ТЕМА 10. БАЗОВЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

10.1. Эмпирическая функция распределения

Под *генеральной совокупностью* понимают все возможные значения параметра, которые могут быть зарегистрированы в ходе неограниченного по времени наблюдения за объектом. Такая совокупность состоит из бесконечного множества элементов. В результате наблюдения за объектом формируется ограниченная по объему совокупность значений параметра x_1, x_2, …, x_n. Такие данные представляют собой *выборку* из генеральной совокупности. Наблюдаемые значения x_i называют *вариантами*, а их количество – объемом выборки n. Для того чтобы по результатам наблюдения можно было делать какие-либо выводы, выборка должна быть *репрезентативной* (представительной), т. е. правильно представлять пропорции генеральной совокупности. Это требование выполняется, если объем выборки достаточно велик, а каждый элемент генеральной совокупности имеет одинаковую вероятность попасть в выборку.

Пусть в полученной выборке значение x_1 параметра наблюдалось n_1 раз, значение x_2 – n_2 раз, значение x_k – n_k раз, $n_1 + n_2 + \ldots + n_k = n$. Совокупность значений, записанных в порядке их возрастания, называют *вариационным рядом*, величины n_i – частотами, а их отношения к объему выборки $n_i = n_i / n$ – *относительными частотами* (частостями). Очевидно, что сумма относительных частот равна единице. Другой формой вариационного ряда является ряд накопленных частот, называемый *кумулятивным рядом*.

Под распределением понимают соответствие между наблюдаемыми вариантами и их частотами или частостями. Пусть n_x – количество наблюдений, при которых случайные значения параметра X меньше x. Частость события $X<x$ равна n_x / n. Это отношение является функцией от x и от объема выборки: $F_n(x) = n_x/n$. Величина $F_n(x)$ обладает всеми свойствами функции распределения:

- $F_n(x)$ – неубывающая функция, ее значения принадлежат отрезку [0 – 1];
- если x_1 – наименьшее значение параметра, а x_k – наибольшее, то $F_n(x)=0$, когда $x <= x_1$, и $F_n(x)= 1$, когда $x > x_k$.

Функция $F^*_n(x)$ определяется по ЭД, поэтому ее называют эмпирической функцией распределения. В отличие от эмпирической функции $F^*_n(x)$ функцию распределения $F(x)$ генеральной совокупности называют теоретической функцией распределения, она характеризует не частость, а вероятность события X<x. При большом объеме наблюдений теоретическую функцию распределения $F(x)$ можно заменить эмпирической функцией $F^*_n(x)$.

Основные свойства функции $F^*_n\ (x)$.

1. 0 $F^*_n(x)$ 1.
2. $F^*_n\ (x)$ - неубывающая ступенчатая функция.
3. $F_n(x) = 0$, x x_1 .
4. $F_n(x) = 1$, $x > x_n$.

Пример 1.1 Задана выборка случайной величины

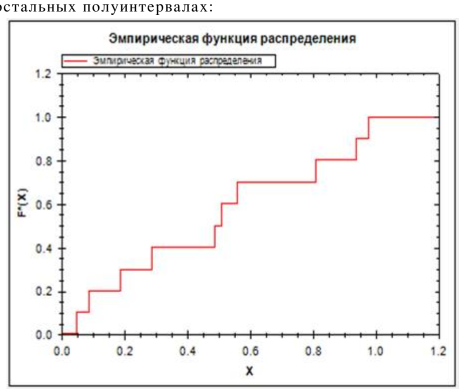
$$X = \{0.56;\ 0.79;\ 0.29;\ 0.56;\ 0.14;\ 1.00;\ 0.98;\ 1.00;\ 0.19;\ 0.08\}.$$

Построить график эмпирической функции распределения $F_n(x)$.

Решение. Вариационный ряд случайной величины имеет вид

$$X_i = \{0.08;\ 0.14;\ 0.19;\ 0.29;\ 0.56;\ 0.56;\ 0.79;\ 0.98;\ 1.00;\ 1.00\}.$$

Выделяем полуинтервалы (-; 0,08], (0,08; 0,14], (0,14; 0,19], …,(1,0; +]. На полуинтервале (-; 0,08] $F_n(x)$= 0/10= 0. При 0,08 < x 0,19 $F_n(x)$= 1/10= 0,1. Аналогично определяем значения $F_n(x)$ на остальных полуинтервалах:



$$F(x) = \begin{cases} 0,\ \text{при}\ -\infty < x \le 0{,}08 \\ 0{,}1,\ \text{при}\ 0.08 < x \le 0{,}14 \\ 0{,}2,\ \text{при}\ 0{,}14 < x \le 0{,}19 \\ 0{,}3,\ \text{при}\ 0{,}19 < x \le 0{,}29 \\ 0{,}4,\ \text{при}\ 0{,}29 < x \le 0{,}56 \\ 0{,}6,\ \text{при}\ 0{,}56 < x \le 0{,}79 \\ 0{,}7,\ \text{при}\ 0{,}79 < x \le 0{,}98 \\ 0{,}8,\ \text{при}\ 0{,}98 < x \le 1{,}0 \\ 1{,}0,\ \text{при}\ 0{,}1 < x < +\infty \end{cases}.$$

Рис. 2.1. График функции $F_n(x)$

Замечание. В каждой точке оси x, соответствующим значениям xi функция $F_n(x)$ имеет скачок. В точке разрыва $F_n(x)$ непрерывна слева и принимает значение, выделенное знаком.

10.2. Гистограмма

При большом объеме выборки (понятие «большой объем» зависит от целей и методов обработки, в данном случае будем считать n большим, если n>40) в целях удобства обработки и хранения сведений прибегают к группированию ЭД в интервалы. Количество интервалов следует выбрать так, чтобы в необходимой мере отразилось разнообразие значений параметра в совокупности и в то же время закономерность распределения не искажалась случайными колебаниями частот по отдельным разрядам. Существуют нестрогие рекомендации по выбору количества M и размера Δ таких интервалов, в частности параметр M рекомендуется выбирать с помощью следующих соотношений:

$$M \approx \mathrm{int}\left(\sqrt{n}\right), \quad n \le 100,$$
$$M \approx \mathrm{int}\left((2\cdots 4)\cdot \lg(n)\right), \quad n > 100, \qquad (1.1)$$

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

Графически статистический ряд отображают в виде гистограммы, полигона и ступенчатой линии. Часто гистограмму представляют как фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиною Δ, а высоты равны $m_i/(n\Delta)$. Такую гистограмму можно интерпретировать как графическое представление эмпирической функции плотности распределения $f_n(x)$, в ней суммарная площадь всех прямоугольников составит единицу. Гистограмма помогает подобрать вид теоретической функции распределения для аппроксимации ЭД.

Полигоном называют ломаную линию, отрезки которой соединяют точки с координатами по оси абсцисс, равными серединам интервалов, а по оси ординат – соответствующим частостям.

Порядок построения гистограммы следующий.

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

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

Ai, Bi - соответственно левая и правая границы i-го интервала ($Ai+1 = Bi$);

$\Delta i = Bi - Ai$ - длина i-го интервала;

mi- количество чисел в выборке, попадающих в i-тый интервал.

При использовании *равноинтервального* метода построения гистограммы параметры Ai, Bi, Δi вычисляются следующим образом:

$$\Delta_i = \Delta = (\hat{x}_n - \hat{x}_1)/M; \quad A_i = \hat{x}_1 + (i-1)\Delta; \quad B_i = A_{i+1}; \qquad i = 1,2,\ldots,M.$$

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 Ai, Bi выбираются таким образом, чтобы в каждый интервал попадало одинаковое количество выборочных значений:

mi = m = n / M.

В этом случае

$$A_1 = \hat{x}_1; \quad B_1 = (\hat{x}_m + \hat{x}_{m+1})/2; \quad A_2 = B_1; \quad A_i = (\hat{x}_{(i-1)m} + \hat{x}_{(i-1)m+1})/2; \quad i = 2,3,\ldots,M$$

3. Вычисляется средняя плотность вероятности для каждого интервала по формуле

$$f_i^* = m_i/(n\cdot\Delta_i).$$

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 Δi и высотой $f_i^* = m_i/(n\cdot\Delta_i)$. Полученная при этом ступенчатая линия называется гистограммой, график которой приблизительно выглядит так, как показано на рис. 1.2.

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.

Достоинства использования гистограммы: простота применения, наглядность.

Рассмотренные представления ЭД являются исходными для последующей обработки и вычисления различных параметров.

Пример 1.2 Дан вариационный ряд выборки случайной величины X (n=100). Построить гистограммы равноинтервальным методом.

$Y_i = \{$0.01; 0.02; 0.03; 0.05; 0.06; 0.08; 0.08; 0.09; 0.11; 0.12; 0.18; 0.19;
0.21; 0.23; 0.24; 0.24; 0.25; 0.26; 0.29; 0.30; 0.30; 0.30; 0.31; 0.32; 0.34;
0.35; 0.35; 0.36; 0.37; 0.41; 0.42; 0.44; 0.50; 0.53; 0.54; 0.54; 0.57; 0.58;
0.58; 0.59; 0.59; 0.59; 0.61; 0.61; 0.62; 0.62; 0.64; 0.65; 0.65; 0.66; 0.66;
0.66; 0.70; 0.72; 0.72; 0.73; 0.73; 0.73; 0.73; 0.74; 0.76; 0.77; 0.78; 0.78;
0.80; 0.81; 0.83; 0.84; 0.85; 0.86; 0.86; 0.87; 0.89; 0.90; 0.90; 0.90; 0.92;
0.93; 0.94; 0.94; 0.94; 0.95; 0.96; 0.97; 0.97; 0.97; 0.98; 0.99; 0.99; 0.99;
0.99; 0.99; 0.99; 0.99; 1.00; 1.00; 1.00; 1.00; 1.00; 1.00}

Разобьем область возможных значений $[X_1, X_n] = [0,1]$ на M непересекающихся интервалов, где M выбирается в соответствии с (1.1)

$M = \sqrt{100} = 10$

При использовании *равноинтервального* метода построения гистограммы параметры Ai, Bi, hi вычисляются следующим образом:

$$h = h_i = \frac{y_n - y_1}{M}; \quad A_i = y_1 + (i-1)h; \quad B_i = A_{i+1}; \quad i = 1,2,\ldots,M.$$

Откуда:

$h_i = h = \dfrac{b-a}{M} = 0.1$

$A_i = [0;\ 0.1;\ 0.2;\ 0.3;\ 0.4;\ 0.6;\ 0.6;\ 0.7;\ 0.8;\ 0.9]$

$B_i = [0.1;\ 0.2;\ 0.3;\ 0.4;\ 0.6;\ 0.6;\ 0.7;\ 0.8;\ 0.9; 1.0]$

Для каждого интервала посчитаем числа v_i - количество чисел в выборке, попадающих в i-тый интервал, и вычислим среднюю плотность вероятности по формуле:

$$f_i^* = \nu_i / (n \cdot h_i).$$

$v_i = [8; 4; 8; 9; 4; 9; 11; 11; 12; 24]$

$f_i^* = [0.8; 0.4; 0.8; 0.9; 0.4; 0.9; 1.1; 1.1; 1.2; 2.4]$

Пример 1.3. Дан вариационный ряд выборки случайной величины X (n=100). Построить гистограммы равноинтервальным методом.

$Y_i = [0.04;\ 0.06;\ 0.06;\ 0.08;\ 0.08;\ 0.09;\ 0.09;\ 0.12;\ 0.12;\ 0.13;\ 0.13;\ 0.14;$
$0.17;\ 0.19;\ 0.20;\ 0.20;\ 0.22;\ 0.22;\ 0.23;\ 0.24;\ 0.26;\ 0.27;\ 0.29;\ 0.29;\ 0.32;$
$0.32;\ 0.32;\ 0.33;\ 0.38;\ 0.38;\ 0.40;\ 0.41;\ 0.42;\ 0.43;\ 0.44;\ 0.45;\ 0.47;\ 0.51;$
$0.52;\ 0.53;\ 0.53;\ 0.53;\ 0.54;\ 0.54;\ 0.55;\ 0.55;\ 0.59;\ 0.60;\ 0.60;\ 0.60;\ 0.61;$
$0.65;\ 0.65;\ 0.70;\ 0.73;\ 0.74;\ 0.75;\ 0.75;\ 0.75;\ 0.76;\ 0.80;\ 0.81;\ 0.81;\ 0.82;$
$0.82;\ 0.86;\ 0.86;\ 0.86;\ 0.86;\ 0.86;\ 0.87;\ 0.88;\ 0.88;\ 0.88;\ 0.89;\ 0.91;\ 0.92;$
$0.92;\ 0.93;\ 0.94;\ 0.94;\ 0.94;\ 0.94;\ 0.96;\ 0.96;\ 0.97;\ 0.98;\ 0.98;\ 0.99;\ 0.99;$
$0.99;\ 0.99;\ 1.00;\ 1.00;\ 1.00;\ 1.00;\ 1.00;\ 1.00;\ 1.00;\ 1.00]$

Разобьем область возможных значений $[Y_1, Y_n] = [0,1]$ на

$M = \sqrt{100} = 10$

В случае применения *равновероятностного* метода количество попаданий в каждый интервал равны:

$v_i = v = \dfrac{n}{M} = \dfrac{100}{10} = 10$

Границы отрезков Ai, Bi вычисляются следующим образом:

$$A_1 = y_1; \quad B_1 = \tfrac{y_v + y_{v+1}}{2}; \quad A_2 = B_1; \quad A_i = \tfrac{y_{(i-1)v} + y_{(i-1)v+1}}{2}; \quad i = 2,3,\ldots,M.$$

$A_i = [0;\ 0.13;\ 0.26;\ 0.4;\ 0.53;\ 0.6;\ 0.8;\ 0.87;\ 0.94;\ 0.99]$

$B_i = [0.13;\ 0.26;\ 0.4;\ 0.53;\ 0.6;\ 0.8;\ 0.87;\ 0.94;\ 0.99; 1.0]$

Откуда:

$h_i = B_i - A_i$

$h_i = [0.13;\ 0.13;\ 0.14;\ 0.13;\ 0.07;\ 0.2;\ 0.07;\ 0.07;\ 0.05; 0.007]$

Для каждого интервала посчитаем числа v_i - количество чисел в выборке, попадающих в i-тый интервал, и вычислим среднюю плотность вероятности по формуле:

$$f_i^* = \nu_i / (n \cdot h_i).$$

$f_i^* = [0.76; 0.76; 0.8; 0.9; 0.4; 0.9; 1.1; 1.1; 1.2; 2.4]$

Рис. 1.2. Гистограмма распределения (равноинтервальный метод)

Рис.1.3. Гистограмма распределения (равновероятностный метод)