Элементы математической статистики

Математическая статистика как наука, ее задачи. Основные понятия

Статистика возникла существенно раньше теории вероятностей. Еще в глубокой древности проводились переписи населения, велись земельные кадастры. Эти операции были связаны с наблюдениями и вычислениями. На протяжении многих веков статистика искала свой математический аппарат и нашла его в теории вероятностей. В результате возник такой раздел математики, как математическая статистика.

Математическая статистика – это раздел математики, изучающий методы сбора, систематизации и обработки результатов наблюдений с целью выявления статистических закономерностей т.е. отыскания законов распределения.

Математическая статистика, как и теория вероятностей, имеет дело с массовыми явлениями. Отличие математической статистики от теории вероятностей в том, что теория вероятностей изучает закономерности случайных явлений на основе абстрактного описания действительности (теоретической вероятностной модели), а математическая статистика оперирует непосредственно результатами наблюдений над случайными явлениями.

Математическая статистика как наука начинается с работ знаменитого немецкого математика Карла Фридриха Гаусса (1777–1855), который на основе теории вероятностей исследовал и обосновал метод наименьших квадратов, созданный им в 1795 г. и примененный для обработки астрономических данных (с целью уточнения орбиты малой планеты Церера). Его именем часто называют одно из наиболее популярных распределений вероятностей – нормальное, а в теории случайных процессов основной объект изучения – гауссовские процессы.

В конце XIX в. – начале XX в. крупный вклад в математическую статистику внесли английские исследователи, прежде всего К. Пирсон (1857–1936) и Р. А. Фишер (1890–1962). В частности, Пирсон разработал критерий «хи-квадрат» проверки статистических гипотез, а Фишер – дисперсионный анализ, теорию планирования эксперимента, метод максимального правдоподобия оценки параметров.

В 30-е годы XX в. поляк Ежи Нейман (1894–1977) и англичанин Э. Пирсон развили общую теорию проверки статистических гипотез, а советские математики академик А. Н. Колмогоров (1903–1987) и член-корреспондент АН СССР Н. В. Смирнов (1900–1966) заложили основы

непараметрической статистики. В сороковые годы XX в. румын А. Вальд (1902–1950) построил теорию последовательного статистического анализа.

Задачи математической статистики

Все задачи математической статистики касаются вопросов обработки наблюдений над массовыми случайными явлениями, но в зависимости от характера решаемого практического вопроса и от объѐма имеющегося экспериментального материала эти задачи могут принимать ту или иную форму.

Укажем некоторые типичные задачи математической статистики, часто встречаемые на практике.

1. Оценивание характеристик изучаемых явлений, объектов, то есть нахождение подходящих значений характеристик, например, параметров распределений случайных величин.
2. Установление детерминированных и вероятностных закономерностей,в частности, определение законов распределения случайных величин.
3. Проверка статистических гипотез.
4. Обеспечение методики сбора, обработки и анализа статистических данных, планирование эксперимента.

К методике обработки статистических данных предъявляются следующие требования: она должна сохранять типичные, характерные черты наблюдаемого явления и отбрасывать все несущественное, второстепенное, случайное.

В экономике, социологии, страховом деле постоянно приходится иметь дело со статистическим материалом, полученным в результате наблюдений, измерений, эксперимента, испытаний. Обработка статистического материала, характеризующего эти явления, и анализ полученных данных в зависимости от цели исследования представляет собой несомненную практическую ценность для описания деятельности обособленных страховых организаций, целью которых является продажа страхового обеспечения. В основном задачи, решаемые экономистами страховых компаний, являются комплексными, так как в условиях уже содержат рассчитанные элементы, связанные с теорией вероятности, обычно это результаты многолетних наблюдений и расчетов специалистов. Например, к таким элементам относят вероятность смерти в определенном возрасте или вероятность наступления другого страхового случая.

Результаты отдельных испытаний в экспериментах, наблюдениях и т.д. обычно считаются независимыми, а условия их проведения – неизменными.

В этом содержится некоторая идеализация, поскольку реально меняются внешние условия, режимы работы аппаратуры и т.д.

Описательная статистика

Генеральная совокупность. Выборка. Выбор

В практике статистических наблюдений различают два вида наблюдений:

- сплошное (изучают все объекты совокупности);
- выборочное (изучается лишь часть объектов совокупности);

Генеральная совокупность – вся подлежащая изучению совокупность объектов.

Выборочная совокупность (выборка) – часть объектов, которая отобрана для непосредственного наблюдения из генеральной совокупности. Обычно выборка составляет 5%-10% от генеральной совокупности.

Использование выборки для построения закономерностей, которым подчинена наблюдаемая случайная величина, позволяет избежать ее сплошного (массового) наблюдения, что часто бывает ресурсоемким процессом, а то и просто невозможным.

Числа объектов в генеральной совокупности и выборке называют их *объемами.* Генеральная совокупность может иметь как конечный, так и бесконечный объем. На практике всю генеральную совокупность изучают сравнительно редко, поскольку если совокупность содержит очень большое число объектов, то провести сплошное обследование невозможно. Тем более если исследование связано с уничтожением объекта или требует больших материальных затрат. В этих случаях изучают выборку.

Примеры.

1. Вся продукция предприятия есть генеральная совокупность, а отдельные экземпляры, подвергнутые контролю, составляют выборку.
2. При изучении продолжительности жизни отдельных слоев населения генеральной совокупностью является все население, а выборкой являются те совокупности, которые подвергались обследованию.
3. При изучении продолжительности телефонного разговора генеральной совокупностью являются все вызовы, а выборкой являются те вызовы, продолжительность которых измерялась.

Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности выносить суждения об ее свойствах в целом.

Основной недостаток выборочного метода – ошибки исследования, называемыми ошибками репрезентативности (представительства).

Однако неизбежные ошибки, возникающие при выборочном методе, могут быть заранее оценены и при правильно организованной выборке сведены к практически незначимым величинам.

Сплошное наблюдение (даже если оно возможно) приводит не только к росту стоимости, трудоемкости, увеличению времени исследования, но и к появлению неустранимых ошибок (т.к. каждое отдельное наблюдение поневоле производится с меньшей точностью).

Требования к выборке. Чтобы по выборке можно было судить о генеральной совокупности, она должна быть *репрезентативной*, т.е. она должна достаточно хорошо воспроизводить генеральную совокупность. Выборка будет обладать таким свойством, если каждый объект генеральной совокупности будет иметь один и тот же шанс быть выбранным, в этом случае выборка является случайной.

Число N объектов генеральной совокупности и число n объектов выборки называют объемами генеральной и выборочной совокупностей соответственно.

На практике применяются различные способы получения выборки. Принципиально эти способы можно подразделить на два вида:

1. Отбор, не требующий расчленения генеральной совокупности на части.

Сюда относятся:

а) простой случайный бесповторный отбор (объекты извлекают по одному из всей генеральной совокупности);

б) простой случайный повторный отбор.

2. Отбор, при котором генеральная совокупность разбивается на части.

Сюда относятся:

а) *типический отбор* (объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части. Например, если детали изготовляют на нескольких станках, то отбор производят не из всей совокупности деталей, произведенных всеми станками, а из продукции каждого станка в отдельности);

б*) механический отбор* (генеральную совокупность «механически» делят на столько групп, сколько объектов должно войти в выборку, а из каждой группы отбирают один объект. Например, если нужно отобрать 20 % изготовленных станком деталей, то отбирают каждую пятую деталь; если требуется отобрать 5 % деталей, то отбирают каждую двадцатую деталь, и т. д.);

в) *серийный отбор* (объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Например, если изделия изготовляются большой группой станков-автоматов, то подвергают сплошному обследованию продукцию только нескольких станков.

Вариационный и статистический ряды

Выборка является труднообозримым множеством. Для дальнейшего изучения выборку подвергают перегруппировке.

Определение. *Вариационным рядом* называется последовательность всех элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются.

Запись вариационного ряда: $x_1, x_2, \ldots, x_n$. Ему соответствует следующая таблица:

i	1	2	3	…	n
x_i	x_1	x_2	x_3	…	x_n

Элементы вариационного ряда x_i называют его вариантами или порядковыми статистиками.

Пример 1. Студенты получили следующие баллы по тесту: 11, 8, 9, 10, 8, 6, 7, 7, 9, 11, 10, 6, 5, 11, 10. Записать статистический и вариационный ряды.

Решение:

11, 8, 9, 10, 8, 6, 7, 7, 9, 11, 10, 6, 5, 11, 10 – это статистический ряд.

Расположим данные в порядке возрастания:

5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 10, 11, 11, 11 – это вариационный ряд.

Представим данный ряд в виде таблицы (с учетом повторений) и в порядке возрастания значений признака, получим ранжированный вариационный ряд.

x_i	5	6	7	8	9	10	11
n_i	1	2	2	2	2	3	3

Здесь x_i – значение признака (варианта), n_i – его частота («вес» значения признака). Сумма всех частот значений признака равна объему выборки: 1+ 2 + 2 + 2 + 2 + 3+ 3=15.

Дискретный статистический ряд

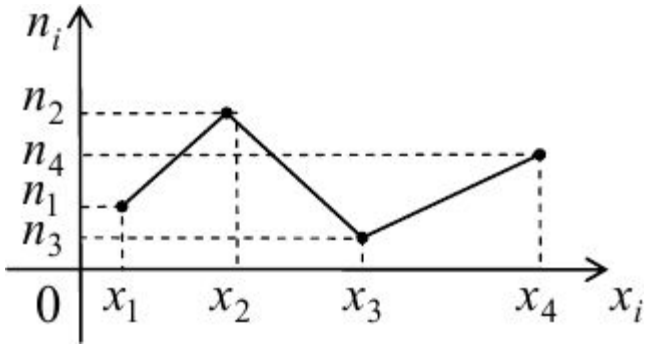
Вариационный ряд называется дискретным, если любые его варианты отличаются на конечную постоянную величину, и называется непрерывным (или интервальным), если его варианты могут отличаться друг от друга на сколь угодно малую величину.

Определение. Дискретным статистическим рядом называется последовательность различных вариант x_i с указанием частот повторения элементов. При этом вместо абсолютных частот n_i можно задавать распределение относительных частот $w_i = \frac{n_i}{n}$, где n объем выборки.

Дискретный статистический ряд можно записать в виде таблицы $\sum_{n} \omega_i = 1$:

x_i	x_1	x_2	…	x_n
ω_i	ω_1	ω_2	…	ω_n

Для наглядного представления эмпирических распределений для переменной дискретного типа строится график, где по оси X откладываются значения переменной, а по оси Y – значения частот. Полученные точки соединяют ломаной линией. Этот график называется полигоном.



Пример 2. Дана выборка, состоящая из чисел 1, 3, 1, 2, 3, 5, 1, 3, 1, 2. Составить вариационный и статистический ряды. Построить полигон относительных частот.

Решение:

Вариационный ряд имеет следующий вид: 1,1,1,1,2,2,3,3,3,5. Объем выборки n = 10.

Статистический ряд приведен в таблице.

x_i	1	2	3	5
n_i	4	2	3	1
ω_i	0,4	0,2	0,3	0,1

Полигон относительных частот имеет вид (рис. 1):

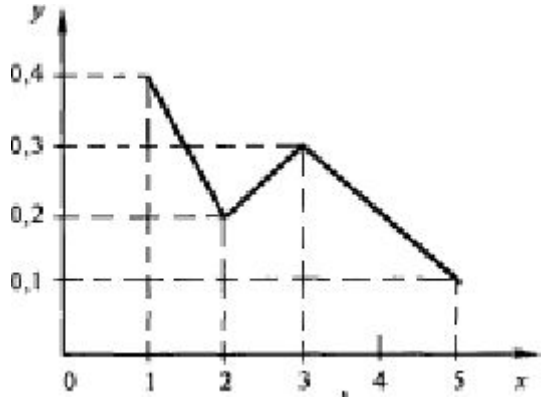


Рис. 1. Полигон относительных частот

Полигон относительных частот дает хорошее представление о распределении частот в выборке.

Элемент, отвечающий наибольшей частоте по сравнению с соседними элементами статистического ряда, называется выборочной **модой** (mod). Так, для выборки из примера 2 получаем, что mod = 1.

Минимальный и максимальный элементы называются крайними, иначе – экстремальными элементами вариационного ряда:

$x_{min} = x_{1;}\ x_{max} = x_n.$

Разность между максимальным и минимальным элементами называется размахом, или широтой выборки:

$$R = x_{max} - x_{min}.$$

Информативность этого показателя невелика. Можно привести много распределений, сильно отличающихся по форме, но имеющих одинаковый размах. Размах вариации используется иногда в практических исследованиях при малых (не более 10) объемах выборки, например, по размаху выборки легко оценить, насколько различаются лучший и худший результаты в группе спортсменов. При больших объемах выборки к его использованию надо относиться с осторожностью.

Интервальный статистический ряд

В случае непрерывного вариационного ряда составляют интервальный статистический ряд, под которым понимают упорядоченную совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданий в каждый из них значений случайной величины.

Как правило, частичные интервалы, на которые разбивается весь интервал варьирования, имеют одинаковую длину и представимы в виде

$[z_i, z_i + h)$ $i = 1, 2, ...,$ где k число интервалов.

Пусть x_{max}, x_{min} – наибольшее и наименьшее значения случайной величины. Длину $h = \frac{x_{max} - x_{min}}{k}$ следует выбирать так, чтобы построенный ряд не был громоздким, но в то же время позволял выявлять характерные изменения случайной величины.

Рекомендуется для нахождения m использовать формулу Старджесса:

$m = 1 + 3{,}322 \lg n$

Если окажется, что k – дробное число, то за длину интервала следует принять либо ближайшую простую дробь, либо ближайшую целую величину. Рекомендуется за начало первого интервала брать величину

$$z_1 = x_{min} - \frac{k}{2}.$$

Частота ω_i представляет собой число наблюдений, попавших в данный интервал.

Интервальный статистический ряд можно записать в виде таблицы

n

$\sum_{i=1} \omega_i = 1$:

x_i	(x_0,x_1)	(x_1,x_2)	…	(x_{k-1},x_k)
ω_i	ω_1	ω_2	…	ω_k

Пусть $\Delta x_i = x_i - x_{i-1}$ - длина *i*-го интервала. Обозначим длину интервала $h_i = \Delta x_i$. Интервальный ряд графически изображается в виде *гистограммы* (рис. 2*)*.

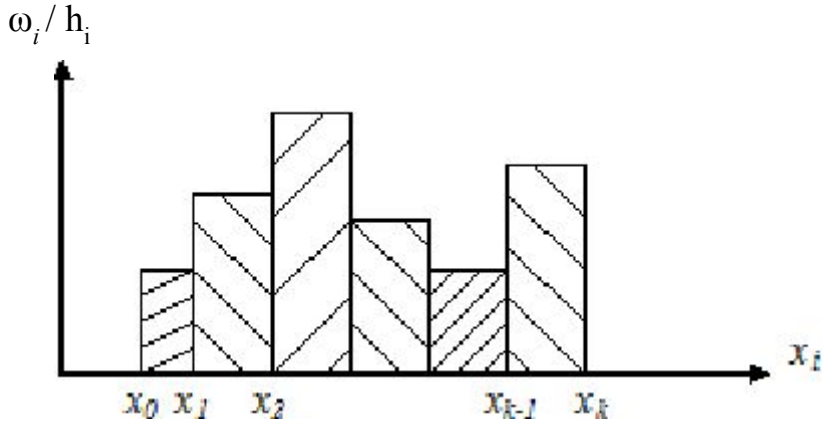


Рис. 2 Гистограмма

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы длиной h_i, а высоты равны ω_i / h_i.

Площадь *i*-го прямоугольника равна частоте ω_i , а площадь гистограммы относительных частот равна 1.

Пример 3. При измерении роста девушек некоторого института была получена следующая выборка (объема n = 30):

178	160	154	183	155	153	167	186	163	155
157	175	170	166	159	173	182	167	171	169
179	165	156	179	158	171	175	173	164	172

Необходимо построить интервальный вариационный ряд и его гистограмму.

Решение:

Найдем по формуле Старджесса оптимальное количество интервалов:

$m = 1 + 3{,}332 \lg n \approx 5{,}59$, *т.е.* $m = 6$.

Так как наибольшая варианта равна 186, а наименьшая 153, то вся выборка попадает в интервал (153; 186).

$z_{\min}^{1} = x_{\min} - \frac{h}{2} = 153 - 3 = 150$ и длина каждого частичного интервала равна

$\frac{186 - 150}{6} = 6.$

Получаем следующие шесть интервалов:

$$[150, 156); [156, 162); [162, 168); [168, 174); [174, 180),$$

а соответствующий интервальный вариационный ряд представлен в таблице.

X	150–156	156–162	162–168	168–174	174–180	180–186
n_i	4	5	6	7	5	3
ω_i	0,13	0,17	0,20	0,23	0,17	0,10

Находим высоты y_i по формуле $\frac{\omega_i}{h_i} = \frac{i}{6}$.

X	150–156	156–162	162–168	168–174	174–180	180–186
ω_i	0,13	0,17	0,20	0,23	0,17	0,10
$\frac{\omega_i}{h_i}$	0,022	0,028	0,033	0,038	0,028	0,017

График построенной гистограммы приведен на рисунке 3.

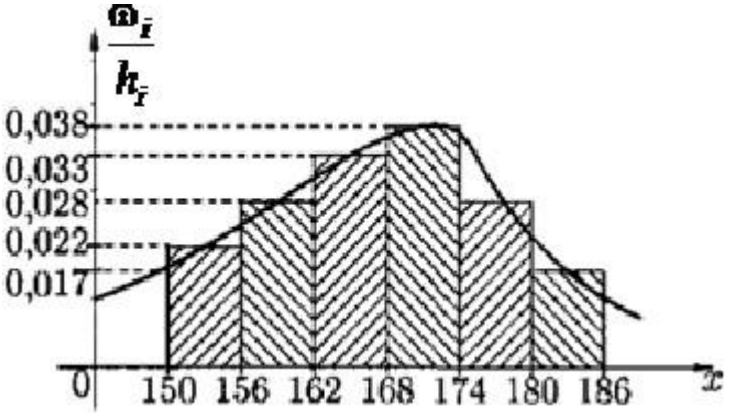


Рис. 3 Гистограмма

Эмпирическая функция распределения

Рассмотрим теперь статистический аналог интегральной функции распределения $F(x)=P(X<x)$ случайной величины X, называемый в статистике *эмпирической функцией распределения*.

Обозначим через n_x число наблюдений, при которых наблюдалось значение признака $X < x$. Число n_x называется *накопленной частотой*, а отношение $\frac{n_x}{n}$ называется *накопленной частостью*. Накопленную частость $\frac{n_x}{n}$ можно получить последовательным суммированием частостей ω_i^* всех вариантов x_i или интервалов (x_{i-1}, x_i), удовлетворяющих условию $X < x$.

Определение. *Эмпирической функцией распределения* (функцией распределения выборки) называется функция $F^*(x)$, определяющая для каждого значения x накопленную частость события $X < x$:

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

где n_x - число вариантов, меньших x ; n - объем выборки.

Эмпирическая функция распределения выборки $F^*(x)$ служит для оценки теоретической функции распределения генеральной совокупности $F(x)=P(X<x)$ и обладает всемисвойствами интегральной функции распределения теории вероятностей:

1. $0 \le F^*(x) \le 1$;
2. $F^*(x)$ – неубывающая функция;
3. $F^*(-\infty) = 0; \quad F^*(\infty) = 1.$

Функция $F^*(x)$ является «ступенчатой», имеются разрывы в точках, которым соответствуют наблюдаемые значения вариант. Величина скачка равна относительной частоте варианты.

В случае построения эмпирической функции распределения для интервального вариационного ряда при ее графическом изображении можно соединить точки графика, соответствующие правым концам интервалов, отрезками прямой. В результате получим непрерывную линию, называемую кумулятивной кривой или *кумулятой.*

Пример 4. Найти эмпирическую функцию по заданному распределению выборки:

x_i	1	5	8
n_i	10	15	25

Решение:

Найдем объем выборки: $n = 10 + 15 + 25 = 50$. Наименьшая варианта равна единице, поэтому $F^*(x) = 0$ при $x \le 1$.

Значение $X < 5$, а именно $x_1 = 1$, наблюдалось 10 раз, следовательно, $F^*(x) = \frac{10}{50} = 0{,}2$ при $1 < x \le 5$.

Значения $x < 8$, а именно: $x_1 = 1$ и $x_2 = 5$, наблюдались $10 + 15 = 25$ раз, следовательно, $F^*(x) = \frac{25}{50} = 0{,}5$ при $5 < x \le 8$.

Так как $x = 8$ – наибольшая варианта, то $F^*(x) = 1$ при $x > 8$.

Напишем искомую эмпирическую функцию:

$$F(x) = \begin{cases} 0 & \text{при} \quad x \le 1, \\ 0{,}2 & \text{при} \quad 1 < x \le 5, \\ 0{,}5 & \text{при} \quad 5 < x \le 8, \\ 1 & \text{при} \quad x > 8. \end{cases}$$

График этой функции изображен на рисунке 4.

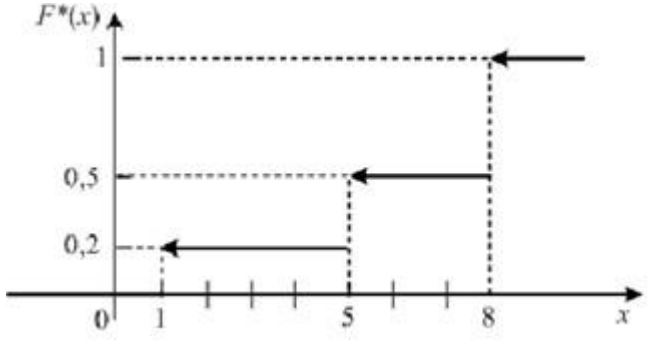


Рис. 4 График эмпирической функции распределения

Пример 5. Данные о количестве пациентов кардиологического отделения больницы приведены в таблице.

62	54	84	59	75	43	49	89	28	49
40	53	18	18	55	51	26	68	76	65
43	39	47	65	55	29	33	42	51	95
85	46	45	42	48	6	73	54	70	56
69	66	33	100	58	42	89	41	36	72
54	50	54	45	48	11	62	33	32	61
36	31	84	61	26	53	64	50	66	63
77	31	84	61	26	53	64	50	66	63
9	30	69	60	9	30	4	27	74	62
19	42	55	79	77	31	92	30	39	96

Найти эмпирическую функцию распределения по данным выборки.

Решение:

Найдем по формуле Старджесса оптимальное количество интервалов:
$m = 1 + 3{,}332 \lg n \approx 7{,}66$, *т.е.* $m = 8$.

Так как наибольшая варианта равна 100, а наименьшая 4, то вся выборка попадает в интервал (4; 100).

Длина каждого частичного интервала равна $\frac{100-4}{8} = 12$

Получаем следующие восемь интервалов:

$$[4,16);[16,28);[28,40);[40,52);[52,64);[64,76);[76,88);[88,100),$$

а соответствующий интервальный вариационный ряд представлен в таблице ниже.

X	4–16	16–28	28–40	40–52	52–64	64–76	75–88	88–100
n_i	5	8	16	20	24	14	7	6
ω_i	0,05	0,08	0,16	0,2	0,24	0,14	0,07	0,06

Гистограмма и эмпирическая функция плотности распределения для полученного интервального ряда представлены на рисунке 5.

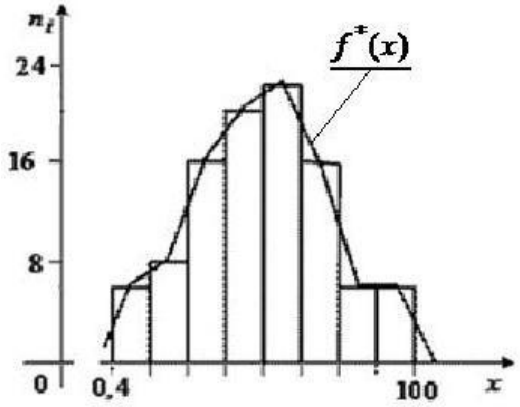


Рис. 5 Гистограмма функции распределения

Запишем эмпирическую функцию распределения.

$$F^*(x)=\begin{cases}0; & x\le 4,\\ 0{,}05; & 4<x\le 16,\\ 0{,}13; & 16<x\le 28,\\ 0{,}29; & 28<x\le 40,\\ 0{,}49; & 40<x\le 52,\\ 0{,}73; & 52<x\le 64,\\ 0{,}87; & 64<x\le 76,\\ 0{,}94; & 76<x\le 88,\\ 1; & x>88.\end{cases}$$

График эмпирической функции распределения и кумулята представлена на рисунке 6.

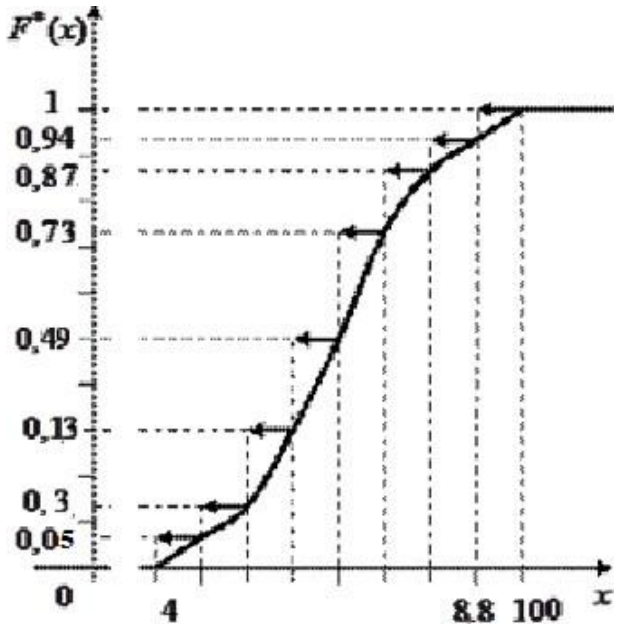


Рис. 6 График эмпирической функции распределения