2.30. Теорема Пирсона, проверка гипотезы о вероятностях в обобщенной схеме Бернулли

1. Рассмотрим последовательность n независимых испытаний с числом k ($k \geq 2$) исходов в каждом испытании (обобщенная схема Бернулли).

Теорема К. Пирсона. Пусть n – число независимых испытаний, результатом каждого испытания является один из k исходов $A_1, \ldots, A_k$ ($k \geq 2$). Вероятности исходов $A_1, \ldots, A_k$ равны, соответственно, $p_1, \ldots, p_k$ и не зависят от номера испытания; все $p_i \neq 0$ и $p_1 + \ldots + p_k = 1$.

Пусть в результате проведения n испытаний исход A_1 наблюдался N_1 раз; … A_i – N_i раз; … ; A_k – N_k раз, при этом $N_1 + \ldots + N_k = n$.

Заметим, что N_i ($i = 1, \ldots, k$) – случайные величины, подчиняющиеся биномиальному распределению с параметрами n и p_i, при этом $M(N_i) = np_i$, $D(N_i) = np_i(1 - p_i)$.

Примем без доказательства утверждение (*теорема Пирсона*):

случайная величина (хи–квадрат) $\mathbb{X}^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$ при $n \to +\infty$

распределена как χ^2_{k-1} (хи-квадрат с $k-1$ степенью свободы):

$$\forall x \;\; P(\mathbb{X}^2 < x) \underset{n \to +\infty}{\to} P(\chi^2_{k-1} < x).$$

Величины N_i называют *наблюдаемыми* частотами, а np_i – *ожидаемыми* частотами.

2. Пусть относительно вероятностей $p_1, \ldots, p_k$ выдвинута *простая* гипотеза H_0: $p_1 = p_1^0, \ldots, p_k = p_k^0$ (альтернативная гипотеза H_1: $\exists i \; (i = 1, \ldots, k) \;\; p_i \neq p_i^0$) и задан уровень значимости α.

В качестве статистики критерия для проверки гипотезы H_0 возьмем $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} = n\left(\sum_{i=1}^{k} \frac{(\frac{N_i}{n} - p_i^0)^2}{p_i^0}\right)$.

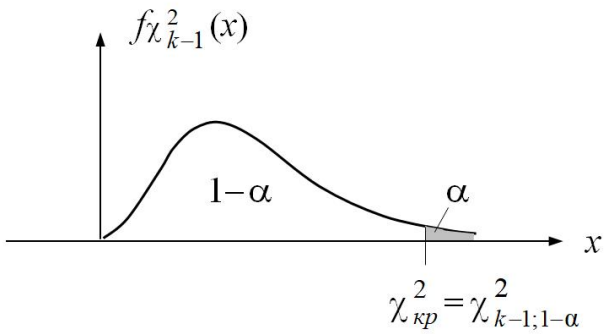
Если гипотеза H_0 верна, то согласно теореме Пирсона имеем:

$$\chi^2 \underset{n \to +\infty}{\sim} \chi^2_{k-1}.$$

Если H_0 – неверна, то хотя бы одна из относительных частот $\frac{N_i}{n}$ сходится по вероятности к величине p_i, отличной от p_i^0:

$\frac{N_i}{n} \underset{n \to +\infty}{\overset{P}{\to}} p_i \neq p_i^0$, поэтому $\chi^2 = n\left(\sum_{i=1}^{k} \frac{(\frac{N_i}{n} - p_i^0)^2}{p_i^0}\right) \underset{n \to +\infty}{\to} \infty$.

Отсюда следует, что гипотеза H_0 должна быть отвергнута, если полученное в опыте значение χ^2 велико:



Таким образом, приходим к правостороннему критерию:

при $\chi^2 \geq \chi^2_{кр}$ – гипотезу H_0 отклоняют;

при $\chi^2 < \chi^2_{кр}$ – гипотезу H_0 принимают.

Пусть в данном эксперименте частоты N_i (случайные величины) приняли конкретные значения n_i, соответственно. Вычисляют

$\chi^2_e = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}$ и решение об отклонении или принятии гипотезы H_0 принимают, сопоставляя значение χ^2_e с критическим числом $\chi^2_{кр} = \chi^2_{k-1;\,1-\alpha}$.

Не следует считать, что при справедливости гипотезы H_0 величина $\mathbb{X}^2$ должна быть близкой к нулю, поскольку результаты наблюдений (измерений) – это реализация *случайной* выборки.

Также необходимо учесть, что применение непрерывного распределения χ^2_{k-1} в качестве аппроксимации распределения дискретной случайной величины $\mathbb{X}^2$ порождает ряд ограничений.

В частности требуется, чтобы n было достаточно велико ($n \geq 50$), не должны быть малыми *ожидаемые* частоты $np_i{}^0$, а также значения *наблюдаемых* частот n_i. Более детально практические рекомендации будут обсуждены далее.

Замечание: статистика Пирсона $\mathbb{X}^2$ от квадрата расстояния между точками в n-мерном пространстве отличается множителями $1/(np_i{}^0)$ слагаемых (весами слагаемых).

2.31. Проверка гипотезы о виде распределения – метод χ^2 для простой гипотезы

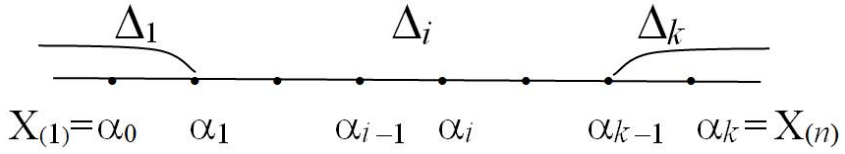
Пусть $X_1, \ldots, X_n$ – выборка из распределения $F_X(x)$ *непрерывной случайной величины* X, объем выборки n достаточно велик и $F_{X_i}(x) = F_X(x)$ для всех $i = 1,..., n$.

Пусть проверяется гипотеза *о виде распределения* H_0: $F_X(x) = F_0(x)$, альтернативная гипотеза H_1: $F_X(x) \neq F_0(x)$ и задан уровень значимости α. Заметим, что гипотеза H_0 – простая, альтернативная гипотеза H_1 – сложная.

Статистика критерия

По вариационному ряду $X_{(1)}, \ldots, X_{(n)}$ построим k промежутков аналогично тому, как это делалось при построении гистограммы, с тем отличием, что в качестве крайних промежутков возьмем полубесконечные: $\Delta_1 = (-\infty; \alpha_1]$,…, $\Delta_i = (\alpha_{i-1}; \alpha_i]$,…, $\Delta_k = (\alpha_{k-1}; +\infty)$.

Число интервалов разбиения k обычно берут таким же, как при построении гистограммы, а именно, применяют либо формулу Старджесса: $k = 1 + 3{,}32 \lg n$, либо формулу: $k = 1{,}72\, n^{1/3}$, а сами промежутки полагают равными (за исключением крайних – полубесконечных):



Число элементов выборки, попавших в i-й промежуток разбиения $\Delta_i = (\alpha_{i-1}; \alpha_i]$ – наблюдаемые частоты N_i (случайные величины).

Обозначим через p_i вероятность $P(X \in \Delta_i)$ для случайной величины X принять значение в промежутке Δ_i.

При справедливости гипотезы H_0 имеем:

$$P(X \in \Delta_1 \,|H_0) = F_0(\alpha_1) = p_1^0;$$

$$P(X \in \Delta_i \,|H_0) = F_0(\alpha_i) - F_0(\alpha_{i-1}) = p_i^0 \ (i=2,...,k-1);$$

$$P(X \in \Delta_k \,|H_0) = 1 - F_0(\alpha_{k-1}) = p_k^0.$$

Проверяемая гипотеза о распределении H_0: $F_X(x) = F_0(x)$, равносильна гипотезе о том, что упомянутые вероятности p_i приняли определенные значения p_i^0. Таким образом, приходим к задаче о проверке *простой гипотезы о вероятностях в обобщенной схеме Бернулли*, рассмотренной в п. 2.30.: H_0: $p_i = p_i^0$ $(i = 1,\ldots,k)$.

Правило принятия решения об отклонении (принятии) проверяемой гипотезы H_0 о виде распределения строится на основе приближения распределения статистики критерия Пирсона $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$ распределением χ^2_{k-1} при больших объемах выборки $(n \geq 50)$.

Практически по реализации выборки $(x_1, x_2, \ldots, x_n)$ получают реализацию вариационного ряда. Отрезок $[x_{(1)}; x_{(n)}]$, где $x_{(1)} = min\ (x_1, x_2, \ldots, x_n) = \alpha_{(0)}$, $x_{(n)} = max\ (x_1, x_2, \ldots, x_n) = \alpha_{(k)}$, содержащий все элементы выборки, разбивают на k равных интервалов. Подсчитывают $\chi^2_e = \sum_{i=1}^{k} \frac{(n_i - np_{ie}^0)^2}{np_{ie}^0}$, где n_i – *наблюдаемые* частоты (значения случайных величин N_i) – число элементов реализации выборки $(x_1, x_2, \ldots, x_n)$, *фактически* попавших в соответствующий интервал Δ_i и вычисляют *ожидаемые* частоты np_{ie}^0.

Решение об отклонении или принятии гипотезы H_0 принимают, сопоставляя значение χ^2_e, соответствующее данной реализации выборки, с критическим числом $\chi^2_{кр} = \chi^2_{k-1;\ 1-\alpha}$.

2.32. Проверка гипотезы о виде распределения – метод χ^2 для сложной гипотезы

Пусть $X_1, \ldots, X_n$ – выборка из распределения непрерывной случайной величины X, функция распределения которой зависит от r неизвестных параметров $F_X(x, \theta_1, \theta_2 \ldots \theta_r)$. В этом случае гипотеза о виде распределения H_0: $F_X(x, \theta_1, \ldots \theta_r) = F_0(x, \theta_1, \ldots \theta_r)$ – сложная.

В функции распределения $F_0(x, \theta_1, \ldots \theta_r)$ неизвестные параметры заменим оценками максимального правдоподобия $\hat{\theta}_{1\text{МП}}, \ldots, \hat{\theta}_{r\,\text{МП}}$. Действуя аналогично процедуре проверки простой гипотезы, рассмотренной в п. 2.31., вычислим вероятности $\hat{p}_i^0 = P(X \in \Delta_i \mid H_0)$:

Δ_1: $\hat{p}_1^0 = F_0(\alpha_1; \hat{\theta}_{1\text{МП}}, \ldots, \hat{\theta}_{r\,\text{МП}})$,

Δ_i: $\hat{p}_i^0 = F_0(\alpha_i, \hat{\theta}_{1\text{МП}}, \ldots, \hat{\theta}_{r\,\text{МП}}) - F_0(\alpha_{i-1}, \hat{\theta}_{1\text{МП}}, \ldots, \hat{\theta}_{r\,\text{МП}})$, $(i = 2, \ldots, k-1)$,

Δ_k: $\hat{p}_k^0 = 1 - F_0(\alpha_{k-1}; \hat{\theta}_{1\text{МП}}, \ldots, \hat{\theta}_{r\,\text{МП}})$.

Гипотезы H_0 и H_1 при этом формулируются следующим образом:

$$H_0: p_i = \hat{p}_i^0 \ (i = 1, \ldots, k); \quad H_1: \exists i \ (i = 1, \ldots, k) \ p_i \neq \hat{p}_i^0 .$$

Доказано (*теорема Фишера*) что при справедливости гипотезы H_0 распределение статистики $\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0}$ при $n \to +\infty$ стремится к распределению случайной величины χ^2_{k-r-1} (с $k-r-1$ степеней свободы): $\forall x \ P\left(\sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i^0)^2}{n\hat{p}_i^0} < x\right) \underset{n \to +\infty}{\to} P(\chi^2_{k-r-1} < x)$,

где r – число параметров, оцениваемых по выборке.

В остальном проверка гипотезы H_0 совпадает с процедурой проверки для случая простой гипотезы, рассмотренной в п. 2.31.:

по реализации выборки $(x_1, x_2, \ldots, x_n)$ вычисляют значение

$$\chi_e^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_{ie}^0)^2}{n\hat{p}_{ie}^0},$$

где n_i – значения *наблюдаемых* частот, фактически полученные в эксперименте; $n\hat{p}_{ie}^0$ – вычисленные *ожидаемые* частоты.

Сопоставляя значение χ_e^2 с $\chi_{кр}^2 = \chi_{k-r-1;1-\alpha}^2$, принимают решения: на уровне значимости α гипотезу H_0 отвергнуть, если $\chi_e^2 \geq \chi_{кр}^2$ или гипотезу H_0 принять, если $\chi_e^2 < \chi_{кр}^2$.

Замечания

Применимость аппроксимации дискретного распределения статистики $\mathbb{X}^2$ непрерывным распределением χ_{k-1}^2 (в случае простой гипотезы) и χ_{k-r-1}^2 (в случае сложной гипотезы), накладывает определенные ограничения на построение упомянутого разбиения. Промежутки разбиения Δ_i следует строить так, чтобы выполнялось условие: "ожидаемое" $n\hat{p}_{ie}^0 \geq 5$. При этом *длины промежутков не должны быть обязательно равными*. Если для какого-либо промежутка $n\hat{p}_{ie}^0 < 5$, или наблюдаемая частота $n_i < 5$, то такой промежуток объединяют с соседним.

Число промежутков k таким образом может сократиться по сравнению с первоначальным и решение об отклонении (принятии) гипотезы H_0 принимают, сравнивая χ_e^2 с $\chi_{кр}^2 = \chi_{k^*-r-1;\,1-\alpha}^2$, где k^* – окончательное число промежутков после объединения.

В то же время, условие "минимальное ожидаемое $n\hat{p}_{ie}^0 \geq 5$" может оказаться слишком жестким – допустимый минимум зависит от числа степеней свободы k.

В специальной литературе по прикладной статистике приводятся также следующие рекомендации. Общее количество промежутков разбиения k должно быть не меньше 8. В каждый интервал разбиения должно попасть не менее 7–10 элементов реализации выборки, причем желательно, чтобы в разные интервалы попало примерно одинаковое число точек.

2.33. Проверка гипотезы о нормальном распределении.

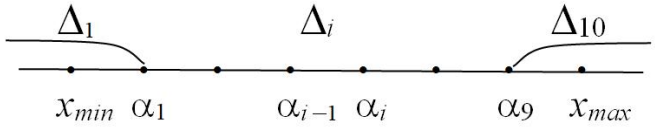
Пример 1_кс

Пусть имеется реализация выборки $x_1, x_2, \ldots, x_{500}$ из некоторого распределения, объем выборки $n = 500$. Вычислены выборочное среднее и исправленное стандартное отклонение, равные соответственно $\bar{x} = 0{,}044289$ и $s^* = 1{,}015813$, определены минимальный и максимальный элементы выборки $x_{(1)} = x_{min} = -2{,}769448$, $x_{(500)} = x_{max} = 2{,}835368$.

На уровне значимости $\alpha = 0{,}05$ необходимо проверить гипотезу H_0 о нормальности распределения, для которого экспериментально получена данная реализация выборки.

Решение

Построим группированный статистический ряд. Число интервалов разбиения примем равным $k = 10$, крайний левый и крайний правый интервалы равными, соответственно, $\Delta_1 = (-\infty;\ \alpha_1]$, $\Delta_{10} = (\alpha_9;+\infty)$.



Неизвестные параметры нормального распределения m и σ заменим значениями их оценок, вычисленными по данной реализации выборки, а именно $m = \hat{m}_e = \bar{x}$, $\sigma = \hat{\sigma}_e = s^*$. Вычислим вероятности $\hat{p}_i^0 = P(X \in \Delta_i \mid H_0)$ для соответствующих интервалов разбиения при справедливости гипотезы H_0: $X \sim N(\bar{x}; s^*)$:

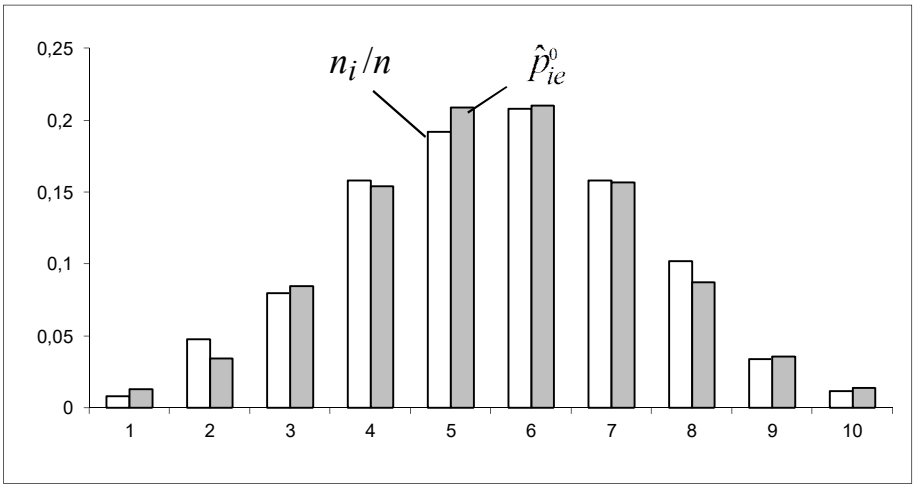
Δ_1 : $\hat{p}_{1e}^0 = \Phi((\alpha_1 - \bar{x})/s^*)$;

Δ_i : $\hat{p}_{ie}^0 = \Phi((\alpha_i - \bar{x})/s^*) - \Phi((\alpha_{i-1} - \bar{x})/s^*)$, $i = 2,\ldots,9$;

Δ_{10}: $\hat{p}_{10e}^0 = 1 - \Phi((\alpha_9 - \bar{x})/s^*)$ ($\Phi(x)$ – функция Лапласа).

Группированные числовые данные и результаты расчетов приведены в таблице и представлены на графике ниже.

i	Интервал $\Delta_i=(\alpha_{i-1};\alpha_i]$		*"Наблюдаемые" частоты* n_i	*"Ожидаемые" значения* $n\hat{p}_{ie}^0$	$\frac{(n_i - n\hat{p}_{ie}^0)^2}{n\hat{p}_{ie}^0}$
1	$-\infty$	-2,209	4	6,635	1,047
2	-2,209	-1,648	24	17,272	2,621
3	-1,648	-1,088	40	42,341	0,129
4	-1,088	-0,527	79	77,126	0,046
5	-0,528	0,033	96	104,401	0,676
6	0,0330	0,593	104	105,029	0,010
7	0,593	1,154	79	78,526	0,002
8	1,154	1,714	51	43,631	1,245
9	1,714	2,275	17	18,013	0,057
10	2,275	$+\infty$	6	7,025	0,149
			500	500	$\chi_e^2 = 5{,}983$



На рисунке выше представлены экспериментальная и теоретическая гистограммы: белые прямоугольники соответствуют наблюдаемым относительным частотам n_i/n, серые – вероятностям $\hat{p}_{ie}^0$ попадания значений выборки в соответствующий интервал.

Критическое число $\chi^2_{кр} = \chi^2_{k-r-1;\,1-\alpha}$ – квантиль порядка $1-\alpha = 0{,}95$ распределения хи-квадрат с числом степеней свободы $k-r-1 = 10-2-1 = 7$: $\chi^2_{k-r-1;\,1-\alpha} = \chi^2_{7;\,0{,}95} = 14{,}067$.

Имеем: $$\chi^2_e = \sum_{i=1}^{10} \frac{(n_i - n\hat{p}^0_{ie})^2}{n\hat{p}^0_{ie}} = 5{,}983 < \chi^2_{7;\,0{,}95} = 14{,}067.$$

Таким образом, гипотеза H_0 о нормальности распределения, из которого получена реализация выборки $x_1, x_2, \ldots, x_{500}$ на уровне значимости $\alpha = 0{,}05$ не противоречит экспериментальным данным и может быть принята. Заметим, что по существу проверялась *гипотеза* H_0: $X \sim N(\bar{x}; s^*)$ *о нормальности распределения с параметрами* $m = \hat{m}_e = \bar{x}$, $\sigma = \hat{\sigma}_e = s^*$.

Замечания

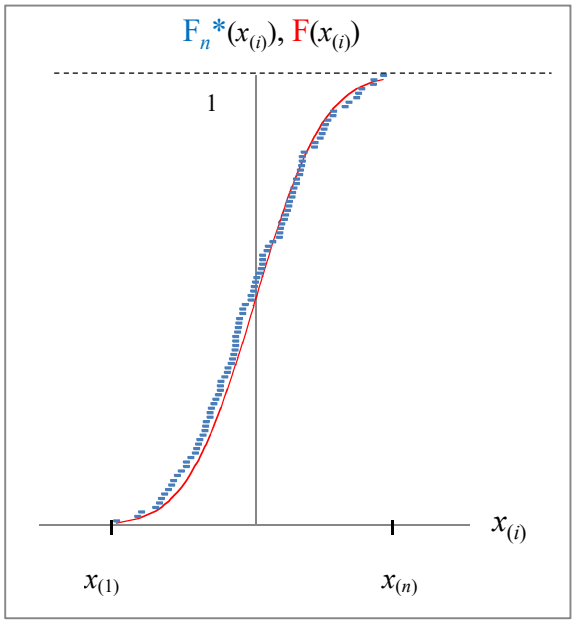
1. О глазомерном методе проверки на нормальность: график КК (Q - Q plot)

Эмпирическая функция распределения $F_n(x)$ при больших n (n – объем выборки) близка к "теоретической" функции распределения случайной величины $F_X(x)$ (теорема Гливенко).

$$\forall \varepsilon > 0 \;\; P\left(\sup_x |F_n(x) - F_X(x)| \le \varepsilon\right) \underset{n \to +\infty}{\to} 1.$$

Пусть $x_{(1)}, \ldots, x_{(n)}$ – данная реализация вариационного ряда, причем $x_{(1)} < x_{(2)}, \ldots, < x_{(n)}$. Реализация эмпирической функции распределения $F_n^*(x)$ – кусочно-постоянная функция, которая в каждой из точек $x_{(i)}$ претерпевает скачок, равный $1/n$: $F_n^*(x_{(i)}) = i \times \frac{1}{n}$;

при $x < x_{(1)}$ $F_n^*(x) = 0$,
при $x > x_{(n)}$ $F_n^*(x) = 1$.



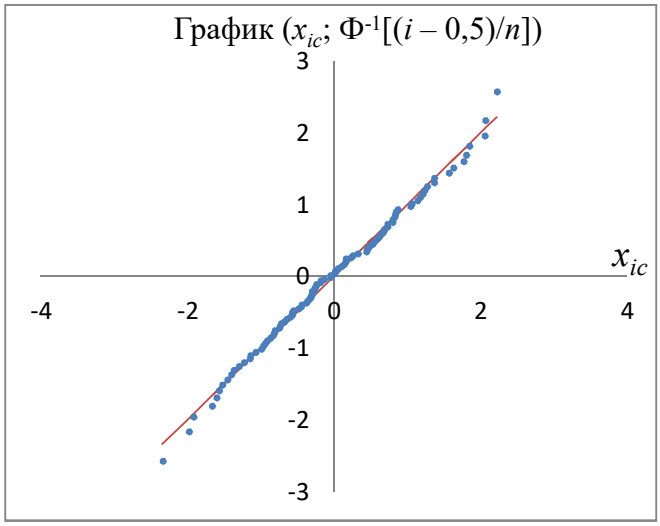
Центрируем и нормируем данные: $x_{ic} = \dfrac{x_{(i)} - \bar{x}}{s^*}$, где $\bar{x}$ – выборочное среднее, s^*– выборочное стандартное отклонение (исправленное).

Подсчитаем значения функции Φ^{-1} в серединах скачков функции $F_n^*(x_{(i)})$, то есть значения $\Phi^{-1}[(i - 0{,}5)/n]$. Применение функции Φ^{-1} к серединам скачков функции $F_n^*(x)$ обусловлено тем, что применить Φ^{-1} к верхним или нижним концам ее скачков нельзя, так как $\Phi^{-1}(0) = -\infty$, а $\Phi^{-1}(1) = +\infty$.

На плоскость $x0y$ нанесем точки с координатами $(x_{ic}, \Phi^{-1}[(i - 0{,}5)/n])$.

В результате получим график "квантиль-квантиль": абсцисса i-й точки – значение $x_{ic} = \dfrac{x_{(i)} - \bar{x}}{s^*}$, (нормированное и центрированное), ордината – соответствующая квантиль стандартного нормального распределения).

Глазомерный метод проверки нормальности состоит в визуальной оценке величины отклонения точек графика от прямой линии: чем лучше ложатся точки на прямую, тем меньше оснований сомневаться в нормальности распределения, из которого получена реализация выборки.



2. Критерий Колмогорова

Критерий предназначен для проверки *простой* гипотезы о виде распределения непрерывной случайной величины X, H_0: $F_X(x) = F_0(x)$, ($F_0(x)$ – непрерывное распределение не содержит неизвестных параметров).

А.Н. Колмогоровым предложена статистика критерия:

$$D_n = \sup_x |F_n(x) - F_0(x)|,$$

где $F_n(x)$ – эмпирическая функция распределения (функция выборки из распределения $F_X(x)$).

Теорема Колмогорова

При справедливости гипотезы H_0 имеет место предельное равенство: $\lim\limits_{n\to\infty} P(\sqrt{n}\, D_n < x) = K(x)$,

где $K(x)$ – функция Колмогорова (табулирована).

Если гипотеза H_0 верна, то $F_n(x)$ при больших n близка к $F_0(x)$ и D_n мало, кроме того, статистика $\sqrt{n}\, D_n$ при больших n распределена приближенно как $K(x)$;
если же гипотеза H_0 неверна, то отклонение $|F_n(x) - F_0(x)|$ при некоторых x конечно и произведение $\sqrt{n}\, D_n$ становится большим при больших n. Таким образом, критерий в этом случае – правосторонний.

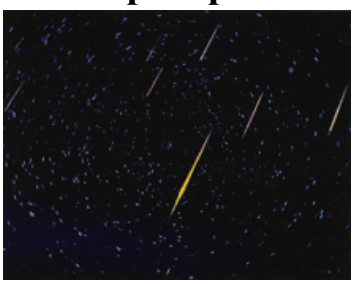
Обозначим квантиль распределения $K(x)$ порядка α через $k_{1-\alpha}$, тогда, если $D_n \geq k_{1-\alpha}$, то гипотеза H_0 отвергается,
в противном случае – H_0 принимается. Описанный критерий относится к случаю *простой* гипотезы H_0, когда распределение $F_0(x)$ полностью задано.

На практике обычно известен вид распределения, но параметры его неизвестны. В этом случае проверяется *сложная* гипотеза о принадлежности выборки нормальному закону, параметры которого оцениваются по этой же выборке. В качестве оценок параметров берут выборочное среднее и выборочное стандартное отклонение и пользуются модификациями критерия Колмогорова для сложных гипотез. Пример такой модификации – критерий Лиллиефорса (Lilliefors); статистика критерия:

$$D_n^* = D_n(\sqrt{n} - 0{,}01 + 0{,}85/\sqrt{n}).$$

Для статистики D_n^* рассчитаны критические числа для разных уровней значимости.

2.34. Проверка гипотезы о распределении Пуассона. Пример 2_кс



Для статистического анализа процесса возникновения метеорных следов в определенной области атмосферы полное время наблюдения разбили на 2550 равных промежутков длительностью $\Delta\tau$, в каждом из которых регистрировалось число обнаруженных метеорных следов.

Результаты приведены в таблице, где наблюдаемые частоты n_i – число промежутков из 2550, в которых было зарегистрировано соответствующее число i $(i=0,1,\ldots,13)$ метеорных следов:

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13
n_i	44	189	375	475	491	430	265	143	83	31	15	7	1	1

На уровне значимости $\alpha=0{,}05$ необходимо проверить гипотезу о том, что случайная величина X – число зарегистрированных метеорных следов за временной промежуток длительностью $\Delta\tau$, подчиняется распределению Пуассона:

$$\mathrm{P}(\mathrm{X}=i)=p_i=\frac{a^i}{i!}e^{-a}\quad(i=0,1,2,\ldots).$$

Решение

Вычислим значение $\hat{a}_e$ оценки неизвестного параметра a распределения Пуассона: $\bar{x}=\frac{1}{2550}\sum_{i=1}^{13} in_i=10264/2550=4{,}025098=\hat{a}_e$.

Поскольку значение n_i в последних двух столбцах исходной таблицы меньше 5, объединим три последних столбца, получим таблицу для расчетов:

i	0	1	2	3	4	5	6	7	8	9	10	11
n_i	44	189	375	475	491	430	265	143	83	31	15	9

Заменим в гипотетическом распределении Пуассона $p_i = \frac{a^i}{i!} e^{-a}$ неизвестный параметр a значением его оценки $\hat{a}_e$, вычисленным по экспериментальной выборке $\hat{a}_e = 4{,}025098$. Таким образом, проверке подлежит гипотеза H_0: $p_i = \hat{p}^0_{ie} = \frac{\hat{a}_e^i}{i!} e^{-\hat{a}_e} \quad i = 0,1,\ldots,11$ (против альтернативы H_1: $\exists i\,(i = 1,\ldots,k) \quad p_i \neq \hat{p}^0_{ie}$), на уровне значимости $\alpha = 0{,}05$. Заметим, что проверяемая гипотеза H_0 – сложная, так как распределение содержит неизвестный параметр a, значение которого заменено значением его оценки $\hat{a}_e = \bar{x}$.

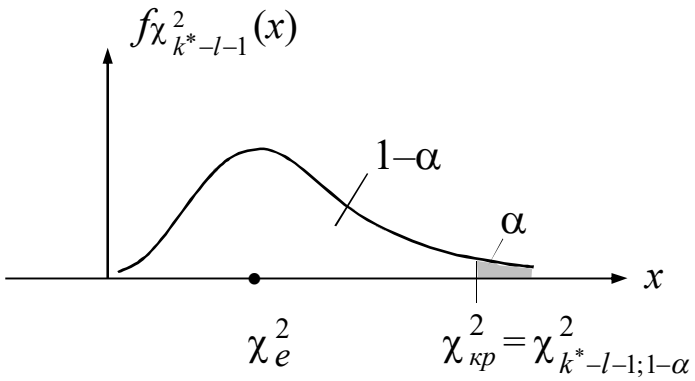
i	n_i	$\hat{p}^0_{ie}$	$n\hat{p}^0_{ie}$	$\frac{(n_i - n\hat{p}^0_{ie})^2}{n\hat{p}^0_{ie}}$
0	44	0,017883	45,54727	0,052561
1	189	0,071959	183,3322	0,175222
2	375	0,144778	368,9651	0,09871
3	475	0,194192	495,0402	0,811265
4	491	0,195353	498,1463	0,10252
5	430	0,157217	401,0176	2,094627
6	265	0,105438	269,0225	0,060145
7	143	0,060611	154,6917	0,883667
8	83	0,030487	77,83116	0,343268
9	31	0,013631	34,80867	0,416735
10	15	0,005485	14,01083	0,069836
≥11	9	0,002969	7,586556	0,263337
	2550	1,0	2550	χ^2_e =5,371893

В таблице выше приведен расчет величины $\chi^2_e = \sum_{i=0}^{11} \frac{(n_i - n\hat{p}^0_{ie})^2}{n\hat{p}^0_{ie}} = 5{,}37$

Заметим, что при расчете “теоретических” вероятностей в случае *целочисленной* случайной величины для крайних значений выборки следует поступать так же, как это делалось для непрерывного распределения, когда крайние интервалы считались полубесконечными. В данном примере вероятность в строке таблицы, обозначенной ≥ 11, равна сумме вероятностей всех значений $i \geq 11$ случайной величины, подчиняющейся распределению Пуассона: $\hat{p}^0_{\geq 11} = 1 - \sum_{i=0}^{10} \hat{p}^0_{ie}$.

Для случая сложной гипотезы статистикой критерия служит χ^2_{k-l-1}, где l – число параметров, оцениваемых по выборке, $k-$ общее число *различных* значений случайной величины X, зарегистрированных в данном эксперименте (аналог числа интервалов разбиения, используемого в методе χ-квадрат при проверке гипотезы о непрерывном распределении). В рассматриваемом примере $k = 14$.

Учтем, что три последних столбца исходной таблицы были объединены в один, поэтому $k^* = k - 2 = 12$, а также то, что неизвестное значение параметра распределения a было заменено значением его оценки $\hat{a}_e$, поэтому число степеней свободы для χ-квадрат равно окончательно $k^* - l - 1 = 12 - 1 - 1 = 10$.



Квантиль порядка 0,95 распределения χ^2 с числом степеней свободы равным 10, равна: $\chi^2_{кр} = \chi^2_{10;0,95} = 18{,}31$.

Поскольку $\chi^2_e = 5{,}37 < 18{,}31 = \chi^2_{кр}$, – нет оснований для отклонения гипотезы H_0. Таким образом, гипотеза H_0 о том, что случайная величина – число метеорных следов, возникающих в выделенной области атмосферы за промежуток времени $\Delta\tau$, подчиняется распределению Пуассона с параметром, равным $\hat{a}_e = 4{,}025098$, не противоречит результатам наблюдений и может быть принята на уровне значимости $\alpha = 0{,}05$.

Пример: *Задача о бомбардировках Лондона*

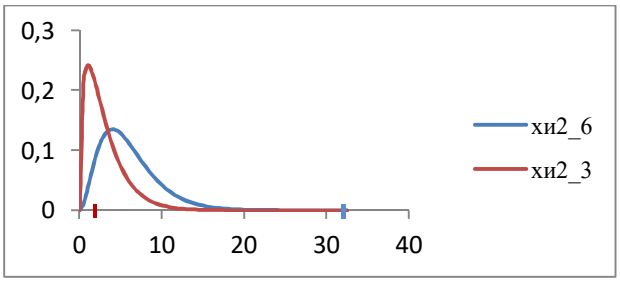
Задача возникла в связи с бомбардировками Лондона во время Второй мировой войны. Для улучшения организации оборонительных мероприятий необходимо было понять цель противника. Для этого карту города разделили сеткой из 24 горизонтальных и 24 вертикальных линий на 576 равных участков. В течение некоторого времени в центре организации обороны города собиралась информация о количестве попаданий снарядов в каждый из участков и были получены следующие данные:

Число попаданий X=i	0	1	2	3	4	5	6	7
Количество участков n_i	229	211	93	35	7	0	0	1

Распределение Пуассона моделирует случайную величину – число однородных событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга. Поэтому гипотеза H_0: стрельба случайна (нет "целевых" участков) равносильна гипотезе: случайная величина X – число зарегистрированных попаданий подчиняется распределению Пуассона.

i	n_i	$i*n_i$	p_i	$n*p_i$	chi2	n_i	p_i	$n*p_i$	chi2
0	229	0	0,393651	226,7427	0,022472	229	0,393651	226,7427	0,022472
1	211	211	0,366997	211,3904	0,000721	211	0,366997	211,3904	0,000721
2	93	186	0,171074	98,53873	0,311325	93	0,171074	98,53873	0,311325
3	35	105	0,053164	30,62228	0,625833	35	0,053164	30,62228	0,625833
4	7	28	0,012391	7,137224	0,002638	8	0,015114	8,705916	0,104296
5	0	0	0,00231	1,330795	1,330795		1	576	1,064646
6	0	0	0,000359	0,206781	0,206781		степ_своб	5-1-1=3	χ^2_e_3
7	1	7	5,4E-05	0,031116	30,16911			p-value=	0,785615
		537	1	576	32,66967				
			степ_своб	8-1-1=6	χ^2_e_6				
$\hat{a}_e = \bar{x} =$.	537/576 =	0,9323		p-value=	1,21E-05				

Расчет по исходной таблице приводит к ошибочному отклонению гипотезы H_0: за счет вклада в χ^2_e значения 30,16911, соответствующего i =7, достигаемый уровень значимости равен *p-value* = 1,21E–05. После объединения малых значений числа попаданий для i =4 –7, гипотеза о распределении Пуассона обоснованно принимается при достигаемом уровне значимости *p-value* = 0,785615.



Замечание

Другой подход к проверке гипотезы H_0 о распределении Пуассона основан на известном свойстве этого распределения: $MX = DX = a$. Если $X_1, \ldots, X_n$ – выборка из распределения Пуассона, то

$S^{*2} \xrightarrow[n\to+\infty]{P} DX = a$ и $\bar{X} \xrightarrow[n\to+\infty]{P} MX = a$, поэтому отношение $S^{*2}/\bar{X}$ должно быть близким к 1 при достаточно больших значениях n. Доказано, что это отношение распределено асимптотически нормально: $S^{*2}/\bar{X} \sim N(1; \sqrt{2a^2/n})$ при $n \to +\infty$.

Если объем выборки n достаточно велик, то распределение центрированной и нормированной случайной величины $((S^{*2}/\bar{X}) - 1)/\sqrt{2a^2/n}$ будет близким к стандартному нормальному распределению $N(0; 1)$.

Заменим в последнем выражении неизвестный параметр a его несмещенной состоятельной асимптотически нормальной оценкой $\hat{a} = \bar{X}$, тогда статистика критерия $U = \sqrt{n}(S^{*2} - \bar{X})/(\bar{X}^2\sqrt{2})$ будет распределена приблизительно нормально при достаточно больших n. Возьмем U в качестве статистики критерия; правило принятия решения при уровне значимости α имеет вид:

при $(|U| \geq u_{1-\alpha/2})$ гипотеза H_0 отвергается,

при $(|U| < u_{1-\alpha/2})$ гипотеза H_0 принимается.

2.35. Выборочный коэффициент корреляции

Пусть $((X_1,Y_1),\dots,(X_n,Y_n))$ – выборка из распределения $F_{XY}(x,y)$ двумерной случайной величины (X,Y), $((x_1,y_1),\dots,(x_n\ y_n))$ – реализация выборки.

В п.1.26. даны определения ковариации и коэффициента корреляции:

$K_{XY} = M((X - MX)\cdot(Y - MY)) = \mathrm{cov}(X,Y)$. – корреляционный момент (ковариация) случайных величин X,Y;

$\rho_{XY} = \dfrac{K_{XY}}{\sigma_X \sigma_Y}$ – коэффициент корреляции.

Числовая характеристика случайной величины – неслучайная величина, ее оценка (функция выборки) – случайная величина, реализация оценки – значение оценки, которое она принимает в точке $((x_1,y_1),\dots,(x_n\ y_n))$ – неслучайная величина (число).

Ниже приведены определения *оценок* указанных числовых характеристик.

$\underline{def}\ \hat{K}_{XY} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ — оценка ковариации K_{XY}.

Преобразуем выражение для $\hat{K}_{XY}$. Раскрывая скобки и вводя обозначение $\dfrac{1}{n}\sum_{i=1}^{n} X_i Y_i = \overline{XY}$, получаем: $\hat{K}_{XY} = \overline{XY} - \bar{X}\bar{Y}$, где $\overline{XY}$ – второй выборочный смешанный начальный момент – оценка второго начального момента $\alpha_{11} = M(XY)$

$\underline{def}\ \hat{\rho}_{XY} = \dfrac{\hat{K}_{XY}}{s_X s_Y}$ – оценка коэффициента корреляции X и Y.

Введем соответствующие обозначения для реализаций указанных оценок, отвечающих данной реализации выборки $(x_1,y_1),\dots,(x_n\ y_n)$ и запишем выражения для реализации оценок выборочной ковариации

$$(\hat{K}_{XY})_e = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = k_{XY}$$

и выборочного коэффициента корреляции

$$(\hat{\rho}_{XY})_e = \frac{k_{XY}}{s_X s_Y} = r_{XY}$$

(s_X и s_Y – выборочные стандартные отклонения).

Доказано, что выборочный коэффициент корреляции $\hat{\rho}_{XY}$ состоятельная оценка коэффициента корреляции ρ_{XY}, обладающая свойствами, аналогичными свойствам ρ_{XY}:

- $|\hat{\rho}_{XY}| \le 1$;
- $Y_i = aX_i + b \ (a \neq 0) \Rightarrow |\hat{\rho}_{XY}| = 1 \ \ \forall i = 1,\ldots,n$

(причем $a > 0 \Rightarrow \hat{\rho}_{XY} = 1$; $a < 0 \Rightarrow \hat{\rho}_{XY} = -1$).

Рассмотрим случай, когда $(X_1,Y_1),\ldots,(X_n,Y_n)$ – *выборка из двумерного нормального распределения.*

Плотность вероятности двумерного нормального распределения имеет вид (см. п.1.30.):

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x-m_X)^2}{\sigma_X^2} - 2r\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right]\right\} (*),$$

$$(-\infty < x < +\infty \ \ -\infty < y < +\infty)$$

где $r = \rho_{XY}$ – коэффициент корреляции.

При $r = \rho_{XY} = 0$ формула $(*)$ имеет вид:

$$f_{XY}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-m_X)^2}{2\sigma_X^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{(y-m_Y)^2}{2\sigma_Y^2}},$$

$$\forall (x,y) \ f_{XY}(x,y) \ \ f_X(x)\cdot f_Y(y) \Leftrightarrow X,Y \text{ – независимы.}$$

Необходимым и достаточным условием независимости компонент *двумерной нормальной случайной величины* (X,Y) является равенство нулю коэффициента корреляции $\rho_{XY} = 0$.

Отсюда следует, что проверка гипотезы о независимости компонент X и Y двумерного нормального распределения, из которого извлечена выборка $((X_1,Y_1),\ldots,(X_n,Y_n))$, сводится к проверке гипотезы о равенстве нулю коэффициента корреляции

$$H_0: \ \rho_{XY} = 0.$$

Доказано, что при справедливости гипотезы H_0 статистика

$$\frac{\hat{\rho}_{XY}}{\sqrt{1-\hat{\rho}_{XY}^2}}\sqrt{n-2}$$

подчиняется распределению Стьюдента с n –2 степенями свободы (n – объем выборки).

Таким образом, в качестве статистики критерия при проверке гипотезы H_0: $\rho_{XY} = 0$ принимают

$$\frac{\hat{\rho}_{XY}}{\sqrt{1-\hat{\rho}_{XY}^2}}\sqrt{n-2} = T_{n-2}$$ (отношение Стьюдента).

Значение статистки критерия, отвечающее данной реализации выборки $(x_1,y_1),\ldots,(x_n,y_n)$: $t_e = \frac{r_{XY}}{\sqrt{1-r^2_{XY}}}\sqrt{n-2}$.

Критическое число $t_{кр} = t_{n-2;1-\alpha/2}$ – квантиль распределения Стьюдента (α – уровень значимости).

Правило принятия решения (критерий) об отклонении или принятии гипотезы H_0 о независимости компонент X и Y двумерного нормального распределения:

$|t_e| \geq t_{кр} \Rightarrow H_0$ отвергается

$|t_e| < t_{кр} \Rightarrow H_0$ принимается

при заданном уровне значимости α.

ЛИТЕРАТУРА

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983, 471 с.
2. Браунли К.А. Статистическая теория и методология в науке и технике.– М.: Наука, 1977. – 408 с., ил.
3. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. – М.: Наука. Гл. ред. Физ-мат. лит.–1988. –(Физико-математическая б-ка инженера). – 480 с.
4. Вероятностные разделы математики. Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д.– СПб.: «Иван Федоров», 2001.– 592 с илл.
5. Ивченко Г.И., Медведев Ю.И. Математическая статистика: Учебное пособие для втузов. 2-е изд. доп. – М.: Высш. шк., 1992. – 304 с., ил.
6. Калинин В.М., Тихомиров С.Р. Лекции по теории вероятностей и математической статистике. СПб., 2002. 89 с.
7. Крамер Г. Математические методы статистики. – М.: Мир, 1975, 648 с.
8. Положинцев Б.И. Теория вероятностей и математическая статистика. Введение в математическую статистику : учеб. пособие.– СПб., Изд-во Политехн. ун-та 2018 .– 96с.
9. Севастьянов Б.А. Вероятностные модели.–М.: Наука. Гл. ред. Физ.-мат. лит., 1992. – (Проблемы науки и техн. Прогресса). –176 с.
10. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. Изд. 3-е, перераб. и доп./ под ред. Фигурнова В.Э. –М.: ИНФРА – М. 2002. –528 с.
11. Чернова Н.И. Теория вероятностей: Учебное пособие / СибГУТИ.— Новосибирск, 2009.—128 с.