

Лекция № 6. ОСНОВНЫЕ ПОНЯТИЕ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

2.1. ОЦЕНКА ЗАКОНА РАСПРЕДЕЛЕНИЯ

Генеральной совокупностью называется множество объектов, из которых производится выборка. Каждый из объектов задает фиксированное значение случайной величины.

Выборка - множество $\{x_1, x_2, \dots, x_n\}$ случайно отобранных объектов (значений) из генеральной совокупности.

Объемом выборки называется число (n) входящих в нее объектов.

Вариационным рядом называется выборка $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n\}$, полученная в результате расположения значений исходной выборки в порядке возрастания. Значения $\hat{x}_i$ называются вариантами.

2.1.1. Эмпирическая функция распределения

Эмпирическая функция распределения определяется формулой

$$F^*(x) = m_{<x} / n, \quad (2.1)$$

где x - аргумент (неслучайная величина, $-\infty < x < +\infty$);

n - объем выборки;

$m_{<x}$ - количество значений в выборке или вариационном ряду, строго меньших x .

При $n \rightarrow \infty$ эмпирическая функция распределения $F^*(x)$ по вероятности сходится к теоретической функции распределения $F(x)$.

Основные свойства функции $F^*(x)$.

1. $0 \leq F^*(x) \leq 1$.

2. $F^*(x)$ - неубывающая ступенчатая функция.

3. $F^*(x) = 0, x \leq \hat{x}_1$.

4. $F^*(x) = 1, x > \hat{x}_n$.

Эмпирическая функция распределения является наилучшей оценкой закона распределения (несмещенной, состоятельной, эффективной). Недостаток функции $F^*(x)$ заключается в ее невысокой наглядности: визуально сложно подобрать типовой закон распределения.

Порядок построения графика функции $F^*(x)$ следующий.

1. Построить вариационный ряд.

2. На числовой оси x выделить полуинтервалы $(A_i, B_i]$, на которых функция $F^*(x)$ не изменяет своего значения. Границы полуинтервалов определяются соседними отличающимися значениями вариационного ряда.

3. На каждом полуинтервале по формуле (2.1) вычисляется значение функции $F^*(x)$.

4. Построить график.

2.1.2. Гистограмма распределения случайной величины

Гистограммой называется оценка плотности распределения вероятности. На практике наиболее часто используются два метода построения гистограммы: равноинтервальный и равновероятностный. Порядок построения гистограммы следующий.

1. Построить вариационный ряд, т.е. расположить выборочные значения в порядке возрастания: $\hat{x}_1 \leq \hat{x}_2 \leq \dots \leq \hat{x}_n$.

2. Вся область возможных значений $[\hat{x}_1, \hat{x}_n]$ разбивается на M непересекающихся и примыкающих друг к другу интервалов.

Из статистических соображений параметр M рекомендуется выбирать с помощью следующих соотношений:

$$M \approx \text{int}(\sqrt{n}), \quad n \leq 100, \quad (2.2)$$

$$M \approx \text{int}((2-4) \cdot \lg(n)), \quad n > 100, \quad (2.3)$$

где $\text{int}(x)$ - целая часть числа x . Желательно, чтобы n без остатка делилось на M .

Введем обозначения параметров:

A_i, B_i - соответственно левая и правая границы i -го интервала ($A_{i+1} = B_i$);

$h_i = B_i - A_i$ - длина i -го интервала;

v_i - количество чисел в выборке, попадающих в i -ый интервал.

При использовании *равноинтервального* метода построения гистограммы параметры A_i, B_i, h_i вычисляются следующим образом:

$$h_i = h = (\hat{x}_n - \hat{x}_1) / M; \quad A_i = \hat{x}_1 + (i-1) \cdot h; \quad B_i = A_{i+1}; \quad i = 1, 2, \dots, M. \quad (2.4)$$

Если при подсчете значений какое-то число в выборке точно совпадает с границей между интервалами, то необходимо в счетчик обоих интервалов прибавить по 0,5.

В случае применения *равновероятностного* метода границы A_i, B_i выбираются таким образом, чтобы в каждый интервал попадало одинаковое количество выборочных значений:

$$v_i = v = n / M. \quad (2.5)$$

В этом случае

$$A_1 = \hat{x}_1; \quad B_1 = (\hat{x}_v + \hat{x}_{v+1}) / 2; \quad A_2 = B_1; \quad A_i = (\hat{x}_{(i-1)v} + \hat{x}_{(i-1)v+1}) / 2; \quad i = 2, 3, \dots, M. \quad (2.6)$$

3. Вычисляется средняя плотность вероятности для каждого интервала по формуле

$$f_i^* = v_i / (n \cdot h_i). \quad (2.7)$$

4. На графике провести две оси: x и $f^*(x)$.

5. На оси x отмечаются границы всех интервалов.

6. На каждом интервале строится прямоугольник с основанием h_i и высотой f_i^* . Полученная при этом ступенчатая линия называется гистограммой, график которой приблизительно выглядит так, как показано на рис. 2.1.

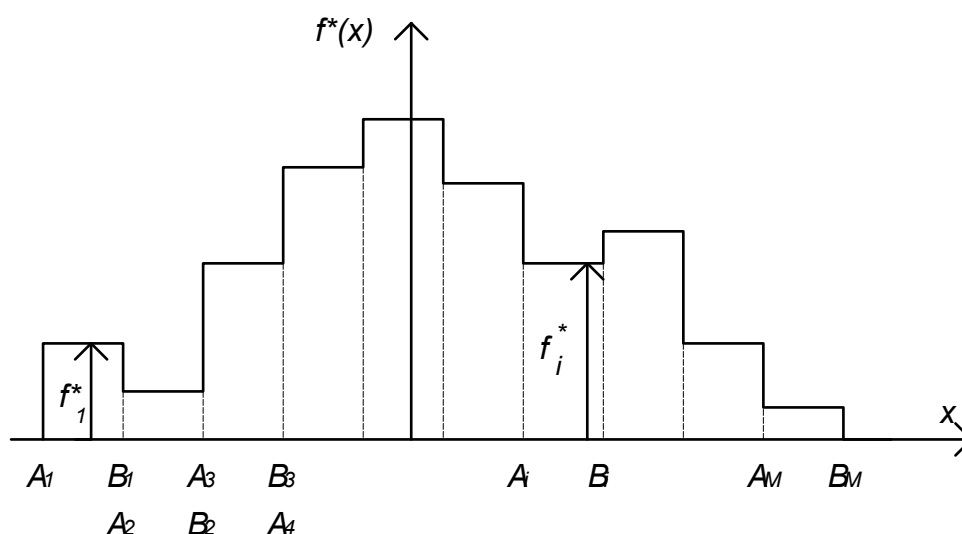


Рис.2.1

Замечания.

1. Суммарная площадь всех прямоугольников равна единице.

2. В равновероятностной гистограмме площади всех прямоугольников одинаковы. По виду гистограммы можно судить о законе распределения случайной величины.

3. Перед построением гистограммы вычисленные значения A_i , B_i , h_i , v_i , f_i^* рекомендуется занести в табл. 2.1.

Таблица 2.1

i	A_i	B_i	h_i	v_i	f_i^*
1					
$\vdots$					
M					

Достоинства использования гистограммы: простота применения, наглядность.

2.2. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ И ЧИСЛОВЫХ ХАРАКТЕРИСТИК

Почти все распределения случайной величины зависят от одного или нескольких параметров. Например, плотность вероятности экспоненциального закона:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

зависит от параметра λ .

Статистической оценкой $\hat{Q}$ параметра Q распределения называется приближенное значение параметра, вычисленное по результатам эксперимента (по выборке).

Статистические оценки делятся на точечные и интервальные. *Точечной* называется оценка, определяемая одним числом. Желательно, чтобы оценка была несмещенной, состоятельной и эффективной.

1. Оценка $\hat{Q}$ называется *несмещенной*, если $M[\hat{Q}] = Q$. Несмещенность - минимальное требование к оценкам.
2. Оценка $\hat{Q}$ называется *состоятельной*, если при увеличении числа n она сходится по вероятности к значению параметра Q :

$$\lim_{n \rightarrow \infty} (P(|\hat{Q} - Q| < \varepsilon)) = 1,$$

где ε - любое положительное число. Несмещенная оценка является состоятельной, если $\lim_{n \rightarrow \infty} D[\hat{Q}] = 0$.

3. Несмещенная оценка $\hat{Q}$ является *эффективной*, если ее дисперсия минимальна по отношению к дисперсии любой другой оценки.

2.2.1. Точечные оценки числовых характеристик

1. Несмещенная состоятельная оценка *математического ожидания*, называемая выборочным средним, вычисляется по формуле

$$\bar{x} = \sum_{i=1}^n x_i / n. \quad (2.8)$$

2. Несмещенная состоятельная оценка *дисперсии* равна

$$S_0^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2.9)$$

3. Смещенная состоятельная оценка *дисперсии*

$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2. \quad (2.10)$$

4. Несмещенная состоятельная оценка дисперсии

$$S_1^2 = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - m_x)^2. \quad (2.11)$$

5. Состоятельная оценка *среднеквадратического отклонения*

$$S_0 = \sqrt{S_1^2}. \quad (2.12)$$

6. Несмещенная состоятельная оценка *корреляционного момента*

$$\hat{K}_{XY} = \frac{1}{n-1} \cdot \sum_{k=1}^n (x_k - \bar{x}) \cdot (y_k - \bar{y}), \quad (2.13)$$

где x_k, y_k - значения, которые приняли случайные величины X, Y в k -м опыте;

$\bar{x}, \bar{y}$ - средние значения случайных величин X и Y соответственно.

7. Состоятельная оценка *коэффициента корреляции*

$$\hat{r}_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (2.14)$$

8. Выборочный *начальный момент* k -го порядка определяется по формуле

$$\hat{\alpha}_k = \frac{1}{n} \cdot \sum_{i=1}^n (x_i)^k. \quad (2.15)$$

9. Выборочный *центральный момент* k -го порядка равен

$$\hat{\mu}_k = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^k. \quad (2.16)$$

10. В случае *неравноточных* измерений несмещенная состоятельная оценка математического ожидания равна

$$\tilde{x} = \sum_{i=1}^n (x_i / D[\xi_i]) / \sum_{i=1}^n (1 / D[\xi_i]), \quad (2.17)$$

где $D[\xi_i]$ - дисперсия случайной величины в i -м опыте.

11. Несмещенная состоятельная и эффективная оценка вероятности в схеме независимых опытов Бернулли:

$$p^* = w = m/n, \quad (2.18)$$

где m - число успешных опытов.

2.2.2. Методы получения оценок параметров распределения

Для вычисления приближенных значений параметров чаще всего применяются методы моментов и максимального правдоподобия.

Суть *метода моментов* заключается в следующем. Пусть имеется выборка $\{x_1, \dots, x_n\}$ независимых значений случайной величины с известным законом распределения $f(x, Q_1, \dots, Q_m)$ и m неизвестными параметрами $Q_1, \dots, Q_m$. Последовательность вычислений следующая:

1. Вычислить значения m начальных и/или центральных теоретических моментов

$$\alpha_k = M[X^k], \quad \mu_k = M[(X - m_x)^k]. \quad (2.19)$$

2. Определить m соответствующих выборочных начальных $\hat{\alpha}_k$ и/или центральных $\hat{\mu}_k$ моментов по формулам (2.15, 2.16).

3. Составить и решить систему из m уравнений, в каждом из которых приравниваются теоретические и выборочные моменты. Каждое уравнение имеет вид $\alpha_k = \hat{\alpha}_k$ или $\mu_k = \hat{\mu}_k$.

Замечание. Часть уравнений может содержать начальные моменты, а оставшаяся часть - центральные. Согласно методу максимального правдоподобия оценки $\hat{Q}_1, \dots, \hat{Q}_m$ получаются из условия максимума по параметрам $Q_1, \dots, Q_m$ положительной функции правдоподобия $L(x_1, \dots, x_n, Q_1, \dots, Q_m)$. Если случайная величина X - непрерывна, а значения x_i независимы, то

$$L(x_1, \dots, x_n, Q_1, \dots, Q_m) = \prod_{i=1}^n f(x_i, Q_1, \dots, Q_m). \quad (2.20)$$

Если случайная величина X - дискретна и принимает независимые значения x_i с вероятностями

$$P(X = x_i) = p_i(x_i, Q_1, \dots, Q_m), \quad (2.21)$$

то функция правдоподобия равна

$$L(x_1, \dots, x_n, Q_1, \dots, Q_m) = \prod_{i=1}^n p_i(x_i, Q_1, \dots, Q_m). \quad (2.22)$$

Система уравнений согласно этому методу может записываться в двух видах:

$$\frac{\partial L(x_1, \dots, x_n, Q_1, \dots, Q_m)}{\partial Q_i} = 0, \quad i = 1, 2, \dots, m \quad (2.23)$$

или
$$\frac{\partial \ln(L(x_1, \dots, x_n, Q_1, \dots, Q_m))}{\partial Q_i} = 0, \quad i = 1, 2, \dots, m. \quad (2.24)$$

Пример 2.1. Пусть x_i - независимые значения случайной величины X , распределенной по экспоненциальному закону, т.е.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Необходимо получить оценку параметра λ методом максимального правдоподобия.

Решение. Функция правдоподобия имеет вид

$$L(x_1, \dots, x_n, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \cdot e^{-\lambda \sum_{i=1}^n x_i}.$$

Тогда
$$\ln(L(x_1, \dots, x_n, \lambda)) = n \cdot \ln(\lambda) - \lambda \cdot \sum_{i=1}^n x_i.$$

Далее записываем уравнение

$$\frac{\partial \ln(L)}{\partial \lambda} = n / \lambda - \sum_{i=1}^n x_i = 0.$$

Отсюда получаем выражение для оценки параметра λ :

$$\hat{\lambda} = n / \sum_{i=1}^n x_i = 1 / \bar{x}.$$

Пример 2.2. Случайная величина X распределена по равномерному закону, т.е.

$$f(x) = \begin{cases} 1/(b-a), & a \leq x < b \\ 0, & x < a \vee x \geq b. \end{cases}$$

Необходимо определить оценки параметров a и b .

Решение. По исходной выборке определяем выборочные моменты $\hat{\alpha}_1$ и по формулам (2.15) и (2.16)

соответственно. Составляем систему их 2-х уравнений: $\alpha_1 = \hat{\alpha}_1$, $\mu_2 = \hat{\mu}_2$.

Здесь

$$\alpha_1 = M[X] = \int_a^b \frac{x}{(b-a)} dx = (a+b)/2,$$

$$\mu_2 = M[(X - m_x)^2] = (b-a)^2 / 12,$$

$$\hat{\alpha}_1 = \sum_{i=1}^n x_i / n, \quad \hat{\mu}_2 = \sum_{i=1}^n (x_i - \hat{\alpha}_1)^2 / n.$$

Подставив данные выражения в систему и решив ее, получим $\hat{a} = \hat{\alpha}_1 - \sqrt{3\hat{\mu}_2}$, $\hat{b} = \hat{\alpha}_1 + \sqrt{3\hat{\mu}_2}$.

2.3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистической гипотезой называется всякое непротиворечивое множество утверждений $\{H_0, H_1, \dots, H_{k-1}\}$ относительно свойств распределения случайной величины. Любое из утверждений H_i называется альтернативой гипотезы. Простейшей гипотезой является двухальтернативная $\{H_0, H_1\}$. В этом случае альтернативу H_0 называют нулевой гипотезой, а H_1 - конкурирующей гипотезой.

Критерием называется случайная величина

$$K = \varphi(x_1, \dots, x_n),$$

которая позволяет принять или отклонить нулевую гипотезу H_0 .

При проверке гипотез можно допустить ошибки 2 родов.

Ошибка первого рода состоит в том, что будет отклонена гипотеза H_0 , если она верна ("пропуск цели").

Вероятность совершить ошибку первого рода обозначается α и называется *уровнем значимости*. Наиболее часто на практике принимают, что $\alpha = 0,05$ или $\alpha = 0,01$.

Ошибка второго рода заключается в том, что гипотеза H_0 принимается, если она неверна ("ложное срабатывание"). Вероятность ошибки этого рода обозначается β .

Критерием согласия называется критерий проверки гипотезы о предполагаемом законе распределения.

2.3.1 Критерий согласия χ^2

Это один из наиболее часто применяемых критериев. Алгоритм проверки гипотезы следующий.

1. Построить гистограмму равновероятностным способом.
2. По виду гистограммы выдвинуть гипотезу

$$H_0 : f(x) = f_0(x),$$

$$H_1 : f(x) \neq f_0(x),$$

где $f_0(x)$ - плотность вероятности гипотетического закона распределения: равномерного, экспоненциального или нормального.

Замечание. Гипотезу об экспоненциальном законе распределения можно выдвигать в том случае, если все числа в выборке положительные.

3. Вычислить значение критерия по формуле

$$\chi^2 = n \sum_{i=1}^M \frac{(p_i - p_i^*)^2}{p_i} = \sum_{i=1}^M \frac{(v_i - np_i)^2}{np_i}, \quad (2.25)$$

где $p_i^* = v_i / n$ - частота попадания в i -й интервал;

p_i - теоретическая вероятность попадания случайной величины в i -ый интервал при условии, что гипотеза H_0 верна.

Формулы для расчета p_i в случае экспоненциального, равномерного и нормального законов соответственно равны.

Экспоненциальный закон

$$p_i = e^{\frac{-A_i}{\bar{x}}} - e^{\frac{-B_i}{\bar{x}}} . \quad (2.26)$$

При этом $A_1 = 0, B_m = +\infty$.

Равномерный закон

$$p_i = (B_i - A_i) / (\hat{x}_n - \hat{x}_1) . \quad (2.27)$$

Нормальный закон

$$p_i = 0,5 \left(\Phi \left(\frac{B_i - \bar{x}}{S_0} \right) - \Phi \left(\frac{A_i - \bar{x}}{S_0} \right) \right) . \quad (2.28)$$

При этом $A_1 = -\infty, B_M = +\infty$.

Замечания. После вычисления всех вероятностей p_i проверить, выполняется ли контрольное соотношение

$$\left| 1 - \sum_{i=1}^M p_i \right| \leq 0,01 . \quad (2.29)$$

4. Из таблицы "Хи-квадрат" Приложения выбирается значение $\chi_{\alpha,k}^2$, где α - заданный уровень значимости ($\alpha = 0,05$ или $\alpha = 0,01$), а k - число степеней свободы, определяемое по формуле

$$k = M - 1 - s .$$

Здесь s - число параметров, от которых зависит выбранный гипотезой H_0 закон распределения. Значения s для равномерного закона равно 2, для экспоненциального - 1, для нормального - 2.

5. Если $\chi^2 > \chi_{\alpha,k}^2$, то гипотеза H_0 отклоняется. В противном случае нет оснований ее отклонить.

2.3.2. Критерий согласия Колмогорова

Последовательность действий при проверке гипотезы следующая.

1. Построить вариационный ряд.

2. Построить график эмпирической функции распределения $F^*(x)$.

3. Выдвинуть гипотезу:

$$H_0 : F(x) = F_0(x) ,$$

$$H_1 : F(x) \neq F_0(x) ,$$

где $F_0(x)$ - теоретическая функция распределения типового закона: равномерного, экспоненциального или нормального. Ниже приведены формулы для расчета $F_0(x)$.

Равномерный закон

$$F_0(x) = \begin{cases} 0, & x < \hat{x}_1 \\ (x - \hat{x}_1) / (\hat{x}_n - \hat{x}_1), & \hat{x}_1 \leq x \leq \hat{x}_n \\ 1, & x > \hat{x}_n \end{cases} \quad (2.30)$$

Экспоненциальный закон

$$F_0(x) = \begin{cases} 1 - e^{-x/\bar{x}}, & x \geq 0 \\ 0, & x < 0 \end{cases} . \quad (2.31)$$

Нормальный закон

$$F_0(x) = 0,5 + 0,5 \cdot \Phi \left(\frac{x - \bar{x}}{S_0} \right) \quad (2.32)$$

4. Рассчитать по формулам (2.30 - 2.32) 10-20 значений и построить график функции $F_0(x)$ в одной системе координат с функцией $F^*(x)$.

5. По графику определить максимальное по модулю отклонение между функциями $F^*(x)$ и $F_0(x)$.
 6. Вычислить значение критерия

$$\lambda = \sqrt{n} \cdot \max |F^*(x) - F_0(x)|. \quad (2.33)$$

7. Принимают тот или иной уровень значимости (чаще всего 0,05 или 0,01). Тогда доверительная вероятность $\gamma = 1 - \alpha$.

8. Из таблицы вероятностей Колмогорова выбрать критическое значение λ_γ .

9. Если $\lambda > \lambda_\gamma$, то нулевая гипотеза H_0 отклоняется, в противном случае - принимается, хотя она может быть неверна.

Достоинства критерия Колмогорова по сравнению с критерием χ^2 : возможность применения при очень маленьких объемах выборки ($n < 20$), более высокая "чувствительность", а, следовательно, меньшая трудоемкость вычислений.

Недостаток: критерий можно использовать в том случае, если параметры $Q_1, \dots, Q_k$ распределения заранее известны, а эмпирическая функция распределения $F^*(x)$ должна быть построена по несгруппированным выборочным данным.

2.4. ОЦЕНКА КОЭФФИЦИЕНТОВ ЛИНЕЙНОЙ РЕГРЕССИИ

Регрессией случайной величины Y на X называется условное математическое ожидание случайной величины Y при условии, что $X = x$:

$$m_Y(x) = M[Y / X = x]. \quad (2.34)$$

Регрессия Y на X устанавливает зависимость среднего значения величины Y от величины X . Если X и Y независимы, то

$$m_Y(x) = m_Y = \text{const.}$$

Простейшим видом регрессии является линейная:

$$m_Y(x) = a_0 + a_1 x.$$

Определение оценок коэффициентов a_0, a_1 осуществляется с помощью метода наименьших квадратов.

Пусть имеется выборка $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, содержащая n пар значений случайных величин X и Y .

Тогда оценки параметров $\hat{a}_0$ и $\hat{a}_1$ вычисляются по следующим формулам:

$$\hat{a}_1 = \frac{K_{XY}^*}{S_0^2(x)}, \quad (2.35)$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}. \quad (2.36)$$

где $\bar{x}, \bar{y}$ - оценки математического ожидания величин X и Y ;

$S_0^2(x)$ - оценка дисперсии величины X ;

$\hat{K}_{XY}$ - оценки корреляционного момента величин X и Y .

Для визуальной проверки правильности вычисления величин $\hat{a}_0, \hat{a}_1$ необходимо построить диаграмму рассеивания и график уравнения регрессии $\bar{y}(x) = \hat{a}_0 + \hat{a}_1 \cdot x$ (рис. 2.2).

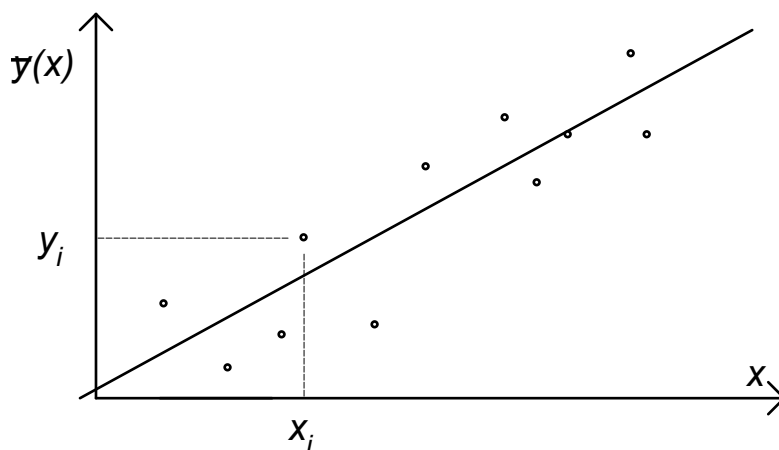


Рис 2.2

Если оценки параметров a_0, a_1 рассчитаны без грубых ошибок, то сумма квадратов отклонений всех точек (x_i, y_i) от прямой $\bar{y}(x) = \hat{a}_0 + \hat{a}_1 \cdot x$ должна быть минимально возможной.

2.5. КОНТРОЛЬНЫЕ ЗАДАЧИ

1. По выборке одномерной случайной величины:

- получить вариационный ряд,
- построить на масштабно-координатной бумаге формата А4 график эмпирической функции распределения $F^*(x)$,
- построить гистограмму равноинтервальным способом,
- построить гистограмму равновероятностным способом,
- вычислить оценки математического ожидания и дисперсии,
- выдвинуть гипотезу о законе распределения случайной величины и проверить ее при помощи критерия согласия χ^2 и критерия Колмогорова ($\alpha = 0.05$).

2. По выборке двумерной случайной величины:

- вычислить оценку коэффициента корреляции,
- вычислить параметры линии регрессии a_0 и a_1 ,
- построить диаграмму рассеивания и линию регрессии.