Критерий Хи-квадарт Пирсона для проверки гипотезы о виде рапспределения

Статистика критерия

$$\chi^2_в = \sum_{l=1}^{m} \frac{\left(\nu_l - np_l\right)^2}{np_l} \Rightarrow^{d} \chi^2\left(m-1-r\right)$$

ν_l - эмпирические частоты

p_l - теоретические вероятности

np_l - теоретические частоты

m - число интервалов разбиения (после объединения)

r - число оцениваемых параметров

Выводы:

Большие значения статистики $\chi^2_в$ говорят о расхождении практики с предположением H_0. Критическое множество имеет вид: $S = (\chi^2_{1-\alpha}(m-1-r); +\infty)$.

Критерий

$\chi^2_в > \chi^2_{1-\alpha}(m-1-r) \rightarrow$ гипотеза H_0 отклоняется

$\chi^2_в < \chi^2_{1-\alpha}(m-1-r) \rightarrow$ гипотеза H_0 принимается

Пример 4

На автомобильном заводе проводится контроль качества сборки двигателей. Одной из ключевых характеристик является допустимое отклонение размера поршня от номинального значения (в микронах). Технологический отдел предполагает, что ошибки производства подчиняются нормальному распределению с математическим ожиданием, близким к нулю (идеальное соответствие номиналу). Проверка этого предположения важна для:

- Настройки системы статистического контроля процесса (SPC)
- Калибровки оборудования
- Оценки доли бракованной продукции

Основная гипотеза: Отклонения размеров поршней от номинала подчиняются нормальному распределению.

$H_0 : X \sim N(a, \sigma)$

Альтернативная гипотеза: Отклонения размеров поршней от номинала не подчиняются нормальному распределению.

$H_1 : X \nsim N(\mu, \sigma^2)$

Проведены замеры 150 случайно выбранных поршней. Данные сгруппированы в интервалы:

Интервал отклонения (микроны)	Наблюдаемая частота (ν_i)
$(-20;-15]$	6
$(-15;-10]$	12
$(-10;-5]$	22
$(-5;0]$	35
$(0;5]$	33
$(5;10]$	25
$(10;15]$	12
$(15;20]$	5

Поскольку параметры нормального распределения a и σ неизвестны, оцениваем их по выборке.

Выборочное среднее (оценка $\hat{a}$):

Для расчета используем середины интервалов.

$$\hat{a} = \overline{X} = \frac{6\cdot(-17.5)+12\cdot(-12.5)+...+5\cdot(17.5)}{150} \approx -0.27 \text{ микрон}$$

Выборочное стандартное отклонение (оценка $\hat{\sigma}$):

$$\hat{\sigma} = \frac{1}{150}\left(6\cdot(-17.5+0.27)^2+126\cdot(-12.5+0.27)^2+...+5\cdot(17.5+0.27)^2\right) \approx 8.24 \text{ микрона}$$

Таким образом, проверяем гипотезу о распределении: $X \sim N(-0.27, 8.24)$.

Для каждого i-го интервала вычисляем вероятность попадания в него при условии, что H_0 верна.

Формулы для расчета вероятностей:

Для интервала $(c_i, d_i]$:

$$p_i = P(c_i < X \le d_i) = \Phi\left(\frac{c_i-\hat{a}}{\hat{\sigma}}\right) - \Phi\left(\frac{d_i-\hat{a}}{\hat{\sigma}}\right)$$

где $\Phi(x)$ — функция стандартного нормального распределения.

Ожидаемая частота:

$n\cdot p_i = 150\cdot p_i$

Результаты расчетов:

Интервал отклонения (микроны)	Теоретичсекая частота (np_i)
$(-\infty;-15]$	5.46
$(-15;-10]$	12.32
$(-10;-5]$	24.23
$(-5;0]$	33.02
$(0;5]$	33.14
$(5;10]$	24.29

(10;15]	12.38
(15;∞)	5.19

Статистика критерия хи-квадрат вычисляется по формуле:

$$\chi^2_в = \sum_{i=1}^{m} \frac{(v_i - np_i)^2}{np_i}$$

где $m = 8$ — количество интервалов.

Проведем вычисления:

$$\chi^2_в = \frac{(6-5.46)^2}{5.46} + \frac{(12-12.32)^2}{12.32} + \frac{(22-24.23)^2}{24.23} + \frac{(35-33.02)^2}{33.02} +$$
$$= \frac{(33-33.14)^2}{33.14} + \frac{(25-24.29)^2}{24.29} + \frac{(12-12.38)^2}{12.38} + \frac{(5-5.19)^2}{5.19} = 0.428$$

Число степеней свободы: $l = m - 1 - r = 8 - 1 - 2 = 5$, где $r = 2$ — количество оцененных параметров (a и σ). Уровень значимости: $\alpha = 0.05$. Критическое значение: По таблице распределения χ^2 для $l = 5$ и $\alpha = 0.05$:

$$\chi^2_{крит} = 11.070 \qquad \chi^2_в = 0.428 < \chi^2_{крит} = 11.070$$

На уровне значимости $\alpha = 0.05$ нет оснований отвергать нулевую гипотезу. Данные не противоречат предположению о нормальном распределении отклонений размеров поршней.

Отклонения размеров поршней от номинала статистически значимо не отличаются от нормального распределения с параметрами $a \approx -0.27$ мкм, $\sigma \approx 8.24$ мкм ($\chi^2 = 0.428$; $l = 5$).

Практические последствия для предприятия:

1. Валидация контроля качества: Можно использовать методы статистического контроля процессов (SPC), основанные на нормальном распределении.
2. Прогнозирование брака: Зная параметры распределения, можно точно оценить долю продукции за пределами допусков (например, рассчитать вероятность отклонения более 20 микрон).
3. Стабильность процесса: Распределение с математическим ожиданием близким к нулю свидетельствует о правильной настройке оборудования.

Это позволяет руководству завода применять нормальную модель для оптимизации производственного процесса и снижения затрат на контроль качества.

Пример 5

В итоге регистрации прихода посетителей выставки получена таблица:

Интервал времени	12–13	13–14	14–15	15–16	16–17
Число посетителей	250	157	99	54	30

С помощью критерия χ^2, проверить гипотезу о том, что время прихода посетителей на выставку распределено по показательному закону. Принять $\alpha = 0{,}1$

Оценим параметр показательного распределения, ранее было установлено что его оценка равна $\hat{\lambda} = \frac{1}{\overline{X}}$. Так как у нас дана уже сгруппированная выборка, вычислим выборочное среднее следующим образом.

$$\overline{X} = \frac{\sum_{i=1}^{k} X_i n_i}{n}$$

Здесь X_i середина каждого интервала, серединой первого интервала будет 0,5 часа, второго 1,5 часа и так далее. Таким образом

$$\overline{X} = \frac{250 \cdot 0{,}5 + 157 \cdot 1{,}5 + \cdots}{590} \approx 1{,}579$$

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{1{,}579} \approx 0{,}633$$

Теперь можно найти теоретические частоты

Для первого интервала от 0 до 1

$$np_1 = 590 \cdot \left(e^{-0 \cdot \hat{\lambda}} - e^{-1 \cdot \hat{\lambda}} \right) = 590 \cdot (1 - 0{,}53) = 276{,}71 \approx 277$$

Для второго интервала от 1 до 2

$$np_1 = 590 \cdot \left(e^{-1 \cdot \hat{\lambda}} - e^{-2 \cdot \hat{\lambda}} \right) = 590 \cdot (0{,}53 - 0{,}281) = 146{,}91 \approx 147$$

Интервал времени	12–13	13–14	14–15	15–16	16–17
Теор. частоты	277	147	78	41	47

Применим критерий Пирсона. Находим статистику Пирсона

$$\chi^2_{\text{в}} = \sum_{i=1}^{5} \frac{\left(\nu_l - np_i\right)^2}{np_i} \approx 12,615$$

Находим квантиль распределения хи квадрат, так как был один оцениваемый параметр то, число степеней свободы $5-1-1=3$, квантиль равен $\chi^2_{0,9}(3) = 6,251$

Таким образом $\chi^2_в > \chi^2_{0,95}(3)$, следовательно гипотеза о показательном распределении отвергается.

Проверка гипотезы о независимости признаков

(таблица сопряженности признаков)

Предположим, имеется большая совокупность объектов, каждый из которых обладает двумя признаками A и B; признак A имеет m уровней: $A_1,\ldots,A_m$, а признак B – k уровней: $B_1,\ldots,B_k$. Пусть уровень A_i встречается с вероятностью $P(A_i)$, а уровень B_j – с вероятностью $P(B_j)$. Признаки A и B независимы, если

$$P(A_i B_j)=P(A_i)P(B_j), i=1,\ldots,m,\ j=1,\ldots,k.$$

Пусть признаки определены на n объектах, случайно извлеченных из совокупности; ν_{ij} – число объектов, обладающих комбинацией $A_i B_j$, $\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}=n$. По совокупности наблюдений $\{\nu_{ij}\}$ (таблица $m\times k$) требуется проверить гипотезу

H_0: признаки A и B независимы .

Задача сводится к случаю с неизвестными параметрами; ими являются вероятности

$$P(A_i), i=1,\ldots,m;\ P(B_j), j=1,\ldots,k,$$

всего $(m-1)+(k-1)$; их оценки:

$$\hat{P}(A_i)=\frac{\sum_{j=1}^{k}\nu_{ij}}{n}=\frac{\nu_{i\cdot}}{n}, \hat{P}(B_j)=\frac{\sum_{i=1}^{m}\nu_{ij}}{n}=\frac{\nu_{\cdot j}}{n}.$$

Тогда статистика χ_B^2 принимает вид: $\hat{P}(A_i)\hat{P}(B_j)=p_{ij}$

$$\chi_B^2=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij}-np_{ij})^2}{np_{ij}}=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{np_{ij}}-2\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}+n\sum_{i=1}^{m}\sum_{j=1}^{k}p_{ij}=$$

$$=n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_{i\cdot}\nu_{\cdot j}}-n=n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_{i\cdot}\nu_{\cdot j}}-1\right)$$

Если гипотеза H_0 верна, то по теореме Фишера χ_B^2 асимптотически распределена по закону хи-квадрат с числом степеней свободы

$$l=mk-1-(m-1)-(k-1)=(m-1)(k-1),$$

и потому, получаем **критерий** уровня значимости α:

Если $\chi_B^2 \geq \chi_{1-\alpha}^2(l)$, то гипотезу H_0 о независимости признаков следует отклонить

Если $\chi_B^2 < \chi_{1-\alpha}^2(l)$, то нет оснований отклонить гипотезу H_0

Замечание 1. Условием применимости критерия является $np_{ij} \geq 4$.

Замечание 2. Ясно, что сформулированный критерий можно применять для проверки независимости двух случайных величин, разбив диапазоны их значений на m и k частей.

Пример 6

Утверждается, что результат действия лекарства не зависит от способа применения. Проверить это утверждение по следующим данным ($\alpha = 0,05$):

Результат\способ	A	B	C	
Благоприятный	11	17	16	44
Неблагоприятный	20	23	19	62
	31	40	35	106

В данном примере признаками являются A – Результат приёма лекарства (2 уровня), B – способ приёма лекарства (3 уровня). Найдём оценку вероятностей каждого из уровней

$$\hat{P}(A_i) = \frac{\sum_{j=1}^{k} \nu_{ij}}{n} = \frac{\nu_{i\cdot}}{n}, \hat{P}(B_j) = \frac{\sum_{i=1}^{m} \nu_{ij}}{n} = \frac{\nu_{\cdot j}}{n}.$$

$$\hat{P}(A_1) = \frac{44}{106}, \hat{P}(A_2) = \frac{62}{106} \qquad \hat{P}(B_1) = \frac{31}{106}, \hat{P}(B_2) = \frac{40}{106}, \hat{P}(B_3) = \frac{35}{106}$$

Найдём вероятности p_{ij}

$$p_{11} = \frac{44}{106}\cdot\frac{31}{106} = \frac{341}{2809}; \quad p_{12} = \frac{44}{106}\cdot\frac{40}{106} = \frac{440}{2809}; \quad p_{13} = \frac{44}{106}\cdot\frac{35}{106} = \frac{385}{2809}$$
$$p_{21} = \frac{62}{106}\cdot\frac{31}{106} = \frac{961}{5618}; \quad p_{22} = \frac{62}{106}\cdot\frac{40}{106} = \frac{620}{2809}; \quad p_{23} = \frac{62}{106}\cdot\frac{35}{106} = \frac{1085}{5618}$$

Теперь найдём статистику Пирсона $\chi_B^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(\nu_{ij} - np_{ij}\right)^2}{np_{ij}}$

$$\frac{\left(11 - 341\cdot\frac{106}{2809}\right)^2}{\left(341\cdot\frac{106}{2809}\right)} + \frac{\left(20 - 961\cdot\frac{106}{5618}\right)^2}{\left(961\cdot\frac{106}{5618}\right)} + \frac{\left(17 - 440\cdot\frac{106}{2809}\right)^2}{\left(440\cdot\frac{106}{2809}\right)} + \frac{\left(23 - 620\cdot\frac{106}{2809}\right)^2}{\left(620\cdot\frac{106}{2809}\right)} + \frac{\left(16 - 385\cdot\frac{106}{2809}\right)^2}{\left(385\cdot\frac{106}{2809}\right)} + \frac{\left(19 - 1085\cdot\frac{106}{5618}\right)^2}{\left(1085\cdot\frac{106}{5618}\right)} = 0.735$$

Преобразуем формулу для вычисления статистики Пирсона

$$\chi_B^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(\nu_{ij} - np_{ij}\right)^2}{np_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\nu_{ij}^2}{np_{ij}} - 2\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij} + n\sum_{i=1}^{m}\sum_{j=1}^{k} p_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\nu_{ij}^2}{n\frac{\nu_{i\cdot}\cdot\nu_{\cdot j}}{n\cdot n}} - 2n + n$$

$$= n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_{i\cdot}\nu_{\cdot j}} - n = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_{i\cdot}\nu_{\cdot j}} - 1\right)$$

Найдём статистику Пирсона используя преобразованную формулу

$$\chi_B^2 = 106\left(\frac{11^2}{44\cdot 31} + \frac{17^2}{44\cdot 40} + \frac{16^2}{44\cdot 35} + \frac{20^2}{62\cdot 31} + \frac{23^2}{62\cdot 40} + \frac{19^2}{62\cdot 35} - 1\right) = 0.735$$

Сравним полученное значение с квантилем распределения хи-квадрат для $\alpha = 0,05$, уровней свободы $(m-1)(k-1) = (3-1)(2-1) = 2$

$$\chi_B^2 0.735 < \chi_{0.95}^2(2) = 5.991$$

Пример 7

Комплектующие одного наименования поступают с трех предприятий. Можно ли считать, что качество не зависит от поставщика? $\alpha = 0,1$.

Результат\поставщик	A	B	C	Всего
Годные	29	38	53	120
Негодные	1	2	7	10
	30	40	60	130

В данном примере признаками являются A – качество деталей (годные или не годные, 2 уровня), B – поставщик (3 уровня).

Найдём статистику Пирсона

$$\chi_B^2 = 130\left(\frac{29^2}{120\cdot 30} + \frac{38^2}{120\cdot 40} + \frac{53^2}{120\cdot 60} + \frac{1^2}{10\cdot 30} + \frac{2^2}{10\cdot 40} + \frac{7^2}{10\cdot 60} - 1\right) = 2.546$$

Сравним полученное значение с квантилем распределения хи-квадрат для $\alpha = 0,1$, уровней свободы $(m-1)(k-1) = (3-1)(2-1) = 2$

$$\chi_B^2 = 2.546 < \chi_{0.9}^2(2) = 4.605$$

Пример 8

Компания "МегаФон-Эконом" разработала три тарифных плана: "Базовый", "Оптимальный" и "Премиум". Маркетологи хотят проверить гипотезу о том, что выбор тарифа зависит от уровня месячного дохода клиента.

1. Формулировка гипотез:

Нулевая гипотеза (H_0): Признаки «Уровень дохода» и «Выбор тарифа» независимы.

$$H_0 : P(\text{Тариф} = i, \text{Доход} = j) = P(\text{Тариф} = i) \cdot P(\text{Доход} = j) \quad \forall i, j$$

Альтернативная гипотеза

H_1 : Признаки «Уровень дохода» и «Выбор тариф» зависимы.

$$H_1 : \exists i, j : P(\text{Тариф} = i, \text{Доход} = j) \neq P(\text{Тариф} = i) \cdot P(\text{Доход} = j)$$

Результаты опроса 300 клиентов:

Уровень дох. ╲ Тариф	Базовый	Оптимальн	Премиум	Σ
Низкий (до 40 тыс)	60	30	10	100
Средний (40-80 тыс)	30	50	20	100
Высокий (свыше 80 тыс)	10	20	70	100
Σ	100	100	100	300

3. Расчет ожидаемых частот (p_{ij})

Если гипотеза H_0 верна, то ожидаемая частота для ячейки (i, j) рассчитывается по формуле:

$$p_{ij} = \frac{(\text{сумма по строке } i) \times (\text{сумма по столбцу } j)}{\text{Общее кол-во } n}$$

Рассчитаем ожидаемые частоты для всех ячеек. Например, для ячейки "Низкий доход / Базовый тариф":

$$p_{11} = \frac{100 \times 100}{300} \approx 33.33$$

Полная таблица ожидаемых частот:

Уровень дох. ╲ Тариф	Базовый	Оптимальн	Премиум	Σ
Низкий (до 40 тыс)	33.33	33.33	33.33	100
Средний (40-80 тыс)	33.33	33.33	33.33	100
Высокий (свыше 80 тыс)	33.33	33.33	33.33	100
Σ	100	100	100	300

Тестовая статистика хи-квадрат Пирсона вычисляется по формуле:

$$\chi^2_в = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij} - n \cdot p_{ij})^2}{n \cdot p_{ij}}$$

где $m = 3$ — число строк, $k = 3$ — число столбцов, ν_{ij} — наблюдаемая частота, p_{ij} — ожидаемая частота.

Произведем расчет для всех ячеек:

$$\chi^2_в = \frac{(60-33.33)^2}{33.33} + \frac{(30-33.33)^2}{33.33} + \frac{(10-33.33)^2}{33.33} + \frac{(30-33.33)^2}{33.33} + \frac{(50-33.33)^2}{33.33} +$$
$$+ \frac{(20-33.33)^2}{33.33} + \frac{(10-33.33)^2}{33.33} + \frac{(20-33.33)^2}{33.33} + \frac{(70-33.33)^2}{33.33} = 114$$

Число степеней свободы:

$$l = (m-1)(k-1) = (3-1)(3-1) = 4$$

Уровень значимости: $\alpha = 0.05$

Критическое значение: По таблице распределения χ^2 для $df = 4$ и $\alpha = 0.05$ находим:

$$\chi^2_{крит} = 9.49$$

$$\chi^2_в = 114.00 > \chi^2_{крит} = 9.49$$

Существует статистически значимая зависимость между уровнем дохода клиентов и выбором тарифа мобильной связи ($\chi^2 = 114.00$; $l = 4$; $p < 0.001$). Анализ таблицы сопряженности показывает, что клиенты с низким доходом предпочитают "Базовый" тариф, а клиенты с высоким доходом — "Премиум". Это позволяет компании направлять рекламные кампании на конкретные доходные группы для повышения эффективности маркетингового бюджета.