Лекция 1

Задачи и понятие матем. статистики. Генеральная и выборочная совокупности. Характеристики выборки. Гистограмма.

Математическая статистика изучает методы сбора и обработки статистических данных с целью получения теоретических и практических выводов вероятностного характера.

Пусть X – СВ, связанная с экспериментом $\mathcal{E}$, с неизвестным законом распределения. Генеральная совокупность значений СВ X — множество всех возможных значений, которые может принимать СВ X. Выборочная совокупность значений СВ X объёма n (или просто выборка объёма n) — совокупность n реализаций СВ X

$$x_1, x_2, \dots, x_n, \qquad (1)$$

полученных в результате проведения серии из n экспериментов $\mathcal{E}$. Задача математической статистики — построение надёжных заключений о генеральной совокупности на основе анализа выборочной совокупности.

В общем случае выборочные значения (1) могут повторяться:

$$\begin{array}{c|cccc} x_i & x_1 & x_2 & \dots & x_k \\ \hline n_i & n_1 & n_2 & \dots & n_k \end{array}, \qquad (2)$$

где $n_1+n_2+\dots+n_k=n$, $x_1<x_2<\dots<x_k$. Ряд $x_1, x_2, \dots, x_k$ называется вариационным рядом,

числа n_i, $i=1,2,\dots,k$, — частотами, а числа $w_i = \frac{n_i}{n}$ — относительными частотами, таблица (2) — распределением выборки. Ясно, что

$$\sum_{i=1}^{k} w_i = 1.$$

Функция распределения СВ X

$$F(x) = P\{X<x\}, \; x \in (-\infty; \infty),$$

может быть оценена (исходя из статистического восприятия вероятности) так:

$$P\{X<x\} \approx \frac{\sum_{i:\, x_i<x} n_i}{n},$$

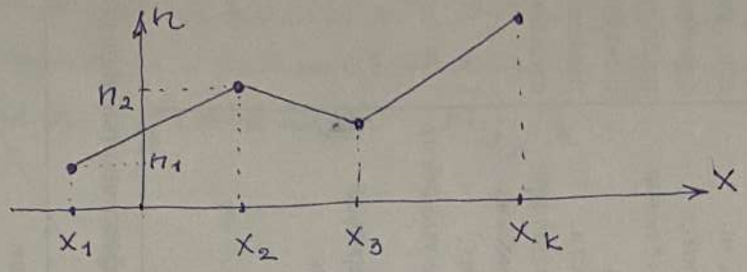
где $\sum_{i:\, x_i<x} n_i$ — сумма всех n_i таких, что ~~i~~ $x_i<x$. Обозначим $n_x = \sum_{i:\, x_i<x} n_i$. По определению, функция

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

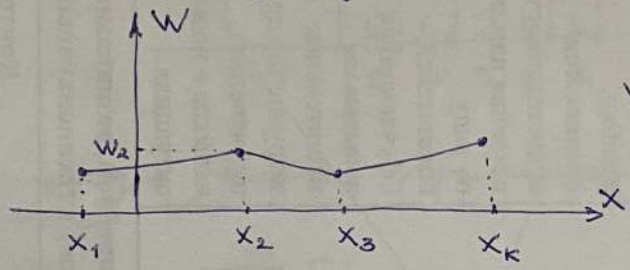
называется эмпирической функцией распределения СВ X, полученной по выборке (2). Очевидны свойства $F^*(x)$:

1) $0 \le F^*(x) \le 1$;
2) $F^*(x)$ — неубывающая функция;
3) поскольку в (2) x_1 — наименьшее, а x_k — наибольшее значение выборки, то $F^*(x)=0$ при $x \le x_1$ и $F^*(x)=1$ при $x > x_k$.

Полигон частот — ломаная, отрезки которой соединяют точки $(x_1,n_1),\dots,(x_k,n_k)$



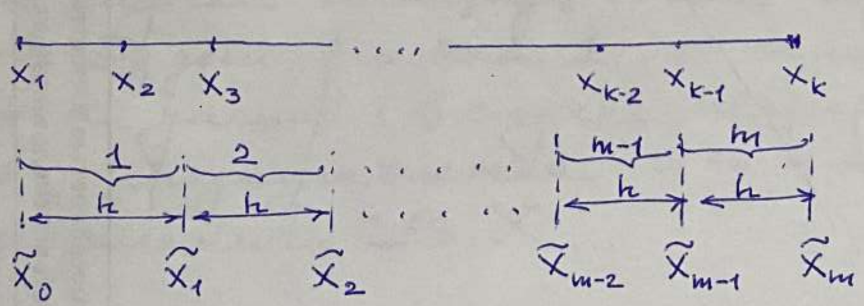
Полигон относительных частот — ломаная, отрезки которой соединяют точки $(x_1,w_1),\dots,(x_k,w_k)$.



$$w_1+w_2+\dots+w_k=1$$

Гистограмма

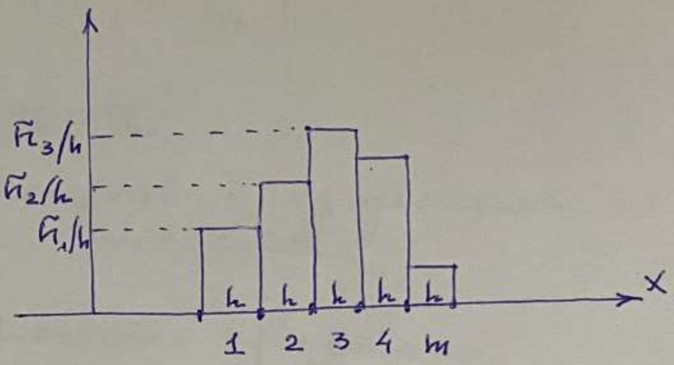
Разобьём все значения $x_1,x_2,\dots,x_k$ наблюдаемой СВ на какое-либо число m частичных интервалов длины h



Обозначим

$$\tilde{n}_j = \begin{cases} \text{Сумма всех } n_i, \\ \text{для которых} \\ \quad \tilde{x}_{j-1} \le x_i < \tilde{x}_j,\ j<m \\ \quad \tilde{x}_{m-1} \le x_i \le \tilde{x}_m \end{cases}$$

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которой служат интервалы длиной h, а высоты равны $\tilde{n}_j/h$:



Площадь всех прямоугольников будет равна

$$\sum_{j=1}^{m} \frac{\tilde{n}_j}{h} \cdot h = n_1+n_2+\dots+n_k = n$$

– общему числу всех значений СВ. Если высоты прямоугольников сделать равными $\tilde{n}_j/nh$, то соответствующая ступенчатая фигура называется гистограммой относительных частот. В этом случае площадь всех прямоугольников будет равна 1, а сама гистограмма даст (в случае непрерывной СВ X) представление о плотности распределения СВ X.

Некоторые характеристики выборки

Рассмотрим выборочное распределение

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

,

$n_1+n_2+\ldots+n_k=n$.

Определим следующие характеристики выборки:

выборочное среднее

$$\overline{x}=\frac{x_1n_1+\ldots+x_kn_k}{n}$$

выборочная дисперсия

$$D_x=\frac{(x_1-\overline{x})^2n_1+\ldots+(x_k-\overline{x})^2n_k}{n}$$

Эмпирический начальный момент порядка s

$$\nu_s=\frac{x_1^sn_1+\ldots+x_k^sn_k}{n}$$

эмпирический центральный момент порядка s

$$\mu_s=\frac{(x_1-\overline{x})^sn_1+\ldots+(x_k-\overline{x})^sn_k}{n}$$

Как видно, выборочное среднее — это эмпирический начальный момент 1-го порядка, а выборочная дисперсия — это эмпирический центральный момент 2-го порядка.

Если есть выборочное распределение для системы двух случайных величин X и Y, т.е. таблица

	$Y=y_1$	$Y=y_2$	…	$Y=y_k$
$X=x_1$	n_{11}	n_{12}	…	n_{1k}
…	…	…	…	…
$X=x_r$	n_{r1}	n_{r2}	…	n_{rk}

,

где n_{ij} – частота пары (x_i, y_j),

то для СВ $Z = f(X,Y)$, где $f(\ldots)$ – заданная функция, выборочное среднее $\overline{Z}$ определяется по формуле

$$\overline{Z} = \frac{\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{k} f(x_i, y_j)\, n_{ij}}{n},$$

где $n = \sum\limits_{i=1}^{r}\sum\limits_{j=1}^{k} n_{ij}$.

В частности, эмпирическая ковариация

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i-\overline{x})(y_j-\overline{y})\,n_{ij},$$

а эмпирический коэффициент корреляции

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{D_X D_Y}}.$$