq \alpha$. Тогава е налице статистически значим резултат. Основната хипотеза се отхвърля, защото са налице основателни статистически доводи за това;

2. За $P-$стойността на критерия за проверка е в сила: $P > \alpha$. Тогава е налице статистически незначим резултат. Основната хипотеза се приема, защото липсват основателни статистически доводи за нейното отхвърляне.

15.2.3 Допълнителни проверки

Когато резултатът от проверката на хипотезата за нормалност е статистически незначим, (т.е. основната хипотеза не се отхвърля) е добре да се направи и допълнителна преценка за близостта на генералното разпределение до нормалното разпределение. Целта на тази проверка е да се установи доколко статистическите характеристики и разпределение на изследвания генерална променлива удовлетворяват характерните за нормалното разпределение свойства. А те са:

математическо очакване (EX) = медиана (me) = мода (mo);

коефициент на асиметрия (S_k) = коефициент на ексцес $(K_u) = 0$.

Затова съответните статистически характеристики трябва приблизително да са равни:

средно $(\overline{x}) \approx$ стат. медиана $(me) \approx$ стат. мода (mo);

стат. коеф. на асиметрия $(S_k) \approx$ стат. коеф. на ексцес $(K_u) \approx 0$.

Забележка 15.3. *Хипотезата за нормалност на разпределението може да се провери с прилагане на χ^2 критерия на Пирсън, който е изложен в предния раздел и се прилага чрез построяване на критичната област $\chi_0^2 \geq \chi_{1-\alpha}(f)$.*

Пример 15.3. *Направени са $n = 500$ измервания на ръста X на 16 годишни девойки. Резултатите са дадени в следната интервалната таблица на честотите (колони $1, 2, 3$).*

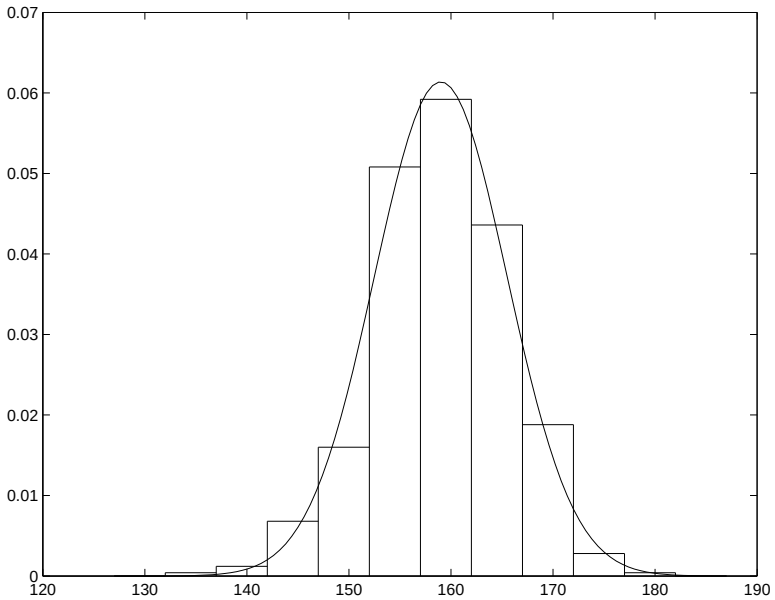
	1	2	3	4	5	6	7
k	$c_{k-1} - c_k$	y_k, cm	n_k	w_k	f_k	$f_k n_k$	$f_k^2 n_k$
1	$132 - 137$	134.5	1	0.002	-5	-5	25
2	$137 - 142$	139.5	3	0.006	-4	-12	48
3	$142 - 147$	144.5	17	0.034	-3	-51	15
4	$147 - 152$	149.5	40	0.080	-2	-80	160
5	$152 - 157$	154.5	127	0.254	-1	-127	127
6	$157 - 162$	159.5	148	0.296	0	0	0
7	$162 - 167$	164.5	109	0.218	1	109	109
8	$167 - 172$	169.5	47	0.094	2	94	188
9	$172 - 177$	174.5	7	0.014	3	21	63
10	$177 - 182$	179.5	1	0.002	4	4	16
			$n = 500$	1.000		$Q_1 = -47$	$Q_2 = 889$

Да се издигне хипотеза за закона за разпределение на с.в. X и да се провери тази хипотеза с ниво на значимост $\alpha = 0.05$.

Понеже хистограмата на относителните честоти има изразена камбановидна форма (фиг. 15.1*), то правим предположение, че X има нормално разпределение. За да оценим параметрите на разпределението, прилагаме метода от точка* 10.2.3*, като изберем фиктивното средно $a = 159.5$ cm и попълним колоните $5 - 9$, където $f_k = (y_k - a)/d$ и $d = c_k - c_{k-1} = 5$ cm е дължината на интервалите. По формулите* $(10.30 - 10.32)$ *намираме, че $EX \approx \overline{x} = 159.03$ cm, $\sigma_X \approx \overline{s} = 6.572$ cm. На фиг.* 15.1 *се вижда,че графиката на плътността на нормално разпределение с параметри $m = 159$ и $\sigma = 6.5$ cm (непрекъсната линия) добре пасва на хистограмата. Затова правим основната хипотеза*

$H_0 : X$ има нормално разпределение $N(159, 6.5)$
с алтернативна хипотеза

$H_1 : X$ няма нормално разпределение $N(159, 6.5)$.



Фиг. 15.1: Хистограма и плътност от пример 15.3

За да приложим критерия χ^2, първо трябва да пресметнем теоретичните вероятности p_k. В условията на хипотезата H_0 случайната

величина X *има функция на разпределение*

$$F_0(x) = 0.5 + \Phi\Big(\frac{x-159}{6.5}\Big),$$

където $\Phi(x)$ *е функцията на Лаплас (Таблица* 3*). Тогава теоретичните вероятности при* $k = 2,3,4,5,6,7,8,9$ *са равни на*

$$p_k = \Phi\Big(\frac{c_k - 159}{6.5}\Big) - \Phi\Big(\frac{c_{k-1} - 159}{6.5}\Big). \tag{15.6}$$

Понеже нормално разпределените случайни величини са неограничени както отгоре така и отдолу, то теоретичните вероятности p_1 *и* p_{10} *са равни на*

$$p_1 = P(X < c_1) = 0.5 + \Phi\Big(\frac{c_1 - 159}{6.5}\Big), \tag{15.7}$$

$$p_{10} = P(X \geq c_9) = 1 - P(X < c_9) = 1 - \Big(0.5 + \Phi\Big(\frac{c_9 - 159}{6.5}\Big)\Big)$$
$$= 0.5 - \Phi\Big(\frac{c_9 - 159}{6.5}\Big). \tag{15.8}$$

За да пресметнем по-лесно p_k *по формулите* (15.6) – (15.8)*, първо пресмятаме числата* $z_k = (c_k - 159)/6.5$ *и* $\Phi(z_k)$*. Резултатите от пресмятанията са дадени в колони* $1-5$ *на следната таблица.*

	1	2	3	4	5	6	7
k	c_k	z_k	$\Phi(z_k)$	p_k	np_k	*обед.*np_k	*обед.*n_k
0	$-\infty$	$-\infty$	-0.5000				
1	137	-3.3846	-0.4996	0.0004	<u>0.20</u>		
2	142	-2.6154	-0.4955	0.0041	<u>2.05</u>		
3	147	-1.8462	-0.4676	0.0279	13.95	16.20	21
4	152	-1.0769	-0.3592	0.1084	54.20	54.20	40
5	157	-0.3077	-0.1208	0.2384	119.20	119.20	127
6	162	0.4615	0.1778	0.2986	149.30	149.30	148
7	167	1.2308	0.3908	0.2130	106.50	106.50	109
8	177	2.0000	0.4772	0.0864	43.20	43.20	47
9	182	2.7692	0.4972	0.0200	10.00	11.40	8
10	$+\infty$	$+\infty$	0.5000	0.0028	<u>1.40</u>		

Забелязваме, че в редовете с $k = 1, 2$ *и* 10 *на колона* 5 *имаме* $np_k < 5$ *и затова обединяваме тези редове съответно със съседните редове* 3 *и* 9 *и получаваме колона* 6 *на обединените теоретични честоти* np_k *и колона* 7 *на обединените честоти* n_k*, където има* $m = 7$ *реда (честоти). Пресмятаме наблюдаваната стойност на критерия*

$$\chi_0^2 = \frac{(21-16.2)^2}{16.2} + \frac{(40-54.2)^2}{54.2} + \frac{(127-119.2)^2}{119.2} + \frac{(148-149.3)^2}{149.3} + \frac{(109-106.5)^2}{106.5} + \frac{(47-43.2)^2}{43.2} + \frac{(8-11.4)^2}{11.4} \approx 7.07\,.$$

Понеже $m = 7$ *и* $r = 2$ *(от извадката са оценени два параметъра* $EX = 159$ *и* $\sigma_X = 6.5$*), то степента на свобода е равна на* $f = 7-2-1 = 4$*. От Таблица* 5 *определяме квантила* $\chi^2_{0.95}(4) = 9.49$*. Понеже* $\chi_0^2 < \chi^2_{0.95}(4)$*, то хипотезата* H_0 *се приема с доверителна вероятност* $p = 0.95$*.*

15.3 Проверка на хипотезата за независимост на две случайни величини

Да предположим, че са направени n съвместни наблюдения (извадка) на статистическите променливи X и Y, в резултат на което те са приели съответно различните стойности $x_1, x_2, \ldots, x_m$ и различните стойности $y_1, y_2, \ldots, y_k$. Да означим с n_{ij} броя на наблюденията, при които

$$X = x_i \text{ и } Y = y_j,\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, k.$$

Ако X и Y са непрекъснати случайни величини, то областта на стойностите на всяка от тях се разбива на краен брой интервали. В този случай x_i и y_j са средите на съответните интервали, а n_{ij} е броят на наблюденията, при които случайната величина X попада в $i-$тия интервал, а случайната величина Y – в $j-$тия интервал. Резултатите от извадката може да се представят в **таблица на спрегнатост на признаците** с размери $m \times k$ (таблица 15.1), в която

$$n_{i*} = \sum_{j=1}^{k} n_{ij},\ i = 1, 2, \ldots, m; \quad n_{*j} = \sum_{i=1}^{m} n_{ij},\ j = 1, 2, \ldots, k.$$

Числата x_i, n_{i*} задават таблицата на честотите на случайната величина X, а числата y_j, n_{*j} – таблицата на честотите на случайната величина Y.

Таблица 15.1: Таблица на спрегнатост на признаците

$X\backslash Y$	y_1	y_2	$\ldots$	y_k	$n_{i*}=\sum_{j=1}^{k} n_{ij}$
x_1	n_{11}	n_{12}	$\ldots$	n_{1k}	n_{1*}
x_2	n_{21}	n_{22}	$\ldots$	n_{2k}	n_{2*}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_m	n_{m1}	n_{m2}	$\ldots$	n_{mk}	n_{m*}
$n_{*j}=\sum_{i=1}^{m} n_{ij}$	n_{*1}	n_{*2}	$\ldots$	n_{*k}	$n_{**}=n$

Проверява се хипотезата H_0, твърдяща, че случайните величини X и Y са независими. Ако тази хипотеза е вярна, то по определение

$$P(X=x_i, Y=y_j) = P(X=x_i)P(Y=y_j) = p_{i*}p_{*j}.$$

Нека $\overline{p}_{i*} = \dfrac{n_{i*}}{n}$ и $\overline{p}_{*j} = \dfrac{n_{*j}}{n}$ са оценки за вероятностите p_{i*} и p_{*j}. Ако хипотезата H_0 е вярна, то **очакваната честота** $\overline{n}_{ij}$, с които случайната величина X приема стойност x_i (попада в $i-$тия интервал), а случайната величина Y приема стойност y_j (попада в $j-$тия интервал), е равна на

$$\overline{n}_{ij} = n\overline{p}_{i*}\overline{p}_{*j} = \frac{n_{i*}n_{*j}}{n}\,.$$

За проверка на хипотезата H_0 с критерия χ^2 се използва статистиката

$$\chi_0^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij}-\overline{n}_{ij})^2}{\overline{n}_{ij}}\,. \tag{15.9}$$

Забележка 15.4. *Вместо формула* (15.9) *е удобно да се използва формулата*

$$\chi_0^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right). \tag{15.10}$$

Ако основната хипотезата H_0 е вярна и са изпълнени условията

$$\overline{n}_{ij} \geq 4, \quad i=1,2,\ldots,m,\ j=1,2,\ldots,k, \tag{15.11}$$

то статистиката (15.9) има разпределение χ^2 с $f=(m-1)(k-1)$ степени на свобода.

След намирането на наблюдаваната стойност χ_0^2 на критерия, се избира ниво на значимост α и от Таблица 5 се определя квантилът

$$\chi^2_{1-\alpha}(f),$$

където $f = (m-1)(k-1)$ е степента на свобода на разпределението χ^2.

Ако $\chi_0^2 < \chi^2_{1-\alpha}(f)$, то основната хипотеза H_0 се приема с доверителна вероятност $p = 1 - \alpha$;

Ако $\chi_0^2 \geq \chi^2_{1-\alpha}(f)$, то основната хипотеза H_0 се отхвърля.

Забележка 15.5. *Ако очакваната честота $\overline{n}_{ij}$ за някоя клетка от таблицата не изпълнява условието $\overline{n}_{ij} \geq 4$, то съответните ред и колона от таблицата трябва да се обединят със съседни редове и колони.*

Забележка 15.6. *Ако $f = (m-1)(k-1) \geq 8$ и $n \geq 40$, то минималната допустима стойност на очакваните честоти $\overline{n}_{ij}$ може да бъде равна на единица.*

15.4 Проверка на хипотезата за еднородност на няколко извадки

В много случаи се налага проверката на хипотезата за еднородност на няколко извадки, или с други думи, хипотезата за това,че тези извадки са получени от една генерална съвкупност. Ако се проверява еднородността на m различни извадки с обеми съответно $n_1, n_2, \ldots, n_m$ и тези извадки може да се запишат в таблица на спрегнатост на признаците с размери $m \times k$ (таблица 15.1), то за проверката се използва същият критерий, както при проверката за независимост на две случайни величини (два признака).

Пример 15.4. *Определен детайл постъпва на склад от три цеха A,B и C. Резултатите от проверката за качество на детайлите е дадена в следната таблица:*

	Цехове			
Резултати	*A*	*B*	*C*	*Сума*
Годни	29	38	53	120
Негодни	1	2	7	10
Сума	30	40	60	130

С ниво на значимост $\alpha = 0.10$ *да се провери хипотезата за независимост на два признака: качеството на детайла и мястото на производство. По формула* (15.10)

$$\chi_0^2 = 130 \cdot \Big(\frac{29^2}{30 \cdot 120} + \frac{38^2}{40 \cdot 120} + \frac{53^2}{60 \cdot 120} + \frac{1^2}{30 \cdot 10} + \frac{2^2}{40 \cdot 10} + \frac{7^2}{60 \cdot 10} - 1 \Big) \approx 2.546,$$

а степента на свобода е равна на $f = (2-1)(3-1) = 2$. *От Таблица* 5 *намираме, че* $\chi^2_{0.90}(2) = 4.61 > 2.546 = \chi_0^2$ (*хипотезата се приема*), *т.е. качеството на изделието не зависи от цеха на производството. Това твърдение означава също, че трите извадки с обеми* $30, 40$ *и* 60, *получени от цеховете* A, B *и* C, *са еднородни.*

15.5 Проверка на хипотезата за равенство на две вероятности

Да предположим, че независимо са проведени две серии от опити, състоящи се съответно от n_{1*} и n_{2*} опита всяка. В първата серия събитието A се е появило n_{11} пъти, а във втората серия – n_{21} пъти. Трябва да се провери хипотезата H_0 за това, че вероятностите p_1 и p_2 за сбъдване на събитието A в двете серии от опити е една и съща, т.е. $H_0 : p_1 = p_2$. Резултатите от двете серии може да се представят във вида на таблица на спрегнатост на признаците с размери 2×2:

	Събития		
Серия	A	$\overline{A}$	Сума
1	n_{11}	n_{12}	n_{1*}
2	n_{21}	n_{22}	n_{2*}
Сума	n_{*1}	n_{*2}	n

Хипотезата $H_0 : p_1 = p_2$ е еквивалентна на хипотезата за това, че

двете извадки са получени от една генерална съвкупност, т.е. те са еднородни. Тази хипотеза се проверява с критерия χ^2 (вж. раздел 15.1).

В този случай за изчисляването на наблюдаваната стойност на статистиката (15.9) е удобно да се използва формулата

$$\chi_0^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1*}n_{2*}n_{*1}n_{*2}}\,. \tag{15.12}$$

Ако $\chi_0^2 < \chi_{1-\alpha}^2(1)$, то хипотезата H_0 се приема с ниво на значимост α.
Ако $\chi_0^2 \geq \chi_{1-\alpha}^2(1)$, то хипотезата H_0 се отхвърля.

Забележка 15.7. *Критерият χ^2 може да се използва, ако $n > 20$ и за всички очаквани честоти $\overline{n}_{ij} = \dfrac{n_{i*}n_{*j}}{n} > 3$, $i, j = 1, 2$. Ако $n \leq 20$, то при пресмятането на χ_0^2 по формула* (15.12) *трябва вместо n да се постави $n - 1$; при това трябва $n_{1*} > 5$ и $n_{2*} > \dfrac{n_{1*}}{3}$.*

Пример 15.5. *Да се провери хипотезата H_0 за равенство на вероятностите за поява на събитието A при две серии от опити по резултатите от долната таблица. Да се приеме $\alpha = 0.05$.*

	Събития		
Серия	A	$\overline{A}$	*Сума*
1	3	10	$n_{1*} = 13$
2	6	17	$n_{2*} = 23$
Сума	$n_{*1} = 9$	$n_{*2} = 27$	$n = 36$

Минималната очаквана честота е равна на

$$\overline{n}_{11} = \frac{n_{1*}n_{*1}}{n} = \frac{13 \cdot 9}{36} = 3.25 > 3, \quad n = 36 > 20.$$

Следователно може да използваме критерия χ^2. По формула (15.12) *пресмятаме $\chi_0^2 = \dfrac{36 \cdot (3 \cdot 17 - 6 \cdot 10)^2}{13 \cdot 23 \cdot 9 \cdot 27} \approx 0.04$. От Таблица 5 определяме $\chi_{0.95}^2(1) = 3.84$. Понеже $\chi_0^2 < \chi_{0.95}^2(1)$, то хипотезата H_0 се приема.*