Глава 6

ПЕРВИЧНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

6.1. Основные понятия

Математическая статистика занимается статистическим анализом результатов опытов или наблюдений, а также построением и проверкой подходящих моделей процессов и систем на основе результатов экспериментов.

Статистический анализ и построение *вероятностных моделей* процессов и систем основаны на том, что измеряемые в процессе опыта или наблюдений физические (или иного смысла) величины X, характеризующие исследуемый процесс или систему, при повторении опытов подвержены некоторому неконтролируемому разбросу $x_1, x_2, \ldots, x_n$.

Этот разброс обусловлен действием случайных неучтенных факторов и ошибками измерений.

Поэтому величина X рассматривается как одномерная случайная величина, а результаты измерения x_1, x_2,…, x_n этой величины, называемые в математической статистике ее основными признаками – как эмпирическая реализация этого математического понятия.

Совокупность всех мыслимых значений, которые может принимать величина X при данном реальном комплексе условий, называют *генеральной совокупностью*.

Распределение признака *X* в генеральной совокупности совпадает с теоретическим распределением вероятностной величины *X*. Последнее называется распределением *генеральной совокупности*, а его параметры – *параметрами генеральной совокупности*.

Генеральная совокупность может быть конечной (всего *N* мыслимых наблюдений) и бесконечной в зависимости от того, конечна или бесконечна совокупность всех мыслимых значений.

Выборка из данной генеральной совокупности – это результаты ограниченного ряда наблюдений

x_1, x_2,…, x_n значений случайной величины X.

На практике при исследовании мы чаще всего имеем дело с выборками, поскольку обследование всей генеральной совокупности бывает слишком трудоемко (когда n – достаточно большое число), либо принципиально невозможно (в случае бесконечной генеральной совокупности).

Число *n* наблюдений, образующих выборку, называют *объемом выборки*.

Таким образом, вместо большой совокупности объектов изучается совокупность объёма, значительно меньшего по количеству объектов ($n << N$).

Результаты, полученные при изучении выборки, распространяются на объекты всей генеральной совокупности. Для этого выборка должна быть репрезентативной (представительной), то есть правильно представлять генеральную совокупность.

Это обеспечивается случайностью отбора.

Виды отбора:

1) простой случайный:
   – повторный;
   – бесповторный;
2) сложный случайный:
   – типический;
   – механический;
   – серийный.

Простой случайный отбор – производится без деления генеральной совокупности на части.

Повторный отбор – отобранный объект возвращается в генеральную совокупность.

Бесповторный отбор – отобранный объект не возвращается в генеральную совокупность.

Сложный случайный отбор – производится после предварительного деления генеральной совокупности на части.

Типический отбор – генеральная совокупность делится на типы, из каждого типа случайно отбираются объекты пропорционально объёму типов.

Механический отбор – генеральная совокупность делится на части механически, из каждой части случайно отбираются объекты.

Серийный отбор – генеральная совокупность делится на серии, и случайным образом отбираются целые серии объектов.

Разность между наибольшим и наименьшим значениями x_i (i=1,…, n) из выборки называется *размахом выборки*.

Взаимно независимые случайные величины имеют одинаковые распределения, а, следовательно, и одинаковые числовые характеристики (математическое ожидание, дисперсию и т.д.)

Основные задачи математической статистики:

1. Определение закона распределения основного признака (наблюдаемой СВ);
2. Нахождение оценок неизвестных параметров распределений и оценок числовых характеристик СВ;
3. Проверка правдоподобия статистических гипотез;
4. Оптимальная организация и проведение экспериментов, и оптимальная обработка результатов эксперимента.

6.2.Статистическое распределение выборки

Наблюдаемые значения x_i (i=1,…,n) называют *вариантами*, а последовательность значений (вариант), записанных в возрастающем порядке – *вариационным рядом*.

Числа наблюдений n_i называют *частотами*, а их отношения к объему выборки $n_i / n = p_i^*$ - *относительными частотами*.

Статистическим распределением выборки называют перечень вариант x_i и соответствующих им частот n_i или относительных частот p_i^*.

При больших объемах выборки n статистическое распределение выборки становится недостаточно наглядным. В этом случае статистические данные представляются в виде *интервального вариационного ряда*, который носит название *статистического ряда*.

Построение статистического ряда:

1. размах выборки разбивается на q конечных (или бесконечных) интервалов $X_j-0{,}5\Delta X_j< x_i< X_j+0{,}5\Delta X_j$, длины которых (размахи) соответственно $h_j=\Delta X_j$, а середины интервалов X_j , где $j=1,\ldots,q$.
2. Количество интервалов выбирается в основном из практических соображений. В частности, рекомендуется, чтобы значение q было не менее 5-10 и более 20-25 и в каждом интервале должно быть не менее 10 значений.

3. В том случае, если полученные из опыта данные группируются вокруг некоторых значений, то желательно, чтобы эти значения не находились вблизи узлов разбиения интервалов. Затем, подсчитываются число значений выборки n_j, попавших в интервал.

Если данные попадают на границы интервалов, то их либо распределяют равномерно по двум соседним интервалам, либо относят только к одному из них (например, к левому).

Выбор количества интервалов существенно зависит от объема выборки. Существуют такие рекомендации по использованию формулы Старджеса

$$q = log_2 n + 1 \cong 3{,}32 ln\ n + 1$$

или других формул, например:

$$q \cong 5 \lg n,\ q \cong \sqrt{n}.$$

Все эти формулы следует рассматривать как нижнюю оценку.

Так как длина интервала h_j может быть большой, а количество численных значений n_j, попавших в него, сравнительно малым, то для сопоставления групп друг с другом вычисляется также величина

$$\overline{p}^{*}_{j} = p^{*}_{j} / \Delta X_j$$

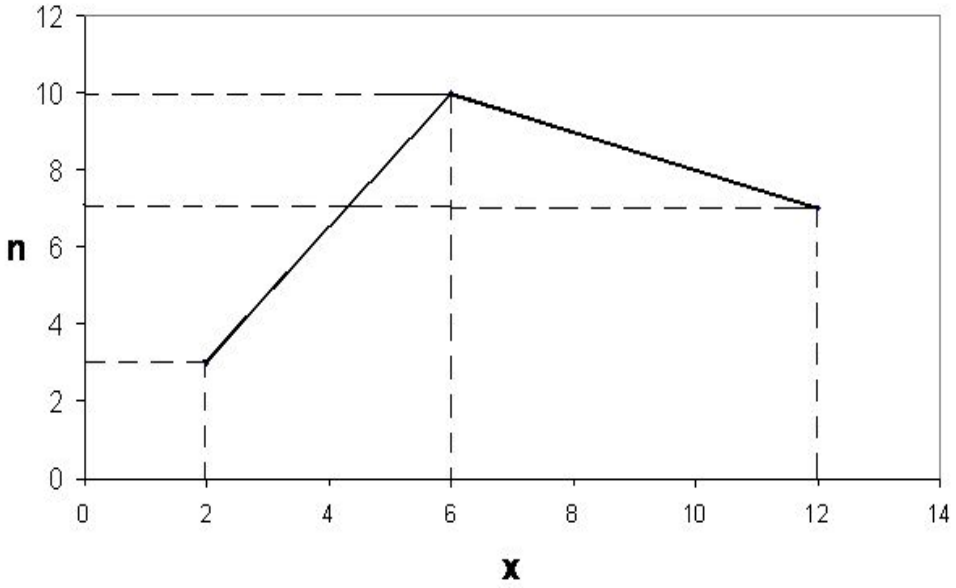
называемая плотностью относительной частоты.

Полученные результаты сводятся в таблицу вида:

Номер интервала	1	2	…	j	…	r
Длина интервала ΔX_j	ΔX_1	ΔX_2	…	ΔX_j	…	ΔX_r
Частота n_j	n_1	n_2	…	n_j	…	n_r
Относит. частота p_j^*	p_1^*	p_2^*	…	p_j^*	…	p_r^*
Плотность относит. частоты $\square p_j^*$	$\square p_1^*$	$\square p_2^*$	…	$\square p_j^*$	…	$\square p_r^*$

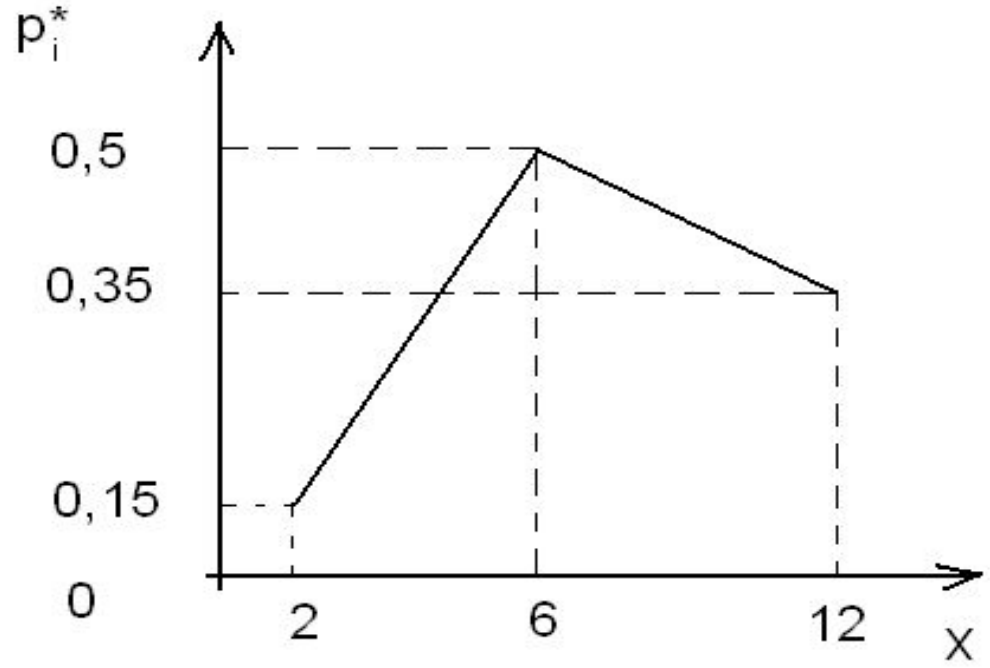
6.3.Полигон частот и гистограмма

Полигоном частот называют ломанную линию, отрезки которой соединяют точки (x_1,n_1), (x_2,n_2), …, (x_n,n_n).

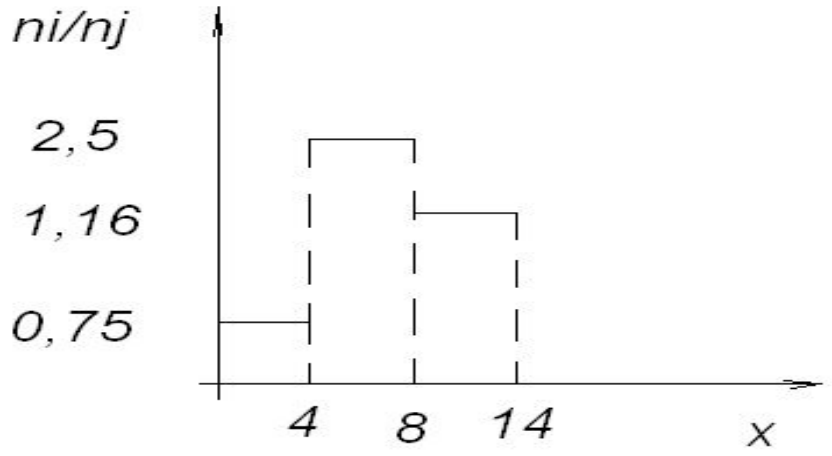


Для построения полигона частот на оси абсцисс откладывают варианты x_i, а по оси ординат – соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

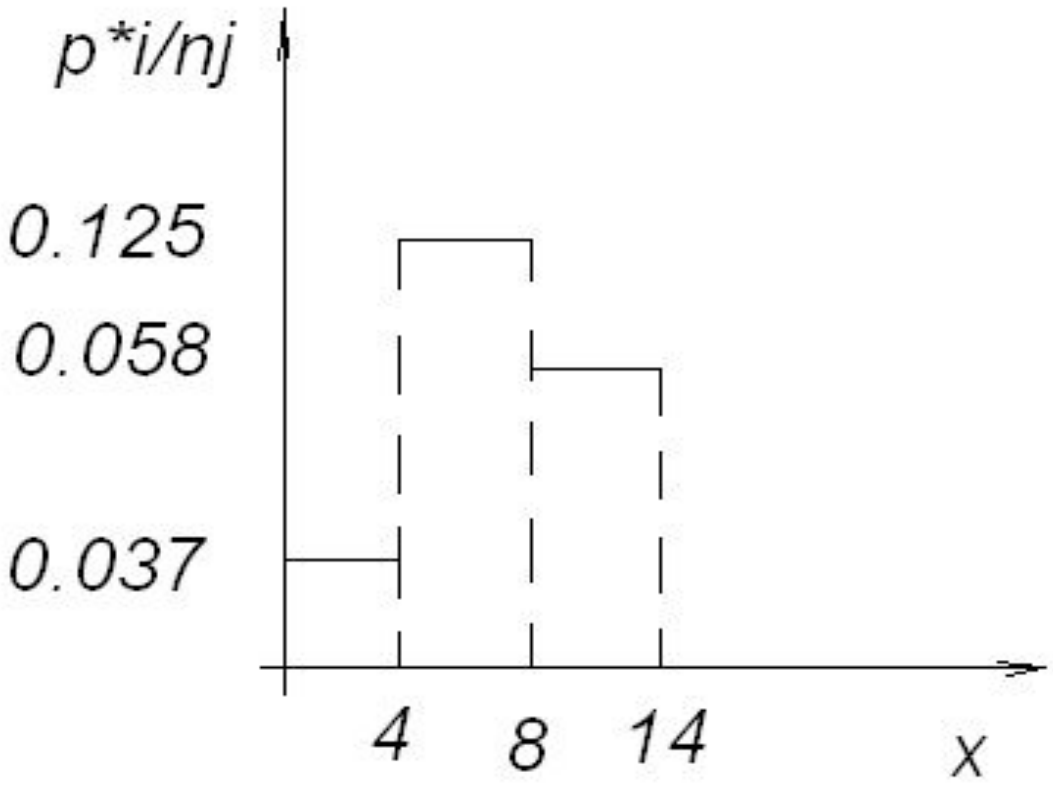
Полигоном относительных частот называют ломанную, отрезки которой соединяют точки (x_1,p^*_1), (x_2,p^*_2), …, (x_n,p^*_n).



Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиною $h_j=\Delta X_j$, а высоты равны отношению n_j / h_j (плотность частоты). Площадь j-го прямоугольника равна n_j – сумме частот j-го интервала. Следовательно, площадь гистограммы частот равна сумме всех частот, т. е. объему выборки.



Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною $h_j = \Delta X_j$, а высоты равны отношению p^*_j / h_j (плотность относительной частоты). Площадь j-го частичного прямоугольника равна p^*_j – сумме относительных частот j-го интервала. Следовательно, площадь гистограммы относительных частот равны сумме всех относительных частот, т.е. единице.


p*i/nj
0.125
0.058
0.037
4
8
14
X

6.4. Эмпирические функции распределения

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющей для каждого значения x частоту события $X<x$, т.е. $F^*(x)=n_x/n$, где n_x – число вариант (значений), меньших x, n – объем выборки. Например, для того чтобы найти $F^*(x')$, надо число вариант, меньших x', разделить на объем выборки $F^*(x') = n_{x'}/n$.

Из т. Бернулли следует, что при неограниченном увеличении n относительная частота события $X < x$, т.е. $F^*(x)$ стремится по вероятности к $F(x)$ этого события, т.к.

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

Эмпирическая (статистическая) функция распределения выборки используется для приближенной оценки теоретической (интегральной) функции распределения генеральной совокупности.

Это подтверждается тем, что $F^*(x)$ обладает всеми свойствами $F(x)$:

1) значения эмпирической функции принадлежат отрезку [0;1];

2) $F^*(x)$ – неубывающая функция;

3) если x_1 – наименьшая варианта, то $F^*(x)=0$ при $x<x_1$;

4) если x_2 – наибольшая варианта, то $F^*(x)=1$ при $x \geq x_2$.

С увеличением объема выборки и количества интервалов. содержащих в пределе одну реализацию случайной величины, гистограмма приближается к плотности распределения исследуемой случайной величины.

Полигон частот является статистическим аналогом ряда распределения случайной величины, а гистограмма – статистическим аналогом плотности распределения.