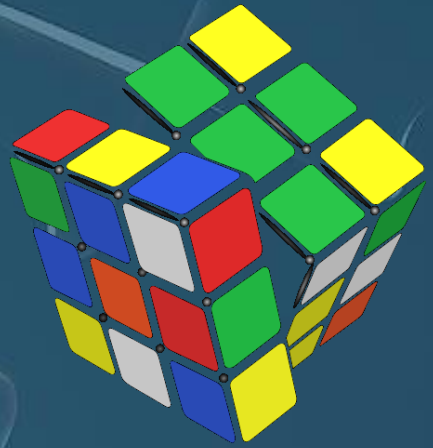
Методы математической статистики



{ выборка из генеральной совокупности - эмпирическая (выборочная) функция распределения – гистограмма – статистические оценки – точечные оценки параметров и их критерии – методы получения оценок параметров – метод моментов – метод наибольшего подобия }

Понятие генеральной совокупности и выборки

- Задачи, решаемые *математической статистикой*, являются, в некотором смысле, обратными задачам теории вероятностей. В вероятностных задачах распределения случайных величин считаются известными. В статистических задачах само распределение считается неизвестным, и целью исследования является получение более или менее достоверной информации об этом распределении, собранной в результате наблюдений.
- Основой статистического анализа являются данные, полученные экспериментатором в результате опыта, например, n повторных измерений некоторой неизвестной величины $X\{x_1,x_2,\ldots,x_n\}$, принимаемых случайной величиной ξ. Это множество называется выборкой из генеральной совокупности Γ_ξ всех значений случайной величины, а количество n – объемом выборки. Эти значения естественно считать реализацией набора из n независимых одинаково распределенных случайных величин с неизвестной функцией распределения $F_\xi(x)$.

  Данные должны быть выбраны из генеральной совокупности случайным образом, их объем достаточно велик. В этом случае выборка называется репрезентативной (представвительной).

Понятие генеральной совокупности и выборки

- Вектор этих данных называют выборкой из генеральной совокупности данных.
  n - мерная случайная величина $X(x_1, x_2, \ldots, x_n)$ с независимыми одинаково распределенными компонентами x_i, $i = 1, 2, .., n$ называется независимой выборкой объема n неизвестного распределения $F_\xi(x)$.

- Любая функция $h = h(x_1, x_2, \ldots\ldots, x_n)$ выборочных значений называется статистикой.

Часто встречается ситуация, когда экспериментатор имеет основания предполагать, что неизвестное распределение принадлежит некоторому семейству распределений $F_\xi(x,\theta)$, зависящему от параметра θ. В этом случае проблема статистического анализа сводится к получению информации об этом неизвестном параметре.

Пример дискретного вариационного ряда

@ Для контроля качества в 40 пробах стали GS50 определялось содержание углерода X (%C) и прочность на разрыв z (Н/мм). Данные оформлены в виде таблицы чисел:

X : 0.3, 0.33, 0.37, 0.36, 0.31, 0.29, 0.34, 0.39, 0.37, 0.38, 0.35, 0.32, 0.39, 0.3, 0.32, 0.32, 0.38, 0.37, 0.38, 0.33, 0.37, 0.33, 0.34, 0.33, 0.3, 0.34, 0.36, 0.33, 0.34, 0.36, 0.29, 0.3, 0.33, 0.32, 0.32, 0.38, 0.37, 0.34, 0.35, 0.36

$X = X(x_1, x_2, \ldots, x_{40})$ – выборка объемом $n = 40$

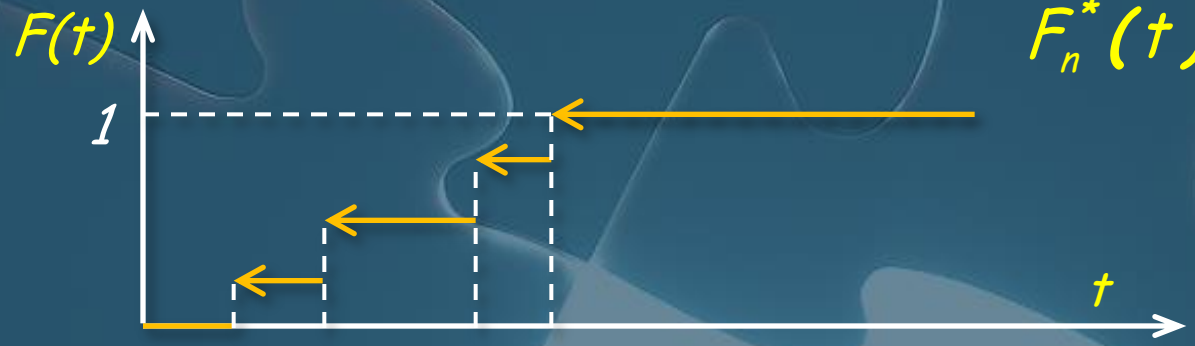
Z : 589, 614, 612, 572, 548, 537, 574, 570, 540, 575, 535, 593, 582, 538, 566, 562, 601, 587, 587, 614, 602, 544, 545, 562, 576, 596, 605, 575, 570, 550, 572, 555, 555, 518, 539, 557, 558, 587, 580, 560

$Z = Z(z_1, z_2, \ldots, z_{40})$ – выборка объемом $n = 40$

Выборочная функция распределения

Пусть $X(x_1, x_2, \ldots, x_n)$ – независимая выборка неизвестного распределения $F_\xi(x)$.

- *Эмпирической (выборочной) функцией распределения* называется функция $F^*_n(t): R \to [0, 1]$, вычисляемая по выборке $X(x_1, x_2, \ldots, x_n)$ как отношение числа элементов выборки, не превосходящих t, к объему выборки:



$$F_n^*(t) = \frac{|\{i : x_i \le t\}|}{n}$$

Теорема Гливенко: В пределе выборочная функция распределения равномерно сходится к теоретической.

$$P\{\sup_{t \in R} |F_n^*(t) - F(t)| \to 0\} = 1$$

Гистограмма

Помимо эмпирических функций распределения, наглядное представление о неизвестном распределении можно получить при помощи *гистограмм*. Пусть $X(x_1, x_2, \ldots, x_n)$ – независимая выборка неизвестного распределения $F_\xi(x)$. Выберем два числа L и R, такими, чтобы все числа x_i попали внутрь интервала $(L, R]$. Разобъем этот интервал его на конечное число меньших интервалов $\Delta_j = r_j - r_{j-1}$

Произведем *группировку* выборки, а именно, для каждого интервала разбиения Δ_j объединим в группу те x_i, которые попали в этот интервал. Пусть n_j - число таких элементов выборки: $n_j = |\{j : x_j \in (r_{j-1}, r_j)\}|, \; j = 1,2,\ldots,k$

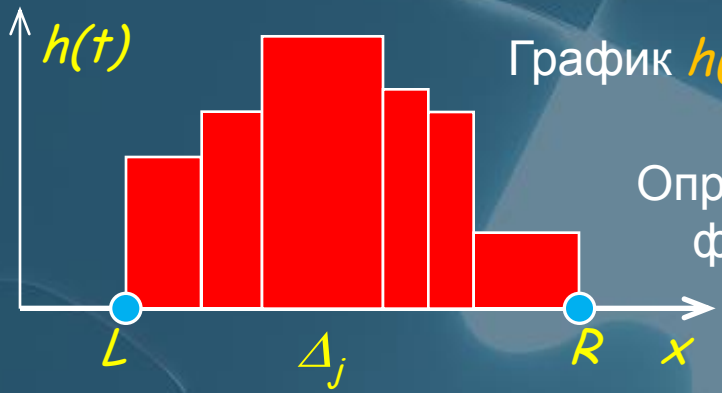


График $h(t)$ - гистограмма

Определим функцию

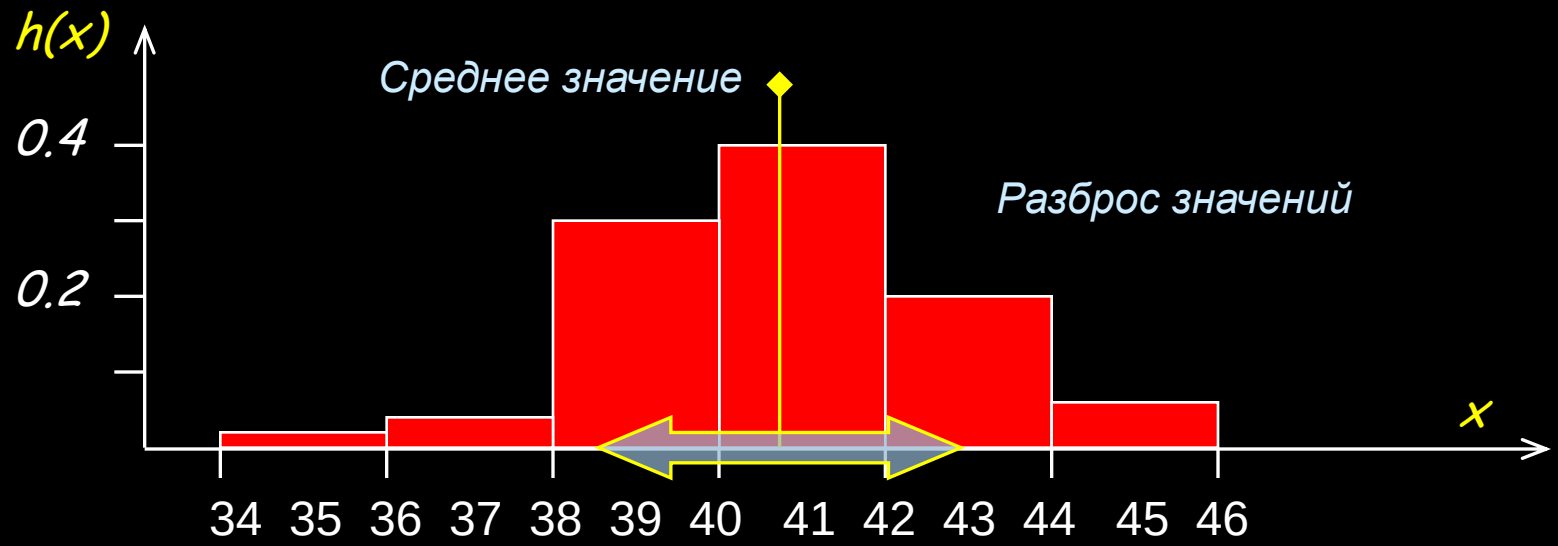
$$h(t) = \begin{cases} 0, & t \le L \\ \frac{n_j}{n}, & t \in (r_j - r_{j-1}], \; j = 1,2,\ldots k \\ 0, & t > R \end{cases}$$

Пример

Вариационный ряд: 34 36 36 37 …38 38 38 ….. 38 …39 40 40 40 41 41 42 42 …44… 45 46

Построить гистограмму

Границы интервалов	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44	44 – 46	
Частоты m_j	2	3	30	40	20	5	$n = 100$



Статистические оценки параметров

Случайная величина X характеризуется рядом числовых параметров: математическим ожиданием, дисперсией, модой, медианой, моментами разных порядков и т.д. Это параметры генеральной совокупности. На основе выборочных данных можно получить статистические оценки этих параметров

- Для оценки *математического ожидания* применяется *выборочное среднее*

$$\tilde{m}_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Для группированной выборки используется формула, в которой все m_j значений выборки, попавшей в j - ый интервал, равны представителю этого интервала (всего их k)

$$\tilde{m}_x = \frac{\sum_{j=1}^{k} z_j m_j}{n}$$

Статистические оценки параметров

- Для оценки *дисперсии* по выборке используется формула

$$\tilde{D}_x = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \tilde{m}_x^2\right)$$

- В случае группированной выборки

$$\tilde{D}_x = \frac{n}{n-1}\left(\frac{1}{n}\sum_{j=1}^{k} z_j^2 - \tilde{m}_x^2\right)$$

Оценка среднеквадратичного отклонения : $\tilde{\sigma}_x = \sqrt{\tilde{D}_x}$

Статистические оценки параметров

- *Модой* любой функция *h (x)* унимодального (одновершинного) распределения является элемент выборки, встречающийся с наибольшей частотой.
- Оценкой *медианы* называют число, которое делит вариационный ряд на две части с равным числом элементов
- Оценки начальных и центральных моментов ***k*** – го порядка вычисляются по формулам :

$$\tilde{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \quad \tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \tilde{m}_x)^k, \quad k = 1, 2, \ldots$$

- Форма распределения случайной величины характеризуется *выборочными коэффициентами асимметрии и эксцесса*

$$\tilde{A}_x = \frac{\tilde{\mu}_3}{\tilde{\sigma}_x^3}, \quad \tilde{E}_x = \frac{\tilde{\mu}_4}{\tilde{\sigma}_x^4} - 3$$

Пример

@ Найти выборочное среднее и дисперсию для группированной выборки:

Границы интервалов	34 – 36	36 – 38	38 – 40	40 – 42	42 – 44	44 – 46	
Частоты m_j	2	3	30	40	20	5	$n = 100$

$$\tilde{m}_x = \frac{1}{100}(35 \cdot 2 + 37 \cdot 3 + 39 \cdot 30 + 41 \cdot 40 + 43 \cdot 20 + 45 \cdot 5) = 40.76$$

$$\tilde{D}_x = \frac{100}{99}\left(\frac{1}{100}(35^2 \cdot 2 + 37^2 \cdot 3 + 39^2 \cdot 30 + 41^2 \cdot 40 + 43^2 \cdot 20 + 45^2 \cdot 5) - (40.76)^2\right) = 4$$

$$\tilde{\sigma}_x = 2$$

Точечные оценки параметров и их критерии

Пусть θ – неизвестный параметр распределения случайной величины.

- Статистика $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$, используемая в приближенном равенстве $\theta \cong \tilde{\theta}$ называется *точечной оценкой неизвестного параметра по выборке*

Какие оценки можно считать хорошими ?

- Оценка называется несмещенной для функци от неизвестного параметра, если

$$M\tilde{\theta}(x_1, x_2, \ldots, x_n) = \theta$$

- Оценка называется эффективной, если при заданном объеме выборки она имеет наименьшую возможную дисперсию

$$D\tilde{\theta}(x_1, x_2, \ldots, x_n) \Rightarrow \min$$

Точечные оценки параметров и их критерии

- Последовательность оценок $\theta \cong \tilde{\theta}^{(n)}$ (соответствующих увеличивающимся в объеме выборкам) называется *состоятельной*, если при росте объемов выборки статистика будет стремиться к истинному значению параметра

$$\forall \varepsilon > 0 \rightarrow P\{| \tilde{\theta}(x_1, x_2, \ldots, x_n) - \theta | < \varepsilon\} \underset{n \to \infty}{\rightarrow} 1$$

то есть $\tilde{\theta}(x_1, x_2, \ldots, x_n) \underset{n \to \infty}{\rightarrow} \theta$

@ *Соответствует ли выборочное среднее отмеченным выше критериям ?*

$$СВ \Rightarrow X(m,D)$$

$$\tilde{m} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1. *Оценка состоятельная, так как выполнены условия теоремы Чебышева*

$$P(|\tilde{m} - m| < \varepsilon) \geq 1 - \frac{D\tilde{m}}{\varepsilon^2} \underset{n\to\infty}{\overset{p}{\to}} 1$$

2. *Оценка несмещенная* $M\tilde{m} = M\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} m}{n} = m$

@

3. Оценка эффективная

$$\tilde{m} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad D\tilde{m} = \frac{\sum_{i=1}^{n} Dx_i}{n^2} = \frac{nD}{n^2} = \frac{D}{n} = \xrightarrow{n\to\infty} 0$$

$$D\tilde{m} = M(\tilde{m} - M\tilde{m})^2 = M(\tilde{m} - m)^2 = M\tilde{m}^2 - m^2 =$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{m^2 n}{n} = \frac{\sum_{i=1}^{n} (x_i^2 - m^2)}{n} = \frac{\frac{1}{n}\sum_{i=1}^{n} M(x_i - m)^2}{n} = \frac{D}{n} = \xrightarrow{n\to\infty} 0$$

Пример

@ *Соответствует ли выборочная дисперсия отмеченным выше критериям ?*

$$СВ \Rightarrow X(m,D) \qquad \tilde{D} = \frac{\sum_{i=1}^{n}(x_i - \tilde{m})^2}{n}$$

1. *Оценка состоятельная*, так как

$$\tilde{D} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \tilde{m}^2 \underset{n\to\infty}{\overset{p}{\to}} Mx^2 - m^2 = D$$

@ 3. Оценка эффективная

$$\tilde{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{m})^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - m) - (\tilde{m} - m)]^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 - \frac{2}{n}(\tilde{m} - m)\sum_{i=1}^{n}(x_i - m) + \frac{n}{n}(\tilde{m} - m)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 - (\tilde{m} - m)^2$$

2. Оценка смещенная ! $M\tilde{D} = \frac{1}{n}M\sum_{i=1}^{n}(x_i - m)^2 - M(\tilde{m} - m)^2 = \frac{nD}{n} - \frac{D}{n} = \frac{n-1}{n}D$

$$\tilde{D}_{\text{несмещенная}} = \left(\frac{n}{n-1}\right)\frac{\sum_{i=1}^{n}(x_i - \tilde{m})^2}{n}$$

Методы получения оценок. Метод моментов.

- Идея метода моментов заключается в приравнивании *теоретических* и *эмпирических моментов*.

Предполагается, что $F_\zeta(x) = F(x,\theta)$ и $M_\theta\zeta$ - конечная величина.

$$M_\theta(x) = \int_{-\infty}^{\infty} x dF(x,\theta) = v_1(\theta) \qquad \tilde{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \int_{-\infty}^{\infty} x d\tilde{F}_n(x) = v_1(\theta)$$

Решая это уравнение получим искомую оценку.

Если нужно оценить k параметров $q_1, q_2, \ldots, q_n$, то нужно найти выражения для моментов k – го порядка, приравнять их соответствующим эмпирическим моментам, и решить полученную систему уравнений.

Преимущества метода: сравнительная простота. Метод часто не дает эффективных оценок .

Метод наибольшего правдоподобия

- При получении оценки естественно найти такое её значение, при котором вероятность реализации выборки $x_1, x_2, \ldots, x_n$ была бы *максимальной*.

Пусть ξ имеет дискретное распределение. Возможные значения параметров: $a_1, a_2, \ldots, a_k$ с соответствующими вероятностями $P_1(\alpha), P_2(\alpha), \ldots P_k(\alpha)$, где α – фиксированное значение параметра. $P(x = a_i) = P_i(\alpha)$.

Пусть в выборке $x_1, x_2, \ldots, x_n$ значения a_j встретились n_j раз ($j = 1, 2, \ldots, k$).

Тогда вероятность при n независимых наблюдениях величины ξ получить выборку $x_1, x_2, \ldots, x_n$ равна

$$P(E) = P_1^{n_1}(\alpha) \cdot P_2^{n_2}(\alpha) \cdot P_3^{n_3}(\alpha) \cdots P_k^{n_k}(\alpha)$$

E – одна из реализаций. Число способов этих реализаций :

$$E = \frac{n!}{n_1!\, n_2!\, n_3! \cdots n_k!}$$

Метод наибольшего правоподобия

$$P = \frac{n!}{n_1!\, n_2!\, n_3! \cdots n_k!}\, p_1^{n_1} \cdot p_2^{n_2} \cdot p_3^{n_3} \cdots p_k^{n_k}$$

$$P = \frac{n!}{n_1!\, n_2!\, n_3! \cdots n_k!}\, L(x_1, x_2, x_3, \ldots, x_n)$$

Функцией правдоподобия называют функцию $L(X, \alpha)$

- $$L(x_1, x_2, x_3, \cdots x_n) = p_1^{n_1}(\alpha) \cdot p_2^{n_2}(\alpha) \cdot p_3^{n_3}(\alpha) \cdots p_k^{n_k}(\alpha)$$

Оценку параметра α будем искать так, чтобы $P = max$ или $L(X, \alpha) = max$.

$\frac{\partial L(X, \alpha)}{\partial \alpha} = 0$ Удобнее брать $\frac{\partial\, ln(L(X, \alpha))}{\partial \alpha} = \frac{1}{L(X, \alpha)} \frac{\partial L(X, \alpha)}{\partial \alpha} = 0$

Решая полученное уравнение или систему уравнений, если параметров больше одного, получим искомые оценки для α.

Преимущества метода: оценки получаются состоятельными, асимптотически эффективными. Оценки однако могут быть смещенными.

Пример

@ *Пусть X_i, $i = 1,2, \dots, n$ – выборка СВ с нормальным распределением. Найти оценки параметров m и D методом наибольшего правдоподобия.*

Функция правдоподобия

$$L(x_1,x_2,\dots x_n,m,D) = L(x_1,x_2,\dots x_n,m,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-m)^2}{2\sigma^2}}$$

$$\ln L = -\frac{n}{2}(\ln(2\pi) + \ln(\sigma^2)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

@ *Необходимое условие экстремума функции* **ln L** :

$$\begin{cases} \dfrac{\partial \ln(L)}{\partial m} = 0 \\ \dfrac{\partial \ln(L)}{\partial \sigma^2} = 0 \end{cases}$$

$$\begin{cases} \dfrac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(x_i - \tilde{m}) = 0 \\ -\dfrac{n}{2\tilde{\sigma}^2} + \dfrac{1}{2(\tilde{\sigma}^2)}\sum_{i=1}^{n}(x_i - \tilde{m})^2 = 0 \end{cases}$$

Решение:

$$\tilde{m} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \tilde{m})^2}{n}$$