Математическая статистика

Предмет математической статистики.

Генеральная и выборочная совокупности

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений.

Двумя основными задачами математической статистики являются:

- определение способов сбора и группировки этих статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т.д.;

б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

Определим основные понятия математической статистики.

Генеральная совокупность– все множество имеющихся объектов.

Выборка – набор объектов, случайно отобранных из генеральной совокупности.

Пример:
Завод выпускает лампочки. Одно из свойств – время работы до сгорания. Если проверять всю продукцию (т.е. генеральную совокупность) - это экономически не выгодно. Поэтому производится выборка и на основе исследования этой выборки делают какие-либо выводы о всей генеральной совокупности. Иногда нельзя проверить всю генеральную совокупность – нельзя вскрыть все консервы, выстрелить все снаряды и т.д.

Объем генеральной совокупности *N* и объем выборки *n* – число объектов в рассматриваемой совокупности.

Виды выборки:

Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;

Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Для того, чтобы по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была **репрезентативной** (представительной).

Вариационный ряд и его характеристики

Пусть интересующая нас случайная величина X принимает в выборке значение $x_1 n_1$ раз, $x_2 - n_2$ раз, …, $x_к - n_к$ раз, причем $\sum_{i-1}^{k} n_k = n$, где n – объем выборки.

Тогда наблюдаемые значения случайной величины $x_1, x_2, \ldots, x_к$ называют **вариантами**, а $n_1, n_2, \ldots, n_к$ – **частотами**.

Если разделить каждую частоту на объем выборки, то получим **относительные частоты** $w_i = \frac{n_i}{n}$.

Последовательность вариант, записанных в порядке возрастания, называют **вариационным** рядом, а перечень вариант и соответствующих им частот или относительных частот – **статистическим рядом**:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
w_i	w_1	w_2	…	w_k

Пример: При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1,1,4,0,1,2,1,2,2,0,5,3,3,1,0,2,2,3,4,1.

Составим вариационный ряд: 0,1,2,3,4,5.

Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	0	1	2	3	4	5
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку**.

Для этого интервал выборки или размах выборки разбивают на κ - непересекающихся интервалов, длина которых для удобства расчетов чаще всего выбирается одинаково.

$$\Delta x = \frac{x_{max} - x_{min}}{k}.$$

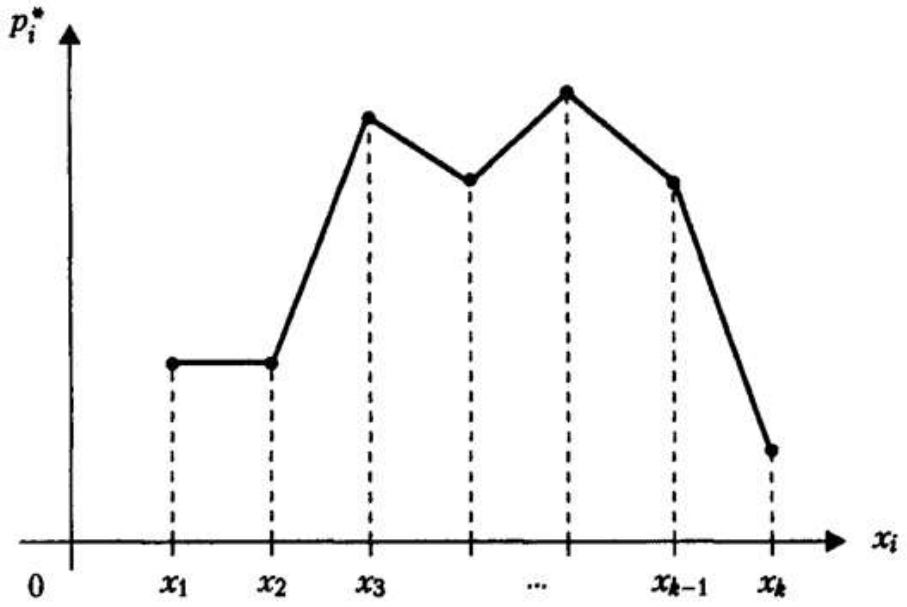
Обычно выбирают k=7-20. Также есть формулы, рекомендующие число интервалов $k=5 \lg n$ или $1+3{,}322 \lg n$ (формула Стерджеса). А затем находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал. Составленная по этим результатам таблица называется

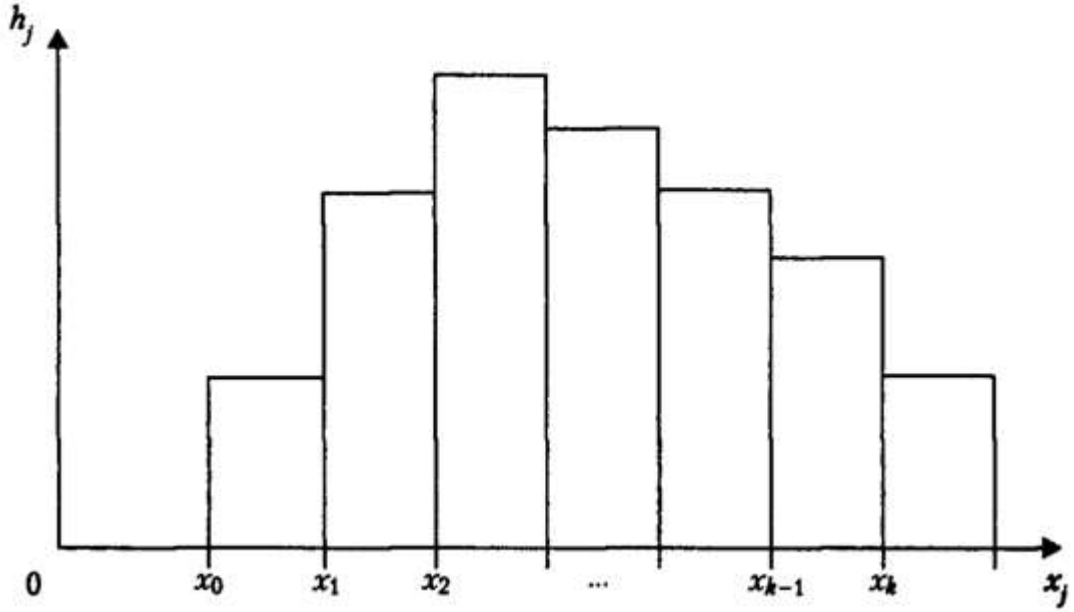
группированным статистическим рядом:

Номера интервалов	1	2	…	k
Границы интервалов	$(a,\ a+h)$	$(a+h,\ a+2h)$	…	$(b-h,\ b)$
Сумма частот вариант, попавших в интервал	n_1	n_2	…	n_k

Для наглядного представления выборки используют ее графическое отображение, используют два вида графика: ***полигон и гистограмма.***

Полигон частот - ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),…, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i– на оси ординат. Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим **полигон относительных частот.**





При построении **гистограммы** над каждым значением x_i строят прямоугольники высота, которых $\frac{n_k}{n}$.

Если вариационный ряд составлен по интервалам, то в качестве x_i следует брать середину интервала.

Очевидно, что величина интервала существенно влияет на вид гистограммы. При малой ширине в каждый интервал попадает незначительное число элементов выборки или даже не попадает ни одного и гистограмма становится сильно «изрезанной» и плохо передаёт основные особенности изучаемого распределения.

Определение: **Выборочной (эмпирической) функцией распределения** называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X<x$. Таким образом,

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число вариант, меньших x, n – объем выборки. Подробнее

$$F^*(x) = \sum_{x_i<x} \frac{n_i}{n} = \begin{cases} 0, x \le x_1 \\ \frac{n_1}{n}, \quad x_1 < x \le x_2 \\ \frac{n_1 + n_2}{n}, \quad x_2 < x \le x_3 \\ \sum_{i=1}^{m-1} \frac{n_i}{n}, \quad x_{m-1} < x \le x_m \\ 1, \quad x > x_m \end{cases}$$

В отличие от эмпирической функции распределения, найденной опытным путем, функцию распределения $F(x)$ генеральной совокупности называют **теоретической функцией распределения**. $F(x)$ определяет вероятность события $X<x$, а $F^*(x)$ – его относительную частоту. При достаточно больших n, как следует из теоремы Бернулли, $F^*(x)$ стремится по вероятности к $F(x)$.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами $F(x)$, а именно:

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

2) $F^*(x)$ – неубывающая функция.

3) Если x_1 – наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшая варианта, то $F^*(x) = 1$ при $x > x_к$.

Пример:

x_i	2	6	10
n_i	12	18	30

n=60 . Тогда

$$F^*(x) = \begin{cases} 0, & x \le 2 \\ \frac{12}{60}, & 2 < x \le 6 \\ \frac{12+18}{60}, & 6 < x \le 10 \\ 1, & x > 10 \end{cases}$$

Одна из задач математической статистики: по имеющейся выборке оценить значения числовых характеристик исследуемой случайной величины.

Определение: **Выборочным средним** $\bar{x}$ (или $\bar{x}_B$) называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_m x_m}{n} = \frac{\sum_{i=1}^{m} n_i x_i}{n},$$

где x_i – варианты, n_i – частоты, n-объём выборки, m- количество вариантов.