38.03.01 Экономика – «Экономика предприятий и организаций»

ЭКОНОМЕТРИКА

Тема 1. Случайные величины. Законы распределения случайных величин

Цель: приобрести практические навыки применения аппарата математической статистики для обработки эмпирических данных.

Основные вопросы, рассматриваемые на занятии:

1) случайные величины;
2) числовые характеристики случайных величин;
3) функция распределения случайной величины;
4) плотность распределения случайной величины;
5) дискретные распределения случайных величин: биномиальное распределение; распределение Пуассона;
6) непрерывные распределения: равномерное распределение; нормальное распределение.

1 Случайные величины

Случайная величина – это величина, принимающая одно из возможных (заранее неизвестных) значений в результате испытаний.

Случайная величина – это действительная функция $\xi = \xi(\omega_i)$, заданная на множестве Ω элементарных событий ω_i $\{\omega_i \in \Omega\}$, так, что любое множество $A = \{\omega_i : \xi(\omega_i) < x\}$ принадлежит алгебре событий *S* (алгебра событий – это множество $S = 2^n$ элементов, на котором определены операции сложения и умножения).

Случайные величины обозначаются: *X*, *Y*, …

Случайные величины бывают: дискретными и непрерывными.

Случайная дискретная величина задаётся отдельными значениями и образует вариационный ряд: $x_1, x_2, .., x_n$.

Непрерывной называют случайную величину, которая может принимать любое значение из некоторого конечного или бесконечного числового интервала (количество возможных значений непрерывной случайной величины – несчетно).

Значения случайной величины образуют статистический ряд.

Для анализа такого ряда используют числовые характеристики описательной статистики:

- математическое ожидание (среднее значение),
- дисперсия (стандартное отклонение).

2 Числовые характеристики случайных величин

2.1 Дискретная случайная величина

Математическое ожидание – среднее значение случайной величины при стремлении количества выборок или количества измерений (иногда говорят — количества испытаний) к бесконечности.

Числовая характеристика, которая определяет возможный средний результат значения случайной дискретной величины X по закону её распределения и соответствующим ему значениям вероятности есть **математическое ожидание** $M(X)$.

Математическое ожидание: $M(X)$: $M[X]=\sum_{i=1}^{n} x_i p_i$.

Закон распределения случайной дискретной величины задаёт соответствие между её возможными значениями и их вероятностями и записывается в виде таблицы:

X	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

$x_1 < x_2 < \ldots < x_n$, и $p_1 + p_2 + \ldots + p_n = 1$.

Пример 1. По закону распределения случайной величины X заданному Таблицей 1:

Таблица 1

X	$x_0 = 0$	$x_1 = 1$	$x_2 = 2$	$x_3 = 3$	$x_4 = 4$
p_i	0,13	0, 34	0,35	0,15	0,02

Определить математическое ожидание.

Решение. Математическое ожидание выразится суммой:

$$M(X) = 0 \cdot 0{,}13 + 1 \cdot 0{,}34 + 2 \cdot 0{,}35 + 3 \cdot 0{,}15 + 4 \cdot 0{,}02 = 1{,}57.$$

Ответ: $M(X) = 1{,}57$.

Основные свойства математического ожидания:

1. Математическое ожидание постоянной величины C есть сама эта величина:
$$M(C) = C.$$
2. Постоянный числовой множитель можно выносить за знак математического ожидания:
$$M(CX) = C\,M(X).$$
3. Математическое ожидание алгебраической суммы случайных независимых величин равно сумме их математических ожиданий:
$$M\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} M(X_i).$$
4. Математическое ожидание линейной функции равно той же линейной функции:
$$M(aX+b) = a\,M(X) + b.$$
5. Математическое ожидание произведения двух независимых случайных величин X,Y равно произведению математического ожидания каждой CB:
$$M(XY) = M(X)\,M(Y).$$

Математическое ожидание это числовая характеристика, относительно которой группируются результаты конкретных испытаний над случайной дискретной величиной. Разброс значений результатов испытаний относительно математического ожидания характеризуется **дисперсией**.

Дисперсия случайной величины – мера разброса значений случайной величины относительно её математического ожидания.

Дисперсией $D(X)$ (или D_x) случайной величины X называется математическое ожидание квадрата разности между CB и её математическим ожиданием:

Дисперсия: $D(X) = M[X - M(X)]^2 = M(X^2) - M^2(X)$

$$D[X] = \sum_{i=1}^{n} (x_i - M[X])^2 p_i.$$

Пример 2. По условию первого примера и полученному значению математического ожидания $M(X) = 1{,}57$ вычислить дисперсию случайной величины.

Решение. $D(X) = (0 - 1{,}57)^2 \cdot 0{,}13 + (1 - 1{,}57)^2 \cdot 0{,}34 + (2 - 1{,}57)^2 \cdot 0{,}35 + (3 - 1{,}57)^2 \cdot 0{,}15 +$
$+ (4 - 1{,}57)^2 \cdot 0{,}02 \approx 0{,}92.$

Ответ: $D(X) \approx 0{,}92$.

Основные свойства дисперсии:

1. Дисперсия постоянной величины равна нулю

$$D(C) = 0,$$

следовательно, если дисперсия равна нулю, то случайная величина $X = C$ постоянная.

2. Дисперсия суммы независимых случайных величин равна сумме дисперсий каждой величины:

$$D(X_1 + X_2 + \ldots + X_n) = D(X_1) + D(X_2) + \ldots + D(X_n),$$

$$D(X_1 + X_2) = D(X_1) + D(X_2)$$

$$D\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} D(X_i).$$

Следовательно: $D(X + C) = D(X)$.

3. За знак дисперсии можно выносить квадрат постоянного множителя:

$$D(CX) = C^2 D(X).$$

4. Дисперсия линейной функции $(a X + b)$ равна произведению квадрата коэффициента a на дисперсию CB:

$$D(a X + b) = a^2 D(X).$$

5. Чем меньше дисперсия, тем выше концентрация исходов конкретных испытаний случайной величины относительно математического ожидания.

Среднеквадратическое отклонение – наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания.

Ещё одной из числовых характеристик, определяющей разброс значений CB и употребляющейся как мера качества статистических оценок, является **среднее квадратичное отклонение** σ.

Средним квадратичным отклонением σ случайной величины X от её математического ожидания называют квадратный корень из дисперсии $D(X)$:

Среднеквадратическое отклонение: $\sigma_X = \sqrt{D[X]}$.

Некоторые свойства среднего квадратичного отклонения:

- среднее квадратичное отклонение постоянной величины равно нулю;
- среднее квадратичное отклонение случайной величины, умноженной на константу K, так же умножается на эту константу:

$$\sigma(KX) = \sqrt{D(KX)} = K\sigma(X);$$

- – среднее квадратичное отклонение $\sigma_{\sum \xi_i}$ суммы случайных элементарных величин $(\xi_1 + \xi_2 + \ldots \xi_n)$ равно корню квадратному из суммы их средних квадратичных отклонений:

$$\sigma_{\sum \xi_i} = \sqrt{\sum_{i=1}^{n} \sigma_i} \ .$$

Пример 3. Используя значения дисперсии из Примера 2, вычислим среднее квадратичное отклонение для *СВ X*, заданной таблицей в Примере 1.

Найдем среднее квадратичное отклонение *СВ X*, если $D(X) \approx 0{,}92$:

$$\sigma(X) = \sqrt{D(X)} = \sqrt{0{,}92} \approx 0{,}956.$$

Ответ: $\sigma(X) = 0{,}956.$

Наиболее полной характеристикой случайной величины является её **закон распределения**.

Аналитически случайная величина задаётся либо функцией распределения, либо плотностью вероятностей.

3 Функция распределения

Функция распределения *F*(x) определяет вероятность того, что случайная величина некоторого случайного события ξ принимает значения меньшее произвольно выбранного значения *x*.

Функцией распределения *случайной величины X* определяется равенством:

$$F(x) = P(\xi < x).$$

С увеличением *x* функция распределения *суммирует* вероятности, является неубывающей и изменяется в пределах $0 \le F(x) \le 1$. Поэтому, иногда, её называют **интегральной функцией распределения**.

Свойства функции распределения *F*(*x*):

1) значения функции распределения изменяются от нуля до единицы

$$0 \le F(x) \le 1, \quad \lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1;$$

2) функция распределения *F*(*x*) *неубывающая функция*, то есть

$$(x_1 < x_2) \Rightarrow F(x_1) \le F(x_2).$$

3) если случайная величина ξ рассматривается на интервале (*a*, *b*), то функция распределения позволяет искать вероятность того, что случайная величина принимает значение из данного интервала по формуле:

$$P(a \le \xi < b) = F(b) - F(a),$$

при этом

$$F(x)=\begin{cases}0, & x\le a\\ 1, & x>b\end{cases}, \text{ то есть}$$

$$\lim_{x\to-\infty}F(x)=0,\ \lim_{x\to+\infty}F(x)=1;$$

4) если случайная величина $\xi \ge x$, то вероятность $P(\xi \ge x) = 1 - F(x)$;

5) вероятность того, что случайная величина ξ примет одно определенное значение, равно нулю, то есть $P(\xi = x) = 0$.

Пример 4. Случайная величина X задана функцией распределения

$$F(x)=\begin{cases}0, \textit{ при } x\le 1,\\ \dfrac{x}{4}+\dfrac{1}{4}, \textit{ при } -1<x\le 3,\\ 1, \textit{ при } x>3.\end{cases}$$

Найти вероятность того, что в результате испытания X примет значение, принадлежащее интервалу (0, 2).

Решение. Так как на интервале (0, 2) $F(x)=\dfrac{x}{4}+\dfrac{1}{4}$, то $F(2)-F(0)=\left(\dfrac{2}{4}+\dfrac{1}{4}\right)-\left(\dfrac{0}{4}+\dfrac{1}{4}\right)=\dfrac{1}{2}$.

Пример 5. Найти функцию распределения случайной дискретной величины X, закон распределения которой задан таблицей:

X	3	4	5	6
P	0,15	0,20	0,3	0, 35

Решение. Так как функция распределения $F(x)$ определяет вероятность того, что CB события X принимает значения меньшее выбранных значений x, то

при $x \le 3$, $F(x) = 0$;

при $3 < x \le 4$, $F(x) = P(x = 3) = 0{,}15$;

при $4 < x \le 5$, $F(x) = P(x = 3) + P(x = 4) = 0{,}35$;

при $5 < x \le 6$, $F(x) = P(x = 3) + P(x = 4) + P(x = 5) = 0{,}65$;

при $x > 6$, $F(x) = P(x = 3) + P(x = 4) + P(x = 5) + P(x = 6) = 1$.

График функции распределения дискретной случайной величины имеет ступенчатый вид:

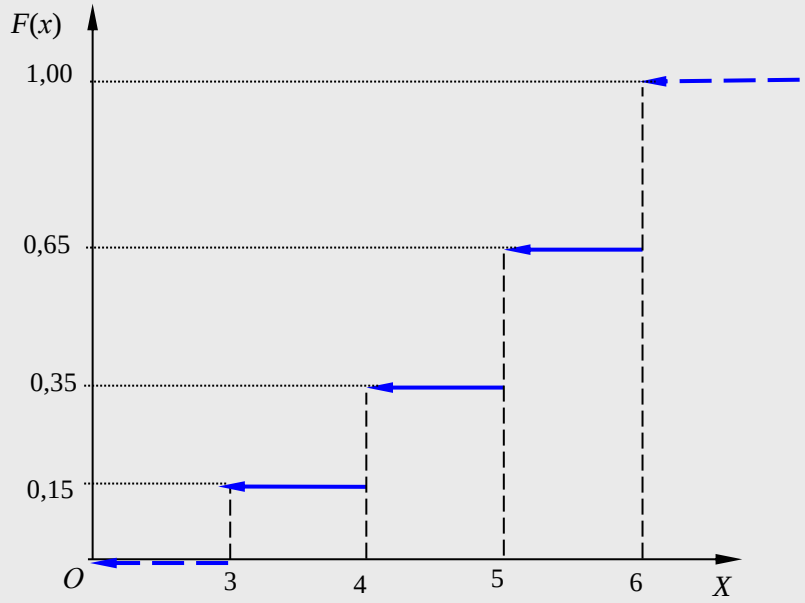


График функции $F(x)$

Аналитический вид функции распределения:

$$F(x)=\begin{cases}0 & \text{если } x\le 3,\\ 0{,}15 & \text{если } 3<x\le 4,\\ 0{,}35 & \text{если } 4<x\le 5,\\ 0{,}65 & \text{если } 5<x\le 6,\\ 1 & \text{если } x>6.\end{cases}$$

2.2 Непрерывная случайная величина

Случайная непрерывная величина на практике встречается редко, так как является абстрактным математическим понятием. Чаще всего, рассматриваются случайные дискретные величины.

Непрерывной называют случайную величину, которая может принимать любое значение из некоторого конечного или бесконечного числового интервала (количество возможных значений непрерывной случайной величины – несчетно).

Математическое ожидание *M* (*X*) случайной непрерывной величины, возможные значения которой принадлежат всей оси *Ox* и *f*(*x*) плотность распределения *СВ X*, определяется интегралом (предполагается абсолютная сходимость интеграла):

$$M(X) = \int_{-\infty}^{+\infty} x\, f(x)\, d\, x\,.$$

Математическое ожидание непрерывной случайной величины X, возможные значения которой принадлежат отрезку [*a, b*], называют определённый интеграл:

$$M(X) = \int_{a}^{b} x\, f(x)\, d\, x.$$

Дисперсией *D*(*X*) случайной непрерывной величины называется значение несобственного интеграла:

$$D(X) = \int_{-\infty}^{+\infty} [x - M(X)]^2\, f(x)\; d\, x.$$

Если возможные значения *X* принадлежат отрезку [*a, b*], то:

$$D(X) = \int_{a}^{b} x^2\; f(x) d\, x - M^2(X)\;.$$

4 Плотность распределения

Исходя из равенства $P(\, a \le \xi < b) = F(b) - F(a)$ можем исследовать, как изменится значение случайной величины, если границы *a* и *b* интервала близки.

Пусть $a = x,\;\; b = x+\Delta x$.

Пусть $F(x)$ непрерывная и дифференцируемая функция распределения случайной величины ξ.

Вычислим вероятность попадания этой случайной величины на интервал $(x,\;\; x+\Delta x)$:

$$P(\, x \le \xi < x + \Delta x) = F(x + \Delta x) - F(x).$$

Возьмем отношение этой вероятности к длине участка Δx, т.е. рассмотрим среднюю вероятность, приходящуюся на единицу длины на этом участке, тогда при $\Delta x \to 0$, перейдем к пределу:

$$\lim_{\Delta x \to 0} \frac{P(x \le \xi < x + \Delta x)}{\Delta x} = F'(x)$$

Введем обозначение:

$$f(x) = F'(x)$$

Функция $f(x)$ – производная функции распределения – характеризует плотность, с которой распределяются значения случайной величины в данной точке. Эта функция называется **плотностью распределения непрерывной случайной величины** ξ.

Свойства плотности вероятности:

1) плотность вероятности есть неотрицательная функция, $f(x) \ge 0$;

2) несобственный интеграл функции $f(x)$ равен единице $\int_{-\infty}^{+\infty} f(x)\,dx = 1,$

$$\int_{-\infty}^{+\infty} f(x\,dx) = P(X < \infty) = F(+\infty) - F(-\infty) = 1 - 0 = 1;$$

3) $F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\,dt;$ $\quad F(x) = F(x) - F(-\infty);$

4) $P(a \le x \le b) = \int_{a}^{b} f(x)\,dx,$ $P(a \le x \le b) = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx,$

если СВ X непрерывная , то $\quad F(x) = P(a \le x \le b) = \int_{a}^{b} f(x)\,dx.$

Кривая, изображающая плотность распределения случайной величины, называется кривой распределения. Плотность распределения, так же как и функция распределения, есть одна из форм закона распределения.

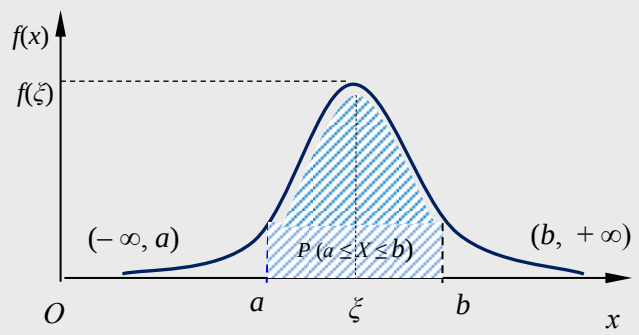


Рисунок 1

На Рисунке 1 площадь криволинейной трапеции, ограниченной графиком кривой плотности вероятности $f(x)$ над отрезком $[a, b]$ определяет вероятность попадания в этот интервал *СВ X*. Так как площадь определяется интегралом

$$S = \int_{a}^{b} f(x)\,dx,$$ тогда

$$\int_a^b f(x)\,dx = P(a \le X \le b) = 1.$$

Пример 5. Случайная величина X задана плотностью распределения функция $f(x)$ которой имеет вид:

$$f(x) = \begin{cases} 0 & \text{если } x \le 0, \\ 0{,}3x & \text{если } 0 < x \le 2, \\ 0 & \text{если } x > 2. \end{cases}$$

Найти математическое ожидание $M(X)$, дисперсию $D(X)$ и среднее квадратическое отклонение $\sigma(X)$.

Решение. Найдем математическое ожидание $M(X)$ *СВ X*:

$$M(X) = \int_0^2 0{,}3x^2\,dx = 0{,}3\frac{x^3}{3}\Big|_0^2 = 2{,}7.$$

Дисперсия случайной непрерывной величины определяется значением несобственного интеграла:

$$D(X) = \int_{-\infty}^{+\infty} [x - M(X)]^2 f(x)\,dx \text{ или } D(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X),$$

$$D(X) = \int_0^2 [x - 2{,}7]^2\, 0{,}3x\,dx = 0{,}3\int_0^2 (x^3 - 5{,}4x^2 + 7{,}29x)\,dx =$$

$$= 0{,}3\left(\frac{x^4}{4} - 5{,}4\frac{x^3}{3} + 7{,}29\frac{x^2}{2}\right)\Big|_0^2 = 1{,}25.$$

Среднее квадратическое отклонение $\sigma(X)$ равно: $\sigma(X) = \sqrt{D(X)} = \sqrt{1{,}25} = 1{,}12$.

Ответ: $M(X) = 2{,}7$; $D(X) = 1{,}25$; $\sigma(X) = 1{,}12$.

Пример 6. Найти вероятность того, что случайная величина X, плотность вероятности которой задана формулой:

$$f(x) = \begin{cases} 0 & \text{если } x \le 0, \\ \dfrac{1}{x^2 - 3} & \text{если } 0 < x \le 3, \\ 0 & \text{если } x > 3. \end{cases}$$

примет значения в интервале (2, 3).

Решение. Искомая вероятность найдется по формуле

$$P(a \le x \le b) = \int_a^b f(x)\,dx;$$

$$P(2 < x < 3) = \int_2^3 \frac{dx}{x^2 - 3} = \frac{1}{2\sqrt{3}} \ln\left|\frac{x - \sqrt{3}}{x + \sqrt{3}}\right|\Bigg|_2^3 = \frac{1}{2\sqrt{3}}\left(\ln\left|\frac{3 - \sqrt{3}}{3 + \sqrt{3}}\right| - \ln\left|\frac{2 - \sqrt{3}}{2 + \sqrt{3}}\right|\right) \approx 0{,}4.$$

Ответ: $P(2 < x < 3) = 0{,}4.$

Пример 7. Функция, плотности распределения вероятности случайной величины X, имеет вид: $f(x) = cx^2 e^{-2x}, \quad 0 \le x < +\infty.$ Найти функцию распределения $F(x)$ *СВ X.*

Решение. Найдем значение коэффициента c из соотношения:

$$\int_0^{+\infty} c\,x^2\,e^{-2x}\,dx = 1 \;\rightarrow\; c = \frac{1}{\int_0^{+\infty} x^2\,e^{-2x}\,dx}.$$

Вычислим интеграл:

$$\int_0^{+\infty} x^2\,e^{-2x}\,dx = -\frac{1}{2}x^2 e^{-2x}\Big|_0^{+\infty} + \int_0^{+\infty} x\,e^{-2x}\,dx =$$

$$= -\frac{1}{4}xe^{-2x}\Big|_0^{\infty} + \frac{1}{4}\int_0^{+\infty} e^{-2x}\,dx = -\frac{1}{4}e^{-2x}\Big|_0^{\infty} = \frac{1}{4}.$$

Коэффициент c = 4, а плотность распределения заданной функции есть $f(x) = 4x^2e^{-2x}$.

Функция распределения $F(x)$ *СВ X* имеет вид:

$$F(X) = 4\int_0^x t^2\,e^{-2t}\,dt = 1 - \frac{4x^2 + 2\cdot 2x + 2}{2}e^{-2x} = 1 - (2x^2 + 2x + 1)e^{-2x}.$$

Ответ: $F(X) = 1 - (2x^2 + 2x + 1)e^{-2x}.$

Начальные и центральные моменты

Понятия математического ожидания и дисперсии являются частными случаями более общего понятия для числовых характеристик случайных величин – *моментов распределения*. Моменты распределения случайной величины вводятся как математические ожидания некоторых простейших функций от случайной величины.

Начальные и центральные моменты являются обобщенными числовыми характеристиками случайных величин.

Начальным моментом *k*-го порядка называют математическое ожидание от случайной величины X^k.

$$\nu_k = M(X^k), \quad k = 1, 2, 3, \ldots$$

$$k = 1, \ \nu_1 = M(X),$$

$$k = 2, \ \nu_1 = M(X^2)$$

и т. д.

Для дискретной случайной величины *X* начальные моменты определяются зависимостью:

$$\nu_k = \sum_{i=1}^{n} x_i^k p_i .$$

Для непрерывной случайной величины, определяются интегралом:

$$\nu_k = \int_{-\infty}^{\infty} x^k \ f(x)\,dx.$$

Если непрерывная величина задана интервалом $X = [a, b]$, то моменты вычисляют по формуле:

$$\nu_k = \int_{a}^{b} x^k \ f(x)\,dx.$$

Центральным моментом *k*-го порядка называют математическое ожидание от случайной величины $(X\text{-}M(X))^k$

$$\mu_k = M(X - M(X))^k, \quad k = 1, 2, 3, \ldots$$

$$k = 1, \ \mu_1 = M(X - M(X)) = 0,$$

$$k = 2, \ \mu_2 = M(X - M(X))^2 = D(X),$$

$$k = 3, \ \mu_3 = M(X - M(X))^3$$

и т.д.

Для дискретной случайной величины *X* центральные моменты вычисляются по формуле:

$$\mu_k = \sum_{i=1}^{n} (x_i - M(X))^k \ p_i .$$

Для непрерывной случайной величины:

$$\mu_k = \int_{-\infty}^{\infty} (x - M(X))^k \ f(x)\,dx.$$

Асимметрия. Центральный момент третьего порядка:

$$\boldsymbol{k = 3,\ \mu_3 = M(X - M(X))^3.}$$

Предположим, что распределение случайной величины симметрично относительно математического ожидания $M[X]$.

Тогда все центральные моменты нечетного порядка равны нулю. Если центральный момент нечетного порядка не равен нулю, то это говорит об асимметричности распределения и чем больше момент, тем больше асимметрия. В качестве характеристики используют центральный момент 3-го порядка. Однако принять этот момент для оценки асимметричности неудобно потому, что его величина зависит от единиц, в которых измеряется случайная величина. Чтобы устранить этот недостаток, μ^3 делят на σ^3 и таким образом получают характеристику.

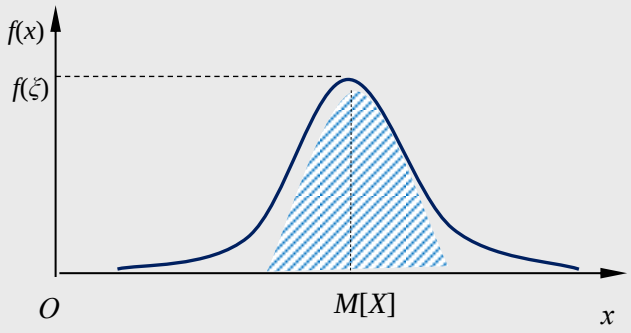
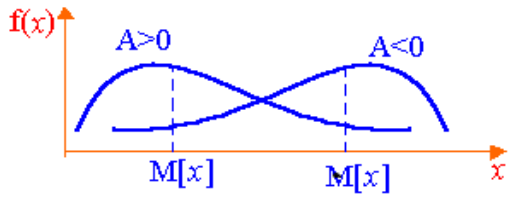


Рисунок 2

Коэффициентом асимметрии A называется величина:

$$\boldsymbol{A = \frac{\mu_3}{\sigma_3}}.$$



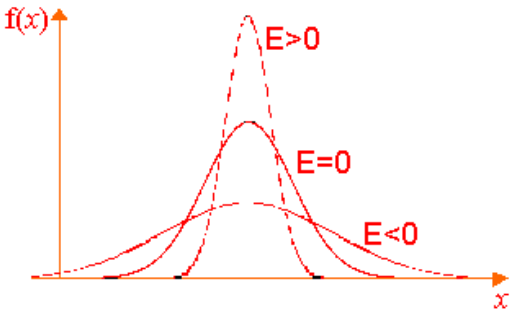
Если коэффициент асимметрии отрицателен, то это говорит о большом влиянии на величину μ_3 отрицательных отклонений. В этом случае кривая распределения более полога слева от $M[X]$. Если коэффициент A положителен, то кривая более полога справа.

Центральный момент 4-го порядка

$$\mu_4 = \boldsymbol{M}(\boldsymbol{X} - \boldsymbol{M}(\boldsymbol{X}))^4$$

служит для оценки так называемого **эксцесса**, определяющего степень крутости островершинности кривой распределения относительно центра нормального распределения

Эксцессом *E* называется величина:



$\boldsymbol{E} = \frac{\mu_4}{\sigma_4} - 3$ Число 3 здесь выбрано потому, что для наиболее распространенного **нормального закона** распределения μ_4/σ_4=3. Поэтому эксцесс служит для сравнения имеющихся распределений с нормальным распределением, у которого эксцесс равен нулю. Это означает, что если у распределения эксцесс положителен, то соответствующая кривая распределения более "островершинная" по сравнению с кривой нормального распределения; если у распределения эксцесс отрицателен, то соответствующая кривая более "плосковершинная".

Пример. ДСВ *X* задана следующим законом распределения:

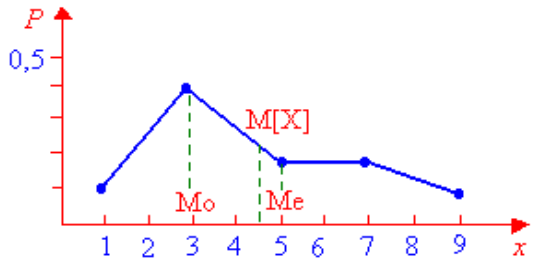


Рис. 5.4

X	1	3	5	7	9
P	0,1	0,4	0,2	0,2	0,1

Найти коэффициент асимметрии и эксцесс.

Решение. Предварительно найдем начальные моменты до 4-го порядка

$$\nu_1 = \boldsymbol{M}[\boldsymbol{X}] = 4{,}6, \qquad \nu_2 = \boldsymbol{M}[\boldsymbol{X}^2] = 26{,}6,$$

$$\nu_3 = \boldsymbol{M}[\boldsymbol{X}^3] = 177{,}4, \quad \nu_4 = \boldsymbol{M}[\boldsymbol{X}^4] = 1293{,}8,$$

Теперь вычислим центральные моменты:

$$\mu_2 = \nu_2 - \nu_1^2 = \boldsymbol{D[X]} = 5{,}44 \quad \Rightarrow$$
$$\sigma = \sqrt{\boldsymbol{D[X]}} = 2{,}33,$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = 4{,}992,$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_3 - 3\nu_1^4 = 1293{,}8 - 4 \cdot 4{,}6 \cdot 177{,}4 +$$
$$+ 6 \cdot (4{,}6)^2 \cdot 177{,}4 - 3 \cdot (4{,}6)^4 = 1293{,}8 - 3264{,}16 + 22522{,}704 -$$
$$- 1343{,}2368 = 20209{,}1072.$$

Таким образом,

$$\boldsymbol{A} = \frac{\mu_3}{\sigma^3} = 0{,}394 \qquad \boldsymbol{E} = \frac{\mu_4}{\sigma^4} - 3 = 20206{,}1072.$$

Пример. НСВ X задана следующей плотностью распределения:

$$\boldsymbol{f(x)} = \begin{cases} 0 & \textbf{при } \boldsymbol{x} < -2, \\ -\boldsymbol{x}^3/4 & \textbf{при } -2 \le \boldsymbol{x} \le 0, \\ 0 & \textbf{при } \boldsymbol{x} > 0. \end{cases}$$

Найти коэффициент асимметрии и эксцесс.

Решение. Предварительно найдем начальные моменты до 4-го порядка

$$\nu_1 = \boldsymbol{M[X]} = -\frac{1}{4}\int_{-2}^{0} \boldsymbol{x}^4 \boldsymbol{dx} = -\frac{1}{4}\frac{\boldsymbol{x}^5}{5}\Big|_{-2}^{0} = -1{,}6,$$

$$\nu_2 = \boldsymbol{M[X^2]} = -\frac{1}{4}\int_{-2}^{0} \boldsymbol{x}^5 \boldsymbol{dx} = -\frac{1}{4}\frac{\boldsymbol{x}^6}{6}\Big|_{-2}^{0} = \frac{8}{3} \approx 2{,}67,$$

$$\nu_3 = \boldsymbol{M[X^3]} = -\frac{1}{4}\int_{-2}^{0} \boldsymbol{x}^6 \boldsymbol{dx} = -\frac{1}{4}\frac{\boldsymbol{x}^7}{7}\Big|_{-2}^{0} = \frac{32}{7} \approx -4{,}57,$$

$$\nu_4 = \boldsymbol{M[X^4]} = -\frac{1}{4}\int_{-2}^{0} \boldsymbol{x}^7 \boldsymbol{dx} = -\frac{1}{4}\frac{\boldsymbol{x}^8}{8}\Big|_{-2}^{0} = 8,$$

Теперь вычислим центральные моменты:

$$\mu_2 = \nu_2 - \nu_1^2 = \boldsymbol{D[X]} \approx 0{,}11 \quad \Rightarrow \quad \sigma = \sqrt{\boldsymbol{D[X]}} = 0{,}33,$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 \approx 0{,}054,$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_3 - 3\nu_1^4 \approx 0{,}1024.$$

Таким образом, $\boldsymbol{A} = \frac{\mu_3}{\sigma^3} \approx 1{,}5$ $\qquad$ $\boldsymbol{E} = \frac{\mu_4}{\sigma^4} - 3 \approx 5{,}46.$

Мода и медиана случайной величины

Наряду с математическим ожиданием, дисперсией и моментами для описания распределения случайной величины применяют также моду и медиану.

Модой *Mo* случайной величины *X* называется наиболее вероятное значение случайной величины.

Термин "наиболее вероятное значение" применим только к ДСВ, в случае НСВ мода совпадает с таким значением случайной величины, при котором плотность распределения имеет максимум. Различают унимодальные (имеющие одну моду), бимодальные (имеющие две моды) и мульти модальные (имеющие несколько мод) распределения. Иногда встречаются распределения, обладающие посередине не максимумом, а минимумом. Такие распределения называются анти модальными.

Мода, например, часто используется при экономических расчетах, когда нужно дать ответ на вопрос, каковы преобладающие в данный момент уровни производительности труда, себестоимость, какой товар имеет наибольший спрос и т.д. В связи с этим вводятся понятия модальная производительность, модальная себестоимость и т.д.

Медианой *Me* случайной величины *X* называется такое ее значение, для которой справедливо равенство

$$P(X<Me) = P(X>Me)$$

т.е. равновероятно, что случайная величина окажется меньше или больше медианы.

С геометрической точки зрения, медиана – это абсцисса точки, в которой площадь, ограниченная кривой распределения, делится пополам. Так как вся площадь должна равняться единице, то функция распределения в этой точке равна 0,5:

$$\boldsymbol{F(Me)} = \frac{1}{2}.$$

О**птимальное свойство медианы:** сумма абсолютных величин отклонений возможных значений случайной величины от медианы, умноженных на соответствующие вероятности, меньше, чем от любой другой величин, т.е. медианы, удовлетворяют условию:

$$\sum_{i=1}^{n} \left| x_i - Me \right| \cdot p_i = min.$$

Это свойство медианы, используется в теории оптимального проектирования. Например, при проектировании остановок, при условии, чтобы общий путь пассажиров был минимальным.

Следует отметить, что если распределение симметрично и унимодально, то математическое ожидание, медиана и мода совпадают.

Кроме моды и медианы иногда используются и другие числовые характеристики. Например,

α-квантилью Q_α ($0<\alpha<1$) случайной величины X называется число, удовлетворяющее неравенствам

$$\boldsymbol{P(X<Q_\alpha)\le\alpha} \quad \boldsymbol{и} \quad \boldsymbol{P(X<Q_\alpha)\le 1-\alpha}\,.$$

Квантили находят самое широкое применение в математической статистике при построении доверительных интервалов и проверке статистических гипотез. Отметим, ½-квантиль совпадает с медианой.

Энтропия $H=H(X)$ дискретной случайной величины X определяется по формуле

$$\boldsymbol{H(X)=-\sum_{i=1}^{n} p_i \log p_i}\,.$$

Энтропия не зависит от значений x_i от случайной величины X, а зависит только от вероятностей p_i, с которыми эти значения принимаются. Энтропия является мерой априорной неопределенности случайной величины. Максимального значения $H_{max}= \log n$ энтропия ДСВ достигает тогда, когда все n возможных значений случайная величина принимает с одной и той же вероятностью $p_i=1/n$, минимальное $H_{min}= 0$ – когда случайная величина принимает единственное значение с вероятностью, равной единице.

Энтропия играет важную роль в теории информации, она в некотором смысле представляет собой минимальный объем памяти, необходимый для записи информации, содержащейся в случайной величине. Поскольку информация записывается обычно в двоичной системе, то основание логарифма берется число 2.

Энтропия $H=H(X)$ непрерывной случайной величины X определяется по формуле

$$\boldsymbol{H(X)=-\int_{-\infty}^{\infty} f(x)\log f(x)dx}\,.$$

Заметим, что в отличие, например, от математического ожидания энтропию НСВ нельзя получить предельным переходом от дискретного случая. Отметим также, что при заданной дисперсии σ^2 максимальную энтропию $\log\sqrt{2\pi e\sigma^2}$ имеет нормально распределенная случайная величина.

Рассмотрим примеры нахождения приведенных характеристик

Пример 8. Найти начальные ν_k и центральные μ_k моменты первых трех порядков случайной величины X, заданной функцией распределения:

$$F(x)=\begin{cases}0 & \text{если } x\le 0\\ \dfrac{x^2}{4} & \text{если } 0<x\le 2\\ 1 & \text{если } x>2\end{cases}.$$

Решение. Найдем плотность распределения $CB\ X$:

$$f(x)=\begin{cases}0 & \text{если } x\le 0\\ 0{,}5x & \text{если } 0<x\le 2\\ 0 & \text{если } x>2\end{cases},$$

По формуле

$$\nu_k=\int\limits_{-\infty}^{+\infty}x^k f(x)\,dx=M(X^k),$$

определяющей начальные моменты ν_k случайной непрерывной величины X, заданной плотностью распределения $f(x)$, найдем ν_1, ν_2, ν_3.

$$\nu_1=\int\limits_0^2 x\,f(x)\,dx=\int\limits_0^2 x\cdot 0{,}5x\,dx=0{,}5\frac{x^3}{3}\Big|_0^2=\frac{4}{3};$$

$$\nu_2=\int\limits_0^2 x^2\,f(x)\,dx=\int\limits_0^2 x^2\cdot 0{,}5x\,dx=0{,}5\frac{x^4}{4}\Big|_0^2=2;$$

$$\nu_3=\int\limits_0^2 x^3\,f(x)\,dx=\int\limits_0^2 x^3\cdot 0{,}5x\,dx=0{,}5\frac{x^5}{5}\Big|_0^2=3{,}2.$$

Центральные моменты $\mu_к$ непрерывной *СВ X*, заданной плотностью распределения $f(x)$, найдем μ_1 , μ_2 , μ_3:

$$\mu_1 = \int_0^2 [x - M(X)] f(x)\ dx = \int_0^2 (x - \frac{4}{3}) \cdot 0{,}5x\ dx = 0;$$

Математическое ожидание отклонения равно нулю, первый центральный момент

$\mu_1 = M(X - M(X)) = 0.$

$$\mu_2 = \int_0^2 [x - M(X)]^2\ f(x)\ dx = \int_0^2 (x - \frac{4}{3})^2 \cdot 0{,}5x\ dx = \frac{2}{9};$$

$$\mu_3 = \int_0^2 [x - M(X)]^3\ f(x)\ dx = \int_0^2 (x - \frac{4}{3})^3 \cdot 0{,}5x\ dx = -\frac{8}{135}.$$

Ответ: $v_1 = 4/3$; $v_2 = 2$; $v_3 = 3{,}2$; $\mu_1 = 0$; $\mu_2 = 2/9$; $\mu_3 = -8/135$.

При решении задач можно пользоваться формулами, выражающие центральные моменты $\mu_к$ через начальные моменты $v_к$:

$\mu_1 = 0$; $\mu_2 = v_2 - (v_1)^2 = 2 - 16/9 = 2/9$;

$\mu_3 = v_3 - 3v_1v_2 + 2(v_1)^2 = 3{,}2 - 3 \cdot 4/3 \cdot 2 + 2\,(4/3)^3 = -8/135$;

$\mu_4 = v_4 - 4v_1v_2 + 6\,(v_1)^2 v_2 - 3(v_1)^4$ и т. д.

5 Законы распределения дискретных случайных величин

5.1 Биномиальное распределение

Рассмотрим систему большого количества n независимых испытаний случайной величины X, в каждом из которых событие A встречается k раз с одинаковой вероятностью $P(A) = p$. В остальных n-k случаях событие A не наступает, а наступает противоположное ему событие $\overline{A}$ с вероятностью q = 1-p, $P(\overline{A}) = q$. Так как наступление события A в каждом из k испытании имеют независимые результаты, то, пользуясь теоремой умножения вероятностей, можно определить наступление вероятности любого одного фиксированного события $p^k q^{n-k}$.

Число вероятностей всех фиксированных событий равно числу сочетаний из n по k, $C_n^k = \frac{n!}{k!(n-k)!}$. Вероятность $P_n(k)$ наступления события A k раз из n испытаний определится по формуле:

$$P_n(k) = P(X = k) = C_n^k p^k q^{n-k},$$

эта формула, называется формулой Бернулли.

Биномиальным называют закон распределения случайной дискретной величины X – числа появления события A в n независимых испытаниях, с вероятностью каждого испытания равным p. Вероятность того, что в результате проведенных n испытаний событие A произошло k раз есть $P_n(k)$:

$$P_n(k) = C_n^k p^k q^{n-k}, \text{ где } q=1-p.$$

Математическое ожидание $M(X)$ и дисперсия $D(X)$ случайной дискретной величины X, распределенной по биномиальному закону, имеет вид:

$$M(X) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = np;$$

$$D(X) = M(X^2) - [M(X)]^2 = \sum_{k=0}^{n} C_n^k p^k q^{n-k} - np^2 = npq.$$

Согласно закону больших чисел, можно с вероятностью близкой к достоверности утверждать, что при большом количестве испытаний относительная частота появления события близка к вероятности его появления в отдельном испытании. Тогда найдется, как угодно малое положительное число ε, для которого выполняется неравенство:

$$B = \left| \frac{k}{n} - p \right| < \varepsilon,$$

вероятность выполнения этого неравенства есть

$$\lim_{n\to\infty} P(B) = \lim_{n\to\infty} P\left(\left| \frac{k}{n} - p \right| \right) = 1.$$

РАСЧЁТНОГРАФИЧЕСКАЯ РАБОТА № 1

Биномиальное распределение

Цель работы – сформировать навыки построения биномиального закона распределения и вычисления числовых характеристик случайной величины средствами Excel.

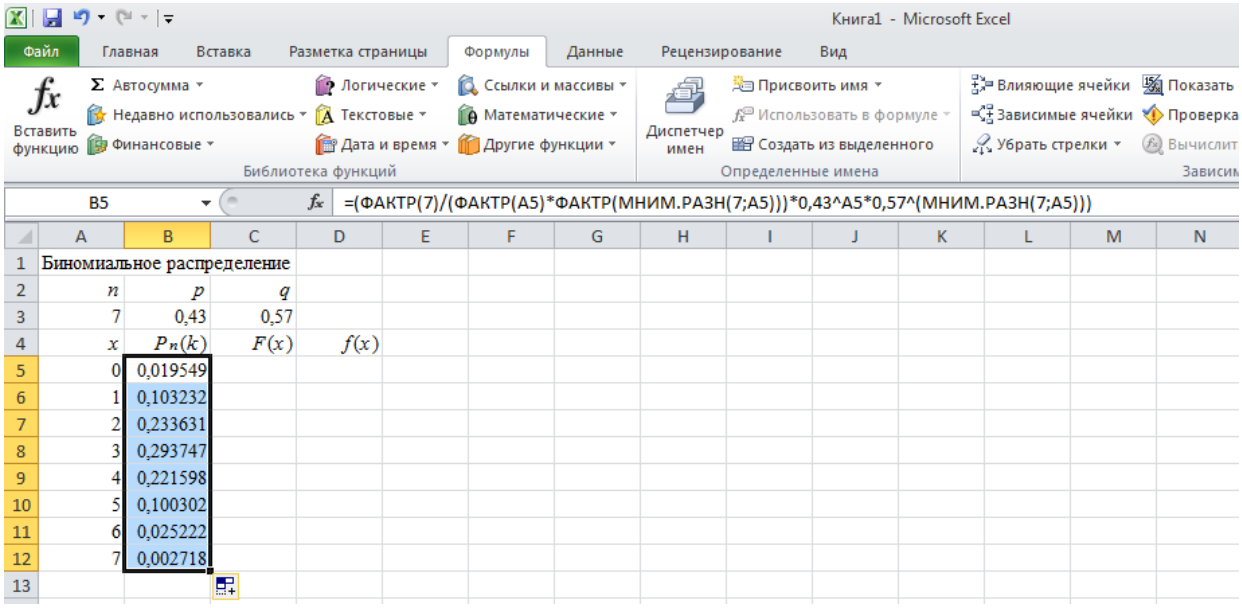
Задание

В серии одинаковых, независимых ***n*** испытаний вероятность успеха равна ***p***:

1) построить ряд распределения;
2) построить многоугольник распределения;
3) составить функцию распределения случайной величины ***X***;
4) записать аналитический вид этой функции, построить график;
5) найти плотность распределения;
6) построить графики функции распределения и гистограмму плотности распределения в Excel;
7) найти математическое ожидание, дисперсию, среднее квадратическое отклонение случайной величины ***X***;
8) вычислить вероятности:
   - трёх успехов $P(X=3)$;
   - хотя бы одного успеха $P(X \geq 1)$;
   - не более четырёх успехов $P(X \leq 4)$;
   - от 2 до 5 успехов $P(2 \leq X \leq 5)$.

1. Построим ряд распределения числа x значений случайной величины X с соответствующими вероятностями $P_n(k)$

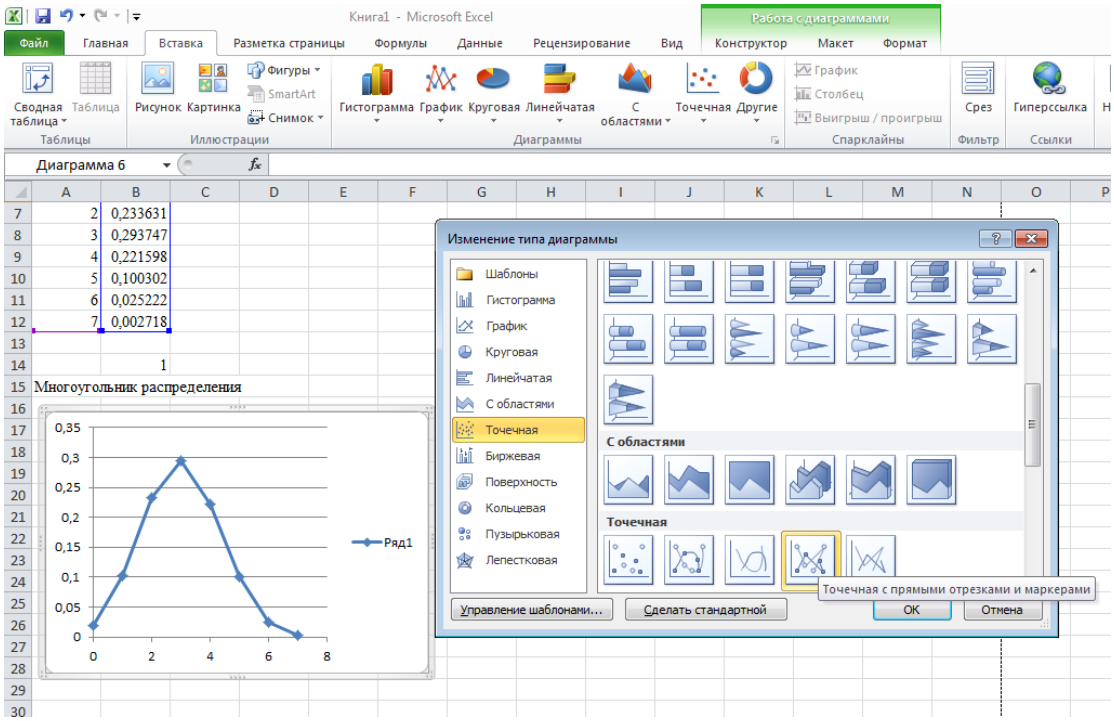
Составим формулу (как показано на рисунке) для расчёта вероятностей $P_n(k)$



	A	B	C	D
1	Биномиальное распределение			
2	n	p	q	
3	7	0,43	0,57	
4	x	$P_n(k)$	$F(x)$	$f(x)$
5	0	0,019549		
6	1	0,103232		
7	2	0,233631		
8	3	0,293747		
9	4	0,221598		
10	5	0,100302		
11	6	0,025222		
12	7	0,002718		
13				

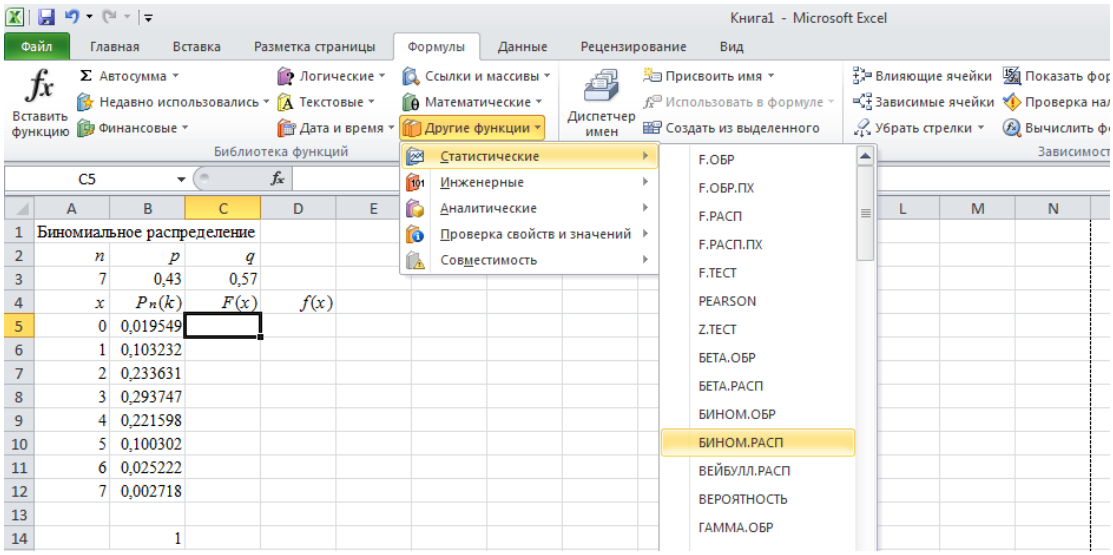
2. Построим многоугольник распределения

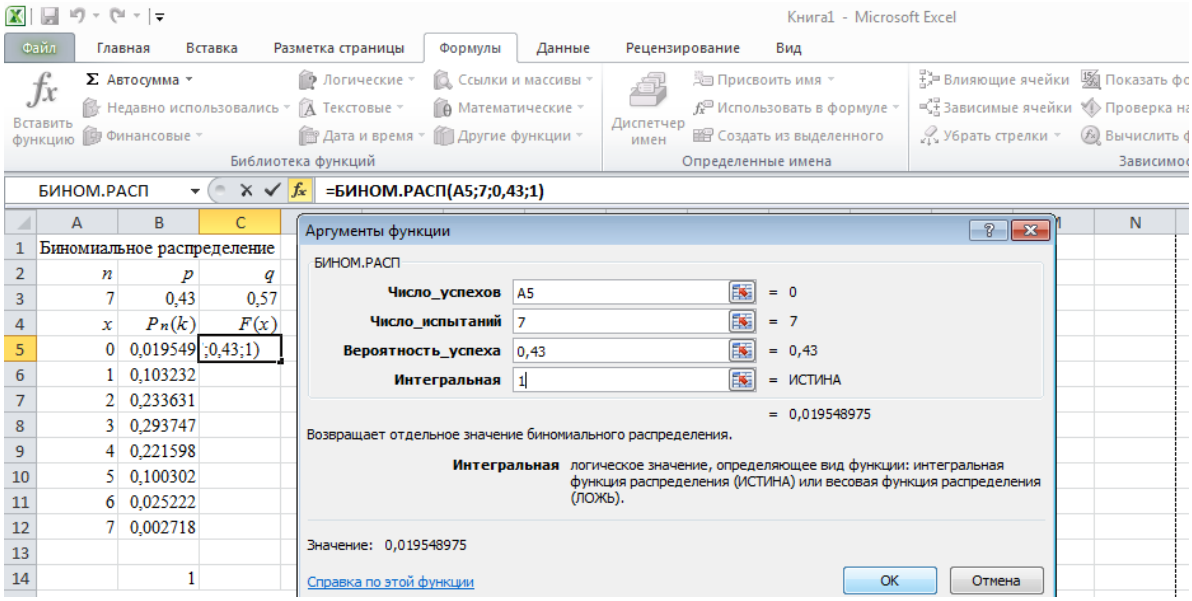
В главном меню выберите закладку **Вставка→График→Все типы диаграмм…→Точечная**, и далее – график с точками, соединенными прямыми линиями. **ОК.**



2. Составим функцию $F(x) = P(X \le x)$ распределения случайной величины X

Воспользуемся функцией **БИНОМ.РАСП()**



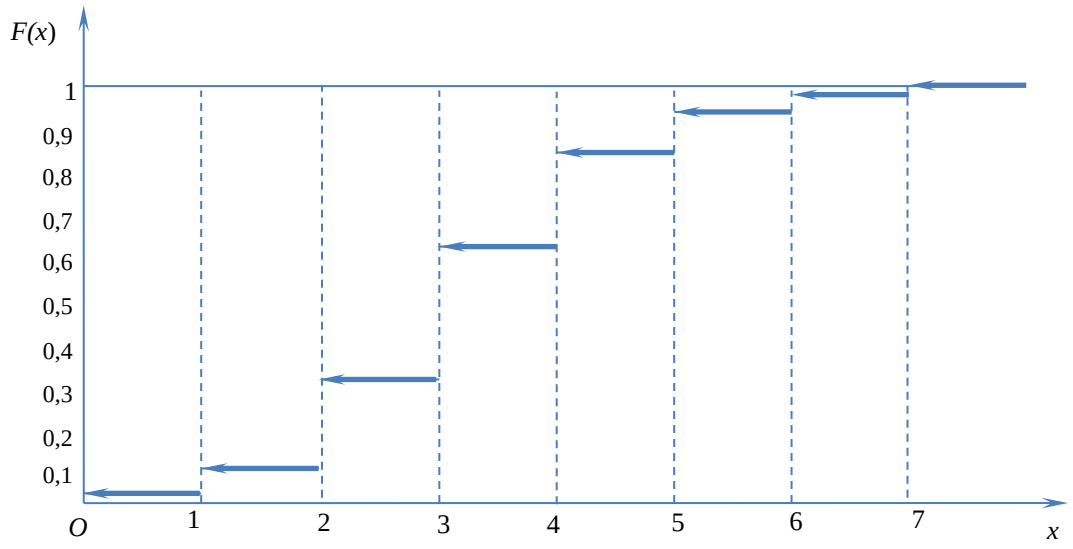


	A	B	C
1	Биномиальное распределение		
2	n	p	q
3	7	0,43	0,57
4	x	$P_n(k)$	$F(x)$
5	0	0,019549	0,019549
6	1	0,103232	0,122781
7	2	0,233631	0,356412
8	3	0,293747	0,650159
9	4	0,221598	0,871757
10	5	0,100302	0,97206
11	6	0,025222	0,997282
12	7	0,002718	1
13			
14		1	

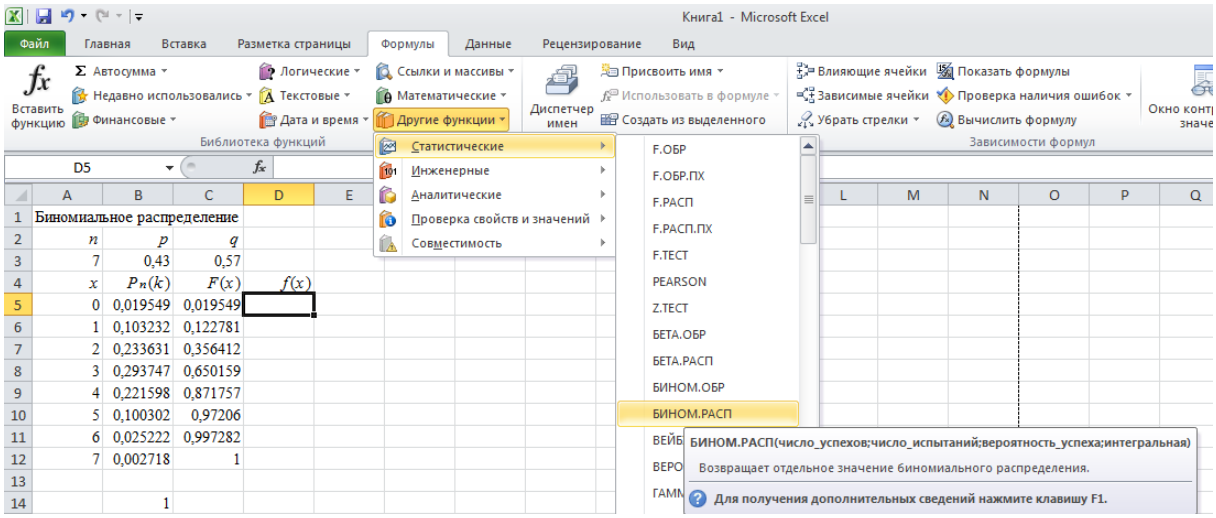
4. Аналитический вид функции распределения:

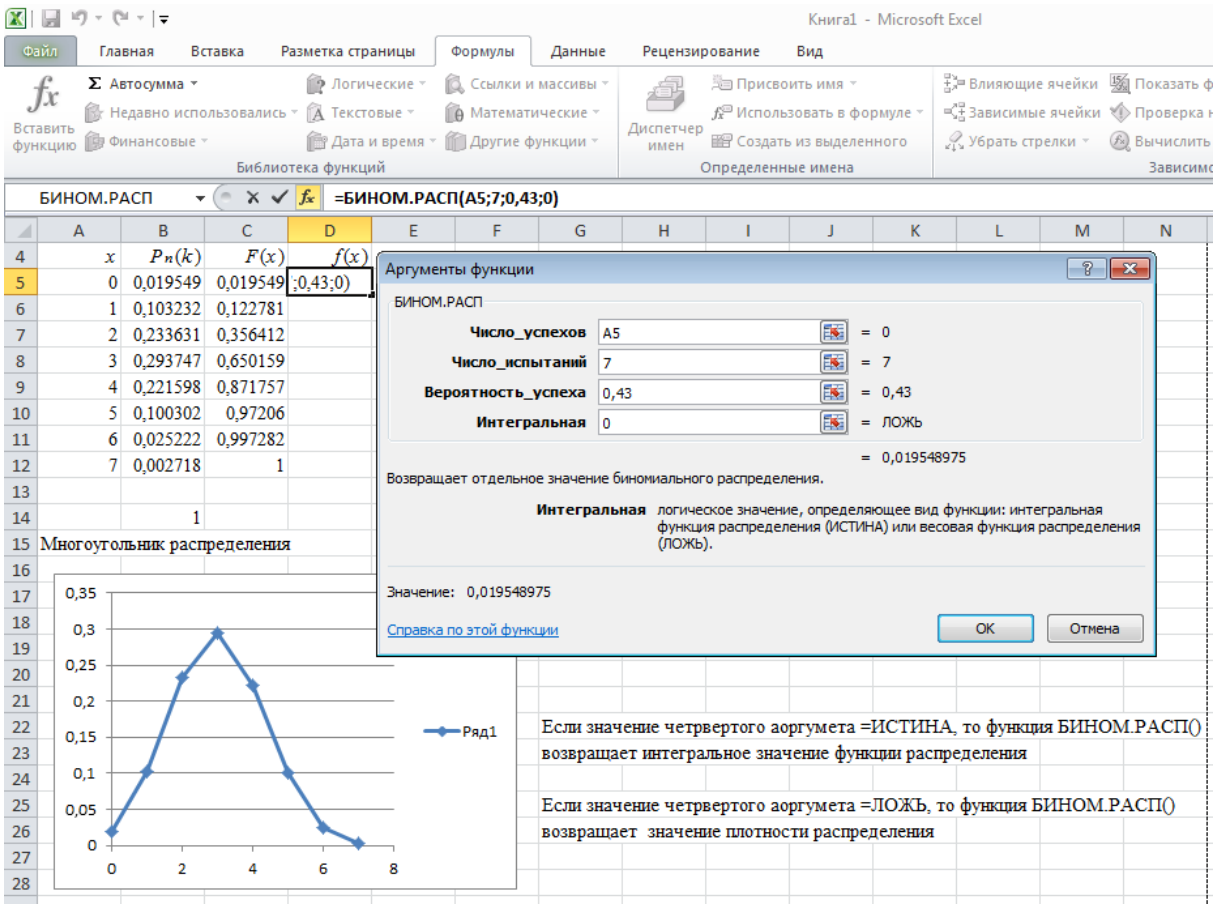
$$F(x)=\begin{cases}0, & x\le 0\\ 0{,}019549, & 0<x\le 1\\ 0{,}122781, & 1<x\le 2\\ 0{,}356412, & 2<x\le 3\\ 0{,}650159, & 3<x\le 4\\ 0{,}871757, & 4<x\le 5\\ 0{,}97206, & 5<x\le 6\\ 0{,}997282, & 6>x\le 7\\ 1, & x>7\end{cases}$$

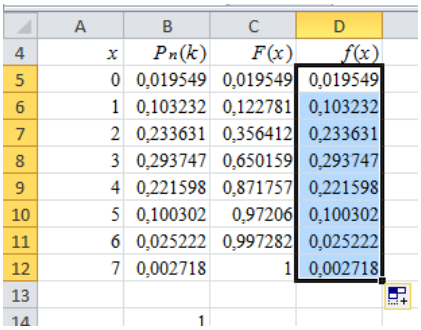
График функции распределения:



5. Найдем плотность распределения



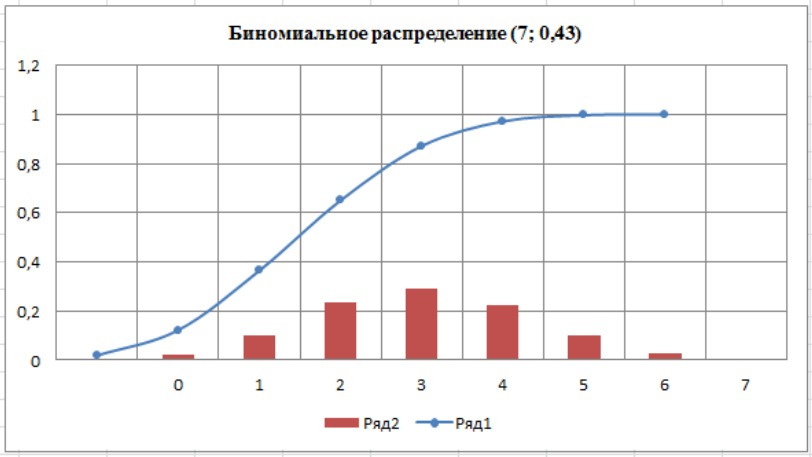




6. Построим графики функции распределения и гистограмму плотности распределения

Для построения интегральной функции распределения используем диаграмму типа ГРАФИК. Для плотности распределения используем диаграмму с группировкой.

x	$F(x)$	$f(x)$
0	0,019549	0,019549
1	0,122781	0,103232
2	0,365412	0,233631
3	0,650159	0,293747
4	0,871757	0,221598
5	0,97206	0,100302
6	0,997282	0,025222
7	1	0,002718



Ряд 1 - график функции распределения.
Ряд 2 - гистограмма плотности распределения.

7. Рассчитаем показатели распределения: математическое ожидание; дисперсию; среднее квадратическое отклонение

Математическое ожидание $M(X)$ и дисперсия $D(X)$ случайной дискретной величины X, распределенной по биномиальному закону, имеет вид:

$$M(X) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = np;$$

$$D(X) = M(X^2) - [M(X)]^2 = \sum_{k=0}^{n} C_n^k p^k q^{n-k} - np^2 = npq.$$

	A	B	C	D	E	F	G	H	I	J
1	Биномиальное распределение									
2	n	p	q							
3	7	0,43	0,57							
4	x	$P_n(k)$	$F(x)$	$f(x)$			$M(X)$	$D(X)$	$\sigma(X)$	
5	0	0,019549	0,019549	0,019549			3,01	1,7157	1,309847	
6	1	0,103232	0,122781	0,103232						
7	2	0,233631	0,356412	0,233631						
8	3	0,293747	0,650159	0,293747						
9	4	0,221598	0,871757	0,221598						
10	5	0,100302	0,97206	0,100302						
11	6	0,025222	0,997282	0,025222						
12	7	0,002718	1	0,002718						
13										

8. Вычислить вероятности:

- трёх успехов $P(X=3)$;
- хотя бы одного успеха $P(X \geq 1)$;
- не более четырёх успехов $P(X \leq 4)$;
- от 2 до 5 успехов $P(2 \leq X \leq 5)$.

1) вероятность появления трех успехов $P(X=3)=0{,}293747$.

1	Биномиальное распределение			
2	n	p	q	
3	7	0,43	0,57	
4	x	$P_n(k)$	$F(x)$	$f(x)$
5	0	0,019549	0,019549	0,019549
6	1	0,103232	0,122781	0,103232
7	2	0,233631	0,356412	0,233631
8	3	0,293747	0,650159	0,293747
9	4	0,221598	0,871757	0,221598
10	5	0,100302	0,97206	0,100302
11	6	0,025222	0,997282	0,025222
12	7	0,002718	1	0,002718

2) определим вероятность появления хотя бы одного успеха $P(X \geq 1)$:

просуммируем B6;B12, $P(X \geq 1) = 0{,}980451$;

3) определим вероятность появления не более четырех успехов, просуммируем B5;B9

$P(X \leq 4) = 0{,}871757$.

4) определим вероятность появления от двух до пяти успехов, просуммируем B7;B10

$P(2 \leq X \leq 5) = 0{,}849278$.

ЗАДАНИЯ ДЛЯ РАСЧЁТНОЙ РАБОТЫ № 1

B	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
n	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
p	0,43	0,3	0,45	0,35	0,55	0,45	0,35	0,47	0,65	0,57	0,47	0,49	0,53	0,56	0,34

Найти вероятности: $P(X=3)$, $P(X \geq 1)$, $P(X \leq 4)$, $P(2 \leq X \leq 5)$.

5.2 Распределение Пуассона

Формула Пуассона:

$P_n(k) = C_n^k p^k q^{n-k} \cong \frac{a^k e^{-a}}{k!}$, где $a = np$ – параметр Пуассона.

Закон Пуассона рассматривается как предельный случай биноминального распределения, когда вероятность p осуществления некоторого события в единичном опыте очень мала ($p \to 0$), но число экспериментов n, производимых в единицу времени, достаточно велико ($n \to \infty$), то произведение $np \to a$ (где a – некоторая положительная постоянная величина).

Выведем формулу Пуассона, учитывая, что $q=1-p$, $p = \frac{a}{n}$

$$P_n(k) = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!}\frac{a^k}{n^k}\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k} = \frac{a^k}{k!}\frac{1}{n^k}\frac{n(n-1)(n-2)\cdot\ldots\cdot(n-k)!}{(n-k)!}\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k} =$$

$$= \frac{a^k}{k!}\frac{1}{n^k}\frac{n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)}{1}\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k} = \frac{a^k}{k!}\frac{1}{n^k}n\,n\left(1-\frac{1}{n}\right)n\left(1-\frac{2}{n}\right)\cdot\ldots\cdot n\left(1-\frac{k-1}{n}\right)\cdot$$

$$\cdot\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k} = \frac{a^k}{k!}\frac{1}{n^k}n^k\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k} =$$

$$= \frac{a^k}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k}.$$

Так как вероятность p осуществления некоторого события в единичном опыте очень мала ($p \to 0$), но число экспериментов n, производимых в единицу времени, достаточно велико ($n \to \infty$), перейдем к пределу:

$$P_n(k) = C_n^k p^k q^{n-k} \cong \frac{a^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\left(1-\frac{a}{n}\right)^n\left(1-\frac{a}{n}\right)^{-k}\right] \cong$$

$$\cong \frac{a^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdot\ldots\cdot\left(1-\frac{k-1}{n}\right)\right]\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^n\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^{-k} \cong$$

$$\cong \frac{a^k}{k!}\cdot 1\cdot\left(\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^{-\frac{n}{a}}\right)^{-a}\cdot 1 \cong \frac{a^k}{k!}\cdot e^{-a}.$$

При большом числе n испытаний и малом значении вероятности p появления события в каждом испытании k используется приближенная формула распределения случайной величины, которая называется **законом распределения Пуассона:**

$$P_n(k) = P(X = k) = \frac{a^k}{k!}e^{-a},$$

где параметр $a = np$ – показывает среднее число появлений событий в n испытаниях.

Математическое ожидание $M(X)$ и дисперсия $D(X)$, случайной дискретной величины, распределенной по закону Пуассона, имеет вид: $M(X) \approx D(X) = np$, то есть распределение Пуассона используется тогда, когда математическое ожидание приближенно равно дисперсии,

$$np \approx npq.$$

Ряд распределения случайной величины X, распределённой по закону Пуассона имеет вид:

x_k	0	1	2	…	k	…
p_k	e^{-a}	$\frac{a}{1!}e^{-a}$	$\frac{a^2}{2!}e^{-a}$	…	$\frac{a^k}{k!}e^{-a}$	…

Убедимся, что последовательность вероятностей, может представлять собой ряд распределения, т.е. что сумма всех вероятностей P_k равна единице.

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a} = e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!}.$$

Разложим функцию e^x в ряд Маклорена:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \text{ положив } x=a,$$

$$e^a = 1 + \frac{a}{1!} + \frac{a^2}{2!} + \ldots = \sum_{k=0}^{\infty} \frac{a^k}{k!},$$

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a} = e^a e^{-a} = 1.$$

Рассмотрим числовые характеристики случайной величины X, распределенной по закону Пуассона

Математическое ожидание

$$M(X) = \sum_{k=0}^{\infty} k \frac{a^k}{k!} e^{-a} = a e^{-a} \sum_{k=0}^{\infty} \frac{a^{k-1}}{(k-1)!} = a e^{-a} e^a = a.$$

Таким образом, параметр a представляет собой математическое ожидание случайной величины X.

Дисперсия

Дисперсией случайной величины X называют математической ожидание квадрата отклонения случайной величины от её математического ожидания:

$$D(X) = M\left[X - M(X)\right]^2,$$

$$D(X) = M(X^2) - M^2(X),$$

$$M(X^2) = \sum_{k=0}^{\infty} k^2 \frac{a^k}{k!} e^{-a} = a \sum_{k=0}^{\infty} k \frac{a^{k-1}}{(k-1)!} e^{-a} = a \sum_{k=0}^{\infty} ((k-1)+1) \frac{a^{k-1}}{(k-1)!} e^{-a} =$$

$$= a \left[\sum_{k=0}^{\infty} (k-1) \frac{a^{k-1}}{(k-1)!} e^{-a} + \sum_{k=0}^{\infty} \frac{a^{k-1}}{(k-1)!} e^{-a} \right] = a\left[a+1\right],$$

$$D(X) = M(X^2) - M^2(X) = a[a+1] - a^2 = a^2 + a - a^2 = a.$$

Таким образом, $M(X)=D(X)=a$, то есть дисперсия случайной величины, распределенной по закону Пуассона, равна ее математическому ожиданию a.

Это свойство распределения Пуассона часто применяют на практике для решения вопроса, правдоподобна ли гипотеза о том, что случайная величина распределена по закону Пуассона. Для этого определяют математическое ожидание и дисперсию случайной величины. Если их значения близки, то гипотеза о том, что случайная величина распределена по закону Пуассона, принимается; резкое различие этих характеристик, свидетельствует против подобной гипотезы.

Среднее квадратическое отклонение

Среднее квадратическое отклонение для случая, когда случайная величина X распределена по закону Пуассона, определяется по формуле: $\sigma(X) = \sqrt{a}$.

Пример. Получен закон распределения количества бракованных изделий в партии:

Количество брака, k_i	0	1	2	3	4	Итого
Количество партий с бракованными изделиями, f_i	604	306	77	12	1	1000

Проверить, соответствует ли эмпирическое распределение распределению Пуассона.

Решение.

Определим среднее число бракованных изделий в партии:

$$a = \frac{604 \cdot 0 + 306 \cdot 1 + 77 \cdot 2 + 12 \cdot 3 + 4 \cdot 1}{1000} = 0{,}5.$$

Определим теоретические значения параметра Пуассона:

$a_i = p_i\, n$

$$a_1 = \frac{0{,}5^0}{0!} e^{-0{,}5} \cdot 1000 \cong 0{,}607 \cdot 1000 \cong 606,$$

$$a_2 = \frac{0{,}5^1}{1!} e^{-0{,}5} \cdot 1000 \cong 0{,}5 \cdot 0{,}607 \cdot 1000 \cong 303,$$

$$a_3 = \frac{0{,}5^2}{2!} e^{-0{,}5} \cdot 1000 \cong 0{,}125 \cdot 0{,}607 \cdot 1000 \cong 76,$$

$$a_4 = \frac{0{,}5^3}{3!} e^{-0{,}5} \cdot 1000 \cong 0{,}021 \cdot 0{,}607 \cdot 1000 \cong 13,$$

$$a_5 = \frac{0{,}5^4}{4!} e^{-0{,}5} \cdot 1000 \cong 0{,}0026 \cdot 0{,}607 \cdot 1000 \cong 2.$$

Количество партий с бракованными изделиями, f_i	604	306	77	12	1	1000
Теоретические значения, f_i	606	303	76	13	2	1000

Таким образом, эмпирическое распределение количества бракованных изделий в партии соответствует распределению Пуассона.

6. Непрерывные распределения случайной величины

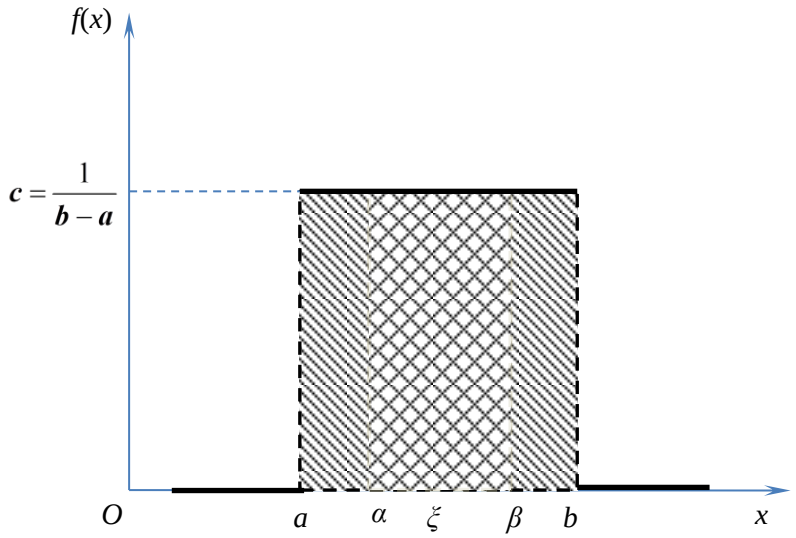
6.1 Равномерное распределение

Равномерное распределение случайной величины является самым простым из всех законов распределения непрерывных случайных величин.

Равномерным распределением непрерывной случайной величины называется распределение, в котором значения x случайной величины X лежат в некотором определённом интервале (a, b) и в пределах этого интервала имеют одинаковую вероятность.

Площадь под кривой плотности распределения равна 1. В нашем случае – это площадь прямоугольника с основанием $(b - a)$ и высотой c, поэтому равномерное распределение иногда называют «прямоугольным».

$$(b-a)c = 1,\ \boldsymbol{c = f(x) = \frac{1}{b-a}}.$$



Это означает, что на данном интервале плотность вероятности постоянна и определяется по формуле: $f(x) = \begin{cases} f(x) = 0, \ \textit{если} \quad x < a; \\ f(x) = \dfrac{1}{b-a}, \ \textit{если} \quad a \le x \le b; \\ f(x) = 0, \ \textit{если} \quad x > b. \end{cases}$

Вероятность попадания случайной величины X, равномерно распределенной на (a, b), на любую часть $(\alpha, \beta) \subset (a, b)$, определяется как площадь дважды заштрихованной области и находится по формуле:

$$P(\alpha < x < \beta) = \frac{\beta - \alpha}{b - a}.$$

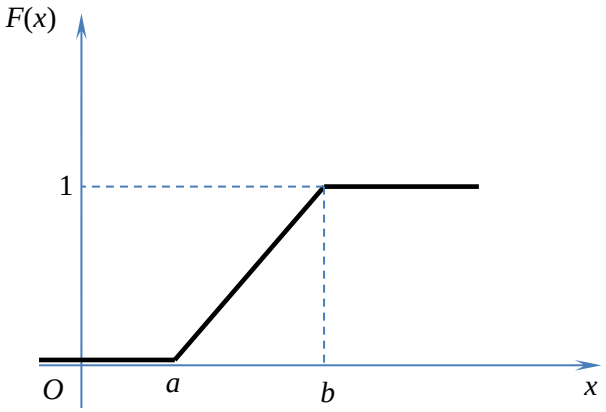
Найдём функцию распределения по известной плотности распределения по формуле:

$$F(x)=\int_{-\infty}^{x} f(x)\,dx,$$

$$F(x)=\int_{a}^{x}\frac{1}{b-a}\,dx=\frac{1}{b-a}\,x\Big|_{a}^{x}=\frac{x-a}{b-a},\ a\le x\le b.$$

Функция распределения $F(x)$ непрерывной случайной величины, равномерно распределённой на интервале (a, b) имеет вид:

$$F(x)=\begin{cases}F(x)=0,\ \textit{если}\ \ x<a;\\ F(x)=\dfrac{x-a}{b-a},\ \textit{если}\ \ a\le x\le b;\\ F(x)=1,\ \textit{если}\ \ x>b.\end{cases}$$



Случайные величины, имеющие равномерное распределение вероятностей, часто встречаются на практике. Например, при снятии показаний измерительных приборов. Ошибка при округлении отсчёта до ближайшего целого деления шкалы является случайной величиной, которая может с постоянной плотностью вероятности принимать любые значения между двумя соседними делениями. Таким образом, данная случайная величина имеет равномерное распределение.

Числовые характеристики случайной величины *X*, равномерно распределённой на интервале (*a*, *b*)

Математическое ожидание

$$\xi=M(X)=\int_{a}^{b} x\,f(x)\,dx=\frac{1}{b-a}\int_{a}^{b} x\,dx=\frac{1}{b-a}\,\frac{x^2}{2}\Big|_{a}^{b}=\frac{b+a}{2}.$$

Величина $\xi=\dfrac{a+b}{2}$, равная середине (a, b), определяет математическое ожидание случайной величины X.

Дисперсия

$$D(X)=\int_a^b (x-M(X))^2 f(x)dx=\int_a^b \left(x-\frac{a+b}{2}\right)^2 \frac{1}{b-a}dx=\frac{1}{b-a}\int_a^b \left(x-\frac{a+b}{2}\right)^2 dx=$$

$$=\left(t=x-\frac{a+b}{2},\ dt=dx;\ t_b=\frac{2b-b-a}{2}=\frac{b-a}{2},\ t_a=\frac{2a-b-a}{2}=-\frac{b-a}{2}\right)=$$

$$=$$

$$\frac{1}{b-a}\frac{t^3}{3}\Bigg|_{-\frac{b-a}{2}}^{\frac{b-a}{2}}=\frac{1}{b-a}\left(\frac{(b-a)^3}{24}+\frac{(b-a)^3}{24}\right)=\frac{1}{b-a}\frac{2(b-a)^3}{24}=\frac{(b-a)^2}{12}.$$

Дисперсия: $D(X)=\frac{(b-a)^2}{12}$.

Среднее квадратическое отклонение (стандартное отклонение):

$$\sigma(X)=\sqrt{D(X)}=\frac{b-a}{2\sqrt{3}}.$$

Пример. Случайная величина *X* распределена равномерно на интервале (*a*, *b*). Найти вероятность того, что в результате опыта она отклонится от своего математического ожидания больше, чем на 3σ.

Решение.

Определим стандартное отклонение:

$$\sigma(X)=\sqrt{D(X)}=\frac{b-a}{2\sqrt{3}}.$$

$$3\sigma=\frac{3(b-a)}{2\sqrt{3}}=\frac{\sqrt{3}(b-a)}{2},$$

При равномерном распределении на участке (*a*, *b*) крайние точки *a* и *b*, ограничивающие участок возможных значений случайной величины, отстоят от её математического ожидания *M* (*X*) на расстояние $\frac{b+a}{2}$, которое меньше, чем $\frac{\sqrt{3}(b-a)}{2}$. Следовательно, вероятность того, что в результате опыта случайная величина *X* отклонится от своего математического ожидания больше, чем на 3σ, равна нулю.

Пример. Светофор работает в двух режимах: 1 минуту горит зелёный; 0,5 минут – красный. Водитель подъезжает к перекрёстку в случайный момент времени. С какой вероятностью он проедет перекрёсток без остановки. Составить закон распределения случайной величины *T* времени ожидания у перекрёстка и найти его числовые характеристики.

Решение.

Период переключения цветов равен 1,5 мин.

Проехать через перекрёсток без остановки можно за интервал (0, 1).

Вероятность $\boldsymbol{P(t \in (0, 1))} = \int_0^1 \boldsymbol{p(\tau)\, d\, \tau} = \frac{2}{3} \int_0^1 \boldsymbol{d\, \tau} = \frac{2}{3}$.

Вероятность остановки $\boldsymbol{q} = 1 - \frac{2}{3} = \frac{1}{3}$. Пассажиру нужно будет ждать не более одной минуты с вероятностью $\frac{1}{3}$.

6.2 Нормальное распределение (распределение Гаусса)

Одним из фундаментальных законов распределения случайной непрерывной величины X, является **закон нормального распределения (распределение Гаусса)**.

Нормальному закону подчиняются только непрерывные случайные величины.

Нормальное распределение случайной величины X является предельным для других распределений, то есть целый ряд различных распределений сводятся к нормальному распределению при бесконечном повторении числа испытаний.

Выведем формулу для нормального распределения

Пусть n и np велики.

Используем формулу Стирлинга:

$$\boldsymbol{n!} = \sqrt{2\pi \boldsymbol{n}} \left(\frac{\boldsymbol{n}}{\boldsymbol{e}}\right)^{\boldsymbol{n}}.$$

$$P_n(k) = C_n^k\, p^k\, q^{n-k} = \frac{n!}{k!(n-k)!}\, p^k\, q^{n-k} = \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}}\, p^k\, q^{n-k} =$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{\left(\frac{n}{e}\right)^n}{\left(\frac{k}{e}\right)^k \left(\frac{n-k}{e}\right)^{n-k}}\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{e^n \frac{k^k}{e^k} \frac{(n-k)^{n-k}}{e^n} e^k}\, p^k\, q^{n-k} =$$

$$= \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}}\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^{n-k+k}}{k^k (n-k)^{n-k}}\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{pn}{k}\right)^k \left(\frac{qn}{n-k}\right)^{n-k}.$$

Ожидаемо, что «пик» значения вероятности $P_n(k)$

будет вблизи $k = np \Rightarrow k - np = \xi \Rightarrow k = \xi + np$.

$n\text{-}k = n - \xi - np = n(1-p) - \xi = nq - \xi$.

$n\text{-}k = nq - \xi$.

$k = \xi + np$.

$$P_n(k) = C_n^k\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{pn}{k}\right)^k \left(\frac{qn}{n-k}\right)^{n-k} = \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}} \left(\frac{pn}{np+\xi}\right)^{np+\xi} \left(\frac{qn}{nq-\xi}\right)^{nq-\xi} =$$

$$= \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}} \left(\frac{1}{1+\frac{\xi}{np}}\right)^{np+\xi} \left(\frac{1}{1-\frac{\xi}{nq}}\right)^{nq-\xi} = \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}}\, \frac{1}{\left(1+\frac{\xi}{np}\right)^{np+\xi} \left(1-\frac{\xi}{nq}\right)^{nq-\xi}}.$$

$$C_n^k\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}}\,\frac{1}{\left(1+\frac{\xi}{np}\right)^{np+\xi}\left(1-\frac{\xi}{nq}\right)^{nq-\xi}}.$$

Используем формулу разложения в ряд Маклорена функции:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^{n-1}\frac{x^n}{n}.$$

$$\mathbf{ln}\left[\left(1+\frac{\xi}{\boldsymbol{np}}\right)^{\boldsymbol{np}+\xi}\left(1-\frac{\xi}{\boldsymbol{nq}}\right)^{\boldsymbol{nq}-\xi}\right] = \mathbf{ln}\left(1+\frac{\xi}{\boldsymbol{np}}\right)^{\boldsymbol{np}+\xi} + \mathbf{ln}\left(1-\frac{\xi}{\boldsymbol{nq}}\right)^{\boldsymbol{nq}-\xi} =$$

$$= (\boldsymbol{np}+\xi)\,\mathbf{ln}\left(1+\frac{\xi}{\boldsymbol{np}}\right) + (\boldsymbol{nq}-\xi)\,\mathbf{ln}\left(1-\frac{\xi}{\boldsymbol{nq}}\right) =$$

$$= (\boldsymbol{np}+\xi)\left[\frac{\xi}{\boldsymbol{np}} - \frac{\xi^2}{2(\boldsymbol{np})^2} + \ldots\right] + (\boldsymbol{nq}-\xi)\left[-\frac{\xi}{\boldsymbol{nq}} - \frac{\xi^2}{2(\boldsymbol{nq})^2} - \ldots\right] =$$

$$= \xi - \frac{\xi^2}{2\boldsymbol{np}} + \ldots + \frac{\xi^2}{\boldsymbol{np}} - \frac{\xi^3}{2(\boldsymbol{np})^2} + \ldots - \xi - \frac{\xi^2}{2\boldsymbol{nq}} - \ldots + \frac{\xi^2}{\boldsymbol{nq}} - \frac{\xi^3}{2(\boldsymbol{nq})^2} + \ldots =$$

$$= \frac{\xi^2}{\boldsymbol{n}}\left(\frac{1}{\boldsymbol{p}}+\frac{1}{\boldsymbol{q}}\right) - \frac{\xi^2}{2\boldsymbol{n}}\left(\frac{1}{\boldsymbol{p}}+\frac{1}{\boldsymbol{q}}\right) + \frac{\xi^3}{2\boldsymbol{n}^2}\left(\frac{1}{\boldsymbol{p}^2}+\frac{1}{\boldsymbol{q}^2}\right) - \ldots$$

$$\ln\left[\left(1+\frac{\xi}{np}\right)^{np+\xi}\left(1-\frac{\xi}{nq}\right)^{nq-\xi}\right] = \frac{\xi^2}{n}\left(\frac{1}{p}+\frac{1}{q}\right) - \frac{\xi^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right) + \frac{\xi^3}{2n^2}\left(\frac{1}{p^2}+\frac{1}{q^2}\right) - \ldots \cong$$

$$\cong \frac{\xi^2}{n}\left(\frac{1}{p}+\frac{1}{q}\right) - \frac{\xi^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right) \cong \left(\frac{\xi^2}{n} - \frac{\xi^2}{2n}\right)\left(\frac{1}{p}+\frac{1}{q}\right) = \frac{\xi^2}{2n}\left(\frac{p+q}{pq}\right) = \frac{\xi^2}{2npq}.$$

$$\ln\left[\left(1+\frac{\xi}{np}\right)^{np+\xi}\left(1-\frac{\xi}{nq}\right)^{nq-\xi}\right] \cong \frac{\xi^2}{2npq}.$$

$$\left(1+\frac{\xi}{np}\right)^{np+\xi}\left(1-\frac{\xi}{nq}\right)^{nq-\xi} \cong e^{\frac{\xi^2}{2npq}}.$$

$$C_n^k\, p^k\, q^{n-k} = \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}}\,\frac{1}{e^{\frac{\xi^2}{2npq}}} = \sqrt{\frac{n}{2\pi(np+\xi))(nq-\xi)}}\;\; e^{-\frac{\xi^2}{2npq}}.$$

Так как $np+\xi \cong np$; $nq-\xi \cong nq$

$$C_n^k\, p^k\, q^{n-k} = \sqrt{\frac{1}{2\pi npq}}\;\; e^{-\frac{\xi^2}{2npq}} = \frac{1}{\sqrt{2\pi npq}}\, e^{-\frac{\xi^2}{2npq}}.$$

Используя формулу $\sigma^2 = npq$ и учитывая, что $k - np = \xi$, окончательно,

$$P_n(k) = C_n^k \, p^k \, q^{n-k} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-a)^2}{2\sigma^2}} .$$

Полученная формула вероятности появления k исходов в n испытаниях называется распределением Гаусса или нормальным распределением.

Функция распределения для нормального закона имеет вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt ,$$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-m)^2}{2\sigma^2}}$ – дифференциальная функция распределения непрерывной случайной величины, где $m = M(X)$.

Функцию распределения по известной плотности распределения для нормального распределения можно найти по формуле:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt.$$

Составленный интеграл относится к «неберущимся», поэтому рассматривают случай, когда: $\boldsymbol{a=M(x)=0}$ **и** $\boldsymbol{\sigma=1}$.

Тогда, $$f(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}}$$

Нормальное распределение с параметрами $\boldsymbol{a=M(x)=0}$ **и** $\boldsymbol{\sigma=1}$ называется **нормированным**, а его функция распределения называется **функцией Лапласа**:

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}} dt.$$

Таким образом, нормальное распределение определяется двумя параметрами: m и σ.

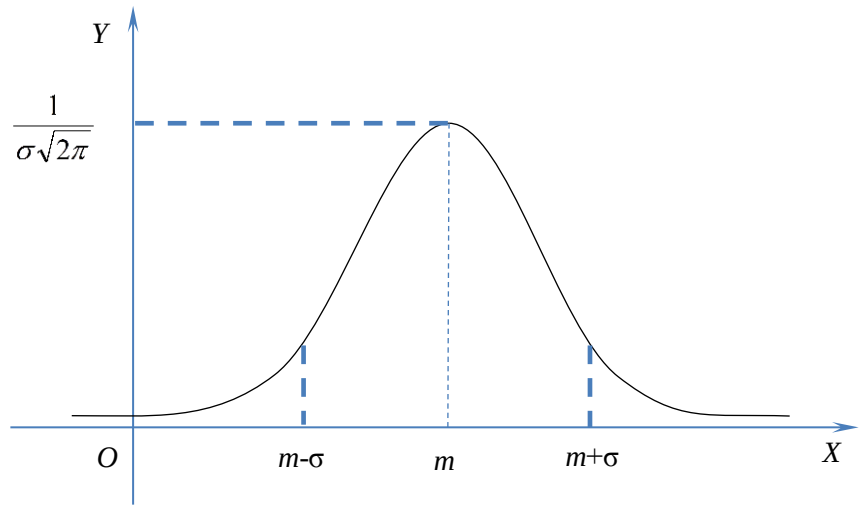
Областью определения этой функции является вся числовая прямая $(-\infty, \infty)$.

Для любого x, $f(x) > 0$, график расположен выше оси Ox.

Ось Ox является горизонтальной асимптотой графика функции, так как

$\lim\limits_{x\to\infty} f(x) = 0.$

График плотности нормального распределения называют **нормальной кривой (кривой Гаусса)**:



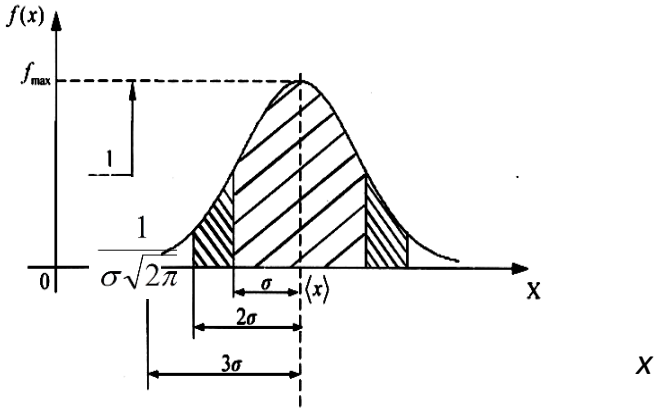
Для практических расчётов используется таблица (ПРИЛОЖЕНИЕ 1) значений функции Лапласа $\Phi(x)$,

$$F(x)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2\sigma^2}}dt=F(x)-0{,}5.$$

Тогда вероятность попадания случайной величины в интервал (a, b) определится по формуле:

$$P(a\le X\le b)=F\left(\frac{b-m}{\sigma}\right)-F\left(\frac{a-m}{\sigma}\right)=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right).$$

Функция нормального распределения имеет вид колокола:
На рисунке показано, что в области $-\sigma < x < \sigma$ на графике сосредоточено 68% площади распределения, в области $-2\sigma < x < 2\sigma$ на графике сосредоточено 95.4% площади распределения, в области $-3\sigma < x < 3\sigma$ на графике сосредоточено 99.7% площади распределения



«правило трех сигм».

Лекция 2

Методы интерполяции и аппроксимации

Лекция 3. Корреляционный-регрессионный анализ

Эконометрика – научная дисциплина, объединяющая совокупности теоретических результатов, методов и моделей, необходимых для количественного описания качественных закономерностей экономической теории.

Этапы эконометрического моделирования:

1) спецификация – представление экономической модели в математической форме;

2) параметризация – оценка параметров построенной модели;

3) верификация – проверка качества модели в целом;

4) прогнозирование или предсказание по построенной модели.

1 Этап спецификации

Спецификация модели – это формулировка вида модели в математической форме, исходя из соответствующей теории связи между рассматриваемыми признаками (переменными).

Признак – показатель, характеризующий некоторое свойство объекта совокупности, рассматриваемый как случайная величина.

Признаки (переменные) делятся на два класса:

- факторные (X_i);
- результативные (Y).

Факторные признаки – это независимые признаки, оказывающие влияние на другие, связанные с ними признаки.

Результативные признаки – это зависимые признаки, которые изменяются под влиянием факторных признаков.

Так, квалификация, стаж работы рабочего – факторные признаки; производительность труда – результативный.

Задача 1

Для анализа зависимости объёма потребления Y домохозяйства от располагаемого дохода X отобрана выборка объёма n=12. Необходимо определить вид зависимости.

На оси OY помещаем ту переменную, соответствующую результативному признаку Y, а на горизонтальной оси OX – переменную-фактор X. Их совместные значения отображаются на координатной плоскости точками.

X	Y
107	102
109	105
110	108
113	110
120	115
122	117
123	119
128	125
136	132
140	130
145	141
150	144

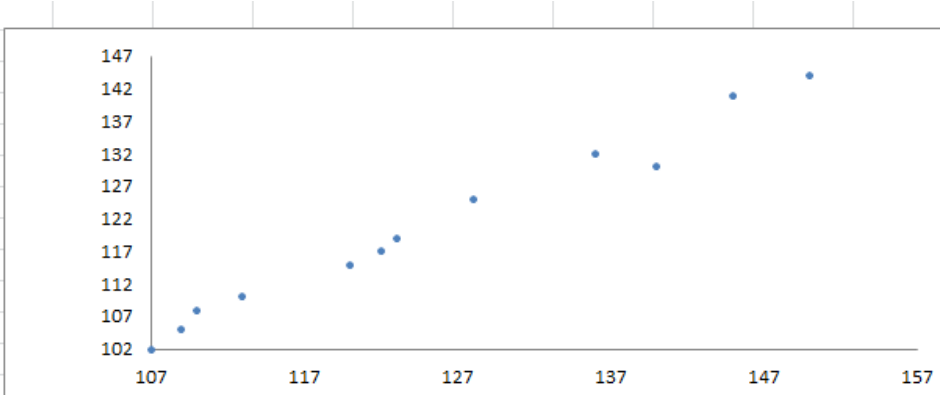


Рисунок 1 – диаграмма рассеяния (корреляционное поле)

Для построения диаграммы рассеивания в MS Excel используем Мастер диаграмм.

На рисунке 1 облако точек вытянуто из нижнего левого в правый верхний угол, это признак положительной зависимости, то есть с ростом доходов, объём потребления растет.

Когда первая переменная принимает высокие значения, вторая принимает низкие, то облако точек вытягивается из верхнего левого угла в нижний правый – это признак отрицательной зависимости.

X	Y
150	102
145	105
140	108
136	110
128	115
123	117
122	119
120	125
113	132
110	130
109	141
107	144

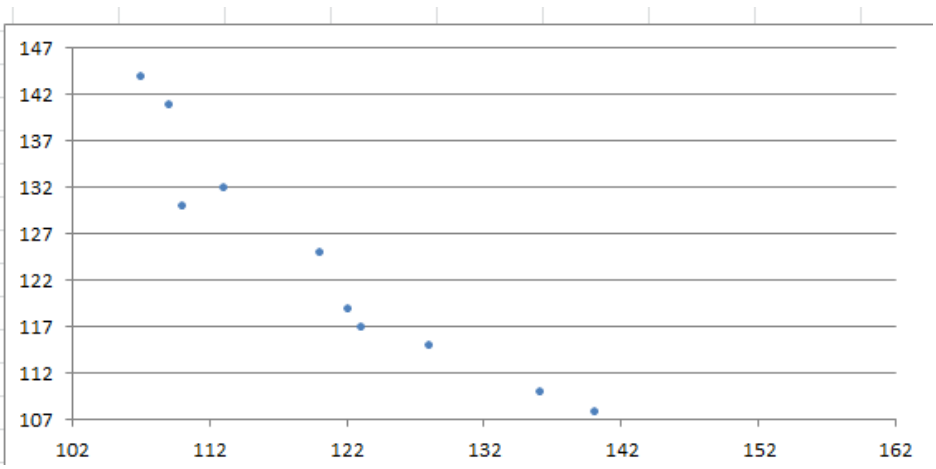


Рисунок 2 – диаграмма рассеяния (корреляционное поле)

Диаграмма рассеивания – это инструмент, позволяющий определить вид и тесноту связи между парами соответствующих переменных (X и Y).

Координаты точек на диаграмме рассеивания соответствуют значениям результативного признака и влияющего на него фактора (факторов). Расположение точек показывает наличие и характер связи между двумя признаками (*X* и *Y*).

Если облако рассеивания бесформенное, значит, никакой связи между двумя переменными нет.

Связи между рассматриваемыми признаками классифицируют по направлению, аналитическому выражению и степени тесноты.

По направлению выделяют связи:

- прямую (рисунок 1);
- обратную (рисунок 2).

Прямой связи соответствуют утверждения:

- высоким (низким) значениям одной переменной соответствуют высокие (низкие) значения другой переменной;
- средним значениям одной переменной соответствуют средние значения другой переменной.

При обратной связи значение результативного признака изменяется под воздействием факторного, но в противоположном направлении (например, с ростом цен, спрос на товар – падает).

Существует две категории зависимостей:

- функциональная;
- корреляционная.

Функциональной называют такую связь, при которой определённому значению факторного признака соответствует одно и только одно значение результативного. Такие связи являются абстракциями, в реальной жизни они встречаются редко, но находят широкое применение в точных науках.

В массовых явлениях проявляются статистические связи, при которых каждому значению независимой переменной *X* соответствует множество значений зависимой переменной *Y*, причем неизвестно заранее, какое именно значение примет *Y*.

Частным случаем статистической зависимости является корреляционная зависимость – связь, при которой каждому значению независимой переменной *X* соответствует определенное математическое ожидание (среднее значение) зависимой переменной *Y*.

Корреляционная связь – это статистическая зависимость между случайными величинами, не имеющими строго функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

По аналитическому выражению выделяют связи линейные и нелинейные.

Линейная связь: $Y = \beta_0 + \beta_1 X$.

Нелинейные связи:

$Y = \beta_0 X^{\beta_1}$ – степенная;

$Y = \beta_0 \beta_1^X$ – показательная;

$Y = e^{\beta_0 + \beta_1 X}$ – экспоненциальная;

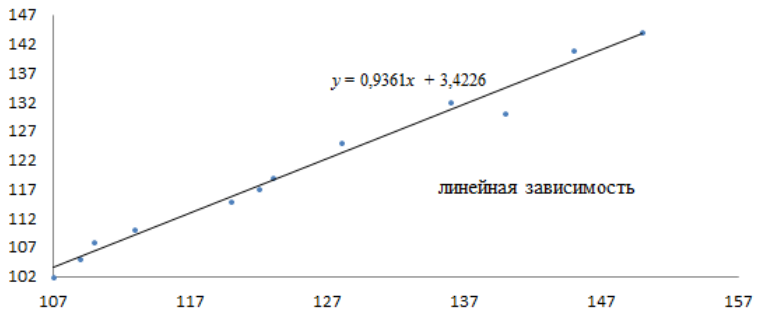
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_n X^n$ – полиномиальная.

Нелинейными являются производственные функции (зависимости между объемом произведенной продукции и основными факторами производства – трудом, капиталом и т.п.), функции спроса (зависимость между спросом на товары или услуги и их ценами или доходом) и др.

Из всех рассмотренных моделей выбирается та, которая лучше соответствует эмпирическим данным и характеру зависимости.

Выбор формулы осуществляется по графическому изображению реальных статистических данных на диаграмме рассеивания.

	X	Y
2	107	102
3	109	105
4	110	108
5	113	110
6	120	115
7	122	117
8	123	119
9	128	125
10	136	132
11	140	130
12	145	141
13	150	144
14		



Соотношение между переменными *X* и *Y* – линейное.

Прямая линия, проведенная через центральную часть скопления точек, дает наиболее подходящую аппроксимацию наблюдаемого соотношения.

Аппроксимация (от лат. *proxima* – ближайшая) или приближение – научный метод, состоящий в замене одних объектов другими более простыми, но близкими к исходным.

На рисунке 1 взаимосвязь между *X* и *Y* близка к линейной, поэтому в качестве зависимости между *X* и *Y* целесообразно выбрать линейную модель:

$$Y = \beta_0 + \beta_1 X .$$

Линейная модель представляет собой геометрическое отображение средних значений анализируемых экономических показателей: объёма потребления *Y* домохозяйства и располагаемого дохода *X* или, другими словами, определяет линию тренда.

Тренд – это направленность изменения экономических показателей, он устанавливает тенденцию экономического роста или спада. В нашем случае – установлена тенденция роста.

Прежде чем перейти к оценке параметров построенной модели, необходимо оценить тесноту связи между рассматриваемыми признаками *X* и *Y*.

Теснота связи определяет меру влияния факторного признака на общую вариацию результативного признака.

Основным методом изучения статистической взаимосвязи является статистическое ***моделирование связи на основе корреляционного и регрессионного анализа***.

Корреляционный анализ не выявляет причину связей между показателями, но устанавливает количественную меру этих связей и подтверждает достоверность наличие связей.

При проведении корреляционного анализа вся совокупность данных рассматривается как множество переменных (факторов), каждая из которых содержит *n* наблюдений.

Коэффициент корреляции Пирсона

Чтобы измерить, как близко находятся точки наблюдения к прямой линии тренда, необходимо вычислить коэффициент корреляции Пирсона, или просто коэффициентом корреляции:

обозначение: r

формула: $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ – эмпирический коэффициент корреляции,

где x_i, y_i – наблюдения, элементы выборки,

$\bar{x}, \bar{y}$ – ожидаемы средние значения,

n – число наблюдений.

Коэффициент корреляции – это инструмент, с помощью которого можно проверить гипотезу о зависимости и измерить силу зависимости двух переменных. Коэффициент корреляции является не идеальным инструментом, он пригоден лишь для измерения силы линейной зависимости.

Коэффициент корреляции может принимать значения от -1 до +1.

Если абсолютное значение коэффициента корреляции находится ближе к 1, то это свидетельствует о сильной связи между переменными, а если ближе к 0, то это говорит о слабой связи или её отсутствии.

Если ближе к -1, то имеет место сильная отрицательная связь. Если его значение равно -1 или +1, то можно говорить о существовании функциональной взаимосвязи между переменными, то есть одну из них можно выразить через другую посредством математической функции.

Таблица 1 – Количественные критерии оценки тесноты связи

Величина коэффициента корреляции	**Характер связи**
До ±3	Практически отсутствует
От ±3 до ±0,5	Слабая
От ±0,5 до ±0,7	Умеренная
От ±0,7 до ±1,0	Сильная

Использование Excel для вычисления коэффициентов корреляции:

ФОРМУЛЫ МАССИВЫ

КОРРЕЛ (массив 1; массив 2),

где:

массив 1 = диапазон данных для первой переменной,

массив 2 = диапазон данных для второй переменной.

Вычислим коэффициент корреляции для данных Задачи 1:

с помощью MS Excel он будет равен 0,991607 .

Используем формулу $r = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$.

Если *X* и *Y* – случайные величины, то *ковариация* является мерой взаимосвязи между двумя величинами и определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$\sigma_{xy}^2 = Cov(x, y) = M\left[(x - \mu_x)(y - \mu_y)\right],$$

где μ_x и μ_y – теоретические средние значения *x* и *y* соответственно.

Для этой формулу составим расчётную таблицу:

	X	Y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$	r
66										
67	X	Y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$	r
68	107	102	-18,25	-18,6667	340,6667	333,0625	348,4444			
69	109	105	-16,25	-15,6667	254,5833	264,0625	245,4444			
70	110	108	-15,25	-12,6667	193,1667	232,5625	160,4444			
71	113	110	-12,25	-10,6667	130,6667	150,0625	113,7778			
72	120	115	-5,25	-5,66667	29,75	27,5625	32,11111			
73	122	117	-3,25	-3,66667	11,91667	10,5625	13,44444			
74	123	119	-2,25	-1,66667	3,75	5,0625	2,777778			
75	128	125	2,75	4,333333	11,91667	7,5625	18,77778			
76	136	132	10,75	11,33333	121,8333	115,5625	128,4444			
77	140	130	14,75	9,333333	137,6667	217,5625	87,11111			
78	145	141	19,75	20,33333	401,5833	390,0625	413,4444			
79	150	144	24,75	23,33333	577,5	612,5625	544,4444			
80	Xср	Yср			сумма	сумма	сумма			
81	125,25	120,6667			2215	2366,25	2108,667	4989633	2233,749	0,991607
82										

Коэффициент корреляции r = 0,991607.

Вывод: так как значение коэффициента корреляции r = 0,991607 достаточно близко к +1, можно судить о сильной положительной связи между переменными: объёмом потребления *Y* домохозяйства и располагаемым доходом *X*.

На основании полученного результата можно говорить о «полезности» факторных признаков при построении уравнения регрессии, которое может использоваться для анализа и прогнозирования.

Корреляция взаимосвязана с регрессией, поскольку первая оценивает силу (тесноту) статистической связи, а регрессия исследует её форму.

Корреляционно-регрессионный анализ для прямолинейной связи при парной корреляции

1 Корреляционно-регрессионный анализ для прямолинейной связи при парной корреляции

Термин «регрессия» был введён Френсисом Галтоном в конце XIX века при анализе зависимости между ростом родителей и ростом детей. Он выяснил, что рост детей у очень высоких родителей в среднем меньше, чем средний рост родителей. У очень низких родителей, наоборот, средний рост детей выше.

Регрессией называется зависимость среднего значения случайной величины результативного признака *Y* от величины факторного *X*.

Регрессионный анализ заключается в определении аналитического выражения связи в виде уравнения регрессии.

Уравнением регрессии – называется уравнение, описывающее корреляционную зависимость между результативным признаком *Y* и одним или несколькими факторными *X*.

При рассмотрении зависимости двух случайных величин говорят о парной регрессии.

Парная регрессия – уравнение связи двух переменных X и Y, причём, изменение первой переменной может служить причиной изменения другой, при этом выделяют переменные:

- объясняющие (независимые) – X;
- объясняемые (зависимые) – Y.

Но такая зависимость не является однозначной, т. е. каждому конкретному значению объясняющей x_i переменной соответствует некоторое вероятностное распределение объясняемой переменной Y. Поэтому анализируют, как объясняющая переменная влияет на зависимую переменную в среднем, что может быть представлено соотношением:

$M(Y \mid X = x_i) = f(x)$ – функция регрессии Y на X – это зависимость между объясняющими переменными X и условным математическим ожиданием (средним значением) зависимой переменной Y, которая строится с целью прогнозирования этого среднего значения $\hat{Y}$ при фиксированных значениях X (x_i – значение независимой переменной в i-м наблюдении), при этом X называется – регрессором, Y – зависимой переменной.

Линейная регрессия представляет собой линейную функцию:

$M(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$,

где β_0 и β_1 – параметры уравнения;

x_i – значение независимой переменной в i-м наблюдении.

В реальности фактические значения y_i зависимой переменной Y не всегда совпадают с соответствующими математическими ожиданиями и могут быть различными при одном и том же значении X, т. е. каждое индивидуальное значение y_i отличается от соответствующего математического ожидания на некоторую величину – ε_i,.

Случайная величина ε_i – называется случайной ошибкой (отклонением, возмущением).

Таким образом, фактическая зависимость должна быть дополнена слагаемым ε_i:

$y_i = M(Y \mid X = x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$ – теоретическая линейная регрессионная модель.

В общем виде $Y = \beta_0 + \beta_1 X + \varepsilon$.

β_0 и β_1 – теоретические параметры (коэффициенты) уравнения;

ε_i – случайная ошибка.

Причины существования случайной ошибки:

1) не включение в регрессионную модель значимых объясняющих переменных;
2) агрегирование переменных. С математической точки зрения агрегирование рассматривается как преобразование исходной модели в модель с меньшим числом переменных;

3) неправильное описание структуры модели;

4) неправильная функциональная спецификация;

5) ошибки измерения.

Так как отклонения ε_i для каждого конкретного наблюдения i – случайны и их значения в выборке неизвестны, то:

1) по наблюдениям x_i и y_i можно получить только оценки b_0 и b_1 параметров β_0 *и* β_1 соответственно;

2) оценки параметров b_0 и b_1 – называются эмпирическими коэффициентами регрессии, носят случайный характер, т. к. соответствуют случайной выборке.

Тогда оценочное уравнение регрессии (построенное по выборочным данным) будет иметь вид: $y_i = b_0 + b_1 x_i + e_i$ – эмпирическая линейная регрессионная модель,

где e_i – оценка теоретического случайного отклонения ε_i .

Задача линейного регрессионного анализа состоит в том, чтобы:

1) получить наилучшие оценки b_0 и b_1;

2) проверить статистические гипотезы о параметрах модели;

3) проверить адекватность модели данным наблюдений.

По выборке данного объёма можно построить эмпирическое уравнение регрессии: $\hat{y}_i = b_0 + b_1 x_i$ (эмпирическая линия регрессии),

где $\hat{y}_i$ – оценка условного математического ожидания $M(Y|X = x_i)$,

b_0 и b_1 – эмпирические коэффициенты регрессии.

Оценки b_0 и b_1 отличаются от истинных значений β_0 и β_1, что приводит к несовпадению эмпирической и теоретической линий регрессии. По различным выборкам из одной и той же генеральной совокупности получают разные значения оценок коэффициентов регрессии.

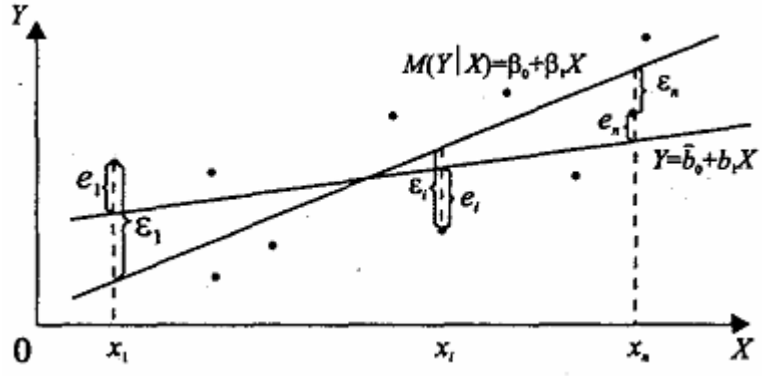


Рисунок 2 – Эмпирическая и теоретическая линии регрессии

Задача состоит в нахождении по выборке данных оценок b_0 и b_1 так, чтобы построенная линия регрессии была **наилучшей** среди всех других прямых.

Оценкам b_0 *и* b_1 уравнения линейной регрессии можно придать экономический смысл.

Коэффициент b_1 уравнения регрессии показывает, на сколько единиц изменится результат *Y* при изменении фактора *X* на 1 единицу.

Коэффициент b_1 определяет тангенс угла наклона прямой относительно положительного направления оси абсцисс.

Связь между *Y* и *X* определяет знак коэффициента регрессии b_1:

если $b_1 > 0$ – прямая связь,

$b_1 < 0$ – обратная связь.

Свободный член b_0 формально показывает прогнозируемый уровень *Y* при величине *X*=0, но только в том случае, если *X* находится близко с выборочными значениями.

Если *X* находится далеко от выборочных значений, то буквальная интерпретация может привести к неверным результатам.

2 Этап параметризации – оценка параметров построенной модели

Для оценки параметров β_0 *и* β_1 – используют метод наименьших квадратов (МНК).

Метод наименьших квадратов

По выборке (x_i, y_i) следует определить оценки b_0 и b_1 эмпирического уравнения регрессии: $\hat{y}_i = b_0 + b_1 x_i$.

Дадим определение *остатка* для каждого наблюдения:

остаток в i-ом наблюдении – это разность между истинным значением переменной y_i в *i*-ом наблюдении и значением $(b_0 + b_1 x_i)$, полученным подстановкой наблюдения x_i в уравнение линейной регрессии, то есть

$$\begin{cases} y_i = b_0 + b_1 x_i + e_i, \\ \hat{y}_i = b_0 + b_1 x_i \end{cases}$$

$y_i - \hat{y}_i = e_i$ – остаток.

Необходимо выбрать такой критерий $Q(b_0, b_1)$, который будет одновременно учитывать величину всех остатков. Как привило, минимизируют суммы квадратов остатков:

$$Q(b_0, b_1) = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Для того чтобы величина $Q(b_0, b_1)$ была минимальной, необходимо, чтобы частные производные равнялись нулю, то есть (см. конспект).

Условия качества оценок линейной регрессии МНК. Теорема Гаусса-Маркова

Качество оценок параметров уравнения регрессии, существенно зависящих от случайной составляющей ε, определяет уровень качества построенной регрессионной модели. Наилучшие результаты оценки параметров линейной регрессии, полученные методом наименьших квадратов, достигаются при выполнении условий *Гаусса - Маркова.*

Условия Гаусса - Маркова

1 *Математическое ожидание случайной составляющей* ε_i *равно нулю для каждого наблюдения*: $M(\varepsilon_i) = 0$.

Случайная составляющая отклонения ε, независимо от знака ($\pm\ \varepsilon$), не оказывает значительного влияния на зависимую переменную Y, следовательно, не должно иметь систематического смещения.

Из условия $M(\varepsilon_i) = 0$ будет следовать выполнение уравнения линейной регрессии $(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$ относительно независимой переменной X и условным математическим ожиданием $M(Y \mid X = x_i)$.

2 *Дисперсия случайных отклонений составляющей* ε_i *постоянна для всех наблюдений*: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma_\varepsilon^2$, $(i \neq j)$. Так как $D(\varepsilon_i) = M[\varepsilon_i - M(\varepsilon_i)]^2$, то, учитывая $M(\varepsilon_i) = 0$, можно записать $D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma_\varepsilon^2$.

Постоянство дисперсии $D(\varepsilon_i)$ не вызывает значительных ошибок случайного отклонения, не зависимо от больших или меньших значений его величин. Это условие является одним из ключевых метода наименьших квадратов.

Постоянство дисперсии случайной составляющей ε_i *называется* ***гомоскедастичностью.***

Непостоянство дисперсии случайной составляющей ε_i *называется* ***гетероскедастичностью.***

3 *Значения случайных отклонений* ε_i *и* ε_j *некоррелированные, то есть независимые друг от друга при* $i \neq j$.

Из этого условия следует выполнение следующих соотношений:

$cov(\varepsilon_i, \varepsilon_j) = \sigma_{\varepsilon i, \varepsilon j} = 0$, если $i \neq j$; $cov(\varepsilon_i, \varepsilon_j) = \sigma_{\varepsilon i, \varepsilon j} = \sigma^2 \neq 0$, если $i = j$.

Если X и Y – случайные величины, то *ковариация* является мерой взаимосвязи между двумя величинами и определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$\sigma_{xy}^2 = Cov(x, y) = M\left[(x - \mu_x)(y - \mu_y)\right],$$

где μ_x и μ_y – теоретические средние значения x и y соответственно.

Выполнение этого условия предполагает отсутствие автокорреляции, то есть систематической связи между случайными отклонениями, $M(\varepsilon_i, \varepsilon_j) = 0$.

4 *Отклонение случайной составляющей* ε_i *распределяются независимо от объясняющих переменных* x_i, $cov\,(x_i, \varepsilon_i) = M\,(x_i, \varepsilon_i) = 0$, *так как* $M(\varepsilon_i) = 0$.

Это условие всегда выполняется, если объясняющие переменные x_i не являются случайными в рассматриваемой регрессионной модели.

Теорема Гаусса-Маркова

Если выполняются четыре условия Гаусса - Маркова, то оценки параметров *линейной регрессионной модели* полученные по МНК будут несмещенными и эффективными среди всех возможных линейных несмещенных оценок.

Действительно, оценки параметров β_0 и β_1 *есть несмещенные* оценки b_0 и b_1, что обеспечивается выполнением 1 - 4 условий Гаусса - Маркова:

$Y = \beta_0 + \beta_1 X + \varepsilon_i$,

так как,

$\beta_0 = M(b_0)$, $\beta_1 = M(b_1)$, $M(\varepsilon_i) = 0$.

С возрастанием объема выборки n увеличивается *состоятельность* (близость оценок b_0 к β_0 и b_1 к β_1), так как дисперсии оценок параметров стремятся к нулю $D\,(b_0) \to 0$, $D\,(b_1) \to 0$.

Эффективность обуславливается тем, что среди всевозможных линейных несмещенных оценок, полученные из условий Гаусса - Маркова оценки, имеют наименьшую дисперсию.

Нарушение хотя бы одного из условий Гаусса - Маркова вызывает нарушение эффективности оценок и к поискам наилучших несмещенных оценок BLUE (*Best Linear Unbiased Estimators*), которые имеют наименьшую дисперсию.

В регрессионном анализе, наряду с условиями Гаусса - Маркова, делается предположение о нормальности распределения случайного отклонения ε, что позволяет определять степень точности полученных оценок параметров.

Условия Гаусса - Маркова рассматриваются как основные предпосылки регрессионного анализа.

Степень точности оценки, параметров уравнения регрессии

Определение степени точности полученных оценок связано с дисперсией случайного отклонения ε, то есть чем меньше дисперсии оценок $D\,(b_0)$ и $D\,(b_1)$, тем выше надежность параметров уравнения регрессии β_0, β_1.

В соответствии со вторым условием Гаусса - Маркова можно считать, что все дисперсии проведенных измерений равны между собой $D(\varepsilon_i) = \sigma_\varepsilon^2 = \sigma^2$.

Так как считаем, что оценки b_0 и b_1 параметров регрессии β_0, β_1, имеют наименьшую (постоянную) дисперсию, то исправленная выборочная дисперсия S^2 является несмещенной оценкой для дисперсии σ^2 случайного теоретического отклонения ε:

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2 = \frac{1}{n-2}\sum_{i=1}^{n}e_i^2 .$$

Дисперсию оценок b_0 и b_1 параметров регрессии β_0, β_1 можно определить по формулам:

$$D(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad D(b_1) \approx S_{b_1}^2 = \frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$D(b_0) = \frac{\sigma^2 \cdot \overline{x^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad D(b_0) \approx S_{b_0}^2 = \frac{S^2 \cdot \overline{x^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = S_{b_1}^2 \cdot \overline{x^2} .$$

Из записанных формул можно сделать следующие выводы и предложения:

1) дисперсии оценок $D(b_1)$ и $D(b_0)$ прямо пропорциональны дисперсии σ^2 случайного теоретического отклонения ε;

2) чем больше разброс значений объясняющей переменной $\sum_{i=1}^{n}([x_i - \bar{x})^2$, то есть, чем шире область изменения дисперсии, тем меньше дисперсия оценок;

3) чем больше величина объема выборки n, тем больше значение $\sum_{i=1}^{n}(x_i - \bar{x})^2$, следовательно, тем меньше дисперсия оценок и тем выше их эффективность.

Ошибки для коэффициентов регрессии вычисляются по формулам: $S_{b_1} = \sqrt{S_{b_1}^2} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ и $S_{b_0} = \sqrt{S_{b_0}^2} = \sqrt{S_{b_1}^2 \cdot \overline{x^2}}$ –

– *величины* S_{b1} и S_{b0} *называются стандартными ошибками коэффициентов регрессии*;

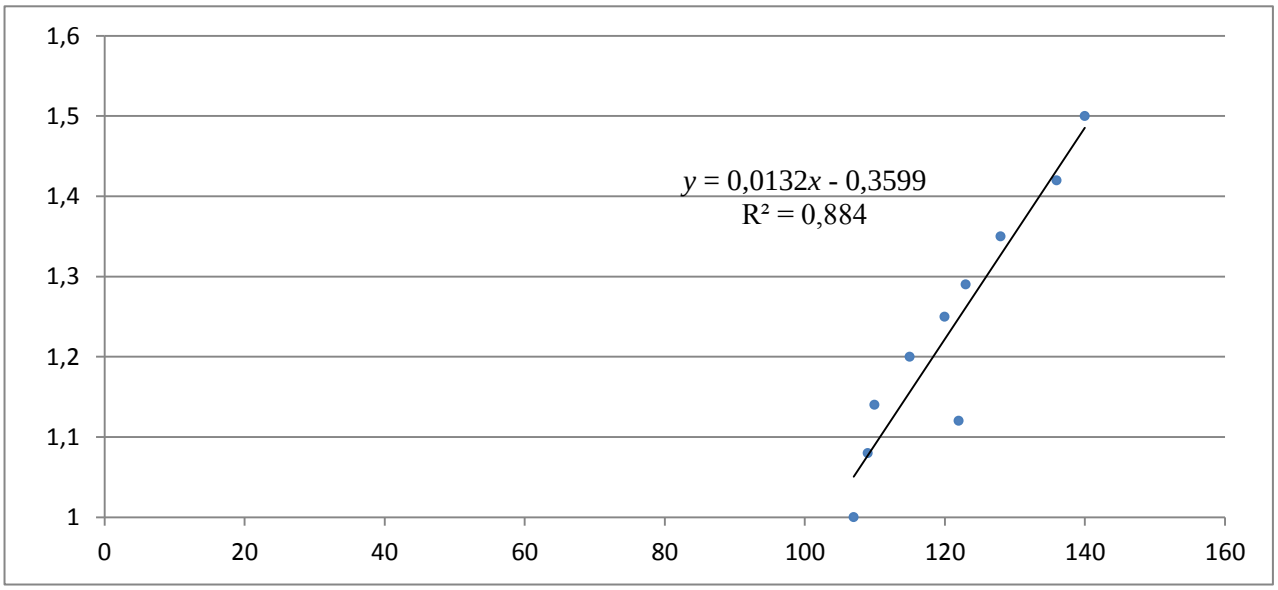
4) корень квадратный из необъясненной дисперсии $S = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}e_i^2}$ – *называется стандартной ошибкой регрессии.*

Рассмотрим пример 1^0

Для анализа количества (в тоннах) использования высококачественного сырья (Y *т*) в зависимости от видов выпускаемой на 10 предприятиях аналогичной продукции (X *штук*) выполнена выборка, представленная в таблице. Определить вид зависимости между Y и X, записать вид функции регрессии; по МНК вычислить коэффициенты b_0 и b_1 уравнения регрессии, являющихся оценками параметров регрессии β_0, β_1; оценить силу линейной зависимости между Y и X.

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,00	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50

Построим поле корреляции, укажем линию тренда и соответствующее уравнение регрессии.



Вычислим коэффициент корреляции

$$r_{xy} = b_1 \frac{S_x}{S_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} = \frac{1{,}5}{\sqrt{113{,}8} \cdot \sqrt{1{,}545 - 1{,}525}} = \frac{1{,}5}{1{,}513} = 0{,}9914\,.$$

Величина коэффициента корреляции близка к 1, поэтому можно сделать вывод о *сильной* линейной зависимости *Y* от *X*.

Выполним верификацию построенной регрессионной модели:

$\hat{y} = 0{,}0132x - 0{,}3599.$

Верификация модели проводится по трем направлениям:

- соответствие модели эмпирическим данным,
- качество оценок параметров уравнения регрессии,
- распределение случайных отклонений.

Анализ соответствия модели эмпирическим данным проводится для того, чтобы определить, в какой степени объясняющая переменная объясняется включенными в модель независимыми переменными.

Проверка статистической значимости уравнения регрессии

Проверка значимости уравнение регрессии сводится к выявлению соответствия математической модели (уравнения регрессии), которая выражает зависимость между исходными экспериментальными данными и соответствует ли выбранное количество объясняющих переменных включенных в уравнение для описания зависимой переменной.

Оценка значимости уравнения регрессии выполняется на основе дисперсионного анализа и *F -распределения Фишера.*

Этот критерий основан на анализе квадратов фактических отклонений зависимой переменной от среднего значения. Часть этого отклонения объясняется включенными в модель переменными, а другая – нет:

$U = U_r + U_e$

Общая сумма квадратов отклонений $U = \sum_{i=1}^{n} (y_i - \bar{y})^2$, как мера отклонений (рассеивания) объясняемой переменной *Y* от среднего значения $\bar{y}$, представляется как *объясненная* или *факторная* сумма квадратов отклонений $U_r = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ и сумма $U_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – *остаточная сумма* квадратов отклонений, как мера разброса, обусловленная неучтенными в модели факторами.

Таким образом:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i-y_i)^2 .$$

Проверка значимости модели по критерию Фишера основана на тестировании гипотезы о том, что объясненная и остаточная дисперсии в расчете на одну степень свободы различаются незначимо.

Степени свободы рассчитываются следующим образом: $k_1=m-1$, $k_2=n-m$, где n – число случайных величин; m – число линейных связей между случайных величин.

n – число точек в выборке;

m – число переменных в уравнении регрессии.

Число степеней свободы – это минимально необходимое число значений зависимой переменной, которых достаточно для получения искомой характеристики выборки и которые могут свободно варьироваться с учетом того, что для этой выборки известны все другие величины, используемые для расчета искомой характеристики.

Термином «число степеней свобод» обозначается число независимых способов описания исследуемой выборки. Так, если значение математического ожидания заранее известно, то число степеней свободы будет равно общему объему n выборки. Если же значение оценивается на опыте как среднее арифметическое $\bar{x}$ этих же n измерений, то число степеней свободы будет равно n-1, так как из общего числа случайных величин вычитается дополнительная связь между всеми элементами выборки, затраченная при определении значения X.

Так как уравнение регрессии строится на основе МНК, запишем соотношение исправленных дисперсий.

Остаточная дисперсия – это общая сумма квадратов отклонений расчетных значений от фактически, разделенная на число наблюдений.

Объясненная дисперсия – доля вариации данных, учитываемая моделью.

$$S^2_{общ} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}; \qquad S^2_r = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{m}; \qquad S^2_e = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-m-1};$$

Отношение факторной (S^2_r / S^2_e) и остаточной дисперсий дает величину F – критерия Фишера:

$$F = \frac{S^2_r}{S^2_e} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \cdot (n-2)}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

с теми же степенями свободы.

Уравнение регрессии признается статистически значимым при уровне значимости α, если фактическое значения F - критерия Фишера больше табличного значения $F_{табл.}$ (α, v_1, v_2):

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \cdot (n-2)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} > F_{(\alpha;1;n-2)}.$$

Проверить значимость уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$, построенного по исходным данным задания 1°.

Решение.

По данным из задания 1^0 составим таблицу и выполним необходимые вычисления:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,00	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$\hat{y}_i$	1,053	1,079	1,092	1,157	1,222	1,248	1,261	1,326	1,380	1,432
$(y_i - \bar{y})^2$	0,055	0,024	0,009	0,0012	0,00022	0,0132	0,003	0,0132	0,034	0,0702
$(\hat{y}_i - y_i)^2$	0, 003	0,000	0,002	0,007	0,001	0,016	0,001	0,0006	0,0016	0,005

$\bar{y} = 1{,}235$; $(y_i - \bar{y})^2 = 0{,}223$; $b_1 = 0{,}013$; $b_0 = -0{,}338$;

$$U = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 0{,}223; \quad U_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0{,}037; \; U_r = U - U_e = 0{,}186.$$

$$F = \frac{U_r(n-2)}{U_e} = \frac{0{,}186 \cdot 8}{0{,}037} = 40{,}21,$$

табличное значение $F_{табл.}$ (0, 05; 1; 8) = 5,32, выполняется неравенство:

$$F = 40{,}21 > F_{(0{,}05;1;8)} = 5{,}32.$$

Уравнение регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$ в целом значимо с уровнем значимости α = 0,05.

Проверка статистической значимости коэффициентов регрессии

Проверка значимости уравнения регрессии сводится к тому, чтобы установить соответствие построенного уравнения регрессии, то есть соответствие построенной математической модели, экспериментальным данным и установить достаточность включенных в уравнение объясняющих переменных для описания зависимой переменной.

Может проводиться проверка: значимости коэффициентов уравнения регрессии или проверка значимости уравнения регрессии.

Проверка статистической значимости коэффициентов линейного уравнения регрессии, осуществляется методами проверки статистических гипотез.

Для проверки значимости коэффициента регрессии b_1 рассматривается следующая статистическая гипотеза с уровнем значимости α:

основная гипотеза $H_0 : \beta_1 = 0$, (коэффициент b_1 незначим);

конкурирующая гипотеза $H_1 : \beta_1 \neq 0$, (коэффициент b_1 значим).

В качестве критерия для проверки гипотезы H_0 берется t – статистика

$$T_{b_1} = \frac{b_1}{S_{b_1}} = \frac{b_1}{\sqrt{S_{b_1}^2}},$$

которая при справедливости гипотезы H_0 имеет распределение Стьюдента с объемом выборки n и числом степеней свободы $\nu = n - 2$, S_{b_1} – стандартная ошибка коэффициента b_1.

Значимость коэффициента b_1 при отклонении гипотезы H_0, с уровнем значимости α, показывает наличие линейной зависимости Y от X, если

$$|T_{b_1}| = \left|\frac{b_1}{S_{b_1}}\right| > t(\alpha/2, n-2).$$

При отклонении конкурирующей гипотезы H_1 коэффициент регрессии b_1 незначим (значения b_1 близки к нулю), в этом случае линейной зависимости Y от X может не быть.

Аналогичным образом для проверки значимости коэффициента b_0 формируется статистические гипотезы:

$H_0 : \beta_0 = 0$, (коэффициент b_0 незначим);

$H_1 : \beta_0 \neq 0$, (коэффициент b_0 значим).

В качестве критерия для проверки основной гипотезы H_0 принимается величина

$$T_{b_0} = \frac{b_0}{S_{b_0}} = \frac{b_0}{\sqrt{S_{b_0}^2}}.$$

Гипотеза H_0 отвергается с уровнем значимости α, если

$$|T_{b_0}| = \left|\frac{b_0}{S_{b_0}}\right| > t(\alpha/2, n-2).$$

Пользуясь данными условия задания 1° и полученными результатами вычислений, проверить статистическую значимость коэффициентов b_1 и b_0 уравнения регрессии.

Составим таблицу и запишем значения вычисленных величин:

i	1	2	3	4	5	6	7	8	9	10
y_i	1,00	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$(y_i - \bar{y})^2$	0,055	0,024	0,009	0,0012	0,00022	0,0132	0,003	0,0132	0,034	0,0702

$\bar{y}$ = 1,235; $\bar{x}$ = 121; $\overline{x^2}$ = 14754; $\overline{y^2}$ = 1,545; $\overline{xy}$ = 150,935;

$(y_i - \bar{y})^2$ = 0,223; b_1 = 0,013; b_0 = – 0,338; $\hat{y}_i = -0{,}338 + 0{,}013x_i$.

Решение. Вычислим S_{b_1} и наблюдаемое значение T_{b_1}:

$$S_{b_1}^2 = \frac{S^2}{n(\overline{x^2} - \bar{x}^2)} = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1)^2}{n(n-2)(\overline{x^2} - \bar{x}^2)} = \frac{0{,}223}{10 \cdot 8 \cdot (14754 - 121^2)} = 0{,}000025;$$

$$S_{b_1} = \sqrt{S_{b_1}^2} = \sqrt{0{,}000025} = 0{,}005;$$

$$T_{b_1} = \frac{b_1}{S_{b_1}} = \frac{0{,}013}{0{,}005} = 2{,}6.$$

Определим критические значения $T_{кр.} = t_{(0{,}025;\ 8)} = 2{,}306$.

Сравним T_{b_1} и $T_{кр}$: $T_{b_1} = 2{,}6 > T_{кр} = 2{,}3$ из сравнения следует, что нулевая гипотеза отвергается в пользу альтернативной, то есть коэффициента уравнения регрессии b_1 статистически значим с уровнем значимости $\alpha = 0{,}05$.

Для проверки значимости коэффициента b_0 вычислим стандартную ошибку S_{b_0} и наблюдаемое значение T_{b_0}:

$$S_{b_0}^2 = \frac{S^2 \sum_{i=1}^{n}(x_i)^2}{n(\overline{x^2} - \bar{x}^2)} = S_{b_1}^2 \cdot \overline{x^2} = 0{,}000025 \cdot 14754 = 0{,}369;$$

$$S_{b_0} = \sqrt{0{,}369} = 0{,}607;$$

$$|T_{b_0} = \left|\frac{b_0}{S_{b_0}}\right| = \left|\frac{-0{,}338}{0{,}607}\right| = |-0{,}556|.$$

Так как | – 0,556| < 2,306, то принимается нулевая гипотеза, следовательно, коэффициент b_0 незначим с уровнем значимости α = 0,05. Это означает, что в данном случае свободным членом b_0 уравнения регрессии можно пренебречь, рассматривая регрессию как $\hat{y} = 0{,}0132x$.

Оценку ***качества уравнения регрессии*** *или соответствующей ей модели дает* ***коэффициент детерминации*** $\boldsymbol{R^2}$.

Напомним, что оценкой и мерой *линейной связи* двух случайных величин X и Y является коэффициент корреляции r_{xy}:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}.$$

В зависимости от тесноты и характера взаимосвязи *СВ* (X, Y) коэффициент может принимать значения от –1 до +1 ($-1 \le r_{xy} \le +1$). С приближением модуля r_{xy} к единице линейная связь между Y и X усиливается и, наоборот, с приближением модуля r_{xy} к нулю линейная связь между Y и X ослабевает.

Если взаимосвязи между *СВ* X и Y *нелинейная*, то тесноту такой связи, задаваемую соотношением $\hat{y} = f(x)$ оценивают с помощью индекса корреляции R:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}}; \quad 0 \le R \le 1,$$

если индекс корреляции R = 1, то связь между *СВ* (X , Y) становится функциональной, зависимость $\hat{y} = f(x)$ согласуется с данными наблюдений.

Коэффициент детерминации дает наиболее эффективную меру качества «подгонки» выстраиваемой регрессионной модели к наблюдаемым значениям объясняемой переменной Y , то есть дает оценку *адекватности* модели.

Коэффициент детерминации, определяется по формуле:

$$R^2 = \frac{U_r}{U} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

или

$$R^2 = 1 - \frac{U_e}{U} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

При рассмотрении парной регрессии коэффициент детерминации совпадает с квадратом коэффициента корреляции $r_{x,y}^2$, выполняется тождество:

$$R^2 = r_{xy}^2 = \left[\frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} \right]^2.$$

Если коэффициент детерминации близок или равен нулю, то зависимость *Y* от *X* незначительна или вообще отсутствует. Если коэффициент детерминации близок или равен единицы, то тем меньше или вообще отсутствует (U_e = 0) ошибка отклонения эмпирических данных от фактических, то есть между *Y* и *X* имеется функциональная линейная зависимость.

Следовательно, коэффициент детерминации может применяться для оценки *качества уравнения регрессии* или соответствующей ей модели.

Для величин, определяющих коэффициент детерминации, вводятся следующие обозначения:

$U = \sum_{i=1}^{n}(y_i - \bar{y})^2$ – *TSS – total sum of squares* – вся дисперсия, характеризующая степень рассеивания функции регрессии около среднего значения $\bar{y}$; общая сумма квадратов отклонения;

$U_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ – *RSS – regression sum of squares* – объясненная часть дисперсии, объясненная сумма квадратов отклонений ;

$U_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – *ESS – error sum of squares* – остаточная сумма квадратов отклонений есть сумма квадратов остатков регрессии, как мера разброса, обусловленная неучтенными в модели факторами. Необъясненная сумма квадратов отклонения.

$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$ — *коэффициент детерминации.*

Рассмотрим некоторые свойства коэффициента детерминации:

1) коэффициент детерминации изменяется от нуля до единицы

$0 \le R^2 \le 1$;

2) при $R^2 = 0$, $U_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0$, то есть полученная регрессия не имеет объясненной части дисперсии;

3) при $R^2 = 1$, $U_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$, значение $\hat{y}_i$ совпадают с значениями y_i, отсутствует зависимость дисперсии от неучтенных факторов $e_i = 0$;

4) чем ближе коэффициент детерминации R^2 к единице, тем меньше среднее отклонение расчетных данных от фактических.

Рассмотрим пример

Определить коэффициент детерминации для уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$.

Решение.

$\bar{y} = 1{,}235$; $(y_i - \bar{y})^2 = 0{,}223$; $b_1 = 0{,}013$; $b_0 = -0{,}338$;

$U = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 0{,}223$; $U_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0{,}037$; $U_r = U - U_e = 0{,}186$.

$$R^2 = 1 - \frac{U_e}{U} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{0{,}037}{0{,}223} = 0{,}834.$$

Значение коэффициента детерминации $R^2 = 0{,}834$ определяет достаточно высокое качество построенного уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$.

Коэффициент детерминации используется для проверки значимости уравнения парной линейной регрессии.

Оценка статистической значимости уравнения парной линейной регрессии выполняется с использованием F - критерия Фишера по заданному уровню значимости α:

$$F_{наб} = \frac{\frac{1}{m}U_r}{\frac{1}{n-m-}U} = \frac{\frac{1}{m}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{(n-m-1)R^2}{m(1-R^2)},$$

$m = 1$ – число степеней свободы v_1 (независимых переменных) в уравнении парной регрессии, $v_2 = n - m - 1 = n - 2$.

Выдвигается основная гипотеза H_0 о статистической *незначимости уравнения регрессии*, которая *отвергается* при выполнении неравенства

$T_{наб.} > T_{крит.}$.

Рассмотрим пример

По полученному значению коэффициента детерминации $R^2 = 0{,}834$, определить статистическую значимость уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$, если n = 10, уровень значимости α = 0,05.

Решение.

Для проверки того, что уравнения регрессии $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = -0{,}338 + 0{,}013\boldsymbol{x}_{\boldsymbol{i}}$ незначимо, рассмотрим следующие статистические гипотезы с уровнем значимости α = 0,05:

основная гипотеза H_0: уравнение регрессии незначимо;

конкурирующая гипотеза H_1: уравнение регрессии значимо.

Запишем F - критерий Фишера по заданному уровню значимости α = 0,05:

$$F_{наб} = \frac{(n-m-1)R^2}{m(1-R^2)} = \frac{8 \cdot 0{,}834}{1 \cdot 0{,}166} = \frac{6{,}672}{0{,}166} = 40{,}19;$$

$$F_{крит.} = F_{(0{,}05;1;\,8)} = 5{,}32.$$

Так как $F_{наб..} = 40{,}19 > F_{крит.} = 5{,}32$, то основная гипотеза H_0 – уравнение регрессии незначимо – *отвергается.* Следовательно, уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$ *значимо.*

Интервальные оценки функции регрессии и коэффициентов β_0 и β_1

Надежность оценок функции регрессии и коэффициентов β_0 и β_1 связана с дисперсией $D(b_1)$ и $D(b_0)$, которые прямо пропорциональны дисперсии случайного теоретического отклонения $D(\varepsilon)$, то $D(\varepsilon)$ фактически есть $D(Y \mid X = x_i)$ переменной Y относительно линии регрессии.

Согласно второму условию Гаусса-Маркова, определяющему качества оценок линейной регрессии МНК, *дисперсия случайных отклонений составляющей ε_i постоянна для всех наблюдений*: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma_\varepsilon^2 = \sigma^2$.

Учитывая, что дисперсия постоянная и $M(\varepsilon_i) = 0$, основываясь на базовых предпосылках МНК можно считать, что отклонения ε_i с нулевым математическим ожиданием подчиняется нормальному распределению

$$\varepsilon_i \sim N(0,\ \sigma^2).$$

Тогда статистики:

$$\boldsymbol{T_{b_1}} = \frac{\boldsymbol{b_1} - \beta_1}{\boldsymbol{S_{b_1}}} = \frac{\boldsymbol{b_1} - \beta_1}{\sqrt{\boldsymbol{S_{b_1}^2}}}; \qquad \boldsymbol{T_{b_0}} = \frac{\boldsymbol{b_0} - \beta_0}{\boldsymbol{S_{b_0}}} = \frac{\boldsymbol{b_0} - \beta_0}{\sqrt{\boldsymbol{S_{b_0}^2}}}$$

имеют распределение Стьюдента с объемом выборки n и
числом степеней свободы $\nu = n - 2$,

S_{b_1} и S_{b_0} – стандартные ошибки коэффициентов b_1 и b_0.

Доверительная вероятность $\gamma = 1 - \alpha$ и число степеней свободы $\nu = n - 2$ по таблицам дают определение критических точек распределения Стьюдента $T_{крит.} = | t_{(\gamma,\nu)} |$,

$$| t_{(\gamma,\nu)} | \le \frac{\boldsymbol{b_1} - \beta_1}{\boldsymbol{S_{b_1}}} \to b_1 - t(\gamma, \nu)\ S_{b_1} \le \beta_1 \le b_1 + t(\gamma, \nu)\ S_{b_1};$$

$$| t_{(\gamma,\nu)} | \le \frac{\boldsymbol{b_0} - \beta_0}{\boldsymbol{S_{b_0}}} \to b_0 - t(\gamma, \nu)\ S_{b_0} \le \beta_0 \le b_0 + t(\gamma, \nu)\ S_{b_0}.$$

Интервалами оценок параметров β_0 и β_1 с надежностью $\gamma = 1 - \alpha$ будут интервалы:

$$\beta_0 : [b_0 - t(\gamma, n-2)\ S_{b_0};\ b_0 + t(\gamma, n-2)\ S_{b_0}];$$

$$\beta_1 : [b_1 - t(\gamma, n-2)\ S_{b_1};\ b_1 + t(\gamma, n-2)\ S_{b_1}].$$

Полученные доверительные интервалы с надежностью $\gamma = 1 - \alpha$ включают (накрывают) параметры β_0 и β_1

Для уравнения парной регрессии $\hat{y}(x) = b_0 + b_1 x_i$ строится доверительный интервал, который определяется возможным средним математическим ожиданием $M(Y \mid X = x)$, но не конкретными значениями зависимой переменной Y, которые отклоняются от $M(Y \mid X = x)$.

Чтобы построить интервал, в который функция попадает с вероятностью γ, уравнение регрессии перепишем в виде:

$$\hat{y}(x) = \bar{y} - b_1(x - \bar{x})$$

и учитывая $\varepsilon_i \sim N(0, \sigma^2)$, можно считать, что $\hat{y}(x)$ подчиняется нормальному распределению. Тогда $M[\hat{y}(x)]$ и $D[\hat{y}(x)]$ тоже зависят от x, а $\hat{y}(x)$ является несмещенной оценкой для функции регрессии:

$M[\hat{y}(x_i)] = M(Y \mid X = x_i) = M[b_0] + M[b_1]\, x_i = \beta_0 + \beta_1 x_i$;

$D[\hat{y}(x)] = D(b_0) + D(b_1)\, x_i^2 + 2cov(b_0, b_1)\, x_i$;

$2cov(b_0, b_1)\, x_i = 2\, M[(b_0 - M(b_0))(b_1 - M(b_1))] = 2\, M[(b_0 - \beta_0)(b_1 - \beta_1)]$;

$2cov(b_0, b_1)\, x_i = 2\, M[(b_0 - \beta_0)(b_1 - \beta_1)] = -2\, \overline{X}$ $D(b_1) = -2\bar{x}\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Учитывая, что

$$D(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad D(b_0) = \frac{\sigma^2 \cdot \overline{x^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad 2\mathrm{cov}(b_0, b_1) = -2\bar{x}\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

получим:

$$D(\hat{y}_i) = \frac{\sigma^2 \cdot \sum_{i=2}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sigma^2 \cdot x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - 2\bar{x}\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

Вместо неизвестной дисперсии σ^2 запишем её несмещенную оценку S^2, получим исправленную дисперсию

$$S_{\hat{y}}^2(x) = S^2\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

Статистикой будет случайная величина

$$T_{\hat{y}}(x) = \frac{\hat{y}(x) - (\beta_0 + \beta_1 x_i)}{S_{\hat{y}}(x)},$$

имеющая распределение Стьюдента с числом степеней свободы $v = n - 2$.

По таблице распределения Стьюдента с числом степеней свободы $v = n - 2$ и уровню значимости α или с надежностью $\gamma = 1 - \alpha$ можно найти критическую точку распределения Стьюдента $T_{крит.} = | t_{(\gamma,v)} | = | t_{(\gamma,n-2)} |$.

Из соотношения, выражающего статистики с доверительной вероятностью $\gamma = 1 - \alpha$ будет выполняться следующее соотношение:

$$T_{\hat{y}}(x) \le | t_{(\gamma,n-2)} |;$$

$$\hat{y}(x) - t_{(\gamma,n-2)} S_{\hat{y}}(x) \le (\beta_0 + \beta_1 x_i) \le \hat{y}(x) + t_{(\gamma,n-2)} S_{\hat{y}}(x);$$

Так как $M[\hat{y}(x_i)] = \beta_0 + \beta_1 x_i$, то интервал:

$$\hat{y}(x) - t_{(\gamma,n-2)} S_{\hat{y}}(x); \quad \hat{y}(x) + t_{(\gamma,n-2)} S_{\hat{y}}(x)$$

есть интервальная оценка (доверительный интервал) для $M(Y \mid X = x_i)$.

Доверительный интервал можно записать в виде:

$$b_0 + b_1 x_i) - t_{(\gamma,n-2)} S_{\hat{y}}(x); (b_0 + b_1 x_i) + t_{(\gamma,n-2)} S_{\hat{y}}(x)$$

или в виде:

$$(b_0 + b_1 x_i) - t_{(\gamma,n-2)} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x - \bar{x})^2}}; (b_0 + b_1 x_i) + t_{(\gamma,n-2)} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x - \bar{x})^2}},$$

S – есть эмпирический стандарт, исправленное среднее квадратическое отклонение.

Пусть y_0 некоторое индивидуальное значение зависимой переменной Y. Доверительный интервал для y_0 зависимой переменной Y будет определяться с учетом рассеяния индивидуальных значений вокруг линии регрессии $M(Y \mid X)$. Тогда следует учитывать изменение оценки дисперсии $S^2_{\hat{y}}(x)$ на величину S^2 (несмещенную оценку S^2 дисперсии σ^2). Несмещенная оценка дисперсии $S^2_{\hat{y}_0}(x)$ запишется в виде:

$$S^2_{\hat{y}_0}(x) = S^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + S^2 = S^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right],$$

соответствующая интервальная оценка

$$\hat{y}(x) - t_{(\gamma,n-2)} S_{\hat{y}_0}(x); \hat{y}(x) + t_{(\gamma,n-2)} S_{\hat{y}_0}(x);$$

$$(b_0 + b_1 x_i) - t_{(\gamma,n-2)} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x - \bar{x})^2}}; (b_0 + b_1 x_i) + t_{(\gamma,n-2)} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x - \bar{x})^2}}.$$

Полученный интервал определяет границы, в пределах которого, находятся не менее $(1-\alpha)$ 100% точек наблюдений $X = x_i$.

Лекция Нелинейная регрессия

Относительно оцениваемых параметров применяются *нелинейные модели* (логарифмическая, полиномиальная, степенная, обратная, экспоненциальная и др.) и *модели, приводимые к линейному виду.*

Нелинейные парные регрессии, определяются видом аналитической зависимости и делятся на два вида:

1) нелинейность по независимой переменной *X*, но линейные по оцениваемым параметрам и регрессии.

К таким функциям относятся:

– *логарифмическая функция* $\boldsymbol{y} = \beta_0 + \beta_1 \cdot \mathbf{ln}\boldsymbol{x} + \varepsilon$;

– *гиперболическая функция, равносторонняя гипербола* $\boldsymbol{y} = \beta_0 + \beta_1 \frac{1}{\boldsymbol{x}} + \varepsilon$;

– *полиноминальная функция* (*степени* k) $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{x}^2 + \cdots + \beta_{\boldsymbol{k}} \boldsymbol{x}^{\boldsymbol{k}} + \varepsilon$;.

2) нелинейность по коэффициентам β , то есть по оценочным параметрам, уравнения регрессии:

– *показательная функция* $\boldsymbol{y} = \beta_0 \cdot \beta_1^{\boldsymbol{X}} \cdot \varepsilon$;

– *степенная функция* $\boldsymbol{y} = \beta_0 \cdot \boldsymbol{x}^{\beta_1} \cdot \varepsilon$;

– *экспоненциальная функция* $\boldsymbol{y} = \beta_0 \cdot \boldsymbol{e}^{\beta_1 \boldsymbol{X}}$;

– *логистическая функция* $y = \frac{\beta_0}{1 + e^{\beta_1 + \delta}} \varepsilon.$

Большинство нелинейных моделей может быть сведено к линейной модели либо заменой переменных, либо путем логарифмирования. Если нелинейные уравнения регрессии невозможно свести к линейному виду *называют внутренне нелинейными* уравнениями.

Нелинейность по независимой переменной X

Построение нелинейной регрессионной модели сводится к оценкам параметров её уравнения регрессии. Оценки параметров регрессий, после линеаризации нелинейного уравнения, выполняется по МНК. Это позволяет получить оценки параметров с минимальным отклонением фактических значений результативного признака y от теоретических значений $\hat{y}_x$.

Вычислив параметры линейного уравнения регрессии с новыми переменными по МНК, возвращаемся к исходному нелинейному уравнению регрессии: $\hat{y}(x) = b_0 + b_1 g(X)$, здесь $g(X)$ есть некоторая нелинейная функция, обычно к таким функциям относятся:

– *логарифмическая функция* $y = \beta_0 + \beta_1 \cdot \ln x + \varepsilon$;

– *гиперболическая функция, равносторонняя гипербола* $y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$;

– *полиноминальная функция* (*степени k*) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$.

Тесноту связи исследуемых нелинейных зависимостей оцениваются индексом корреляции ρ_{xy} $(0 \le \rho_{xy} \le 1)$:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}}.$$

Качество построенной модели определяют величиной коэффициента детерминации R^2 или средняя ошибка аппроксимации $\overline{A}$, которая показывает среднее отклонение расчетных значений от фактических величин, (допустимый предел $\overline{A} \le (\overline{8;10})\%$):

$$\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right| \cdot 100\%.$$

Рассмотрим преобразования регрессионных моделей вида

$$Y = \beta_0 + \beta_1 \cdot \ln x + \varepsilon$$

и приведение их к линейному виду с помощью введения новых переменных.

Уравнение регрессии этой модели имеет вид:

$$\hat{y} = b_0 + b_1 \cdot \ln x.$$

Логарифмическую модель $\ln Y = \ln \beta_0 + \beta_1 \cdot \ln X + \ln \varepsilon$ парной регрессии можно обобщать на число более двух переменных, переходя к моделям множественной регрессии:

$$\ln Y = \ln \beta_0 + \beta_1 \cdot \ln X_1 + ..., \beta_i \cdot \ln X_i + ..., \beta_n \cdot \ln X_n + \ln \varepsilon,$$

для модели с функцией Кобба-Дугласа:

$$\ln Y = \ln \beta_0 + \beta_1 \cdot \ln X_1 + \beta_2 \cdot \ln X_2 + \ln \varepsilon,$$

где коэффициенты β_1, β_2, есть *эластичности* переменной Y по переменным X_1 и X_2; для модели с функцией Энгеля $Y = AX^{\beta}$:

$$\ln Y = \ln \beta_0 + \beta \cdot \ln X + \ln \varepsilon.$$

В логарифмической модели зависимая и объясняющая переменные, не являясь линейными, заданы в логарифмическом виде с учетом случайной погрешности ε.

Уравнение

$$\ln Y = \ln\beta_0 + \beta_1 \cdot \ln X_1 + ..., \beta_i \cdot \ln X_i + ..., \beta_n \cdot \ln X_n + \ln\varepsilon$$

является линейным относительно натуральных логарифмов величин, входящих в уравнение. Вводя замену величин: $ln\ Y = Y^{\circ}$, $\ln\beta_0 = \beta_0^{\circ}$; $\ln X_i = X_i^{\circ}$, $\varepsilon^{\circ} = \ln\varepsilon$, получаем линейную регрессионную модель:.

$$Y^{\circ} = \beta_0^{\circ} + \beta_1 X_1^{\circ} + ... + \beta_i X_i^{\circ} + ... + \beta_n X_n^{\circ} + \varepsilon^{\circ}.$$

Например, экономическая зависимость моделируется функцией Энгеля

$$Y = AX^{\beta}\varepsilon,$$

прологарифмировав обе части функции, получаем уравнение, линейное относительно логарифмов:

$$\ln Y = \ln\beta_0 + \beta \cdot \ln X + \ln\varepsilon,$$

вводя замену, имеем линейную регрессионную модель:

$$Y^{\circ} = \beta_0^{\circ} + \beta X^{\circ} + \varepsilon^{\circ}.$$

Полученной модели соответствует уравнение линейной регрессии:

$$\hat{y}^{\circ} = b_0^{\circ} + b_1 x^{\circ},$$

где $\hat{y}^{\circ} = \ln Y$; $\ln\beta_0 = b_0^{\circ}$; $\ln X = x^{\circ}$.

Для построения модели, определяющей темпы роста (прироста) показателей (рост инфляции от объёма денежной массы; рост затрат какого-либо ресурса и т. д.) обычно используются *полулогарифмические модели вида*:

$$\ln Y = \ln\beta_0 + \beta \cdot X + \ln\varepsilon; \quad Y = \ln\beta_0 + \beta \cdot \ln X + \ln\varepsilon.$$

Например, зависимость сложного темпа прироста r процентной ставки вклада Y в некоторый момент времени t (Y_t) от первоначальной величины банковского вклада Y_0, определяется зависимостью:

$$Y_t = Y_0(1+r)^t \varepsilon_t,$$

здесь ε_t – случайная величина, которая может появиться в силу возможного изменения процентной ставки.

Прологарифмировав обе части функции, получаем уравнение, линейное относительно логарифмов:

$$\ln Y_t = \ln Y_0 + t \cdot \ln(1+r) + \ln\varepsilon_t,$$

вводя обозначения $ln\ Y_0 = \beta_0$, $ln\ (1+ r) = \beta$, $ln\ \varepsilon_t = \varepsilon_t^{\circ}$, получаем *полулогарифмическую модель вида*:

$$\ln Y_t = \beta_0 + \beta t + \varepsilon_t^{\circ},$$

её можно свести к линейной вводя обозначение

$$ln\ Y_t = Y^{\circ};\ \text{тогда}\ Y^{\circ} = \beta_0 + \beta t + \varepsilon_t^{\circ}.$$

Темп прироста вклада Y определяется коэффициентом β, умножив β на 100 получим процентный прирост вклада.

Нелинейность по коэффициентам регрессии

Если в уравнение взаимосвязь переменных нелинейным образом зависит от коэффициентов регрессии, то такие уравнения являются нелинейными по коэффициентам β_0 и β_1.

Построение модели показательной функции

$$\boldsymbol{Y} = \beta_0 \cdot \beta_1^{\boldsymbol{x}} \cdot \varepsilon$$

выполняется в следующем порядке.

Прологарифмируем обе части этой функции, данная модель примет вид:

$$\mathbf{ln}(\boldsymbol{Y}) = \mathbf{ln}(\beta_0) + \boldsymbol{x}\,\mathbf{ln}(\beta_1) + \mathbf{ln}\varepsilon\,.$$

$$\boldsymbol{Y}^{\circ} = \beta_0^{\circ} + \boldsymbol{x}\beta_1^{\circ} + \varepsilon^{\circ}\,.$$

Для получения параметров модели значения *Y* следует заменить значениями логарифмов, а значения *x* не изменять, то есть задавать так же как до логарифмирования:

$\boldsymbol{Y}^{\circ} = \mathbf{ln}\boldsymbol{Y}$; $\mathbf{ln}\beta_0 = \beta_0^{\circ}$; $\mathbf{ln}\beta_1 = \beta_1^{\circ}$, $\varepsilon^{\circ} = \mathbf{ln}\varepsilon$, получим

$$\boldsymbol{Y}^{\circ} = \beta_0^{\circ} + \boldsymbol{x}\beta_1^{\circ} + \varepsilon^{\circ}\,.$$

Полученной модели соответствует уравнение линейной регрессии:

$$\hat{\boldsymbol{y}}^{\circ} = \boldsymbol{b}_0^{\circ} + \boldsymbol{b}_1^{\circ}\,\boldsymbol{x}\,,$$

где $\hat{\boldsymbol{y}}^{\circ} = \mathbf{ln}\boldsymbol{Y}$; $\mathbf{ln}\beta_0 = \boldsymbol{b}_0^{\circ}$; $\mathbf{ln}\beta_1 = \boldsymbol{b}_1^{\circ}$.

Коэффициенты $\boldsymbol{b}_0^{\circ}$ и $\boldsymbol{b}_1^{\circ}$ вычисляются с использованием МНК. Коэффициенты $\boldsymbol{b}_0^{\circ}$ и $\boldsymbol{b}_1^{\circ}$ будут эффективными оценками β_0, β_1, но при этом может возникать некоррелированность отклонения ε_i, тогда, отклонения ε_i исходной модели должны иметь логарифмически нормальное распределение

$$\varepsilon^{\circ} = ln\ \varepsilon \sim N(0,\ \sigma^2).$$

При переходе к линейной модели с помощью логарифмирования получаем измененные значения параметров, далее, выполняя обратные преобразования, поучаем параметры исходного уравнения регрессии.

Построение модели *степенной функции*

$$\boldsymbol{Y} = \beta_0 \cdot \boldsymbol{X}^{\beta_1} \cdot \varepsilon$$

выполняется в следующем порядке.

Путем логарифмирования проведем линеаризацию переменных:

$$\mathbf{ln}\boldsymbol{Y} = \mathbf{ln}\beta_0 + \beta_1 \mathbf{ln}\boldsymbol{X} + \mathbf{ln}\varepsilon,$$

введем замену логарифмов величины:

$$ln\ Y = Y^{\circ},\ \mathbf{ln}\beta_0 = \beta_0^{\circ};\quad \mathbf{ln}\boldsymbol{X} = \boldsymbol{X}^{\circ},\ \varepsilon^{\circ} = \mathbf{ln}\varepsilon,$$

относительно введенных величин получаем линейную регрессионную модель: $\boldsymbol{Y}^{\circ} = \beta_0^{\circ} + \beta_1\, \boldsymbol{X}^{\circ} + \varepsilon^{\circ}$.

Полученной модели соответствует уравнение линейной регрессии:

$$\hat{\boldsymbol{y}}^{\circ} = \boldsymbol{b}_0^{\circ} + \boldsymbol{b}_1\, \boldsymbol{x}^{\circ},$$

где $\hat{\boldsymbol{y}}^{\circ} = \mathbf{ln}\boldsymbol{Y}$; $\mathbf{ln}\beta_0 = \boldsymbol{b}_0^{\circ}$; $\mathbf{ln}\boldsymbol{X} = \boldsymbol{x}^{\circ}$.

Коэффициенты $\boldsymbol{b}_0^{\circ}$ и $\boldsymbol{b}_1^{\circ}$ вычисляются с использованием МНК. Коэффициенты $\boldsymbol{b}_0^{\circ}$ и $\boldsymbol{b}_1^{\circ}$ будут эффективными оценками β_0, β_1, но при этом может возникать некоррелированность отклонения ε_i, тогда, отклонения ε_i исходной модели должны иметь логарифмически нормальное распределение

$$\varepsilon^{\circ} = ln\ \varepsilon \sim N(0, \sigma^2).$$

При переходе к линейной модели с помощью логарифмирования получаем измененные значения параметров, далее, выполняя обратные преобразования, поучаем параметры исходного уравнения регрессии.

Пример выполнения заданий лабораторной работы по темам:

Линейная и нелинейная парная регрессия

I. Парная линейная регрессия

II. Нелинейная парная регрессия

Задание I. По данным выборки, представленной в таблице № 1 определить:

1) вид зависимости между Y и X,

2) вычислить коэффициенты b_0 и b_1 уравнения регрессии, являющихся оценками параметров регрессии β_0, β_1;

3) построить уравнение регрессии;

4) оценить силу линейной зависимости между Y и X;

5) проверить статистическую значимость коэффициентов b_1 и b_0 уравнения регрессии с уровнем значимости $\alpha = 0{,}05$;

6) проверить значимость построенного уравнения регрессии с уровнем значимости $\alpha = 0{,}05$;

7) определить коэффициент детерминации и оценить статистическую значимость уравнения регрессии;

8) построить интервальные оценки для коэффициентов регрессии β_0 и β_1 с надежностью $\gamma = 0{,}95$ и интервальную оценку для построенного уравнения регрессии с уровнем значимости $\alpha = 0{,}05$;

Задание II. По данным выборки, представленной в таблице №1 используя параметры построенного линейно регрессионного уравнения, построить и рассчитать параметры модели:

1) степенной функции;

2) показательной функции;

3) оценить построенные модели через ошибку аппроксимации и F – критерий Фишера.

4) определить, какая из построенных регрессионных моделей, лучше описывает взаимосвязь Y, X.

Таблица №1 результатов выборки:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50

I. **Выполнение задания лабораторной работы по теме:**

Парная линейная регрессия

1.1. *Определить вида зависимости между Y и X.*

Построим корреляционное поле:

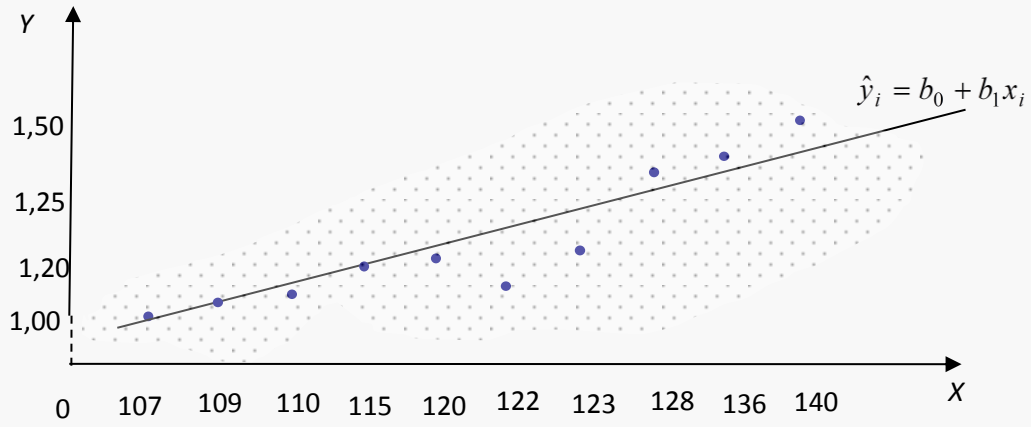


Рисунок 0.1. Корреляционное поле зависимости Y от X

Расположение точек корреляционного поля зависимости *Y* от *X* дает возможность сделать вывод, что ***зависимость линейная*** $\hat{y}_i = b_0 + b_1 x_i$.

1.2. *Вычислить коэффициенты b_0 и b_1 уравнения регрессии.*

Вычислим коэффициенты b_0 и b_1, для этого по заданной выборке найдем значения:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i;$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1210}{10} = 121; \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{147548}{10} = 14754{,}8; \quad \overline{y^2} = 1{,}545;$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{12{,}35}{10} = 1{,}235; \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1509{,}35}{10} = 150{,}935;$$

$$b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{150{,}935 - 121 \cdot 1{,}235}{14754{,}8 - 121^2} = \frac{150{,}935 - 149{,}435}{14754{,}8 - 14641} = \frac{1{,}5}{113{,}8} = 0{,}013;$$

$$b_0 = \bar{y} - b_1 \bar{x} = 1{,}235 - 0{,}013 \cdot 121 = 1{,}235 - 1{,}573 = -0{,}338.$$

b_1 = 0,013; b_0 = – 0,338.

1.3. *Построить уравнение регрессии.*

Запишем уравнение регрессии и построим прямую линию регрессии на корреляционном поле:

$$\hat{y}_i = b_0 + b_1 x_i;\ \ \hat{y}_i = -0{,}338 + 0{,}013 x_i.$$

Уравнение регрессии: $\hat{y}_i = -0{,}338 + 0{,}013 x_i$

1.4. *Оценить силу линейной зависимости между Y и X*

Чтобы оценить силу линейной зависимости между Y и X, вычислим коэффициент корреляции:

$$r_{xy} = b_1 \frac{S_x}{S_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} = \frac{1{,}5}{\sqrt{113{,}8} \cdot \sqrt{1{,}545 - 1{,}525}} = \frac{1{,}5}{1{,}513} = 0{,}9914.$$

Величина коэффициента корреляции близка к единицы, можно сделать вывод о ***сильной линейной зависимости Y от X.***

1.5. *Проверить статистическую значимость коэффициентов b_1 и b_0 уравнения регрессии.*

Пользуясь построенным уравнением регрессии

$$\hat{y}_i = -0{,}338 + 0{,}013 x_i,$$

вычислим y_i для каждого значения независимой переменной x_i, затем вычислим значения $(y_i - \bar{y})^2$:

$\bar{y}$ = 1,235; $\bar{x}$ = 121; $\overline{x^2}$ = 14754; $\overline{y^2}$ = 1,545; $\overline{xy}$ = 150,935; $(y_i - \bar{y})^2$ = 0,223.

Составим таблицу № 2 и запишем значения вычисленных величин:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$(y_i - \bar{y})^2$	0,055	0,024	0,009	0,0012	0,00022	0,0132	0,003	0,013	0,034	0,0702

Проверим гипотезу о статистической значимости коэффициентов b_1 и b_0.

$H_0: \beta_0 = 0$, (коэффициент b_0 незначим);

$H_1: \beta_0 \neq 0$, (коэффициент b_0 значим).

Вычислим S_{b_1} и наблюдаемое значение T_{b_1}:

$$S_{b_1}^2 = \frac{S^2}{n(\overline{x^2} - \bar{x}^2)} = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1)^2}{n(n-2)(\overline{x^2} - \bar{x}^2)} = \frac{0{,}223}{10 \cdot 8 \cdot (14754 - 121^2)} = 0{,}000025;$$

$$S_{b_1} = \sqrt{S_{b_1}^2} = \sqrt{0{,}000025} = 0{,}005;$$

$$T_{b_1} = \frac{b_1}{S_{b_1}} = \frac{0{,}013}{0{,}005} = 2{,}6.$$

Определим критические значения $T_{кр.} = t_{(0{,}025;\ 8)} = 2{,}306$. Сравним T_{b_1} и $T_{кр}$: $T_{b_1} = 2{,}6 > T_{кр} = 2{,}3$ из сравнения следует, что нулевая гипотеза отвергается в пользу альтернативной.

Коэффициента уравнения регрессии b_1 статистически значим с уровнем значимости $\alpha = 0{,}05$.

Для проверки значимости коэффициента b_0 вычислим стандартную ошибку S_{b_0} и наблюдаемое значение T_{b_0}:

$$S_{b_0}^2 = \frac{S^2 \sum_{i=1}^{n} (x_i)^2}{n(\overline{x^2} - \bar{x}^2)} = S_{b_1}^2 \cdot \overline{x^2} = 0{,}000025 \cdot 14754 = 0{,}369;$$

$$S_{b_0} = \sqrt{0{,}369} = 0{,}607;$$

$$|T_{b_0} = \left| \frac{b_0}{S_{b_0}} \right| = \left| \frac{-0{,}338}{0{,}607} \right| = |-0{,}556| .$$

Так как $|-0{,}556| < 2{,}306$, то принимается нулевая гипотеза. ***Коэффициент b_0 незначим с уровнем значимости*** $\alpha = 0{,}05$.

Таким образом, свободным членом построенного уравнения регрессии $b_0 = -0{,}338$ можно пренебречь.

1.6. *Проверить значимость построенного уравнения регрессии с уровнем значимости* $\alpha = 0{,}05$.

Пользуясь построенным уравнением регрессии

$$\hat{y}_i = -0{,}338 + 0{,}013 x_i ,$$

вычислим $\hat{y}_i$ для каждого значения независимой переменной x_i, затем вычислим значения $(\hat{y}_i - y_i)^2$:

$\bar{y} = 1{,}235$; $(y_i - \bar{y})^2 = 0{,}223$; $b_1 = 0{,}013$; $b_0 = -0{,}338$;

$U = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 0{,}223$; $U_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0{,}037$; $U_r = U - U_e = 0{,}186$.

Дополним таблицу №2 значениями $\hat{y}_i$ и $(\hat{y}_i - y_i)^2$:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$(y_i - \bar{y})^2$	0,055	0,024	0,009	0,0012	0,00022	0,0132	0,003	0,013	0,034	0,0702
$\hat{y}_i$	1,053	1,079	1,092	1,157	1,222	1,248	1,261	1,326	1,380	1,432
$(\hat{y}_i - y_i)^2$	0, 000	0,000	0,002	0,007	0,001	0,016	0,001	0,0006	0,0016	0,005

Уравнение регрессии признается статистически значимым при уровне значимости $\alpha = 0{,}05$, если фактическое значения F – критерия Фишера больше табличного значения $F_{табл.}(\alpha, \nu_1, \nu_2)$. Вычислим критерий Фишера:

$$F = \frac{U_r (n-2)}{U_e} = \frac{0{,}186 \cdot 8}{0{,}037} = 40{,}21,$$

табличное значение $F_{табл.}$ (0, 05; 1; 8) = 5,32, выполняется неравенство:

$$F = 40{,}21 > F_{(0{,}05;1;8)} = 5{,}32.$$

Уравнение регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$ ***в целом значимо с уровнем значимости*** α = 0,05.

1.7. *Определить коэффициент детерминации и оценить статистическую значимость уравнения регрессии.*

По полученным данным, определим коэффициент детерминации для уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$:

$$R^2 = 1 - \frac{U_e}{U} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{0{,}037}{0{,}223} = 0{,}834.$$

Значение коэффициента детерминации $R^2 = 0{,}834$ определяет достаточно ***высокое качество построенного уравнения регрессии*** $\hat{y}_i = -0{,}338 + 0{,}013x_i$.

Для проверки того, что уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$ незначимо, рассмотрим следующие статистические гипотезы с уровнем значимости α = 0,05:

основная гипотеза H_0: уравнение регрессии незначимо;

конкурирующая гипотеза H_1: уравнение регрессии значимо.

Запишем F – критерий Фишера по заданному уровню значимости α = 0,05:

$$F_{наб} = \frac{(n-m-1)R^2}{m(1-R^2)} = \frac{8 \cdot 0{,}834}{1 \cdot 0{,}166} = \frac{6{,}672}{0{,}166} = 40{,}19;$$

$$F_{крит.} = F_{(0{,}05;1;8)} = 5{,}32.$$

Так как $F_{наб..} = 40{,}19 > F_{крит.} = 5{,}32$, то основная гипотеза H_0 – уравнение регрессии незначимо *отвергается.* Следовательно, уравнения регрессии $\hat{y}_i = -0{,}338 + 0{,}013x_i$ ***значимо***.

Средняя ошибка аппроксимации оценит качество модели:

$$\bar{A} = \frac{1}{n}\sum_{1}^{n}\left|\frac{y - \hat{y}}{y}\right| 100\% = \frac{1}{10} \cdot 0{,}44 \cdot 100\% = 4{,}4\%$$

Качество построенной модели ***хорошее***, средняя ошибка аппроксимации меньше 10%.

1.8. *Построить интервальные оценки для коэффициентов регрессии* β_0 и β_1 *с надежностью* γ = 0,95 *и интервальную оценку для построенного уравнения регрессии с уровнем значимости* α = 0,05.

Пользуясь полученными результатами выполненных вычислений, построим интервалы оценки β_0 и β_1 с надежностью γ = 0,95:

$b_0 = -0{,}338$; $b_1 = 0{,}013$; $S_{b_0} = 0{,}607$; $S_{b_1} = 0{,}005$; $|t_{(0{,}95;,8)}| = 2{,}31$.

Построим интервал для оценочного коэффициента b_0 :

$$[b_0 - t(\gamma,\ n-2)\ S_{b_0};\ b_0 + t(\gamma,\ n-2)\ S_{b_0}],$$

$$[-0{,}338 - 2{,}31\cdot 0{,}607;\quad -0{,}338 + 2{,}31\cdot 0{,}607];$$

$$b_0 \in [-1{,}74;\ 1{,}06].$$

Так как нижняя граница интервала отрицательная, то при положительной верхней границе интервала, оцениваемый параметр b_0 считается нулевым и β_0
– незначим, точный интервал β_0 *можно не определять.*

Интервал для оценочного коэффициента b_1:

$$[b_1 - t(\gamma,\ n-2)\ S_{b_1};\quad b_1 + t(\gamma,\ n-2)\ S_{b_1}],$$

$$[0{,}013 - 2{,}31\cdot 0{,}005;\quad 0{,}013 + 2{,}31\cdot 0{,}005];$$

$$b_1 \in [0{,}00145;\ 0{,}02455].$$

Значит, $\beta_0 \subset [-1{,}74;\ 1{,}06]$; $\beta_1 \subset [0{,}00145;\ 0{,}02455]$.

Для *интервальной оценки функции регрессии* построенного уравнения $\hat{y}_i = -0{,}338 + 0{,}013x_i$ получены $b_0 = -0{,}338$; $b_1 = 0{,}013$; $|t_{(0{,}95;,8)}| = 2{,}31$, задано $X = 140$, вычислим

$$(X - \bar{x})^2 = (140-121)^2 = 361;\quad \sum_{i=1}^{n}(x_i - \bar{x}_B)^2 = 1138;$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_B)^2} = \sqrt{\frac{1}{9}\cdot 1138} = 11{,}24;$$

$$(b_0 + b_1 X_i) - t_{(\gamma,n-2)} S\sqrt{\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = -0{,}338 + 0{,}013\cdot 140 - 2{,}31\cdot 11{,}24\sqrt{\frac{1}{10} + \frac{361}{1138}} = -15{,}62;$$

$$(b_0 + b_1 X) + t_{(\gamma,n-2)} S\sqrt{\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum_{i=1}^{n}(x - \bar{x})^2}} = -0{,}338 + 0{,}013\cdot 140 + 2{,}31\cdot 11{,}24\cdot 0{,}646 = 19{,}25.$$

Запишем ***доверительный интервал для среднего значения Y при X* = 140**:

$$(-15{,}62;\ 18,\ 25).$$

Доверительный интервал зависимой переменной Y будет определяться с учетом рассеяния индивидуальных значений вокруг линии регрессии $M(Y \mid X)$. Определим интервал, в пределах границ которого, находятся все возможные значения y_0 зависимой переменной Y, то есть не менее 95% точек наблюдений $X = x_i$, а за пределами интервала не более 0,05%.

$$(b_0 + b_1 X_i) - t_{(\gamma,n-2)} S\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = -29{,}42;$$

$$(b_0 + b_1 X_i) + t_{(\gamma,n-2)} S\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 32{,}39.;$$

Доверительный интервал, включающий все возможные случайные значения y_0 некоторой индивидуальной зависимой переменной Y :

$$(-29{,}42;\ 32{,}39).$$

Полученный интервал определяет границы, в пределах которого, находятся не менее $(1-\alpha)$ 100% точек наблюдений $X = x_i$.

Вычислим коэффициент эластичности

$$\overline{Э} = b_1 \frac{\overline{x}}{\overline{y}} = 0{,}013\frac{121}{1{,}235} = 1{,}27.$$

Средний коэффициент эластичности показывает, что по совокупности результат Y от своей средней величины изменится на 1,27 процента при изменении фактора X на 1 процент.

II. **Выполнение задания лабораторной работы по теме:**

Нелинейная парная регрессия

Выполнение задания лабораторной работы по теме:

Нелинейная парная регрессия

Задание II. По данным выборки, представленной в таблице №1 используя параметры построенного линейно регрессионного уравнения, построить и рассчитать параметры модели:

1) степенной функции;

2) показательной функции;

3) оценить построенные модели через ошибку аппроксимации и индекс корреляции;

4) определить, какая из построенных регрессионных моделей, лучше описывает взаимосвязь *Y, X.*

2.1 *Построить и рассчитать параметры модели степенной функции*

Рассмотрим функциональную зависимость $\boldsymbol{y = b_0 x^{b_1}}$.

Рассчитаем параметры степенной модели $\boldsymbol{Y = \beta_0 \cdot X^{\beta_1} \cdot \varepsilon}$.

Путем логарифмирования проведем линеаризацию переменных:

$$\boldsymbol{\ln Y = \ln \beta_0 + \beta_1 \ln X + \ln \varepsilon},$$

введем замену величин: $ln\ Y = Y^{\circ}$, $\boldsymbol{\ln \beta_0 = \beta_0^{\circ}}$; $\boldsymbol{\ln X = X^{\circ}}$, $\boldsymbol{\varepsilon^{\circ} = \ln \varepsilon}$,

получаем линейную регрессионную модель: $\boldsymbol{Y^{\circ} = \beta_0^{\circ} + \beta_1 X^{\circ} + \varepsilon^{\circ}}$.

Оценкой полученной модели будет уравнение линейной регрессии:

$$\boldsymbol{\hat{y}^{\circ} = b_0^{\circ} + b_1 x^{\circ}},$$

где $\hat{y}^{\circ} = \ln Y$; $\ln \beta_0 = b_0^{\circ}$; $\ln X = x^{\circ}$.

Составим таблицу

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$\ln y = y^{\circ}$	0,049	0,077	0,131	0,182	0,223	0,113	0,254	0,300	0,350	0,405

Найдем значения:

$$\sum_{i=1}^{10} y_i^{\circ} = 2{,}084; \qquad \overline{y^{\circ}} = \frac{1}{n}\sum_{i=1}^{n} y_i^{\circ} = \frac{1}{10} \cdot 2{,}084 = 0{,}208;$$

$$\sum_{i=1}^{n} (x_i^{\circ})^2 = 229{,}52; \qquad \overline{(x^{\circ})^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i^{\circ})^2 = \frac{229{,}52}{10} = 22{,}952;$$

$$\overline{x^{\circ} y^{\circ}} = \frac{1}{n}\sum_{i=1}^{n} (x_i^{\circ} \cdot y_i^{\circ}) = \frac{10{,}071}{10} = 1{,}007;$$

Вычислим значения коэффициента линейной регрессии b_0°, b_1

$$b_1 = \frac{\overline{x^\circ y^\circ} - \bar{x}^\circ \bar{y}^\circ}{\overline{x^{\circ 2}} - (\bar{x}^\circ)^2} = \frac{1{,}007 - 4{,}79 \cdot 0{,}208}{22{,}952 - 4{,}79^2} = \frac{0{,}0106}{0{,}008} = 1{,}3;$$

$$b_0^\circ = \bar{y}^\circ - b_1 \overline{x^\circ} = 0{,}208 - 1{,}3 \cdot 4{,}79 = -6{,}0 \ .$$

По полученным значениям запишем линейное уравнение:

$$\hat{y}^\circ = -6{,}0 + 1{,}3\, x^\circ,$$

выполним потенцирование:

$$\hat{y} = e^{-6} x^{1{,}3} = 0{,}0025 x^{1{,}3}.$$

Уравнение степенной регрессии имеет вид: $\hat{y} = 0{,}0025 x^{1{,}3}$.

Определим теоретическое значение результата $\hat{y}_x$ и вычислим необходимые величины для определения индекса корреляции ρ_{xy}.

Результаты вычислений запишем в таблицу:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
$\ln x = x^\circ$	4,67	4,69	4,70	4,74	4,79	4,80	4,81	4,85	4,91	4,94
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$\ln y = y^\circ$	0,049	0,077	0,131	0,182	0,223	0,113	0,254	0,300	0,350	0,405
$\hat{y}_x$	1,086	1.113	1,126	1,193	1,261	1,288	1,302	1,371	1,484	1,541
$y - \hat{y}$	–0,036	–0,033	0,014	0,007	–0,011	–0,168	–0,012	–0,021	–0,084	–0,041
$(y - \hat{y}_x)^2$	0,0013	0,0011	0,0002	0.00005	0,00012	0,028	0,00014	0,00044	0,0041	0,0017
$\left\lvert \frac{y - \hat{y}}{y} \right\rvert$	–0,034	–0,031	0,123	0,006	–0,0088	–0,15	–0,009	–0,0155	–0,059	–0,027

Вычислим: $\sum_1^{10} (y - \hat{y}_x)^2 = 0{,}03749$; $(y_i - \bar{y})^2 = 0{,}223$;

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}} = \sqrt{1 - \frac{0{,}03749}{0{,}223}} = 0{,}912.$$

Критерий оценок тесноты связи определяет индекс корреляции, находящийся в интервале $0{,}7 \le |\rho_{xy}| = 0{,}0{,}912 < 1{,}0$ это говорит о сильной связи *н*елинейной зависимости, заданной уравнением $\hat{y} = 0{,}0025 x^{1{,}3}$.

Вычислим среднюю ошибку аппроксимации $\overline{A}$:

$$\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right| \cdot 100\% = \frac{1}{10} \left| -0{,}2053 \right| \cdot 100\% = 2{,}05\% \ .$$

Средняя ошибка аппроксимации говорит о наличии незначительных отклонений расчетных значений от фактических величин.

следовательно, среднее отклонение расчетных значений от фактических значительно меньше в регрессионной линейной модели $Y = \beta_0 + \beta_1 X + \varepsilon$.

Средняя ошибка аппроксимации построенного линейного уравнения $\hat{y}_i = -0{,}338 + 0{,}013 x_i$ регрессии – $\overline{A}_{лин.} = 4{,}4\%$; для степенного уравнения регрессии $\hat{y} = 0{,}0025 x^{1{,}3}$ – $\overline{A}_{степ.} = 2{,}05\%$, следовательно, среднее отклонение расчетных значений от фактических значительно меньше в регрессионной степенной модели.

Можно сделать вывод: степенная модель $Y = \beta_0 \cdot X^{\beta_1} \cdot \varepsilon$ описывает взаимосвязь *Y, X* лучше, чем линейная модель $Y = \beta_0 + \beta_1 X + \varepsilon$.

2.2 *Построить и рассчитать параметры модели показательной функции*

Рассмотрим функциональную зависимость $y = b_0 \cdot b_1^X$.

Рассчитаем параметры показательной модели $Y = \beta_0 \cdot \beta_1^{\,X} \cdot \varepsilon$.

Путем логарифмирования проведем линеаризацию переменных:

$$\ln Y = \ln \beta_0 + X \ln \beta_1 + \ln \varepsilon,$$

введем замену величин: ln $Y = Y°$, $\ln \beta_0 = \beta_0^\circ$; $\ln \beta_1 = \beta_1^\circ$, $\varepsilon^\circ = \ln \varepsilon$,

получаем линейную регрессионную модель: $Y^\circ = \beta_0^\circ + X\beta_1^\circ + \varepsilon^\circ$.

Оценкой полученной модели будет уравнение линейной регрессии:

$$\hat{y}^\circ = b_0^\circ + x b_1^\circ,$$

где $\hat{y}^\circ = \ln Y$; $\ln \beta_0 = b_0^\circ$; $\ln \beta_1 = b_1^\circ$.

Вычислим необходимые величины, используя ранее проведенные расчеты:

i	1	2	3	4	5	6	7	8	9	10
x_i	107	109	110	115	120	122	123	128	136	140
y_i	1,05	1,08	1,14	1,20	1,25	1,12	1,29	1,35	1,42	1,50
$\ln y = y^\circ$	0,049	0,077	0,131	0,182	0,223	0,113	0,254	0,300	0,350	0,405
$x \cdot y°$	5,243	8,393	14,41	20,93	26,76	13,78	31,24	38,40	47,60	56,70
$\hat{y}_x$	1,059	1.0817	1,0929	1,150	1,211	1,237	1,249	1,315	1,428	1,489
$y - \hat{y}$	– 0,009	-0,0017	--0,047	0,005	0,039	–0,117	0,041	0,035	–0,008	0,011
$\left\lvert \frac{y - \hat{y}}{y} \right\rvert$	– 0,008	-0,0016	-0,412	0,004	0,0312	–0,104	0,0317	0,0259	-0,0056	0,0073

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1210}{10} = 121; \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{147548}{10} = 14754{,}8;$$

$$\overline{y^\circ} = \frac{1}{n}\sum_{i=1}^{n} y_i^\circ = \frac{1}{10} \cdot 2{,}08 = 0{,}208; \quad \overline{x\,y^\circ} = \frac{1}{n}\sum_{i=1}^{n} (x \cdot y_i^\circ) = \frac{263{,}45}{10} = 26{,}345;$$

Вычислим значения коэффициента линейной регрессии b_0°, b_1°

$$b_1^\circ = \frac{\overline{x\,y^\circ} - \bar{x}\,\bar{y}^\circ}{\overline{x^2} - (\bar{x})^2} = \frac{26{,}345 - 121 \cdot 0{,}208}{14754{,}8 - 121^2} = \frac{1{,}177}{113{,}8} = 0{,}0103;$$

$$b_0^\circ = \bar{y}^\circ - b_1^\circ \bar{x} = 0{,}208 - 0{,}0103 \cdot 121 = -1{,}0419 .$$

По полученным значениям запишем линейное уравнение:

$$\hat{y}^\circ = -1{,}0419 + 0{,}0103x , \quad \hat{y}_x = -1{,}0419 \cdot 0{,}0103^x$$

$$\ln Y = \ln\beta_0 + X \ln\beta_1$$

$$e^Y = e^{-1{,}0419} e^{0{,}0103X} \rightarrow Y = 0{,}352 \cdot 0{,}01^X$$

Определим теоретическое значение результата $\hat{y}_x$ и вычислим необходимые величины для определения индекса корреляции ρ_{xy}.

Вычислим: $\sum_1^{10} (y - \hat{y}_x)^2 = 0{,}0206$; $(y_i - \bar{y})^2 = 0{,}223$;

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}} = \sqrt{1 - \frac{0{,}0206}{0{,}223}} = 0{,}952.$$

Критерий оценок тесноты связи определяет индекс корреляции, находящийся в интервале $0{,}7 \le |\rho_{xy}| = 0{,}0{,}952 < 1{,}0$ это говорит о сильной связи нелинейной зависимости, заданной уравнением $e^Y = e^{-1{,}0419} e^{0{,}0103X} \rightarrow Y = 0{,}352 \cdot 0{,}01^X$.

Вычислим среднюю ошибку аппроксимации $\bar{A}$:

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right| \cdot 100\% = \frac{1}{10} |-0{,}4311| \cdot 100\% = 4{,}31\% .$$

Средняя ошибка аппроксимации построенного показательного уравнения $e^Y = e^{-1{,}0419} e^{0{,}0103X} \rightarrow Y = 0{,}352 \cdot 0{,}01^X$ составляет $\bar{A}_{показ.} = 4{,}31\%$, примерно такой же показатель тесноты связи как в линейной модели, линейного уравнения $\hat{y}_i = -0{,}338 + 0{,}013x_i$ регрессии – $\bar{A}_{лин.} = 4{,}1\%$; для степенного уравнения регрессии $\hat{y} = 0{,}0025x^{1{,}3}$ – $\bar{A}_{степ.} = 2{,}05\%$, следовательно, среднее отклонение расчетных значений от фактических значительно меньше в регрессионной степенной модели.

Можно сделать вывод: степенная модель $Y = \beta_0 \cdot X^{\beta_1} \cdot \varepsilon$ описывает взаимосвязь Y, X лучше, чем линейная и показательная модели.

III **Контрольная задача:**

По семи регионам Краснодарского края за 2004 год проанализированы следующие показатели: расходы на приобретение продовольственных товаров (*y* %) от общих расходов и средняя заработная плата одного работающего в час (*x руб/час.*). Результаты анализа приведены в таблице:

i	1	2	3	4	5	6	7
y%	68,8	61,2	59,9	56,7	55,0	54,3	49,3
x руб/час.	45,1	59,0	57,2	61,8	58,8	47,2	55,2

Требуется рассчитать параметры степенной модели $\boldsymbol{Y} = \beta_0 \cdot \boldsymbol{X}^{\beta_1} \cdot \varepsilon$.