Оценка параметров распределения по эмпирическим данным

(Ахметов С.К.)

Определения

Генеральная совокупность – это совокупность всех возможных значений СВ

Выборка – это конечный набор значений СВ, полученный в результате наблюдений

Репрезентативная выборка – это выборка, которая достаточно полно характеризует генеральную совокупность

Задача статистических методов – определить свойства СВ в целом на основании анализа выборки

Статистические оценки (m_x^*, σ_x^*, D_x^* и т.д.) – это числовые характеристики СВ, полученные по эмпирическим данным.

Требования к свойствам статистических оценок

1. ***Оценка*** $G^* = f(x_1, x_2, x_3, \ldots, x)$ **–** неизвестного параметра ***G*** называется **состоятельной**, если по мере роста числа наблюдений ***n*** она стремиться к оцениваемому значению ***G,*** т.е.

$$\lim_{n \to \infty} \mathbf{P}\left\{|G - G^*| < \varepsilon\right\} = 1$$

ε – сколь угодно малое число

2. Несмещенность. Оценка $G^* = f(x_1, x_2, x_3, \ldots, x)$ – неизвестного параметра ***G*** называется несмещенной, если при любом объеме выборки ***n*** результат ее осреднения по всем возможным выборкам данного объема приводит к точному (истинному) значению оцениваемого параметра, т.е., т. е. $\mathbf{M}[G^*] = G$

Несмещенность означает отсутствие систематической погрешности при оценивании параметра

3. ***Эффективность. Оценка*** $G^* = f(x_1, x_2, x_3, \ldots, x)$ – называется эффективной, если среди всех оценок параметра ***G*** она обладает наименьшей мерой случайного разброса относительно истинного значения оцениваемого параметра, т.е. $D[G^*] = D_{min}$

Эффективная оценка имеет минимальную случайную погрешность.

Эмпирические кривые обеспеченностей

Эмпирическая кривая обеспеченности - это функция обеспеченности, построенная по эмпирическим данным.

Возможны два способа построения эмпирической кривой обеспеченности.

Первый способ – при наличии большого числа наблюдений

Второй способ – при наличии небольшого числа наблюдений

Последовательность построения эмпирической кривой обеспеченности при большом числе наблюдений

1. ***Определяется амплитуда*** (размах ***R***) колебаний ***СВ***

2. ***Разбивается амплитуда колебаний на k равных интервалов.*** Величина ***k*** примерно рассчитывается по формуле $k \approx 5\ln(n)$

3. ***Определяется длина расчетного интервала по формуле*** $l = R/k$

В левой границы первого интервала принимается значение большее или равное максимальному значению СВ. Тогда значение правой границы первого интервала будет равно разнице между левой границей и длиной расчетного интервала. Левая граница каждого последующего интервала должна быть меньше на 1 соответствующей правой границы интервала

Последовательность построения эмпирической кривой обеспеченности при большом числе наблюдений (продолжение)

4. Подсчитывается число случаев попадания СВ в каждый интервал *(по этим данным можно построить график эмпирических частот)*

5. Рассчитывается относительная частота попадания СВ в каждый интервал по формуле

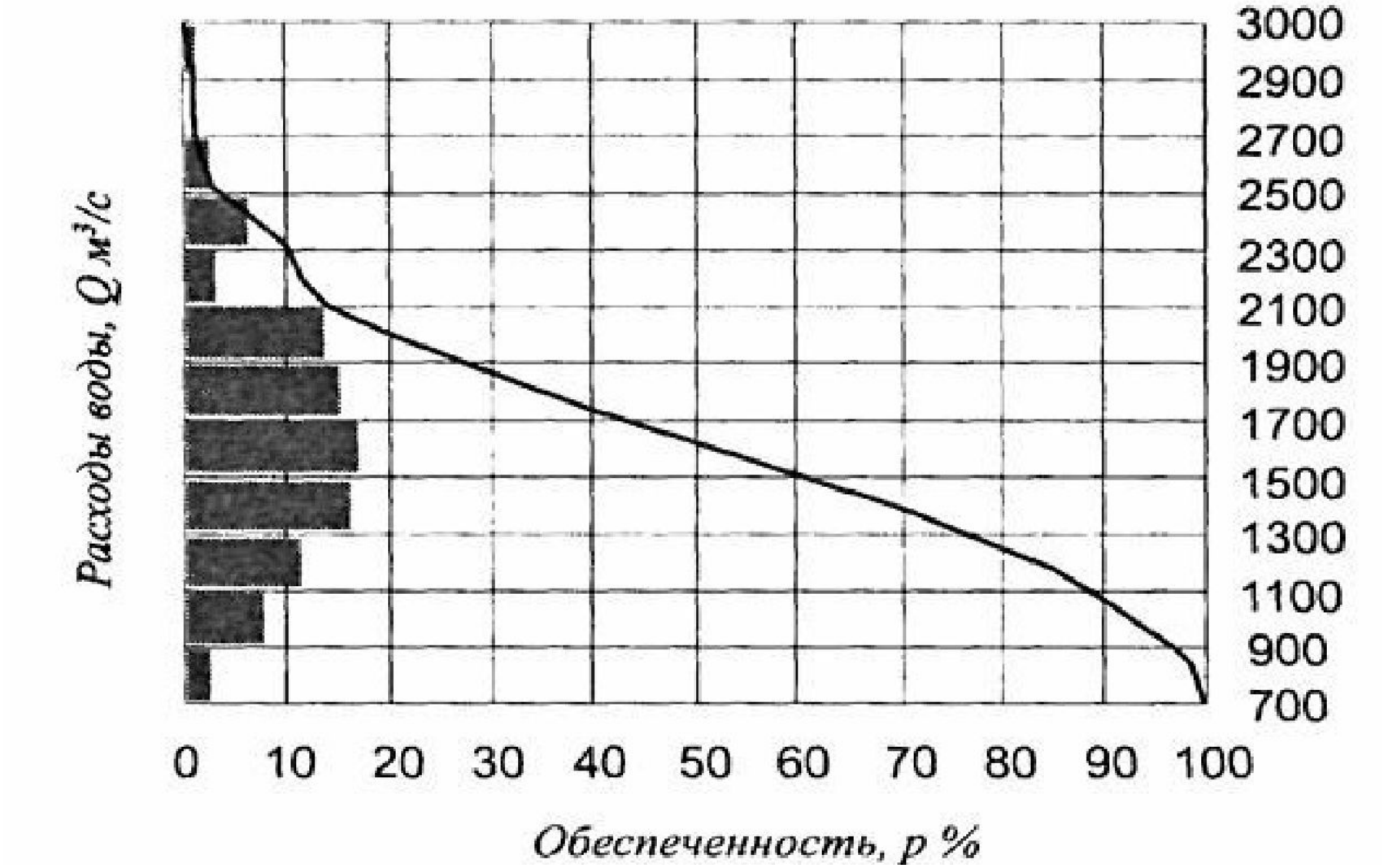
$$p^* = (m_i/n) \cdot 100\%$$

где m_i – число попаданий СВ в i – й интервал

Значения p^* последовательно суммируются и умножаются на 100% (по этим данным можно построить график эмпирической кривой обеспеченности)

В гидрологии значения ***СВ*** принято откладывать по вертикали, а значение вероятности ***p*** – по горизонтали.

Гистограмма эмпирических частот и эмпирическая кривая обеспеченностей



Последовательность построения эмпирической кривой обеспеченности при небольшом числе наблюдений

1. Ряд наблюдений располагается в убывающем или возрастающем порядке. В гидрологии – в убывающем порядке.

2. Приближенно вычисляется обеспеченность по формуле:

$$p_m = \mathbf{P}\{X \geq x_m\} \approx (m/n)100\ \%$$

где m - порядковый номер x_m в ранжированном ряду;

p_m – обеспеченность (в%) m – ного члена ранжированного ряда

При расчете обеспеченности последнего члена ряда по этой формуле получится, что $p_m = (n/n)100 = 100\%$

Чтобы этого избежать используют другие формулы

Формула Хансена $p_m = ((m-0.5)100)/n;$ Формула Крицкого – Менкеля (Вейбула) $p_m = 100m/(n+1);$ Формула Чегодаева $p_m = (100(m-0.3))/(n+0.4)$, Универсальная формула Грингортена

$$p_m = (100(m-a))/(n+1-2a)$$

при определенных значениях a – получаются все перечисленные формулы. Сам Грингортен предложил определять a по длине ряда (по таблицам).

Методы расчета оценок параметров распределения

Для построения аналитической кривой нужно оценить по эмпирическим данным параметры распределения. Обычно, это ***МО, СКО, C_v, C_s*** или ***C_v/C_s***

Эти методы расчета условно делят на аналитические, графоаналитические и графические.

К числу аналитических методов относятся метод моментов и метод наибольшего правдоподобия.

Метод моментов

При методе расчет теоретических моментов заменяется на расчет эмпирических моментов. При этом вместо $N \to \infty$ берется конечное число значений **СВ - *n***, а теоретическая вероятность p_i заменяется на расчетную $p_i = 1/n$.

Эмпирический начальный α_s^* и центральный μ_s^* моменты ***S*** го порядка определяются по формулам

$$\alpha_s^* = \frac{1}{n}\sum_{i=1}^{n} x_i^s \qquad \mu_s^* = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^s$$

Тогда математическое ожидание МО можно вычислить по формуле

$$m_x^* = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Эта оценка состоятельная и несмещенная.

Дисперсия – это второй центральный момент, поэтому ее можно вычислить так

$$D^* = S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Эта оценка состоятельная, но смещенная. Поэтому для расчетов используется формула

где S_H^2 и S_C^2 – соответственно несмещенная и смещенная оценки дисперсии;

n/(n-1) – поправка на смещенность

Метод моментов

Поэтому для расчетов используется формула

$$S_H^2 = \frac{n}{n-1} S_C^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

где S_H^2 и S_C^2 – соответственно несмещенная и смещенная оценки дисперсии;

n/(n-1) **–** поправка на смещенность

Тогда σ^* и C_v^* определяются по формулам

$$\sigma^* = S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$C_v^* = \frac{S}{\bar{x}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (k_i - 1)^2}$$

где $k_i = x_i / x_{ср.}$ – модульный коэффициент

Несмещенная оценка C_s^* определяется как

$$C_{s,H}^* = \frac{n \sum_{i=1}^{n} (k_i - 1)^3}{(n-1)(n-2)(C_v^*)^3}$$

Преимущество: метод не зависит от закона распределения СВ

Недостаток: при больших значениях C_v^* (больше 0.5), достоверность оценок ощутимо снижается.

Метод наибольшего правдоподобия (МНП)

Для нахождения оценок методом наибольшего правдоподобия нужно, прежде всего, построить функцию правдоподобия

Для этого делаются следующие последовательные шаги:

1. Заданная аналитическая функция распределения вероятности логарифмируется. Берется натуральный логарифм.

2. Создается функция правдоподобия путем интегрирования прологарифмированной функции плотности вероятности

3. Затем для каждого параметра распределения создается своя функция правдоподобия путем дифференцирования полученной функции распределения по требуемому параметру и приравнивается к нулю, чтобы найти ее максимум.

4. Из полученных уравнений находятся оценки, например МО и СКО.

МНП для нормальной функции распределения

Функция плотности вероятности для нормального распределения

$$f(x) = [1/(\sigma_x\sqrt{2\pi})]\exp[-(x-m_x)^2/(2\sigma_x^2)]$$

здесь m_x и σ_x - искомые параметры распределения.

1. Найдем логарифм этой функции

$$\ln f(x) = -\ln\sigma - 0{,}5\ln(2\pi) - \left(1/2\sigma^2\right)(x_i - m_x)^2$$

2 Проинтегрируем ее

$$L = \sum_{i=1}^{n}\ln f(x) = -\sum_{i=1}^{n}\ln\sigma - 0{,}5\sum_{i=1}^{n}\ln(2\pi) - (1/2\sigma^2)\sum_{i=1}^{n}(x_i - m_x)^2$$

3. Далее находим уравнение правдоподобия для оценки параметра m_x

$$\frac{\partial L}{\partial m_x} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m_x) = 0$$

получаем, что $$m_x^* = \frac{1}{n}\sum_{i=1}^{n}x_i$$

МНП для нормальной функции распределения

4. Аналогично находим уравнение правдоподобия для оценки σ_x

$$\frac{\partial L}{\partial \sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - m_x)^2 - \frac{n}{\sigma} = 0$$

Отсюда получим, что

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m_x)^2}{n}}$$

То есть для нормального распределения оценки параметров, полученные МНП, совпадают с моментными оценками.

Для других функций распределения плотности вероятности система уравнений правдоподобия получается сложной. Поэтому используются численные методы решений, а на их основе строятся номограммы для практического применения МНП.

МНП для кривой Крицкого - Менкеля

Для кривой Крицкого и Менкеля параметры C_s и C_v определяются с помощью специально разработанных номограмм как функций вспомогательных статистик λ_2 и λ_3.

$$\lambda_2 = \frac{\sum_{i=1}^{n} \lg k_i}{n-1}, \qquad \lambda_3 = \frac{\sum_{i=1}^{n} k_i \lg k_i}{n-1}$$

Считается, что оценки трех - параметрического гамма – распределения, полученные таким путем, являются состоятельными, эффективными и несмещенными.

При использовании МНП нужно помнить, что наибольший вес придается средним членам выборки, в отличие от метода моментов, где наибольший вклад вносят крайние члены выборки. Однако, это свойство МНП проявляется в случае выборок с большим рядом.

Недостаток: нужно знать аналитическое выражение функции распределения заранее.

СПАСИБО ЗА ВНИМАНИЕ!