Занятие 4. Критерий хи-квадрат проверки гипотез.

Домашнее задание:

Глава 19, задачи: 282, 286, 291, 299, 303, 274 (дисперсионный анализ).

Задача 4.1.

Эксперимент заключается в подбрасывании некоторой монеты n раз и фиксировании количеств выпавших гербов ν_1 и решек ν_2. В результате проведения эксперимента с $n = 4040$ наблюдалось $y_1 = 2048$ выпадений герба и $y_2 = 1992$ выпадений решки.

Требуется проверить гипотезу о симметричности монеты для уровня значимости $\alpha = 0.05$.

Решение:

1) В условиях задачи совокупность наблюдаемых величин (ν_1, ν_2) имеет полиномиальное распределение $\Pi(p_1, p_2; n)$, где $p_1 + p_2 = 1$:

$$P\{\nu_1 = k_1, \nu_2 = k_2\} = \frac{n!}{k_1! k_2!} p_1^{k_1} p_2^{k_2},$$

$$k_1 + k_2 = n,$$

где p_1 – вероятность выпадения герба и p_2 – вероятность выпадения решки.

2) У симметричной монеты вероятности выпадения герба и решки совпадают:

$$p_1 = p_1^0 = 0.5,$$

$$p_2 = p_2^0 = 0.5.$$

Если гипотезу о симметричности монеты считать основной гипотезой H_0, тогда H_0 соответствует множество полиномиальных распределений $\wp_0$, состоящее из одного полиномиального распределения $\Pi(p_1^0, p_2^0; n)$.

3) Требование проверить гипотезу о симметричности H_0 фактически означает следующее: на основании одной полученной реализации наблюдения (y_1, y_2) требуется сделать обоснованное заключение о том, имеет ли совокупность величин (ν_1, ν_2) полиномиальное распределение $\Pi(p_1^0, p_2^0; n)$ или какое-либо другое полиномиальном распределение.

Для проверки гипотезы H_0 можно использовать критерий хи-квадрат со статистикой критерия $X_n^2(\nu_1, \nu_2 \mid p_1^0, p_2^0)$:

$$X_n^2(\nu_1, \nu_2 \mid p_1^0, p_2^0) = \frac{(\nu_1 - np_1^0)^2}{np_1^0} + \frac{(\nu_2 - np_2^0)^2}{np_2^0},$$

которая для реализации (y_1, y_2) принимает значение:

$$X_n^2(y_1, y_2 \mid p_1^0, p_2^0) = \frac{(y_1 - np_1^0)^2}{np_1^0} + \frac{(y_2 - np_2^0)^2}{np_2^0} =$$

$$= \frac{(2048 - 2020)^2}{2020} + \frac{(1992 - 2020)^2}{2020} = 2\frac{28^2}{2020} \approx 0.776.$$

4) Если гипотеза о симметричности H_0 верна, то есть наблюдение (ν_1, ν_2) имеет полиномиальное распределение $\Pi(p_1^0, p_2^0; n)$, тогда распределение статистики $X_n^2(\nu_1, \nu_2 \mid p_1^0, p_2^0)$ при больших n приближенно совпадает с распределением хи-квадрат с

одной степенью свободы $\chi^2(1)$. Отсюда следует, что в критической области $\Gamma_\alpha = (h_\alpha; \infty)$ пороговое значение h_α является квантилью распределения $\chi^2(1)$ уровня $1-\alpha$:

$$h_\alpha \approx 3.8415$$

5) Поскольку значение статистики $X_n^2(y_1, y_2 \mid p_1^0, p_2^0) \approx 0.776$ не принадлежит критической области $\Gamma_\alpha = (3.8415; \infty)$, то основная гипотеза H_0 принимается.

Ответ:

Гипотеза о симметричности принимается.

Задача 4.2.

В результате опроса $n = 100$ человек об уровне зарплаты, были получены следующие результаты:

Номер уровня	1	2	3	4	5
Уровень зарплаты (тысячи рублей)	[0; 10)	[10; 20)	[20; 30)	[30; 40)	[40; 50)
Количество человек	16	21	25	22	16

Определить наименьший уровень значимости отклонения гипотезы о равномерном распределении зарплаты по уровням.

Решение:

1) В данной задаче наблюдаемой является совокупность случайных величин $(\nu_1, \nu_2, \nu_3, \nu_4, \nu_5)$, в которой величина ν_i $(i = \overline{1,5})$ является количеством человек, уровень зарплаты которых принадлежит уровню с номером i. Пусть вероятности p_i $(i = \overline{1,5})$ есть вероятности того, что отдельно взятый человек имеет зарплату, принадлежащую уровню с номером i, тогда совокупность величин $(\nu_1, \nu_2, \nu_3, \nu_4, \nu_5)$ имеет полиномиальное распределение $\Pi(p_1, ..., p_5; n)$:

$$P\{\nu_1 = k_1, \nu_2 = k_2, \nu_3 = k_3, \nu_4 = k_4, \nu_5 = k_5\} = \frac{n!}{k_1! k_2! k_3! k_4! k_5!} p_1^{k_1} p_2^{k_{21}} p_3^{k_3} p_4^{k_4} p_5^{k_5}.$$

2) Пусть гипотеза о равномерном распределении количеств человек по уровням зарплаты является основной гипотезой H_0. Равномерно распределение количеств означает, что отдельно взятый человек с равной вероятностью имеет один из уровней зарплаты, другими словами гипотеза H_0 утверждает, что все вероятности p_i равны между собой:

$$p_1 = p_1^0 = 0.2,$$
$$\ldots,$$
$$p_5 = p_5^0 = 0.2.$$

Легко видеть, что основной гипотезе H_0 соответствует множество распределений $\wp_0$ совокупности величин $(\nu_1, \nu_2, \nu_3, \nu_4, \nu_5)$, состоящее из одного полиномиального распределения $\Pi(p_1^0, ..., p_5^0; n)$.

3) Для проверки гипотезы H_0 можно использовать критерий хи-квадрат, со статистикой критерия $X_n^2(\nu_1, ..., \nu_5 \mid p_1^0, ..., p_5^0)$:

$$X_n^2(\nu_1, ..., \nu_5 \mid p_1^0, ..., p_5^0) = \sum_{k=1}^{5} \frac{(\nu_k - np_k^0)^2}{np_k^0},$$

которая для реализации $(y_1, ..., y_5) = (16, 21, 25, 22, 16)$ принимает значение:

$$X_n^2(y_1, ..., y_5 \mid p_1^0, ..., p_5^0) = \sum_{k=1}^{5} \frac{(y_k - np_k^0)^2}{np_k^0} =$$

$$= \frac{(-4)^2 + 1^2 + 5^2 + 2^2 + (-4)^2}{2020} = \frac{62}{20} = 3.1 .$$

Если гипотеза о симметричности H_0 верна, тогда распределение статистики $X_n^2(\nu_1,..., \nu_5 \mid p_1^0,..., p_5^0)$ при больших n приближенно совпадает с распределением хи-квадрат с четырьмя степенями свободы $\chi^2(4)$, поэтому наименьший уровень значимости отклонения основной гипотезы H_0 является вероятностью $\alpha_{\min}$:

$$\alpha_{\min} = P\{X_n^2(\nu_1,..., \nu_5; p_1^0,..., p_5^0) \geq X_n^2(n_1,..., n_5; p_1^0,..., p_5^0)\} =$$
$$= P\{X_n^2(\nu_1,...,\nu_5; p_1^0,..., p_5^0) \geq 3.1\} \approx P\{\chi_4^2 \geq 3.1\} \approx 0.54 .$$

Поскольку полученное значение вероятности не является малым, то можно утверждать, что гипотеза о равномерности согласуется с полученной реализацией $(y_1,..., y_5) = (16,21,25,22,16)$. Любой критерий хи-квадрат с уровнем значимости не более 0.54 принимает основную гипотезу.

Ответ:

Наименьший уровень значимости отклонения гипотезы о равномерности примерно равен 0.54.

Задача 4.3.

Расстояние $m = 1$ метр измеряется $n = 50$ раз некоторым прибором, не имеющим систематической ошибки, в результате измерений, которые следует считать независимыми, получены числовые значения $(x_1,..., x_n)$. Проверить гипотезу о нормальном распределении ошибки измерения прибора, используя критерий хи-квадрат с количеством интервалов $k = 4$.

Решение:

1) Числовые значения $(x_1,..., x_n)$, полученные при измерении расстояния n раз, можно считать реализацией выборки $(\xi_1,..., \xi_n)$, в которой каждая случайная величина ξ_i имеет некоторое неизвестное распределение и представима в виде:

$$\xi_i = m + \varepsilon_i ,$$

где случайные величины ε_i являются случайными ошибками измерения прибора.

Основная гипотеза, обозначим её H_0, утверждает, что все ошибки ε_i имеют некоторое нормальное распределение $N(0,\sigma^2)$ или, что то же самое, распределение случайных величин ξ_i является нормальным $N(m,\sigma^2)$ с известным значением $m = 1$ и неизвестным параметром σ^2.

2) Для проверки основной гипотезы с помощью критерия хи-квадрат необходимо предварительно провести разбиение всей числовой оси на четыре интервала L_1, L_2, L_3 и L_4, причем желательно выбрать интервалы таким образом, чтобы вероятности попадания случайных величин ξ_i в каждый интервал были примерно одинаковыми:

$$P\{\xi_i \in L_1\} \approx P\{\xi_i \in L_2\} \approx P\{\xi_i \in L_3\} \approx P\{\xi_i \in L_4\} \approx \frac{1}{4}.$$

Легко видеть, что если значение $y_{0.25}$ является квантилью распределения $N(m,\sigma^2)$ уровня $\frac{1}{4}$:

$$\Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{4}$$

и в качестве интервалов выбраны интервалы:

$$L_1 = (-\infty ; y_{0.25}) ,\ L_2 = (y_{0.25}; m) ,\ L_3 = (m; m + m - y_{0.25}) ,\ L_3 = (m + m - y_{0.25}; \infty) ,$$

тогда,

$$P\{\xi_i \in L_1\} = P\{\xi_i \in L_2\} = P\{\xi_i \in L_3\} = P\{\xi_i \in L_4\} = \frac{1}{4},$$

действительно,

$$P\{\xi_i \in L_1\} = P\{\xi_i \in (-\infty\,; y_{0.25})\} = \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{4},$$

$$P\{\xi_i \in L_2\} = P\left\{\xi_i \in (y_{0.25}\,; m)\right\} = \Phi\left(\frac{m - m}{\sigma}\right) - \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{2} - \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{4},$$

$$P\{\xi_i \in L_3\} = P\left\{\xi_i \in (m; m + m - y_{0.25})\right\} = \Phi\left(\frac{m + m - y_{0.25} - m}{\sigma}\right) - \Phi\left(\frac{m - m}{\sigma}\right) = \Phi\left(\frac{m - y_{0.25}}{\sigma}\right) - \frac{1}{2} =$$

$$= 1 - \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) - \frac{1}{2} = \frac{1}{2} - \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{4},$$

$$P\{\xi_i \in (m + m - y_{0.25}\,; \infty\} = 1 - \Phi\left(\frac{m + m - y_{0.25} - m}{\sigma}\right) = 1 - \Phi\left(\frac{m - y_{0.25}}{\sigma}\right) =$$

$$= 1 - \left(1 - \Phi\left(\frac{y_{0.25} - m}{\sigma}\right)\right) = \Phi\left(\frac{y_{0.25} - m}{\sigma}\right) = \frac{1}{4}.$$

К сожалению, значение параметра σ^2 неизвестно, поэтому для построения интервалов разбиения предварительно оценим величину σ^2 с помощью, например, выборочной дисперсии $\hat{\mu}_2(\xi_1,..., \xi_n)$:

$$\hat{\mu}_2(\xi_1,..., \xi_n) = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - m)^2 .$$

В качестве квантили $y_{0.25}$ можно использовать приближенное значение $\hat{y}_{0.25}$, удовлетворяющее равенству:

$$\Phi\left(\frac{\hat{y}_{0.25} - m}{\sqrt{\hat{\mu}_2(x_1,..., x_n)}}\right) = \frac{1}{4}.$$

Откуда,

$$\hat{y}_{0.25} = m + \sqrt{\hat{\mu}_2(x_1,..., x_n)}\,\Phi^{-1}\left(\frac{1}{4}\right),$$

Таким образом, для проверки гипотезы в качестве интервалов можно использовать следующие интервалы:

$$\hat{L}_1 = (-\infty\,; \hat{y}_{0.25}),\ \hat{L}_2 = (\hat{y}_{0.25}\,; m),\ \hat{L}_3 = (m; m + m - \hat{y}_{0.25}),\ \hat{L}_4 = (m + m - \hat{y}_{0.25}\,; \infty).$$

3) На основе выборки $(\xi_1,..., \xi_n)$ построим совокупность величин $(\nu_1, \nu_2, \nu_3, \nu_4)$, в которой каждая случайная величина ν_j является количеством случайных величин ξ_i, принадлежащих интервалу L_j:

$$\nu_j = \sum_{i=1}^{n} I(\xi_i, \hat{L}_j),\ j = \overline{1, k},$$

$$I(\xi_i, \hat{L}_j) = \begin{cases} 0, & \xi_i \notin \hat{L}_j \\ 1, & \xi_i \notin \hat{L}_j \end{cases}.$$

Поскольку совокупность величин $(\xi_1,..., \xi_n)$ является выборкой, то величины $(\nu_1, \nu_2, \nu_3, \nu_4)$ имеют полиномиальное распределение $\Pi(p_1, p_2, p_3, p_4; n)$ с неизвестными вероятностями p_j ($j = \overline{1, k}$).

Основная гипотеза H_0, утверждающая нормальное распределение случайных величин ξ_i , утверждает, что величины $(\nu_1, \nu_2, \nu_3, \nu_4)$ имеют полиномиальное распределение $\Pi(p_1^0(\sigma), p_2^0(\sigma), p_3^0(\sigma), p_4^0(\sigma); n)$, или иначе неизвестные вероятности p_j равны соответствующим значениям функций $p_j^0(\sigma)$ при некотором значении параметра σ^* :

$$H_0: \ p_j = p_j^0(\sigma^*), \ j = \overline{1,k},$$

где функции $p_j^0(\sigma)$ являются вероятностями:

$$p_j^0(\sigma) = P\{\xi_i \in \hat{L}_j\},$$

которые вычисляются в соответствии с утверждаемым гипотезой H_0 нормальным распределением величин ξ_i . В соответствии с основной гипотезой H_0 и выбранными интервалами $\hat{L}_j$ вероятности $p_j^0(\sigma)$ имеют вид:

$$p_1^0(\sigma) = P\{\xi_i \in \hat{L}_1\} = P\{\xi_i \in (-\infty; \hat{y}_{0.25})\} = \Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right),$$

$$p_2^0(\sigma) = P\{\xi_i \in \hat{L}_2\} = P\{\xi_i \in (\hat{y}_{0.25}; m)\} = \frac{1}{2} - \Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right),$$

$$p_3^0(\sigma) = P\{\xi_i \in \hat{L}_3\} = P\{\xi_i \in (m; m + m - \hat{y}_{0.25})\} == \frac{1}{2} - \Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right),$$

$$p_4^0(\sigma) = P\{\xi_i \in \hat{L}_4\} = P\{\xi_i \in (m + m - \hat{y}_{0.25}; \infty)\} = \Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right).$$

Легко видеть, что вероятности $p_j^0(\sigma)$ зависят от параметра σ сложным образом и дальнейшее решение задачи с параметром σ окажется весьма сложным. Тем не менее, как не трудно заметить, вероятности зависят от параметра σ через выражение $\Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right)$, поэтому можно ввести новый параметр γ :

$$\gamma = \Phi\left(\frac{\hat{y}_{0.25} - m}{\sigma}\right),$$

и считать, что вероятности p_j^0 зависят от параметра γ :

$$p_1^0(\gamma) = \gamma,$$

$$p_2^0(\gamma) = \frac{1}{2} - \gamma,$$

$$p_3^0(\gamma) = \frac{1}{2} - \gamma,$$

$$p_4^0(\gamma) = \gamma.$$

Согласно теореме Фишера в случае, если основная гипотеза H_0 верна, то статистика критерия хи-квадрат $X_n^2(\nu_1, \nu_2, \nu_3, \nu_4 \mid \hat{\gamma})$:

$$X_n^2(\nu_1, \nu_2, \nu_3, \nu_4 \mid \hat{\gamma}) = \sum_{j=1}^{k} \frac{(\nu_i - np_j^0(\hat{\gamma}))^2}{np_j^0(\hat{\gamma})}$$

имеет распределение хи-квадрат с числом степеней свободы $k - 1 - d$ (где $d = 1$ размерность параметра), если в качестве значения параметра γ используется МП-оценка $\hat{\gamma}(\nu_1, \nu_2, \nu_3, \nu_4)$, построенная по функции правдоподобия $L(y_1, y_2, y_3, y_4 \mid \gamma)$ величин $(\nu_1, \nu_2, \nu_3, \nu_4)$ в предположении, что гипотеза H_0 верна. Если гипотеза H_0 верна, тогда совокупность величин $(\nu_1, \nu_2, \nu_3, \nu_4)$ имеет полиномиальное распределение $\Pi(p_1^0(\sigma), p_2^0(\sigma), p_3^0(\sigma), p_4^0(\sigma); n)$ и функция правдоподобия $L(y_1, y_2, y_3, y_4 \mid \gamma)$ имеет вид:

$$L(y_1, y_2, y_3, y_4 \mid \gamma) = \frac{n!}{y_1! y_2! y_3! y_4!}\left[p_1^0(\gamma)\right]^{y_1}\left[p_2^0(\gamma)\right]^{y_2}\left[p_3^0(\gamma)\right]^{y_3}\left[p_4^0(\gamma)\right]^{y_4}.$$

Уравнение правдоподобия для нахождения МП-оценки $\hat{\gamma}$ имеет вид:

$$\left.\frac{\partial}{\partial\gamma}\ln L(\nu_1, \nu_2, \nu_3, \nu_4 \mid \gamma)\right|_{\gamma=\hat{\gamma}} = 0,$$

$$\left.\frac{\partial}{\partial\gamma}\ln\left(\frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!}\left[p_1^0(\gamma)\right]^{\nu_1}\left[p_2^0(\gamma)\right]^{\nu_2}\left[p_3^0(\gamma)\right]^{\nu_3}\left[p_4^0(\gamma)\right]^{\nu_4}\right)\right|_{\gamma=\hat{\gamma}} = 0$$

$$\frac{\partial}{\partial\gamma}\left(\ln\frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!} + \nu_1\ln p_1^0(\hat{\gamma}) + \nu_2\ln p_2^0(\hat{\gamma}) + \nu_3\ln p_3^0(\hat{\gamma}) + \nu_4\ln p_4^0(\hat{\gamma})\right) = 0$$

$$\frac{\partial}{\partial\gamma}\left(\ln\frac{n!}{\nu_1!\nu_2!\nu_3!\nu_4!} + \nu_1\ln\hat{\gamma} + \nu_2\ln\left(\frac{1}{2} - \hat{\gamma}\right) + \nu_3\ln\left(\frac{1}{2} - \hat{\gamma}\right) + \nu_4\ln\hat{\gamma}\right) = 0$$

$$\nu_1\frac{1}{\hat{\gamma}} + \nu_2\frac{(-1)}{\frac{1}{2} - \hat{\gamma}} + \nu_3\frac{(-1)}{\frac{1}{2} - \hat{\gamma}} + \nu_4\frac{1}{\hat{\gamma}} = 0,$$

$$\frac{\nu_1\left(\frac{1}{2} - \hat{\gamma}\right) - \nu_2\hat{\gamma} - \nu_3\hat{\gamma} + \nu_4\left(\frac{1}{2} - \hat{\gamma}\right)}{\hat{\gamma}\left(\frac{1}{2} - \hat{\gamma}\right)} = 0,$$

$$\nu_1\left(\frac{1}{2} - \hat{\gamma}\right) - \nu_2\hat{\gamma} - \nu_3\hat{\gamma} + \nu_4\left(\frac{1}{2} - \hat{\gamma}\right) = 0, \ \hat{\gamma} \neq 0, \frac{1}{2},$$

$$\hat{\gamma}(\nu_1, \nu_2, \nu_3, \nu_4) = \frac{\frac{1}{2}(\nu_1 + \nu_4)}{\nu_1 + \nu_2 + \nu_3 + \nu_4}, \ \hat{\gamma} \neq 0, \frac{1}{2}.$$

Полученное выражение для МП-оценки следует подставить в выражение для статистики критерия хи-квадрат $X_n^2(\nu_1, \nu_2, \nu_3, \nu_4 \mid \hat{\gamma})$ и вычислить значение статистики на основе результатов эксперимента.

4) Предположим, что полученные в результате эксперимента числа $(x_1,\ldots, x_n)$ таковы, что выборочная дисперсия $\hat{\mu}_2(\xi_1,\ldots, \xi_n)$ величины σ^2 имеет значение $\hat{\mu}_2(x_1,\ldots, x_n)$:

$$\hat{\mu}_2(x_1,\ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = 0.0225,$$

тогда для $\hat{y}_{0.25}$ будет получено следующее значение:

$$\hat{y}_{0.25} \approx 1 + \sqrt{0.0225}\cdot(-0.675) \approx 0.9,$$

и следовательно для проверки гипотезы можно использовать следующие интервалы:

$$\hat{L}_1 = (-\infty; 0.9),\ \hat{L}_2 = (0.9; 1),\ \hat{L}_3 = (1; 1.1),\ \hat{L}_4 = (1.1; \infty).$$

Далее, пусть числовые значения $(x_1,\ldots, x_n)$ таковы, что количество чисел x_i принадлежащих интервалу $\hat{L}_1$ равно $y_1 = 12$, интервалу $\hat{L}_2$ – $y_2 = 11$, интервалу $\hat{L}_3$ – $y_3 = 11$, интервалу $\hat{L}_4$ – $y_4 = 16$. Считая вектор (y_1, y_2, y_3, y_4) реализацией совокупности величин $(\nu_1, \nu_2, \nu_3, \nu_4)$ вычислим значение МП-оценки $\hat{\gamma}(\nu_1, \nu_2, \nu_3, \nu_4)$:

$$\hat{\gamma}(y_1, y_2, y_3, y_4) = \frac{\frac{1}{2}(y_1 + y_4)}{y_1 + y_2 + y_3 + y_4} = \frac{\frac{1}{2}(12 + 16)}{12 + 11 + 11 + 16} = \frac{14}{50} = 0.28 .$$

Теперь подставим вычисленное значение оценки в выражения для вероятностей p_j^0:

$$p_1^0(\hat{\gamma}) = 0.28 , \ p_2^0(\hat{\gamma}) = 0.22 , \ p_3^0(\gamma) = 0.22 , \ p_4^0(\gamma) = 0.28 ,$$

и вычислим значение статистики критерия $X_n^2(y_1, y_2, y_3, y_4 \mid \hat{\gamma})$:

$$X_n^2(y_1, y_2, y_3, y_4 \mid \hat{\gamma}) = \sum_{j=1}^{k} \frac{(y_i - np_j^0(\hat{\gamma}))^2}{np_j^0(\hat{\gamma})} =$$

$$= \frac{(12 - 50 \cdot 0.28)^2}{50 \cdot 0.28} + \frac{(11 - 50 \cdot 0.22)^2}{50 \cdot 0.22} + \frac{(11 - 50 \cdot 0.22)^2}{50 \cdot 0.22} + \frac{(16 - 50 \cdot 0.28)^2}{50 \cdot 0.28} =$$

$$= \frac{(12 - 14)^2}{14} + \frac{(11 - 11)^2}{11} + \frac{(11 - 11)^2}{11} + \frac{(16 - 14)^2}{14} = \frac{4 + 4}{14} = \frac{8}{14} .$$

Если основная гипотеза H_0 верна, то распределение статистики $X_n^2(\nu_1, \nu_2, \nu_3, \nu_4 \mid \hat{\gamma})$ при больших n имеет распределение хи-квадрат с числом степеней свободы $k - 1 - d$ (где $d = 1$ размерность параметра), тогда наименьший уровень значимости отклонения основной гипотезы есть вероятность $\alpha_{\min}$:

$$\alpha_{\min} = P\left\{ X_n^2(\nu_1, \nu_2, \nu_3, \nu_4 \mid \hat{\gamma}) \ge \frac{8}{14} \right\} \approx P\left\{ \chi_2^2 \ge \frac{8}{14} \right\} \approx 0.75 .$$

Полученное значение вероятности не является малым, поэтому основная гипотеза принимается.

Ответ:

Наименьший уровень значимости отклонения гипотезы о нормальном распределении ошибки примерно равен 0.75.

Задача 4.4.

Некоторое лекарство может применяться тремя различными способами и приводить к двум различным результатам. После процедуры применения лекарства были собраны следующие данные о результатах:

Способ применения	1	2	3
Результат			
1 (нет)	11	17	16
2 (есть)	20	23	19

Определить наименьший уровень значимости отклонения гипотезы о том, что результат применения лекарства не зависит от способа его применения.

Решение:

1) Будем считать, что в процедуре применения лекарства некоторым случайным образом выбирался способ применения, и при выборе происходило в точности одно из событий B_j – «выбран способ применения j» ($j = 1,2,3$). Применение лекарства выбранным способом случайным образом приводило к двум различным результатам, которые будем представлять двумя событиями A_i – «результат применения i» ($i = 1,2$).

В такой постановке требуется проверить гипотезу о том, что события A_i (результат применения) и события B_j (способ применения) попарно независимы.

2) Исходное наблюдение ν представляет собой матрицу:

$$\nu = \begin{pmatrix} \nu_{11} & \nu_{12} & \nu_{13} \\ \nu_{21} & \nu_{22} & \nu_{23} \end{pmatrix},$$

в которой каждая случайная величина ν_{ij} обозначает количество произошедших событий $A_i B_j$ («результат применения i при выбранном способе j»). Если применение лекарства в каждом отдельном случае носит независимый характер, то совместное распределение величин ν_{ij} является полиномиальным распределением $\Pi(p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}; n)$, в котором каждая вероятность p_{ij} является неизвестной вероятностью наступления события $A_i B_j$ и n является общим количеством применений лекарства.

Основная гипотеза о попарной независимости событий H_0 утверждает, что полиномиальное совместное распределение величин ν_{ij} имеет вид $\Pi(p_1^0 q_1^0, p_1^0 q_2^0, p_1^0 q_3^0, p_2^0 q_1^0, p_2^0 q_2^0, p_2^0 q_3^0; n)$ при некоторых вероятностях $p^0 = (p_1^0, p_2^0)$ и $q^0 = (q_1^0, q_2^0, q_3^0)$, или иначе, что все неизвестные вероятности p_{ij} представимы в виде произведения вероятностей $p_i^0 q_j^0$:

$$H_0: \; p_{ij} = p_i^0 q_j^0$$

Для проверки основной гипотезы H_0 может применяться критерий хи-квадрат проверки сложной гипотезы, в котором в качестве параметра можно выбрать, например, вектор вероятностей (p_1^0, q_1^0, q_2^0) (остальные вероятности вычисляются через вероятности параметра: $p_2^0 = 1 - p_1^0$, $q_3^0 = 1 - q_1^0 - q_2^0$).

В соответствии с процедурой проверки по критерию хи-квадрат необходимо вычислить МП-оценку параметра (p_1^0, q_1^0, q_2^0) и вероятности p_2^0 и q_3^0:

$$\hat{p}_1^0(\nu) = \frac{\nu_{11} + \nu_{12} + \nu_{13}}{\sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}}, \quad \hat{p}_2^0(\nu) = \frac{\nu_{21} + \nu_{22} + \nu_{23}}{\sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}},$$

$$\hat{q}_1^0(\nu) = \frac{\nu_{11} + \nu_{21}}{\sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}}, \quad \hat{q}_2^0(\nu) = \frac{\nu_{12} + \nu_{22}}{\sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}}, \quad \hat{q}_3^0(\nu) = \frac{\nu_{13} + \nu_{23}}{\sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}},$$

и значение статистики критерия:

$$X_n^2(\nu \mid p_1^0, q_1^0, q_2^0) = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(\nu_{ij} - n\hat{p}_i^0 \hat{q}_j^0)^2}{n\hat{p}_i^0 \hat{q}_j^0},$$

$$n = \sum_{i=1}^{2}\sum_{j=1}^{3} \nu_{ij}$$

3) В данном случае реализацией матрицы величин ν является матрица y:

$$y = \begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{pmatrix} = \begin{pmatrix} 11 & 17 & 16 \\ 20 & 23 & 19 \end{pmatrix}.$$

В соответствии с реализацией y вычислим значения вероятностей:

$$\hat{p}_1^0(y) = \frac{y_{11} + y_{12} + y_{13}}{\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}} = \frac{11 + 17 + 16}{106} \approx 0.4151,$$

$$\hat{p}_2^0(y) = \frac{y_{21} + y_{22} + y_{23}}{\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}} = \frac{20 + 23 + 19}{106} \approx 0.5849,$$

$$\hat{q}_1^0(y) = \frac{y_{11} + y_{21}}{\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}} = \frac{11 + 20}{106} \approx 0.2925 \quad ,$$

$$\hat{q}_2^0(y) = \frac{y_{12} + y_{22}}{\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}} = \frac{17 + 23}{106} \approx 0.3774 \quad ,$$

$$\hat{q}_3^0(y) = \frac{y_{13} + y_{23}}{\sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}} = \frac{16 + 19}{106} \approx 0.3301 \quad .$$

С учетом полученных значений вероятностей, значение статистики критерия:

$$X_n^2(y \mid p_1^0, q_1^0, q_2^0) = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(y_{ij} - n\hat{p}_i^0 \hat{q}_j^0)^2}{n\hat{p}_i^0 \hat{q}_j^0} \approx 0.7346 \quad .$$

Если основная гипотеза H_0 верна, то распределение статистики $X_n^2(\nu \mid p_1^0, q_1^0, q_2^0)$ при больших n приближенно совпадает с распределением хи-квадрат с числом степеней свободы $(k-1)(m-1)$, где $k = 2$ – количество событий A_i и $m = 3$ – количество событий B_j. Таким образом, наименьший уровень значимости отклонения основной гипотезы H_0 является вероятность $\alpha_{\min}$:

$$\alpha_{\min} = P\left\{X_n^2(\nu \mid p_1^0, q_1^0, q_2^0) \geq 0.7346\right\} \approx P\left\{\chi_2^2 \geq 0.7346\right\} \approx 0.6926$$

Полученное значение вероятности не является малым, поэтому гипотеза о независимости принимается.

Ответ:

Наименьший уровень значимости отклонения гипотезы о независимости примерно равен 0,7.

Задача 4.5.

В двух группах людей проведено исследование о распределении по группам крови, в результате исследования получены следующие данные:

Группа крови	1	2	3	4
Группа людей				
1	121	120	79	33
2	118	95	121	30

Проверить гипотезу об однородности распределения людей в двух группах по группам крови.

Решение:

1) Будем считать, что в процессе исследования каждой группы людей поочередно проводится анализ крови каждого человека, причем результат анализа заранее не известен, и потому представляется некоторой случайной величиной, принимающей значения 1, 2, 3 или 4.

Если у человека из первой группы людей группа крови оказывается равной j ($j = \overline{1,4}$), то будем считать, что произошло событие A_{1j}. Если у человека из группы людей 2 группа крови оказывается равной j ($j = \overline{1,4}$), то будем считать, что произошло событие A_{2j}. Причем, вероятности p_{ij} $(i = \overline{1,2}\ \ j = \overline{1,4})$ событий A_{ij} не известны.

При такой постановке задачи требуется проверить гипотезу, для которой введем обозначение H_0, о том, что вероятности p_{ij} событий A_{ij} при всех j попарно равны:

$$H_0: \ p_{1j} = p_{2j}, \ j = \overline{1,4} \ .$$

2) Совокупность наблюдаемых величин ν представляет собой матрицу:

$$\nu = \begin{pmatrix} \nu_{11} & \nu_{12} & \nu_{13} & \nu_{14} \\ \nu_{21} & \nu_{22} & \nu_{23} & \nu_{24} \end{pmatrix},$$

в которой, каждая случайная величина ν_{ij} равна количеству произошедших событий A_{ij}. Если группа крови каждого отдельного человека не зависит от группы крови остальных человек в этой же группе людей, то совместное распределение случайных величин ν_{1j} является полиномиальным распределением $\Pi(p_{11}, p_{12}, p_{13}, p_{14}; n_1)$ (где $n_1 = \nu_{11} + \nu_{12} + \nu_{13} + \nu_{14}$), а совместное распределение случайных величин ν_{2j} является полиномиальным распределением $\Pi(p_{21}, p_{22}, p_{23}, p_{24}; n_2)$ (где $n_1 = \nu_{21} + \nu_{22} + \nu_{23} + \nu_{24}$). Если группа крови каждого отдельного человека также не зависит и от групп крови человек в другой группе людей, то случайные векторы $(\nu_{11}, \nu_{12}, \nu_{13}, \nu_{14})$ и $(\nu_{21}, \nu_{22}, \nu_{23}, \nu_{24})$ независимы и совместное распределение всех случайных величин ν_{ij} является произведением двух полиномиальных распределений $\prod_{i=1}^{2} \Pi(p_{i1}, p_{i2}, p_{i3}, p_{i4}; n_i)$.

Основная гипотеза об однородности H_0 утверждает, что совместное распределение случайных величин ν_{ij} является произведением $\prod_{i=1}^{2} \Pi(p_1^0, p_2^0, p_3^0, p_4^0; n_i)$ при некотором векторе вероятностей $(p_1^0, p_2^0, p_3^0, p_4^0)$ ($p_1^0 + p_2^0 + p_3^0 + p_4^0 = 1$), или иначе, что существует вектор вероятностей $p^0 = (p_1^0, p_2^0, p_3^0, p_4^0)$ такой, что при каждом фиксированом j неизвестные вероятности p_{ij} равны между собой и равны вероятности p_j^0:

$$H_0 : \; p_{1j} = p_{2j} = p_j^0, \; j = \overline{1,4}.$$

3) Для проверки основной гипотезы может быть использован критерий со статистикой:

$$X_n^2(\nu_{11}, \ldots, \nu_{24} \mid \hat{p}^0) = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(\nu_{ij} - n_i \hat{p}_j^0)^2}{n_i \hat{p}_j^0},$$

где n_1 и n_2 - количества людей в первой и второй группах, а вектор $\hat{p}^0(\nu) = (\hat{p}_1^0(\nu), \hat{p}_2^0(\nu), \hat{p}_3^0(\nu))$ ($\hat{p}_4^0(\nu) = 1 - (\hat{p}_1^0(\nu) + \hat{p}_2^0(\nu) + \hat{p}_3^0(\nu))$) является МП-оценкой вектора параметров $p^0 = (p_1^0, p_2^0, p_3^0)$:

$$\hat{p}_j^0(\nu) = \frac{\nu_{1j} + \nu_{2j}}{n_1 + n_2}.$$

4) В данном случае реализацией матрицы величин ν является матрица y:

$$y = \begin{pmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \end{pmatrix} = \begin{pmatrix} 121 & 120 & 79 & 33 \\ 118 & 95 & 121 & 30 \end{pmatrix}.$$

Количество людей в первой группе $n_1 = 353$, во второй – $n_2 = 364$. В соответствии с реализацией y вычислим значения вероятностей:

$$\hat{p}_1^0(y) = \frac{y_{11} + y_{21}}{n_1 + n_2} = \frac{121 + 118}{353 + 364} \approx 0.3333,$$

$$\hat{p}_2^0(y) = \frac{y_{12} + y_{22}}{n_1 + n_2} = \frac{120 + 95}{353 + 364} \approx 0.2999,$$

$$\hat{p}_3^0(y) = \frac{y_{13} + y_{23}}{n_1 + n_2} = \frac{79 + 121}{353 + 364} \approx 0.2789,$$

$$\hat{p}_4^0(y) = \frac{y_{14} + y_{24}}{n_1 + n_2} = \frac{33 + 30}{353 + 364} \approx 0.0879 \ .$$

С учетом полученных значений вероятностей, значение статистики критерия:

$$X_n^2(y_{11},..., y_{24} \mid \hat{p}^0) = \sum_{i=1}^{2} \sum_{j=1}^{4} \frac{(\nu_{ij} - n_i \hat{p}_j^0)^2}{n_i \hat{p}_j^0} \approx 11.74 \ .$$

Если основная гипотеза H_0 верна, то распределение статистики $X_n^2(y_{11},..., y_{24}; \hat{p}^0)$ при больших n приближенно совпадает с распределением хи-квадрат с числом степеней свободы $(k-1)(m-1)$, где $k = 4$ – количество событий A_{1j} и $m = 2$ – количество групп людей. Таким образом, наименьший уровень значимости отклонения гипотезы H_0 является вероятность $\alpha_{\min}$:

$$\alpha_{\min} = P\left\{X_n^2(y_{11},..., y_{24}; \hat{p}^0) \ge 11.74\right\} \approx P\left\{\chi_3^2 \ge 11.74\right\} \approx 0.0083$$

Полученное значение вероятности является малым, поэтому гипотезу об однородности следует отклонить, считая, что реализация y наблюдаемых величин противоречит гипотезе.

Ответ:

Наименьший уровень значимости отклонения гипотезы об однородности примерно равен 0,0083.

Задача 4.6.

Несимметричная монета имеет вероятность выпадения герба $p_1 = 0.6$. Какое количество бросаний монеты n нужно сделать, чтобы критерий хи-квадрат с уровнем значимости $\alpha = 0.02$ отклонил гипотезу о симметричности монеты с вероятностью не менее чем $P_r = 0.9$?

Решение:

1) Прежде всего, построим критерий хи-квадрат для проверки гипотезы о симметричности монеты с заданным уровнем значимости $\alpha = 0.02$.

Предположим, что в результате n бросаний несимметричной монеты происходит случайное количество ν_1 выпадений герба и случайное количество ν_2 выпадений решки ($\nu_1 + \nu_2 = n$). Таким образом, совокупность наблюдаемых величин ν образована двумя случайными величинами ν_1 и ν_2, $\nu = (\nu_1, \nu_2)$, причем распределение совокупности ν является полиномиальным $\Pi(p_1, p_2; n)$, где $p_1 = 0.6$ и $p_2 = 1 - p_1 = 0.4$.

Гипотеза о симметричности монеты H_0, которую примем в качестве основной, утверждает, что наблюдение ν имеет полиномиальное распределение $\Pi(p_1^0, p_2^0; n)$, где $p_1^0 = p_2^0 = 0.5$.

Критерий хи-квадрат проверки гипотезы H_0 со статистикой:

$$X_n^2(\nu_1, \nu_2 \mid p_1^0, p_2^0) = \sum_{i=1}^{2} \frac{(\nu_i - np_i^0)^2}{np_i^0},$$

отклоняет основную гипотезу H_0, если значение статистики $X_n^2(\nu_1, \nu_2 \mid p_1^0, p_2^0)$ оказывается больше некоторого порогового значения h_α:

$$X_n^2(\nu_1, \nu_2; p_1^0, p_2^0) \ge h_\alpha \ ,$$

где h_α определяется в соответствии с заданным уровнем значимости α как квантиль распределения хи-квадрат с одной степенью свободы уровня $1-\alpha$:

$$h_\alpha = Q_{\chi_1^2}(1-\alpha) \ .$$

Вероятность отклонения основной гипотезы H_0, при условии, что она не верна, по условию задачи должна быть не меньше, чем заданное значение P_r:

$$P\left\{X_n^2(\nu_1,\nu_2 \mid p_1^0, p_2^0) \geq Q_{\chi_1^2}(1-\alpha) \mid \Pi(p_1, p_2; n)\right\} \geq P_r ,$$

откуда следует найти число n.

2) Известно, что при больших n распределение статистики $X_n^2(\nu_1,\nu_2 \mid p_1^0, p_2^0)$, при условии, что основная гипотеза H_0 не верна, приближенно совпадает с нецентральным распределением хи-квадрат с одной степенью свободы и параметром нецентральности $a(n) = n\sum_{i=1}^{2} \frac{(p_i - p_i^0)^2}{p_i^0}$:

$$X_n^2(\nu_1,\nu_2; p_1^0, p_2^0) \sim \chi^2(1, a(n)).$$

Откуда вероятность отклонения основной гипотезы H_0, при условии, что она не верна, может быть приближенно вычислена с помощью распределения $\chi^2(1, a(n))$:

$$P\left\{X_n^2(\nu_1,\nu_2; p_1^0, p_2^0) \geq Q_{\chi_1^2}(1-\alpha) \mid \Pi(p_1, p_2; n)\right\} \approx P\left\{\chi_1^2(a(n)) \geq Q_{\chi_1^2}(1-\alpha)\right\},$$

где $\chi_1^2(a(n))$ - случайная величина имеющая распределение $\chi^2(1, a(n))$. По определению нецентрального распределения хи-квадрат, случайная величина $\chi_1^2(a(n))$ может быть представлена с помощью случайной величины ξ имеющей стандартное нормальное распределение и смещения $a(n)$:

$$\chi^2(1, a(n)) = (\xi + \sqrt{a(n)})^2 ,$$

$$\xi \sim N(0,1).$$

Таким образом, вероятность отклонения основной гипотезы H_0, при условии, что она не верна, приближенно представима в виде:

$$P\left\{X_n^2(\nu_1,\nu_2; p_1^0, p_2^0) \geq Q_{\chi_1^2}(1-\alpha) \mid \Pi(p_1, p_2; n)\right\} \approx P\left\{(\xi + \sqrt{a(n)})^2 \geq Q_{\chi_1^2}(1-\alpha)\right\},$$

где последнюю вероятность следует сделать не меньшей, чем P_r:

$$P\left\{(\xi + \sqrt{a(n)})^2 \geq Q_{\chi_1^2}(1-\alpha)\right\} \geq P_r .$$

3) Заметим, что неравенство $(\xi + \sqrt{a(n)})^2 \geq Q_{\chi_1^2}(1-\alpha)$ выполняется тогда и только тогда, когда $\xi + \sqrt{a(n)} \leq -\sqrt{Q_{\chi_1^2}(1-\alpha)}$ либо $\xi + \sqrt{a(n)} \geq \sqrt{Q_{\chi_1^2}(1-\alpha)}$, причем неравенства представляют собой несовместные события, откуда следует равенство для вероятностей:

$$P\left\{(\xi + a(n))^2 \geq Q_{\chi_1^2}(1-\alpha)\right\} = P\left\{\xi + \sqrt{a(n)} \leq -\sqrt{Q_{\chi_1^2}(1-\alpha)}\right\} + P\left\{\xi + \sqrt{a(n)} \geq \sqrt{Q_{\chi_1^2}(1-\alpha)}\right\} =$$

$$= P\left\{\xi \leq -\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} + P\left\{\xi \geq \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\},$$

и неравенство для вероятности отклонения следовательно приобретает вид:

$$P\left\{\xi \leq -\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} + P\left\{\xi \geq \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \geq P_r .$$

Легко видеть, что при увеличении n параметр нецентральности $a(n)$ возрастает:

$$a(n) = n\sum_{i=1}^{2} \frac{(p_i - p_i^0)^2}{p_i^0} \xrightarrow[n\to\infty]{} \infty ,$$

откуда следует, что значение выражения $-\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}$ убывает и вероятность $P\left\{\xi \leq -\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\}$ стремиться к нулю:

$$-\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)} \xrightarrow[n\to\infty]{} \infty$$

$$P\left\{\xi \le -\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \xrightarrow[n\to\infty]{} \infty \ ,$$

поэтому первой вероятностью в неравенстве можно пренебречь:

$$P\left\{\xi \le -\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} + P\left\{\xi \ge \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \ge P\left\{\xi \ge \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \ge P_r \ ,$$

и рассматривать неравенство:

$$P\left\{\xi \ge \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \ge P_r \ ,$$

$$1 - P\left\{\xi < \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \ge P_r \ ,$$

$$P\left\{\xi < \sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right\} \le 1 - P_r \ .$$

Случайная величина ξ имеет стандартное нормальное распределение $N(0,1)$, поэтому вероятность слева вычисляется с помощью функции распределения стандартного нормального распределения:

$$\Phi\left(\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)}\right) \le 1 - P_r \ ,$$

откуда,

$$\sqrt{Q_{\chi_1^2}(1-\alpha)} - \sqrt{a(n)} \le \Phi^{-1}(1 - P_r) \ ,$$

$$\sqrt{a(n)} \ge \sqrt{Q_{\chi_1^2}(1-\alpha)} - \Phi^{-1}(1 - P_r) \ ,$$

$$\sqrt{n\sum_{i=1}^{2}\frac{(p_i - p_i^0)^2}{p_i^0}} \ge \sqrt{Q_{\chi_1^2}(1-\alpha)} - \Phi^{-1}(1 - P_r) .$$

Подставляя в последнее неравенство числовые значения, получим неравенство:

$$\sqrt{n\left(\frac{(0.6-0.5)^2}{0.5} + \frac{(0.4-0.5)^2}{0.5}\right)} \ge \sqrt{Q_{\chi_1^2}(1-0.02)} - \Phi^{-1}(1 - 0.9) ,$$

$$\sqrt{n\frac{0.02}{0.5}} \ge \sqrt{5.3824} - (-1.28) ,$$

$$\sqrt{n\frac{1}{25}} \ge 2.32 + 1.28 \ ,$$

$$\sqrt{n}\frac{1}{5} \ge 3.6 \ ,$$

$$n \ge (5 \cdot 3.6)^2 \ ,$$

$$n \ge 324 \ .$$

Ответ:

Нужно сделать не менее 324 бросаний.