

Презентація на тему : Гіпотези про вид розподілу.

Підготував: Батюк М.М

План:

- Поняття статистичний гіпотези.
- Перевірка гіпотези про закон розподілу.
- Критерії згоди Пірсона , Колмогорова .

Означення статистичної гіпотези і задача про її статистичну перевірку

Дані вибірових спостережень часто становлять основу для прийняття одного з кількох альтернативних рішень (продукція може бути бракованою або якісною, точність обробки виробу в межах норми або нижча від норми і т. д.). Із загальнометодологічного погляду тут йдеться про висунення деякої гіпотези, яку відхиляють або приймають після проведення деякого експерименту. Якщо цей експеримент має статистичний (стохастичний) характер, кажуть, що гіпотеза є *статистичною*.

Означення. *Статистичною називають гіпотезу про властивості генеральної сукупності, що перевіряється на основі вибірки.*

Статистичними гіпотезами можуть бути, наприклад, такі твердження: розподіл імовірностей випадкової величини є нормальний; розподіл імовірностей випадкової величини є пуассонівський; у нормальному розподілі випадкової величини параметри $\alpha = 20$ і $\sigma = 1,5$; у показниковому розподілі випадкової величини параметр $\alpha = 5$; випадкові величини X і Y незалежні і т. п.

У математичній статистиці виділяють два основні *типи статистичних гіпотез*:

- *гіпотези про закон розподілу ймовірностей випадкової величини (ознаки генеральної сукупності);*
- *гіпотези про значення параметрів розподілу випадкової величини (ознаки генеральної сукупності).*

*Статистичні гіпотези першого типу називають **непараметричними**, а другого типу – **параметричними**.*

Означення. *Основною (нульовою) називають висунуту гіпотезу і позначають $0 H$.*

Альтернативною (конкуруючою) називають гіпотезу, яка повністю або частково логічно заперечує нульову гіпотезу, і позначають $1 H$.

Наприклад, математичне сподівання нормально розподіленої випадкової величини $\alpha = 10$ – основна гіпотеза; математичне сподівання нормально розподіленої випадкової величини $\alpha \neq 10$ – альтернативна гіпотеза. Записують це так: $0 H : \alpha = 10; 1 H : \alpha \neq 10$.

Задача про статистичну перевірку статистичних гіпотез формулюється так. Розглядають деяку гіпотезу про те, що розподіл імовірностей деякої випадкової величини має той чи інший вигляд, або параметри розподілу мають ті чи інші значення.

Задача полягає у тому, щоб на основі вивчення статистичних даних (вбірки) підтвердити справедливість висунутої гіпотези чи спростувати її. При цьому вказується також імовірність того, що прийняте рішення є правильним або помилковим. Проблема зменшення ймовірності того, що прийняте рішення є помилковим, є також однією із задач математичної статистики. У результаті статистичної перевірки гіпотези може бути прийняте одне з двох правильних рішень: 1) гіпотеза приймається і вона істинна; 2) гіпотеза відхиляється і вона неістинна.

Поряд із тим у результаті статистичної перевірки статистичної гіпотези можуть бути допущені помилки (прийняті неправильні рішення) двох типів: 1) гіпотеза відхиляється, але вона істинна (помилка першого роду); 2) гіпотеза приймається, але вона неістинна (помилка другого роду).

Виявляється, що помилка першого роду має вагоміші наслідки, ніж помилка другого роду.

Щоб застрахувати себе від помилки першого роду або принаймні звести до мінімуму ризик її допущення, вводиться спеціальне число α , яке виражає ймовірність відхилення правильної гіпотези.

Означення. Ймовірність допущення помилки першого роду називають рівнем значущості і позначають через α .

Число α задають наперед і найчастіше його вибирають рівним 0,1; 0,05; 0,01. Якщо $\alpha = 0,05$, то це означає, що ймовірність допустити помилку першого роду є мала, а саме – ми ризикуємо її допустити у 5-ти випадках зі 100.

Перевірка гіпотези про закон розподілу. Критерій згоди Пірсона
Критерієм згоди називають статистичний критерій перевірки гіпотези
 про закон розподілу ймовірностей випадкової величини (ознаки генеральної сукупності). Є кілька критеріїв згоди: критерій Колмогорова, критерій Смірнова, критерій Пірсона та ін. Розглянемо критерій згоди Пірсона (критерій χ^2), який ґрунтується на порівнянні емпіричних і теоретичних частот.

Нехай висунуто гіпотезу H_0 : *випадкова величина X розподілена за законом A .*

Здійснивши вибірку обсягу n , знаходять і записують у вигляді таблиці

інтервальний статистичний розподіл частот:

$[z_{i-1}, z_i)$	$[z_0, z_1)$	$[z_1, z_2)$	...	$[z_{m-1}, z_m)$
n_i	n_1	n_2	...	n_m

$$\sum_{i=1}^m n_i = n.$$

Оскільки перевіряється гіпотеза про те, що розподіл ознаки X генеральної сукупності описується певною (конкретною) функцією розподілу $F(x)$, або, що те ж саме, щільністю розподілу $p(x)$, то для кожного інтервалу $1 [z_{i-1}, z_i)$ і $z_{i-1} < z_i$ – можна визначити теоретичні ймовірності і p попадання значень випадкової величини X у цей інтервал, а отже, і теоретичні частоти $n'_i = np_i$.

Для обчислення ймовірностей i p використовують

$$p_i = P\{z_{i-1} \leq X < z_i\} = F(z_i) - F(z_{i-1}) = \int_{z_{i-1}}^{z_i} p(x) dx, \quad i = \overline{1, m}. \quad (1)$$

Зазначимо, що для обчислення ймовірностей p_1 і p_m у формулах (1) покладають, відповідно $z_0 = -\infty$ і $z_m = +\infty$. Тоді $\sum_{i=1}^m p_i = 1$.

Отримані результати обчислень зручно записати у формі таблиці:

$[z_{i-1}, z_i)$	$[z_0, z_1)$	$[z_1, z_2)$	...	$[z_{m-1}, z_m)$
n_i	n_1	n_2	...	n_m
p_i	p_1	p_2	...	p_m
$n'_i = np_i$	n'_1	n'_2	...	n'_m

Згідно з критерієм Пірсона для перевірки гіпотези H_0 вводиться випадкова величина (статистика) K :

$$K = \chi^2 = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i},$$

де m – кількість груп у статистичному розподілі вибірки; n_i – емпірична частота ознаки X в i -тій групі; $n'_i = np_i$ – теоретична частота; p_i – ймовірність того, що значення X належить i -тій групі.

Відомо, що при $n \rightarrow \infty$ закон розподілу статистики K прямує до закону розподілу χ^2 з $k = m - r - 1$ ступенями вільності, де m – кількість груп у статистичному розподілі вибірки; r – кількість параметрів гіпотетичного розподілу A (наприклад, $r = 2$ для нормального розподілу, $r = 1$ для розподілу Пуассона, $r = 0$ для рівномірного розподілу).

Для критерію χ^2 будують правосторонню критичну область за правилом:

$$P\{\chi^2 > \chi_{кр}^2\} = \alpha. \quad (2)$$

За заданим рівнем значущості α і кількістю ступенів вільності k із таблиці критичних точок розподілу χ^2 (в якій дано розв'язки рівняння (2)) знаходять критичну точку $k_{кр} = \chi_{кр}^2(\alpha, k)$.

На підставі даних вибірки, записаних у таблиці, обчислюють емпіричне значення критерію Пірсона:

$$K_{емп} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i}.$$

Порівнюємо значення $K_{емп}$ і $k_{кр}$: якщо $K_{емп} \geq k_{кр}$, то гіпотезу H_0 відхиляють; якщо ж $K_{емп} < k_{кр}$, то гіпотезу H_0 приймають.

Застосування критерію χ^2 вимагає дотримання таких умов: 1) експериментальні дані мають бути незалежними, тобто вибірка має бути випадковою; 2) обсяг вибірки має бути достатньо великим (практично не меншим ніж 50 одиниць), а частота кожної групи – не меншою за 5. Якщо остання умова не виконується, то проводиться попереднє об'єднання нечисленних груп.

Критерій згоди Пірсона дає відповідь на питання, чи розбіжність між емпіричними і теоретичними частотами зумовлена випадковістю, чи вона є значущою. Як і будь-який інший критерій він *не доводить справедливості* гіпотези H_0 , а лише дозволяє встановити на прийнятному рівні значущості *узгодженість чи неузгодженість* гіпотези H_0 з даними спостережень.

Приклад 1. У таблиці наведено покази 500 навмання вибраних механічних годинників, виставлених у вітринах годинникових магазинів. Для рівня значущості $\alpha = 0,05$ перевірити гіпотезу H_0 про те, що покази годинників рівномірно розподілені на проміжку $[0; 12)$.

i	0	1	2	3	4	5	6	7	8	9	10	11
n_i	41	34	54	39	49	45	41	33	37	41	47	39

$$\sum_{i=0}^{11} n_i = 500.$$

Тут i – номер проміжка від i -ої години до $(i+1)$ -ої, $i = \overline{0,11}$; n_i – кількість годинників, покази яких потрапляють в i -ий проміжок.

Розв'язання. Для рівномірного розподілу щільність розподілу ймовірностей визначається функцією:

$$p(x) = \begin{cases} 0, & x \notin [0; 12); \\ 1/12, & x \in [0; 12). \end{cases}$$

Для $i = \overline{0, 11}$ обчислюємо:

$$p_i = P\{X \in [i, i+1)\} = \int_i^{i+1} \frac{dx}{12} = \frac{1}{12}; \quad n'_i = np_i = 500 \cdot \frac{1}{12} = 41,7.$$

$$K_{\text{емп}} = \sum_{i=0}^{11} \frac{(n_i - n'_i)^2}{n'_i} = \frac{1}{41,7} \left((41 - 41,7)^2 + \dots + (39 - 41,7)^2 \right) = 9,99.$$

Кількість ступенів вільності $k = m - r - 1 = 12 - 0 - 1 = 11$. За таблицею критичних точок розподілу χ^2 знаходимо: $k_{\text{кр}} = \chi^2_{\text{кр}}(0,05; 11) = 19,7$. Отже,

$K_{\text{емп}} < k_{\text{кр}}$, тому гіпотезу H_0 приймаємо.

Приклад 2. У таблиці наведено емпіричні частоти n_i та теоретичні частоти n'_i , обчислені, виходячи з гіпотези H_0 про нормальний розподіл генеральної сукупності. Для рівня значущості $\alpha = 0,01$ перевірити гіпотезу H_0 про нормальний розподіл генеральної сукупності.

n_i	8	16	40	72	36	18	10
n'_i	6	18	36	76	39	18	7

Розв'язання. Обчислимо емпіричне значення критерію Пірсона ($m = 7$):

$$K_{\text{емп}} = \sum_{i=1}^7 \frac{(n_i - n'_i)^2}{n'_i} = \frac{(8-6)^2}{6} + \frac{(16-18)^2}{18} + \dots + \frac{(10-7)^2}{7} = 3,06.$$

За таблицею критичних точок розподілу χ^2 для рівня значущості $\alpha = 0,01$ і кількості ступенів вільності $k = m - r - 1 = 7 - 2 - 1 = 4$ ($r = 2$ для нормального розподілу) знаходимо: $k_{\text{кр}} = \chi^2_{\text{кр}}(0,01; 4) = 13,3$. Отже, $K_{\text{емп}} < k_{\text{кр}}$, тому гіпотезу H_0 приймаємо.

Критерий Колмогорова. Уже говорилось (параграф 3 гл. 1), что в силу теоремы Гливенко–Кантелли эмпирическая функция распределения $F^*(x)$ представляет собой состоятельную оценку теоретической функции распределения $F(x)$. Поэтому можно сравнить эмпирическую функцию распределения $F^*(x)$ с гипотетической $F_0(x)$ и, если мера расхождения между ними мала, то считать справедливой гипотезу H_0 .

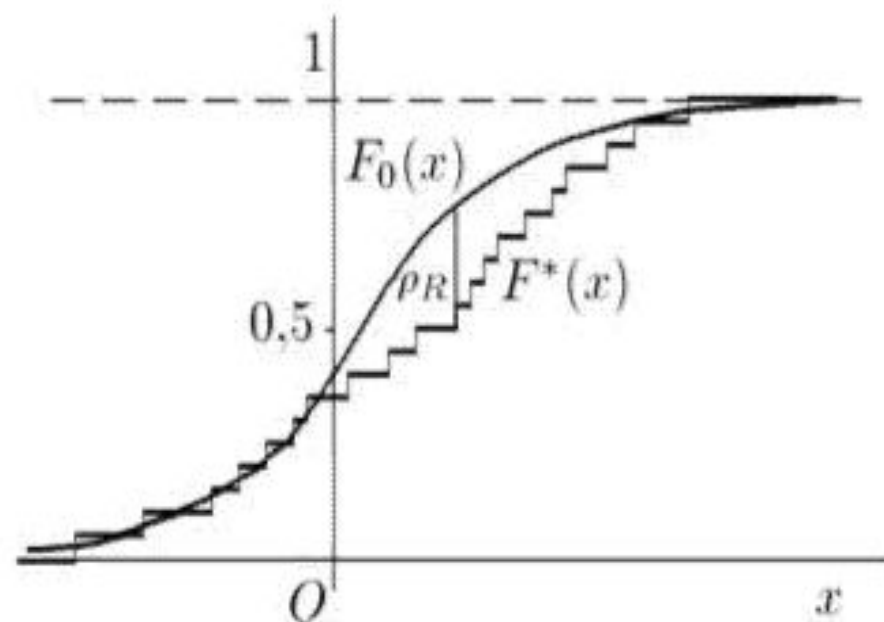


Рис. 6

Наиболее естественной и простой из таких мер (будем предполагать, что $F_0(x)$ — непрерывная функция) является равномерное расстояние

$$\begin{aligned}\rho_R &= \rho_R(F^*(x), F_0(x)) = \\ &= \sup_{-\infty < x < \infty} |F^*(x) - F_0(x)|\end{aligned}$$

Однако при построении критерия Колмогорова более удобно пользоваться нормированным расстоянием $\rho = \sqrt{n} \rho_R$.

Итак, рассмотрим статистику

$$\rho = \rho(X_1, \dots, X_n) = \sqrt{n} \sup_{-\infty < x < \infty} |F^*(x) - F_0(x)|.$$

Критерий Колмогорова предписывает принять гипотезу H_0 , если $\rho < C$, и отвергнуть в противном случае, где C — критическое значение критерия.

Если гипотеза H_0 справедлива, то распределение статистики ρ не зависит от гипотетической функции распределения $F_0(x)$ (доказательство этого факта следует из инвариантности статистики критерия Колмогорова относительно монотонных преобразований, в частности преобразования $g(x) = F_0^{(-1)}(x)$, где $F_0^{(-1)}(x)$ — обратная к $F_0(x)$ функция; преобразование $g(x)$ приводит выборку $X_1, \dots, X_n$ к равномерно распределенной на отрезке $(0,1)$).

При $n \rightarrow \infty$ распределение статистики ρ сходится к распределению Колмогорова и критическое значение C при большом объеме выборки практически совпадает с $(1 - \alpha)$ -квантилью $k_{1-\alpha}$ распределения Колмогорова.

При практической реализации критерия Колмогорова сначала по выборке $X_1, \dots, X_n$ составляют вариационный ряд $X_1^*, \dots, X_n^*$. Затем находят $F_0(X_i^*)$ и определяют значения статистики ρ по формуле

$$\rho = \sqrt{n} \left[\max_{1 \leq i \leq n} \left| F_0(X_i^*) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right].$$

Наконец, сравнивают полученное значение ρ с критическим значением C для заданного уровня значимости α и принимают или отвергают гипотезу H_0 .

Пример . Проверим с помощью критерия Колмогорова гипотезу H_0 о том, что проекция X вектора скорости молекулы водорода на ось координат распределена по нормальному закону. Проверку произведем

для уровня значимости $\alpha = 0,05$. Параметры нормального закона не заданы, значит, мы имеем дело со сложной гипотезой H_0 и сначала должны оценить среднее ϑ_1 и дисперсию ϑ_2 . Поскольку мы будем пользоваться критерием Колмогорова, хотелось бы оценки ϑ_1^* и ϑ_2^* неизвестных параметров ϑ_1 и ϑ_2 выбрать таким образом, чтобы они доставляли минимальное значение статистики критерия Колмогорова

$$\rho = \sqrt{n} \left[\max_{1 \leq i \leq n} \left| \Phi(X_i^*; \vartheta_1, \vartheta_2) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right],$$

где $X_1^*, \dots, X_n^*$ — вариационный ряд выборки $X_1, \dots, X_n$,

а $\Phi(x; m, \sigma^2) = \Phi((x - m)/\sigma)$ — функция распределения нормального закона с параметрами m, σ^2 . Однако искать минимум ρ как функции от ϑ_1 и ϑ_2 — весьма сложная в вычислительном плане задача, так как $\Phi(x)$ даже не выражается в элементарных функциях. Поэтому в качестве оценок ϑ_1^* и ϑ_2^* используем оценки максимального

правдоподобия $\vartheta_1^* = m^* = 0,082$ и $\vartheta_2^* = \sigma^{2*} = 1,34$

Теперь с помощью критерия Колмогорова будем проверять простую гипотезу $H_0 : F(x) = F_0(x) = \Phi(x; \vartheta_1^*, \vartheta_2^*) = \Phi(x; 0,082, 1,34)$.

Вычислив сначала $Y_i^* = (X_i^* - \vartheta_1^*)/\sqrt{\vartheta_2^*}$ и воспользовавшись равенством $\Phi(X_i^*; \vartheta_1^*, \vartheta_2^*) = \Phi((X_i^* - \vartheta_1^*)/\sqrt{\vartheta_2^*})$, последовательно находим затем значения $\Phi(Y_i^*)$, $F_i^* = (2i - 1)/(2n)$ и $S_i = |\Phi(Y_i^*) - F_i^*|$

Наконец, определяя значение статистики критерия Колмогорова

$$\rho = \sqrt{n} \left[\max_{1 \leq i \leq n} \left| \Phi(X_i^*; \vartheta_1, \vartheta_2) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right] = 0,07\sqrt{50} = 0,49$$

(максимальное значение S_i равно 0,06) и сравнивая его с 0,95-квантилью распределения Колмогорова $k_{0,95} = 1,36$ видим, что $\rho < k_{0,95}$.

Значит, мы должны принять гипотезу H_0 и считать распределение проекции вектора скорости молекулы водорода нормальным. $\square$