Тема 6. КРИТЕРИЙ ХИ-КВАДРАТ ПИРСОНА ПРОВЕРКИ ГИПОТЕЗ

6.1. Простая гипотеза о вероятностях

Полагаем, что результатом **одного наблюдения** может быть m различных вариантов (не обязательно числовых, любой физической природы)

$A_1, \ldots, A_m$.

$p_1, \ldots, p_m$ - **истинные**, но **неизвестные** вероятности, $\sum_{i=1}^{n} p_i = 1$.

n **независимых** наблюдений:

$\nu_1, \ldots, \nu_m$ - числа появлений исходов в n наблюдениях, $\sum_{i=1}^{m} \nu_i = n$.

Гипотеза H о значениях вероятностях:

$p_1^0, \ldots, p_m^0$ — **гипотетические** (или теоретические) значения вероятностей, $p_i^0 > 0$, $i = \overline{1, m}$, $\sum_{i=1}^{n} p_i^0 = 1$.

Требуется по результатам наблюдений $\nu_1, \ldots, \nu_m$ проверить **гипотезу** H:

$$H\colon\ p_i = p_i^0,\ i = \overline{1, m}.$$

Оценки для истинных вероятностей $\hat{p}_1, \ldots, \hat{p}_m$:

$$\hat{p}_1 = \frac{\nu_1}{n},\ \ldots,\ \hat{p}_m = \frac{\nu_m}{n}.$$ относительные частоты появления

Мерой расхождения между **гипотетическими** и **эмпирическими** вероятностями принимается СВ

$$X^2 = n\sum_{i=1}^{m} p_i^0\left(\frac{\hat{p}_i - p_i^0}{p_i^0}\right)^2 = \sum_{i=1}^{m}\frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{m}\frac{\nu_i^2}{np_i^0} - n, \qquad \textbf{(6.1)}$$

- усредненное с **весами** p_i^0 значение **квадрата относительного отклонения** значений $\hat{p}_i$ от p_i^0. Статистика X^2 называется статистикой **хи-квадрат** Пирсона.

Процедура проверки гипотезы: если величина X^2 приняла **«слишком большое»** значение, т.е. если X^2 слишком велико, т.е.

$$X^2 \geq h, \qquad \textbf{(6.2)}$$

то гипотеза *H* **отклоняется**. Область тех значений статистики X^2, при которых гипотеза отклоняется, **называется критической**.

Если же $X^2 < h$,

то будем говорить, что «**наблюдения не противоречат гипотезе».**

На вопрос, что означает **«слишком большое»** значение, отвечает теорема К. Пирсона.

Теорема 6.1 (К. Пирсон). Если гипотеза H **верна** и $p_i^0 > 0$ (т.е. нулей нет), $i = \overline{1, m}$, то при $n \to \infty$ распределение статистики X^2 **асимптотически** подчиняется **хи-квадрат** распределению χ^2_{m-1} с $m-1$ степенями свободы, т.е.

$$\mathbf{P}\{X^2 < x \,|\, H\} \xrightarrow[n\to\infty]{} F_{m-1}(x) \equiv \mathbf{P}\{\chi^2_{m-1} < x\}.$$

Порог h выберем из условия:

$$\text{P}\{\text{ошибиться} \,|\, H \text{ верна}\} = \alpha,$$

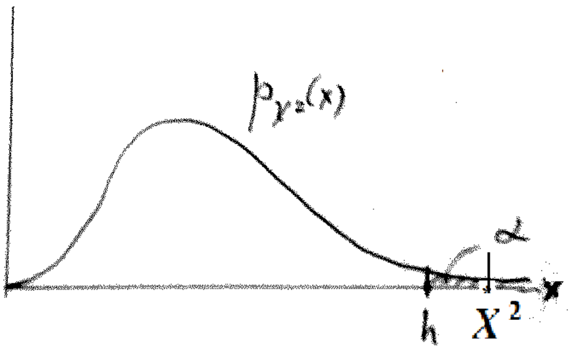
α называется **уровнем значимости**)

$$\text{P}\{\text{отклонить } H \,|\, H \text{ верна}\} \equiv \text{P}\{X^2 \geq h \,|\, H\} \approx \text{P}\{\chi^2_{m-1} \geq h\} = \alpha, \qquad \textbf{(6.3)}$$

откуда

$$h = Q(1-\alpha, m-1) \qquad \textbf{(6.3a)}$$

— квантиль уровня $(1-\alpha)$ **хи-квадрат** распределения χ^2_{m-1} с $(m-1)$ степенями свободы.



Процедура сравнения с порогом h может быть записана иначе:

$$\text{P}\{\chi^2_{m-1} \geq X^2\} \leq \alpha. \qquad \textbf{(6.4)}$$

Это выполняется тогда и только тогда, когда выполняется (**6.2**).

Рис. 10. Выбор критического значения

Уровень значимости 0.001 считается более высоким, чем 0.01

Замечание 6.1об условиях. Теорему Пирсона можно применять, если **все ожидаемые частоты** p_i^0 удовлетворяют неравенствам:

$$np_i^0 \geq 10, \quad i = \overline{1, m}. \qquad \textbf{(6.5a)}$$

Если $m \geq 10$, то достаточно выполнения следующих неравенств:

$$\nu_i \geq 4, \quad i = \overline{1, m}. \qquad \textbf{(6.5б)}$$

Если условия (**6.5**) не выполняются, необходимо некоторые исходы A_i объединять.

Свойство состоятельности критерия. Описанное решающее правило, обладает важным свойством**:**

если гипотеза H **неверна**, то с ростом числа наблюдений оно отклоняет гипотезу с вероятностью, стремящейся к 1: **мощность критерия** $W(p)$:

$$W(p) \equiv \text{P}\{\text{отклонить } H \,|\, \bar{H}: p, p \neq p^0\} \xrightarrow[n\to\infty]{} 1$$

Важную характеристику (мощность $W(p)$) можно определить **приближенно**, опираясь на следующую теорему.

Теорема 6.2. Если гипотеза H **неверна**, то при $n \to \infty$ распределение статистики X^2 сходится к распределению $\chi^2_{m-1}(a)$ **— нецентральному хи-квадрат** с числом степеней свободы $(m-1)$ и **параметром нецентральности** a, причём

$$a = n\sum_{i=1}^{m} \frac{(p_i - p_i^0)^2}{p_i^0}.$$

Для простого запоминания заметим, что эта формула получается из формулы (**6.1**) заменой наблюдаемых частот ν_i на истинные np_i.

СПРАВКА о нецентральном распределении хи-квадрат $\chi^2_k(a)$

Пусть СВ $\alpha_1, \alpha_2, ..., \alpha_k$ **независимы** и **нормально** распределены по закону $N(0,1)$. Составим случайную величину

$$(\alpha_1 + a_1)^2 + (\alpha_2 + a_2)^2 + ... + (\alpha_k + a_k)^2,$$

где $a_1, a_2, ..., a_k$ — произвольные **постоянные**. Нетрудно увидеть, что распределение этой случайной величины зависит не от k параметров $a_1, a_2, ..., a_k$, а только от суммы их квадратов:

$$a = \sum_{i=1}^{k} a_i^2.$$

Это означает, что составленную случайную величину можно представить в виде:

$$\chi^2_k(a) = (\alpha_1 + \sqrt{a})^2 + \alpha_2^2 + ... + \alpha_k^2.$$

Выпишем первые два момента этой СВ:

$$\mathbf{M}\chi^2_k(a) = k + a \quad \text{и} \quad \mathbf{D}\chi^2_k(a) = 2k + 4a.$$

При увеличении k (согласно **ЦПТ**) распределение СВ $\chi^2_k(a)$ асимптотически нормально:

$$\chi^2_k(a) \sim N(k + a, 2k + 4a).$$

Если воспользоваться этим приближением, то для **функции мощности** будет выполняться приближённое равенство:

$$W(p) \equiv \mathrm{P}\{X^2 \geq h \,|\, \bar{H} : p, p \neq p^0\} \approx P\{\chi^2_{m-1}(a) \geq h\} \approx . \, 1 - \Phi\left(\frac{h-(m-1+a)}{\sqrt{2(m-1)+4a}}\right)$$

Существует более точное

приближение для $\chi^2_k(a)$ **— приближение Патнайка**.

Оно основано на приближении $\chi^2_k(a)$ величиной $c\chi^2_m$,

$$\chi^2_k(a) \approx c\chi^2_m, \qquad c=? \quad m=?$$

где c и m **подбираются** из условий равенства **первых двух моментов** этих СВ: $M\chi_k^2(a) \approx Mc\chi_m^2$, $D\chi_k^2(a) \approx Dc\chi_m^2$,

$$k + a = cm \quad \text{и} \quad 2k + 4a = c^2 \cdot 2m.$$

Решая полученную систему уравнений относительно переменных c и m, получаем:

$$c = \frac{k + 2a}{k + a}, \quad m = \frac{(k + a)^2}{k + 2a}.$$

Следовательно функция распределения СВ $\chi_k^2(a)$ приближенно равна:

$$\mathbf{P}\{\chi_k^2(a) < x\} \approx \mathbf{P}\{c\chi_m^2 < x\} = \mathbf{P}\{\chi_m^2 < x/c\}.$$

Пример 6.1. Известный естествоиспытатель, француз Бюффон, при 4040 бросаниях монеты получил $\nu_1 = 2048$ появлений герба и $\nu_2 = 1992$ появлений цифры. Совместимо ли это с гипотезой о том, что монета **правильная** и вероятность выпадения герба при одном подбрасывании равна $p = \frac{1}{2}$?

Решение. Дано: $n = 4040$, $\nu_1 = 2048$, $\nu_2 = 1992$, H: $p = p^0 = 1/2$.

Опишем **процедуру проверки гипотезы**:

если **статистика Пирсона** X^2 принимает достаточно **большое** значение, т.е. $X^2 \geq h$,

то гипотеза H **отклоняется**, причём порог h определяется из условия:

$$\mathbf{P}(\text{отклонить гипотезу } H \mid \text{гипотеза } H \text{ верна}) = \alpha = 0{,}05.$$

Определяем значение **статистики** X^2. В нашем случае $p_1^0 = p^0 = 1/2$, $p_2^0 = 1 - p^0 = 1/2$, поэтому

$$np_1^0 = np_2^0 = 2020.$$

Следовательно,

$$X^2 = \frac{(\nu_1 - np_1^0)^2}{np_1^0} + \frac{(\nu_2 - np_2^0)^2}{np_2^0} = \frac{(2048 - 2020)^2}{2020} + \frac{(1992 - 2020)^2}{2020} = 2 \cdot \frac{28^2}{2020} \approx 0{,}776.$$

Найдём **критический порог** h:

$$\mathbf{P}(\text{отклонить } H \mid H \text{ верна}) = \mathbf{P}\{X^2 \geq h \mid H\} \approx \mathbf{P}\{\chi_1^2 \geq h\} = \mathbf{P}\{\xi_0^2 \geq h\} =$$

$$= 1 - \mathbf{P}\{|\xi_0| < \sqrt{h}\} = \alpha, =0.05$$

где $\xi_0 \sim N(0,1)$. Получили уравнение

$$\mathbf{P}\{|\xi_0| < \sqrt{h}\} = 1 - \alpha \equiv 0{,}95, \quad \Rightarrow \quad 2\Phi(\sqrt{h}) - 1 = 0{,}95,$$

где $\mathbf{\Phi}(x)=\frac{1}{2\pi}\int\limits_{-\infty}^{x} e^{-z^2/2}dz$ — функция распределения стандартного нормального закона $N(0;1)$. Отсюда получаем

$$\sqrt{h}=1{,}96 \text{ и } h=3{,}84.$$

Следовательно процедура **не отклоняет** гипотезу H. Делаем вывод, что

«наблюдения не противоречат гипотезе H».

Итак, процедура вынесения решения такова: если

$$X^2 \geq h = 3{,}84,$$

то гипотеза H о симметрии монеты отклоняется.

В противном случае говорят, что **«наблюдения не противоречат гипотезе H»**.

Значение h получено, исходя из вероятности ошибки при истинности H, равной $\alpha = 0{,}05$.

Заметим, что эта процедура справедлива при любом (достаточно большом) числе n наблюдений.

Продолжим анализ этой ситуации. Нам всегда нужно знать,

хорошо ли процедура ОТКЛОНЯЕТ гипотезу, ЕСЛИ ОНА НЕ ВЕРНА.

На этот вопрос отвечает **МОЩНОСТЬ КРИТЕРИЯ**.

Пусть $n=350$

и **предположим**, что монета **несимметрична**, а именно, истинная вероятность равна:

$$p_1 = \mathrm{P}(\Gamma) = 0{,}6, \quad p_2 = \mathrm{P}(Ц) = 1 - p_1 = 0{,}4, \; p \equiv (p_1, p_2).$$

С какой **вероятностью** наша процедура **отклонит** гипотезу о **симметрии** монеты?

Если вероятности отличаются от теоретических, то статистика Пирсона асимптотически распределена **по нецентральному хи-квадрат** с параметром **нецентральности** a:

$$a=\sum_{i=1}^{2}\frac{(np_i-np_i^0)^2}{np_i^0}=n\sum_{i=1}^{2}\frac{(p_i-p_i^0)^2}{p_i^0}.$$

Используя эту теорему, определяем значения **мощности** $W(p)$ в точке p, т.е. вероятность

$$W(p)=\mathrm{P}\{\text{отклонить } H \mid \bar{H}: p \neq p^0,\, p=(0{,}6;\ 0{,}4)\}=$$
$$=\mathrm{P}\{X^2 \geq h \mid \bar{H}: p \neq p^0\} \approx \mathrm{P}\{\chi_1^2(a) \geq h\},$$

значение параметра **нецентральности**: $a = n\cdot 2\cdot\frac{(0{,}6-0{,}5)^2}{0{,}5} = n\cdot 4\cdot 0.01 = n\cdot 0.04 = 14.$

Обозначив $\xi_0 \sim N(0,1)$, вычисляем нужную вероятность:

$$P\{\chi_1^2(a) \ge h\} = P\{(\xi_0+\sqrt{a})^2 \ge h\} = 1 - P\{-\sqrt{h}-\sqrt{a} \le \xi_0 \le \sqrt{h}-\sqrt{a}\} \approx$$
$$\approx 1 - P\{\xi_0 \le \sqrt{h}-\sqrt{a}\} = 1-\Phi(\sqrt{h}-\sqrt{a}) = 1-\Phi(-1{,}8) \approx 0{,}96.$$

Здесь учтено, что $-\sqrt{h}-\sqrt{a} = -1{,}96-3{,}74 = -5{,}7$ и $\Phi(-5{,}7)\approx 0.$

Задача. Сколько потребуется бросаний, n=?, несимметричной монеты $p = 0{,}6$, чтобы с вероятностью $0{,}9$ гипотеза о симметрии монеты была бы отклонена критерием **хи-квадрат**, имеющем уровень значимости 0,02?

Формализуем задачу: дана вероятность $p = 0{,}6$. Найти $n = ?$

Условия:

$$W(p) \equiv \mathbf{P}\{\text{отклонить гипотезу } H \mid p\} = 0{,}9, \qquad \textbf{(6.6)}$$
$$\alpha \equiv \mathbf{P}\{\text{отклонить гипотезу } H \mid p^0\} = 0{,}02. \qquad \textbf{(6.7)}$$

Решение. Процедура: пусть ν_1 — число выпавших гербов, а ν_2 — число выпавших цифр.

$$X^2 = \sum_{i=1}^{2} \frac{(\nu_i - np_i)^2}{np_i}, \text{ где } p_1 = p_2 = p^0 = \frac{1}{2}.$$

Если $X^2 \ge h$, то гипотеза H отклоняется.

Имеем систему двух уравнений:

$$\alpha \equiv P\{\text{отклонить } H \mid p^0\} \approx P\{\chi_1^2 \ge h\} = 0{,}02 \quad \Rightarrow \quad h = h(\alpha) = 5{,}4. \qquad \textbf{(6.8а)}$$
$$W(p) \equiv P\{\text{отклонить } H \mid p\} \approx P\{\chi_1^2(a) \ge h\} = 0{,}9, \qquad \textbf{(6.8б)}$$

где a — параметр **нецентральности**, равный

$$a = \sum_{i=1}^{2}\frac{(np_i - np_i^0)^2}{np_i^0} = n\sum_{i=1}^{2}\frac{(p_i-p_i^0)^2}{p_i^0}. = n\cdot 2\cdot\frac{0{,}1^2}{0{,}5} = \frac{n}{25} \quad \Rightarrow \quad n = 25a.$$

Учитывая, что $\chi_1^2(a) = (\xi_0+\sqrt{a})^2$, где $\xi_0 \sim N(0,1)$, для $\beta(p)$ имеем:

$$W(p) = \mathbf{P}\{\chi_1^2(a) \ge h\} = \mathbf{P}\{(\xi_0+\sqrt{a})^2 \ge h\} = 0{,}9 \quad \Leftrightarrow$$
$$\Leftrightarrow \quad \mathbf{P}\{-\sqrt{h}-\sqrt{a} < \xi_0 < \sqrt{h}-\sqrt{a}\} = 0{,}1 \quad \Leftrightarrow$$
$$\Leftrightarrow \quad \mathbf{\Phi}(\sqrt{h}-\sqrt{a}) - \mathbf{\Phi}(-\sqrt{h}-\sqrt{a}) = 0{,}1.$$

Поскольку $\mathbf{\Phi}(-\sqrt{h}-\sqrt{a}) \approx 0,$ получаем приближённое равенство

$$\mathbf{\Phi}(\sqrt{h}-\sqrt{a}) \approx 0{,}1,$$

откуда:

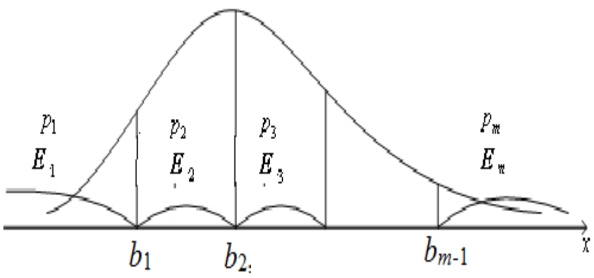
$$\sqrt{h}-\sqrt{a} \approx \mathbf{\Phi}^{-1}(0{,}1) = -1{,}3 \quad \Rightarrow$$
$$\Rightarrow \quad \sqrt{a} \approx \sqrt{h}+1{,}3 = \sqrt{5{,}4}+1{,}3 \approx 2{,}3+1{,}3 = 3{,}6.$$

Далее, $n = 25a = 25\cdot 3{,}6^2 = 324$. Итак, чтобы наш критерий с уровнем значимости $\alpha = 0{,}02$ при вероятности выпадения герба $0{,}6$ отклонял гипотезу о симметрии с вероятностью 0,9, потребуется приближенно 324 бросания.

6.2. Простая гипотеза о распределении

Пусть $\xi \equiv (\xi_1, \xi_2, ..., \xi_n)$ — выборка, по которой нужно проверить гипотезу H **о распределении**:

H: **функция распределения наблюдений** $\xi_j, j = 1,...,n,$ **равна** $F(x)$.



Область значений СВ ξ_i разбиваем на *m* промежутков $E_1, E_2, ..., E_m$, где $m \le \frac{n}{10}$, (в каждом не менее 10) и определяем вероятности

$$P\{\xi_j \in E_i\} = p_i^0, i = 1,...,m.$$

Удобно задать вероятности одинаковыми, равными $p_i^0 = 1/m$. Тогда границы промежутков дискретизации $b_i = Q\,\frac{i}{m}$ есть квантили уровня $\frac{i}{m}$:

$$P\{\xi_j < b_i\} = F(b_i) = \frac{i}{m}.$$

Вычисляется статистика Пирсона

$$X^2 = \sum_{i=1}^{m} \frac{(\nu_i - np_i^0)^2}{np_i^0},$$

Если гипотеза H **верна**, то статистика X^2 **приближенно** распределена по закону χ^2_{m-1}.

Если $X^2 \ge h,$ то гипотеза H отклоняется.

Порог h определяется из знакомого условия:

$$P(\text{откл. } H \mid H \text{ верна}) = \alpha \quad \Rightarrow h \approx Q(1 - \alpha, m - 1).$$

6.3. Сложная гипотеза о распределении

Пусть $\xi \equiv (\xi_1, \xi_2, ..., \xi_n)$ — выборка: H: $\xi_j \sim F(x; a)$. Например, гипотеза H — о пуассоновости … или H — о нормальности …

Последовательность действий

1. Оценить параметр a методом МП: a^* — МП-оценка. Далее как в предыдущем случае.

2. Выбрать m промежутков дискретизации, полагая параметр a, равным a^*. Определить вероятности $p_1^0, ..., p_m^0$, используя оценки a^*.

3. Определить $\nu_1, \ldots, \nu_m$.

4. Вычислить $\tilde{X}^2 = X^2\ \nu_1, \ldots, \nu_m, p_1^0(a^*), \ldots, p_m^0(a^*)$.

5. Записать правило принятия решения:

если $\tilde{X}^2 \geq h$, то гипотеза H отклоняется,

где порог h находится с помощью следующей теоремы.

Теорема 6.3. Если гипотеза H верна, то при $n \to \infty$ функция распределения статистики $\tilde{X}^2$ сходится к функции распределения СВ χ^2_{m-1-k}, т.е.

$$\mathbf{P}\{\tilde{X}^2 < x \mid H\} \xrightarrow[n\to\infty]{} \mathbf{P}\{\chi^2_{m-1-k} < x\}.$$

Поэтому из равенства $\mathbf{P}\{\chi^2_{m-1-k} < x\} = \alpha$ находим

$$h = Q(1-\alpha; m-1-k).$$

Замечание. Существует множество статистических задач, в которых используется нормальность наблюдений. Чтобы применить соответствующие процедуры, нужна проверка гипотезы о нормальности наблюдений. В лабораторной работе 5 эта задача рассматривается

6.4. Гипотеза о независимости двух бинарных признаков (таблица сопряженности признаков 2×2)

Пусть имеем n наблюдений над n объектами, выбранными случайно из большой совокупности. Любой объект характеризуется двумя признаками A и B (каждый признак присутствует или отсутствует в объекте). Имея наблюдения, нужно ответить на вопрос, влияют ли эти признаки друг на друга.

Проверяется гипотеза H о **независимости признаков** A и B. Если обозначить $p_A = \mathbf{P}\{A\}$, $p_B = \mathbf{P}\{B\}$, то гипотеза о независимости сводится к следующему: H:

$$\mathbf{P}\{AB\} = p_A p_B, \tag{6.9}$$

т. е. вероятность встретить сочетание признаков A и B равна произведению вероятностей встретить A и встретить B. В результате анализа на присутствие признаков могут появиться события

$$AB,\ A\bar{B},\ \bar{A}B,\ \bar{A}\bar{B}.$$

Пусть эти комбинации появились $\nu_1, \nu_2, \nu_3, \nu_4$ число раз соответственно. Гипотеза о независимости A и B сводится к гипотезе о том, что вероятности этих событий таковы

$$p_1^0 = p_A p_B,\ p_2^0 = p_A(1-p_B),\ p_3^0 = (1-p_A)p_B,\ p_4^0 = (1-p_A)(1-p_B).$$

Вероятности p_A и p_B — два неизвестных параметра. Задача сводится к проверке **сложной гипотезы о вероятностях**: оцениваются вероятности p_A, p_B, и оценки подставляются в выражение для статистики X^2.

Приведём окончательные конкретные рабочие формулы. Для этого представим данные в следующей таблице 2×2:

	A	$\bar{A}$	
B	ν_{11}	ν_{12}	$\nu_{1\bullet} = \nu_{11} + \nu_{12}$
$\bar{B}$	ν_{21}	ν_{22}	$\nu_{2\bullet} = \nu_{21} + \nu_{22}$
	$\nu_{\bullet 1} = \nu_{11} + \nu_{21}$	$\nu_{\bullet 2} = \nu_{12} + \nu_{22}$	n

Если провести необходимые выкладки, получим следующую статистику:

$$\tilde{X}^2 \equiv n \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{\bullet 1}\nu_{\bullet 2}\nu_{1\bullet}\nu_{2\bullet}}, \tag{6.10}$$

где в числителе стоит квадрат определителя матрицы, а в знаменателе — произведение частных сумм (точка в обозначениях означают суммирование по соответствующему индексу).

Пример 6.2. 1000 человек классифицировали по признаку дальтонизма. По приведённым ниже данным проверить, **есть ли зависимость между наличием дальтонизма ($Д$, $\bar{Д}$) и полом (*М*, *Ж*) человека.** Принять уровень значимости (вероятность ошибочно отклонить гипотезу) равным $\alpha = 0{,}003$.

Решение. Это задача о независимости бинарных признаков: **проверяем гипотезу H о независимости дальтонизма от пола.**

	$Д$	$\bar{Д}$	
М	$\nu_{11} = 38$	$\nu_{12} = 442$	$\nu_{1\bullet} = 480$
Ж	$\nu_{21} = 6$	$\nu_{22} = 514$	$\nu_{2\bullet} = 520$
	$\nu_{\bullet 1} = 44$	$\nu_{\bullet 2} = 956$	$n = 1000$

Вычисляем по (**6.10**) статистику $\tilde{X}^2$ Пирсона:

$$\tilde{X}^2 = 1000 \cdot \frac{(38 \cdot 514 - 442 \cdot 6)^2}{480 \cdot 520 \cdot 956 \cdot 44} \approx 27{,}1.$$

При **истинности** гипотезы H статистика $\tilde{X}^2$ подчиняется **приближенно** закону хи-квадрат χ_1^2 и, если

$$\tilde{X}^2 \geq h = Q(1 - \alpha, 1),$$

то гипотеза H **отклоняется**. Поскольку в примере $27{,}1 > h = 3^2 = 9$, то гипотеза H **отклоняется**, т.е. выносим решение: **признаки зависимы.**

Можно выносить решение иначе: вычисляем условную вероятность статистике $\tilde{X}^2$ принять значение **не менее** 27,1, если гипотеза **верна**:

$$\mathbf{P}\{\tilde{X}^2 > 27{,}1 \mid H\} \approx \mathbf{P}\{\chi_1^2 > 27{,}1\} = \mathbf{P}\ |\xi_0| > \sqrt{27}\ \approx 10^{-7}.$$

Эта вероятность слишком мала, чтобы признать H. Гипотеза **отклоняется** с очень **высоким уровнем значимости**.

6.5. Гипотеза о независимости признаков (таблица сопряженности признаков $k \times m$)

Предполагается, что каждый из двух признаков A и B имеет несколько уровней:

$$A_1, A_2, \ldots, A_m \quad \text{и} \quad B_1, B_2, \ldots, B_k.$$

Пусть $\mathbf{P}(A_i) = q_i$ и $\mathbf{P}(B_j) = r_j$.

Результаты n наблюдений записываются в таблицу $k \times m$, при этом

$(A_i\ B_j)$ наблюдалось ν_{ij} раз

Проверяется гипотеза о независимости признаков, т.е.

$$p_{ij} = \mathbf{P}\{A_i B_j\} = \mathbf{P}\{A_i\}\mathbf{P}\{B_j\} = q_i r_j,$$

где q_i и r_j — **неизвестные параметры** общим числом $k + m - 2$. Статистика Пирсона после подстановки оценок принимает вид:

$$\tilde{X}^2 = n\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\nu_{ij}^2}{\nu_{i,\bullet}\cdot\nu_{\bullet,j}} - 1\right].$$

Если гипотеза **верна**, то $\tilde{X}^2 \sim \chi_f^2$ **асимптотически**, где $f = (k-1)\times(m-1)$ — **число степеней свободы**.

Пример 6.3. Утверждается, что результат (плюс или минус) действия лекарства зависит от способа (A, B, C) его применения. Проверить это утверждение по следующим данным.

	A	B	C	
–	11	17	16	44
+	20	23	19	62
	31	40	35	106

1-й вариант вынесения решения. Вычислим статистику Пирсона при $\alpha = 0{,}05$:

$$\tilde{X}^2 = 106\left[\frac{11^2}{31\cdot 44} + \frac{17^2}{40\cdot 44} + \ldots + \frac{23^2}{40\cdot 62} + \frac{19^2}{35\cdot 62} - 1\right] \approx 0{,}63.$$

Если $\tilde{X}^2 > h = Q(1-\alpha, f = 2) = 5{,}99 \approx 6,$

то гипотеза H **отклоняется**.

Поскольку $\tilde{X}^2 < h$, то гипотеза H **не отклоняется**, т.е. наблюдения **не противоречат** гипотезе о независимости результата от способа применения.

2-й эквивалентный **вариант** вынесения решения. Определим вероятность получить **«столь же большое»**, как полученное $\tilde{X}^2$. Если вероятность мала (не превосходит α), то гипотеза H **отклоняется**, т.е., если

$$\mathbf{P}\{\chi_2^2 \geq \tilde{X}^2\} \leq \alpha, \text{ то гипотеза } H \textbf{ отклоняется}.$$

В нашем случае $\mathbf{P}\{\chi_2^2 \geq \tilde{X}^2 = 0{,}63\} \approx 0{,}3 > \alpha$, следовательно наблюдения **не противоречат** гипотезе.

6.6. Проверка гипотезы об однородности выборок

Теоретическое введение.

Пусть имеется две выборки, как всегда для критерия Пирсона, в частотном виде.

	A_1	…	A_m	
Выборка 1	ν_{11}	…	ν_{1m}	$n_1 = \nu_{1\bullet}$
Выборка 2	ν_{21}	…	ν_{2m}	$n_2 = \nu_{2\bullet}$
	$\nu_{\bullet 1}$	…	$\nu_{\bullet m}$	$n = n_1 + n_2$

Для **первой выборки** неизвестные вероятности равны $q_1, \ldots, q_m$,

для **второй выборки** неизвестные вероятности равны $r_1, \ldots, r_m$.

Проверяется гипотеза о равенстве вероятностей:

$$H:\ q_i = r_i = p_i,\ i = \overline{1, m}.$$

При гипотезе H имеем вектор p **неизвестных вероятностей**, которые оцениваются по минимуму статистики X^2, и после подстановки получаем решающую статистику

$$\tilde{X}^2 = n\left[\sum_{i=1}^{2}\sum_{j=1}^{m}\frac{\nu_{ij}^2}{\nu_{i,\bullet}\cdot\nu_{\bullet,j}} - 1\right]. \tag{6.11}$$

Если гипотеза H верна, то $\tilde{X}^2 \sim \chi^2_f$, где $f = (m-1)$. Потому, если $\tilde{X}^2 \geq h = Q(1-\alpha, f)$, то гипотеза H **об однородности двух выборок отклоняется**.

При проверке гипотезы об однородности k выборок, формула остаётся **верной**, но в первой сумме число 2 заменяется на k, и число степеней свободы

$$f = (k-1)(m-1).$$

Заметим, что формула имеет такой же вид, что и для теста проверки гипотезы о независимости 2-х признаков (таблица $k \times m$ сопряжённости признаков).

Пример 6.4. Два дня подряд записывали уровень шума радиоприёмника. В первый день из 120 замеров в течение одного часа превышение эталонного уровня шума произошло 8 раз, а во второй день из 120 замеров — 12 раз. Проверить гипотезу о том, что вероятность превышения уровня шума не изменилась. Данные записаны в таблицу.

	Π	$\bar{\Pi}$	
1-й день	$\nu_1 = 8$	$n_1 - \nu_1 = 112$	$n_1 = 120$
2-й день	$\nu_2 = 12$	$n_2 - \nu_2 = 108$	$n_2 = 120$
	20	220	$n = n_1 + n_2 = 240$

Здесь «Π» обозначает превышение эталонного уровня шума, а «$\bar{\Pi}$» — отсутствие превышения. Гипотеза об однородности выборок H: $p_j = q_j$,. $j = 1, 2$.

Вычислим статистику $\tilde{X}^2$:

$$\tilde{X}^2 = 240\left[\frac{8^2}{20\cdot120} + \frac{112^2}{220\cdot120} + \frac{12^2}{20\cdot120} + \frac{108^2}{220\cdot120} - 1\right] \approx 0{,}87.$$

Её значение $0{,}87 = \tilde{X}^2$ меньше … $h = Q(1-\alpha, 1) = 3{,}84$ (при $\alpha = 0{,}05$). Следовательно, наблюдения н**е противоречат** гипотезе о равенстве вероятностей.

Пример 6.5. Проверка гипотезы **об однородности** 3-х выборок (условных)

	A	B	C	n_i
Выборка 1	6	10	14	30
Выборка 2	10	10	10	30
Выборка 3	14	10	6	30

	30	30	30	90

Объёмы выборок **одинаковы**. В первой выборке значения возрастают, во второй — не меняются, в третьей — убывают.

Проверяем гипотезу *H* **об однородности трёх выборок**, т.е. о равенстве вероятностей. Вычислим статистику Пирсона:

$$\tilde{X}^2 = 90\left[\frac{6^2}{30\cdot 30}+\frac{10^2}{30\cdot 30}+\ldots+\frac{10^2}{30\cdot 30}+\frac{6^2}{30\cdot 30}-1\right] = 6,4.$$

Если гипотеза H **верна**, то

$$\tilde{X}^2 \sim \chi^2_{2\cdot 2} = \chi^2_4 \qquad \text{и} \qquad \mathbf{P}\{\chi^2_4 \geq 6,4\} \approx 0,18$$

(различие между наблюдаемыми и ожидаемыми частотами есть, но при однородности *H* это значение **вполне вероятно получить**; эта вероятность 0,18). Следовательно, наблюдения **не противоречат** гипотезе H. Вроде бы разные распределения, однако **наблюдений маловато** для **отклонения гипотезы** об однородности.

Если все данные таблицы увеличить в k раз, то $\tilde{X}^2$ увеличится в k раз. будет не 6,4, а $k\cdot 6,4$. При $k = 2$ статистика $\tilde{X}^2$ увеличится в два раза, и гипотеза H будет отклонена с минимальным уровнем значимости:

$$\mathbf{P}\{\chi^2_4 \geq 12,8\} \approx 0,015.$$

Получить такое различие слишком маловероятно: 0,015. **Гипотезу отклоним**.

ДОМАШНЕЕ ЗАДАНИЕ

6.1 [2]. При 120 бросаниях игральной кости шестёрка выпала 40 раз. Согласуется ли этот результат с утверждением, что кость правильная? Принять $\alpha = 0,05$.

6.2 [2]. Число выпадений герба при 20 подбрасываниях двух монет распределились следующим образом:

Количество гербов	0	1	2
Число подбрасываний	**4**	**8**	**8**

Согласуются ли эти результаты с предположениями о симметричности монет и независимости результатов подбрасывания? Принять $\alpha = 0,05$.

6.3 [2]. Ниже приводятся данные о фактических объёмах сбыта (в условных единицах) в пяти районах:

Район	1	2	3	4	5
Фактический объём	**110**	**130**	**70**	**90**	**100**

сбыта					

Согласуются ли эти результаты с предположениями о том, что сбыт продукции в этих районах должен быть одинаковым? Принять $\alpha = 0{,}01$.

6.4 [2]. На экзамене студент отвечает только на один вопрос по одной из трёх частей курса. Анализ вопросов, заданных 60 студентам, показал, что 23 студента получили вопросы из первой, 15 — из второй и 22 — из третьей частей курса.

Можно ли считать, что студент, идущий на экзамен, с равной вероятностью получит вопрос по любой из трёх частей курса? Принять $\alpha = 0{,}10$.

6.5 [2]. Во время второй мировой войны на Лондон упало 537 самолётов-снарядов. Вся территория Лондона была разделена на 576 участков площадью по 0,25 км^2. Ниже приведены числа n_k участков, на которые упало k снарядов:

k	0	1	2	3	4	5 и более
n_k	229	211	93	35	7	1

Согласуются ли эти данные с гипотезой о том, что число снарядов, упавших на каждый из участков, имеет распределение Пуассона? Принять $\alpha = 0{,}05$.

6.6 [2]. Для определения зависимости цвета волос жителей от их местожительства были обследованы три группы людей из районов A, B и C. Свидетельствуют ли приводимые ниже результаты обследования о зависимости цвета волос жителей от их местожительства? Принять $\alpha = 0{,}05$.

Район	Цвет волос		
	Рыжий	Светлый	Тёмный
A	2	9	9
B	3	6	21
C	15	15	20

6.7 [2]. В течение месяца завод поставил предприятию 200 корпусов, из которых 3 оказались дефектными. В следующий месяц было поставлено 850 корпусов, из которых 7 оказались дефектными. Изменилась ли доля дефектных корпусов в поставках завода? Принять $\alpha = 0{,}01$.

Ответ на вопрос

Б) Теперь найдём **точный доверительный** интервал. В качестве **оценивающей** статистики используем сумму $\zeta = \sum_{i=1}^{n} \xi_i$. Выясним закон распределения суммы. Одно слагаемое ξ_i подчиняется **показательному** закону с **плотностью**

$$p(x_i; a) = \begin{cases} ae^{-ax_i}, \text{ если } x_i \geq 0; \\ 0, \quad \text{если } x_i < 0, \end{cases} \quad \text{где } i = \overline{1, n}, \ a > 0.$$

Мы знаем, что закон распределения **хи-квадрат** χ_2^2 с **двумя** степенями свободы есть **показательный** закон с параметром $a = \frac{1}{2}$:

$$\alpha_1^2 + \alpha_2^2 = \chi_2^2 \sim E\ a = \frac{1}{2}, \text{ где } \alpha_1, \alpha_2 \sim N(0;1).$$

Если умножить ξ_i на $2a$, то получится СВ $2a\xi_i$ с плотностью

$$p(x_i) = \begin{cases} \frac{1}{2} e^{-x_i/2}, \text{ если } x_i \geq 0; \\ 0, \quad \text{если } x_i < 0, \end{cases}$$

т.е. СВ χ_2^2. Тогда СВ $2a\zeta = \sum_{i=1}^{n} 2a\xi_i$ есть сумма $2a\zeta = \sum_{i=1}^{n} \chi_{2i}^2 = \chi_{2n}^2$. Итак, **центральная статистика** $\varphi(\xi; a) = 2a\sum_{i=1}^{n} \xi_i$ имеет распределение χ_{2n}^2.