

Статистический анализ экспериментальных данных

При выполнении измерений экспериментатор пытается определить значение той или иной величины. И как только начинаются измерения, он сталкивается с интересной ситуацией: если использовать достаточно точные приборы, то можно увидеть, что повторное измерение одной и той же величины приводит иногда к результатам, **слегка отличающимся** от результатов первоначального измерения.

Это явление характерно как для простых, так и для сложных измерений.

Почему существует разброс, откуда берется изменение?

Ответ на этот вопрос очевиден: условия проведения эксперимента все время меняются, и в условиях реального эксперимента от них избавиться невозможно.

Мы «обречены» выполнять измерения величин, которые никогда не остаются постоянными.

Поэтому постановка вопроса о значении некоторой величины может быть некорректной, нужна постановка такого вопроса, который отражал бы это свойство изменчивости.

Решение состоит в том, чтобы характеризовать физическую величину **не одним значением, а вероятностью найти в эксперименте то или иное значение.**

Для этого вводится функция, называемая **распределением вероятности** обнаружения физической величины, которая показывает, какие значения чаще встречаются в эксперименте.

Далее мы увидим, что функция распределения в большинстве экспериментов является достаточно простой и имеет две характеристики.

Первая – **среднее значение физической величины,**

вторая – показывает **область вокруг этой средней величины,** в которой сосредоточено большинство результатов эксперимента.

Она характеризует **ширину этого распределения** и называется погрешностью.

Эта ширина имеет строгую интерпретацию в терминах теории вероятностей, т.е. можно указать, с какой вероятностью мы должны обнаружить истинное значение в заданной области вокруг измеренного среднего значения. Назовем эту погрешность **естественной**.

Для экспериментатора построение функции распределения требует проведения **многократных** (бесконечного числа) **измерений**, что бывает дорого и никому не нужно.

Поэтому приходится ограничиваться **конечным числом измерений**, что приносит **дополнительную погрешность**.

Возникает и другая проблема: в каждом эксперименте присутствует **измерительный прибор**, который вносит изменения в начальную функцию распределения, приводя к **дополнительной (приборной) погрешности**.

Разделение погрешности на естественную и приборную достаточно **условное**, оно позволяет лучше понять природу погрешности.

Экспериментатор должен всегда задавать себе два вопроса:

Как измерить физическую величину, т.е. как определить ее характеристики— среднюю и ширину, и

до какой степени удастся разумно уменьшить погрешность эксперимента?

Поэтому важно понимать взаимосвязь между тремя составляющими погрешности:

- естественную погрешность можно уменьшить, изменяя условия проведения эксперимента,
- погрешность, связанную с конечностью числа измерений — увеличивая их число,
- приборную — используя более точные методы и инструменты измерений.

Вместе с тем невозможно **уменьшить погрешность до нуля**. Для нее существует **нижний предел**, оценка которого – **принципиальный физический вопрос**.

Поэтому нашей задачей является определить те экспериментальные методы, которые адекватны желаемой и достижимой точности.

В зависимости от желаемой точности могут возникнуть различные ситуации:

- если мы хотим получить порядок измеряемой величины, то и погрешность должна оцениваться грубо;
- если мы хотим получить точность порядка нескольких процентов, тогда необходимо и более аккуратно определять погрешности;
- если необходимо получить точность, сравнимую с точностью эталонных измерений, то проблема определения погрешности может стать более важной и сложной, чем проблема измерения самой величины.

Кроме указанных в эксперименте могут иметь место и другие источники ошибок, которые вызывают так называемые **систематические ошибки**.

Выявление их и анализ **намного сложнее**, чем случайных.

Можно указать три основных источника систематических ошибок:

- методика, выбранная для проведения эксперимента,
- плохая работа измерительных приборов, и, наконец,
- ошибки самого экспериментатора.

Поскольку отклик из-за влияния неконтролируемых факторов является **случайной величиной**, то при обработке результатов эксперимента широко **используется аппарат теории вероятности и математической статистики**, поэтому необходимо **напомнить** необходимые понятия и определения этого раздела математики.

КРАТКИЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Случайные величины и параметры их распределений

Поскольку из-за влияния неконтролируемых факторов отклик – это всегда случайная величина, при обработке результатов эксперимента широко используется аппарат теории вероятностей и математической статистики, поэтому напомним некоторые основные понятия и определения этого раздела математики.

Случайное событие – событие, реализацию которого при определенном комплексе условий невозможно заранее предсказать.

Например, реализацию такого события, как пять аварий на шахте в течение месяца, невозможно предсказать заранее, поскольку аварий может быть и три, и семь, и четыре, и т.д.

Случайная величина – величина, которая может принимать какое-либо значение из установленного множества и с которой связано вероятностное распределение.

Случайная величина может быть дискретной или непрерывной.

Дискретная случайная величина – случайная величина, которая может принимать значения только из конечного или **счетного множества действительных чисел**.

Непрерывная случайная величина - случайная величина, которая может принимать **любые значения** из конечного или бесконечного интервала.

Если при фиксированном наборе уровней всех контролируемых факторов провести n измерений отклика X , то в результате будет получен ряд хотя и близких, но отличающихся друг от друга значений:

$$x_i \quad i=(1, 2, \dots, n,) \quad (1)$$

где x_i – i -е измерение величины x ;

x_1, x_2, \dots, x_n – реализация случайной величины x .

Пример 2.1. В результате изучения работы шахты на протяжении полутора лет было зарегистрировано следующее количество аварий в течение каждого месяца (табл. 1).

Таблица 1

Число аварий на шахте по месяцам (общее число наблюдений $n = 18$)

Месяц	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Число аварий	3	4	3	5	5	5	6	4	6	5	5	2	4	6	7	5	6	7

В данном примере число аварий на шахте в течение месяца - это дискретная случайная величина.

В первом из $n = 18$ измерений этой величины было получено значение $x_1 = 3$, во втором - $x_2 = 4$ и т.д., до $x_{18} = 7$

Приведенные в табл. 1 значения - это реализация такой случайной величины, как число аварий на шахте в течение месяца.

Каждому значению дискретной случайной величины X (любому из событий A , когда случайная величина X принимает какое-либо строго определенное значение x), можно поставить в соответствие следующее отношение:

$$W=m/n \quad (2)$$

где m – число наблюдений, в которых дискретная случайная величина X оказалась равна x ;

n – общее количество наблюдений.

Величину W называют частотой реализации события A .

В примере 1, в шести наблюдениях: $i = 4, 5, 6, 10, 11$ и 16 , количество аварий на шахте в течение месяца X оказалось равным пяти ($X = 5$), следовательно, частота реализации такого события, как пять остановок, равна $6/18 = 0,33$.

Частоты реализаций для других событий (две, три, четыре и т.д. аварии) приведены в табл. 2.

Таблица 2

Число аварий x	2	3	4	5	6	7
Количество наблюдений m , в которых реализовалось событие $X = x$	1	2	3	6	4	2
Частота реализации, $W = m / n$	0,06	0,11	0,17	0,33	0,22	0,11

Если продолжить наблюдения за работой шахты в течение еще полутора лет, то, конечно же, совершенно не обязательно, что на протяжении следующих восемнадцати месяцев пять аварий будет снова зарегистрировано ровно в 6 случаях из 18 наблюдений, а частота реализации этого события опять окажется равной 0,33.

Однако при возрастании числа повторений одного и того же комплекса условий частота реализации такого события, как, например, пять аварий в течение месяца, будет принимать все **более и более устойчивое значение.**

Так, если подсчитать частоту реализации данного события за 36 месяцев, то она уже практически не будет отличаться от того значения, которое затем можно будет получить за четыре с половиной года (при условии, что за все это время наблюдений в работе шахты не произойдет никаких существенных изменений).

Предел, к которому стремится отношение m/n при неограниченном возрастании числа опытов n , называется вероятностью случайного события.

Вероятность $P(A)$ события A – число от нуля до единицы, которое представляет собой предел частоты реализации события A при неограниченном числе повторений одного и того же комплекса условий.

Для дискретной случайной величины можно указать вероятность, с которой она принимает каждое из своих возможных значений конечного или счетного множества действительных чисел.

Для непрерывной случайной величины задают вероятность ее попадания в один из заданных интервалов области ее определения (поскольку вероятность того, что она примет какое-либо конкретное свое значение, стремится к нулю).

Полностью свойства случайной величины описываются **законом ее распределения**, под которым понимают связь между **возможными значениями случайной величины и соответствующими им вероятностями**.

Распределение случайной величины – функция, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

В математике используют два способа описания распределений случайных величин:

интегральный (функция распределения)
и дифференциальный (плотность распределения).

Функция распределения $F(x)$ – функция, определяющая для всех действительных x вероятность того, что случайная величина X принимает значение не больше, чем x .

$$F(x) = P(X \leq x). \quad (3)$$

Функция распределения $F(x)$ имеет следующие свойства:

1. Ее ордината, соответствующая произвольной точке x_1 , представляет собой вероятность того, что случайная величина X будет меньше, чем x_1 , т.е. $F(x_1) = P(X \leq x_1)$.

2. Функция распределения принимает значение, заключенное между нулем и единицей:

$$0 \leq F(x) \leq 1 \quad (4)$$

3. Функция распределения стремится к нулю при неограниченном уменьшении x и стремится к единице при неограниченном возрастании x , то есть

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1. \quad (5)$$

$x \rightarrow -\infty$

$x \rightarrow +\infty$

4. Функция распределения представляет собой монотонно возрастающую кривую, то есть

$$F(x_2) > F(x_1), \text{ если } x_2 > x_1. \quad (5a)$$

5. Ее приращение на произвольном отрезке $(x_1; x_2)$ равно вероятности того, что случайная величина X попадет в данный интервал:

$$F(x_2) - F(x_1) = P(X \leq x_2) - P(X \leq x_1) = P(x_1 < X \leq x_2). \quad (6)$$

Рассмотрим, какие особенности имеют функции распределения дискретных случайных величин.

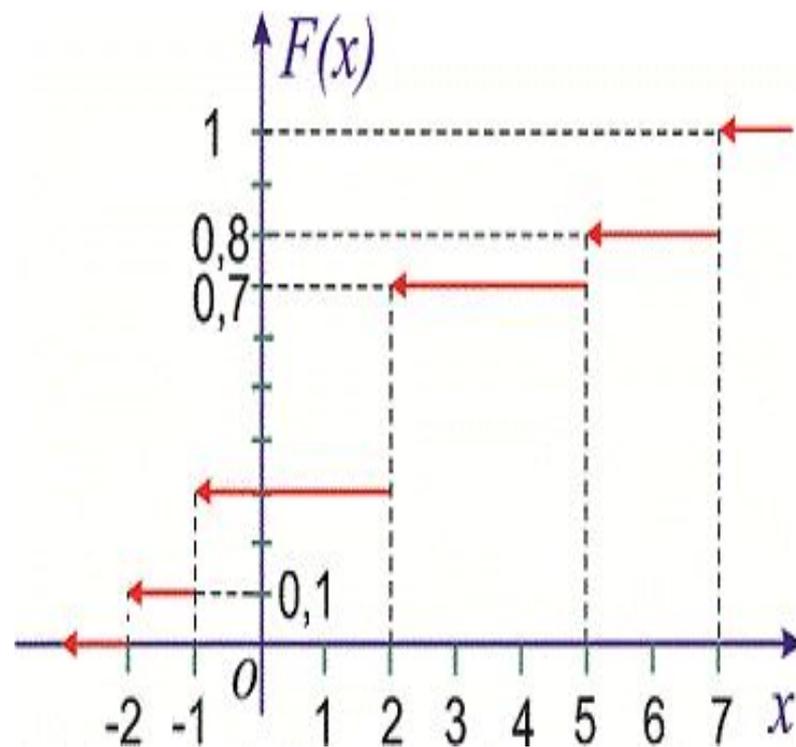
Пусть X – дискретная случайная величина, принимающая возможные значения x_1, x_2, \dots, x_n с вероятностями p_1, p_2, \dots, p_n .
Функция распределения вероятностей этой случайной величины X равна

$$F(x) = P(X \leq x) = \sum_{x_k} p_k$$

где производится суммирование вероятностей всех возможных значений случайной величины X , меньших чем x .

Такая функция всегда разрывная, ступенчатая (рис.): от $-\infty$ до x_1 включительно функция равна нулю, в точке x_1 происходит скачок на величину p_1 , и функция остается постоянной до x_2 включительно и т.д., то есть возможным значениям случайной величины соответствуют скачки функции, равные вероятностям этих значений.

Последний скачок на p_n происходит в точке x_n , и функция равна единице от x_n до $+\infty$. Таким образом, сумма всех скачков равна единице.



Плотность распределения $f(x)$ – первая производная (если она существует) функции распределения

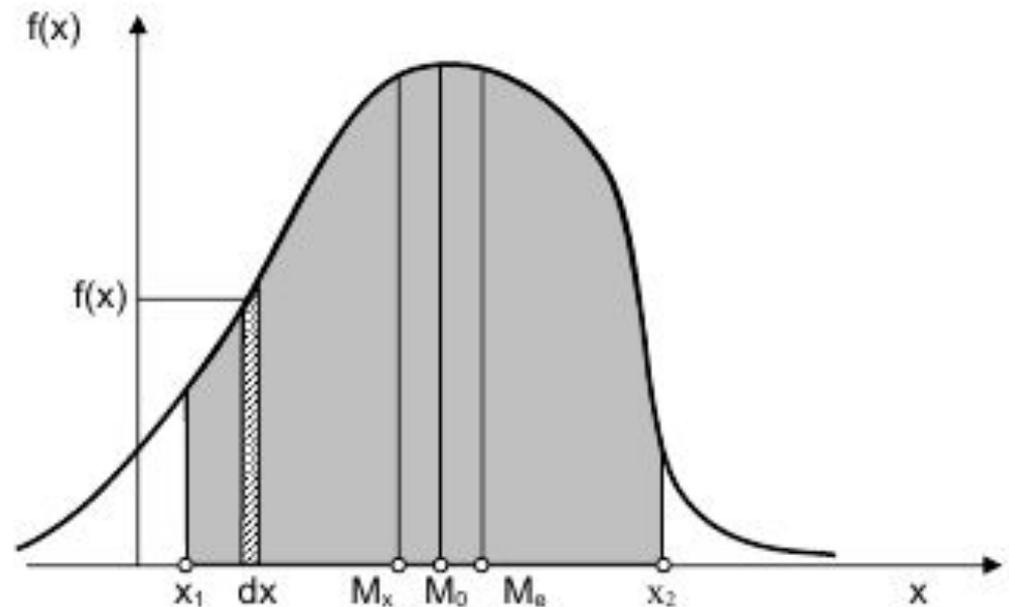
$$f(x) \stackrel{(7)}{=} \frac{dF(x)}{dx}$$

Плотность функции распределения $f(x)$ имеет следующие свойства

1 Плотность распределения вероятностей является неотрицательной функцией, т.е.

$$f(x) \geq 0. \quad (8)$$

Это свойство справедливо, так как $F(x)$ есть неубывающая функция.



2. Функция распределения случайной величины X равна определенному интегралу от плотности распределения вероятностей в пределах

$$F(x) = \int_{-\infty}^x f(x) dx. \quad (9)$$

3. Вероятность события, состоящая в том, что случайная величина X примет значение, заключенное в полуинтервале $[x_1, x_2]$, равна определенному интегралу от плотности распределения вероятностей на этом полуинтервале:

$$P(x_1 < X < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx.$$

4. Интеграл плотности распределения в бесконечно большом интервале $(-\infty, +\infty)$ равен единице:

$$P(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(x) dx = 1,$$

так как попадание случайной величины в интервал $-\infty < X < +\infty$ есть достоверное событие

В большинстве случаев при обработке экспериментальных данных, основываясь на тех или иных **предположениях (гипотезах)** относительно свойств исследуемой случайной величины, удается записать функцию ее распределения (а следовательно, и плотность распределения как первую производную от функции распределения) с точностью до некоторых неизвестных параметров.

Например, для случайной величины, которая удовлетворяет так называемому нормальному закону распределения (закону распределения Гаусса), функцию распределения можно записать в виде

$$F(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^x e^{-\frac{(x-M_x)^2}{2\sigma_x^2}} dx,$$

для случайной величины, имеющей, например, распределение Вейбула (используемое для описания результатов многих экспериментов), функция распределения определяется следующим выражением:

$$\begin{aligned} F(x) &= 1 - e^{-\left(\frac{(x-x_H)}{c}\right)^b}, \text{ при } X > x_H, \\ \hline F(x) &= 0, \quad (13) \quad \text{при } X \leq x_H. \end{aligned}$$

В функциях (12) и (2.13) константы Mx , σ_x^2 и c , b , x_H являются параметрами распределений, причем первое из этих двух выражений относится к двухпараметрическому виду закона распределения, а второе, соответственно, – к трехпараметрическому

Параметр распределения – постоянная, от которой зависит функция распределения.

Следовательно, если известен вид функции распределения (каким-либо образом установлено, что случайная величина не противоречит тому или иному закону распределения), то для того, чтобы однозначно охарактеризовать случайную величину, достаточно задать только лишь параметры ее распределения.

Важнейшими параметрами распределения, задающими случайную величину X , являются ее **математическое ожидание** M_x (характеризует центр рассеивания) и **дисперсия** σ_x^2 (характеризует степень рассеивания).

Математическое ожидание M_x – среднее взвешенное по вероятностям значение случайной величины.

Для дискретной случайной величины математическое ожидание определяется выражением

$$M_x = \sum x_i p_i, \quad (14)$$

где x_i – значения дискретной случайной величины, а $p_i = P(X = x_i)$.

Если в условиях примера 1 предположить, что $p_i \approx W_i$ (см. табл. 2), то для математического ожидания такой дискретной случайной величины, как

число остановок доменной печи в течение месяца, можно получить следующее значение:

$$M_x = 2 \cdot 0,06 + 3 \cdot 0,11 + 4 \cdot 0,17 + 5 \cdot 0,33 + 6 \cdot 0,22 + 7 \cdot 0,11 = 4,87.$$

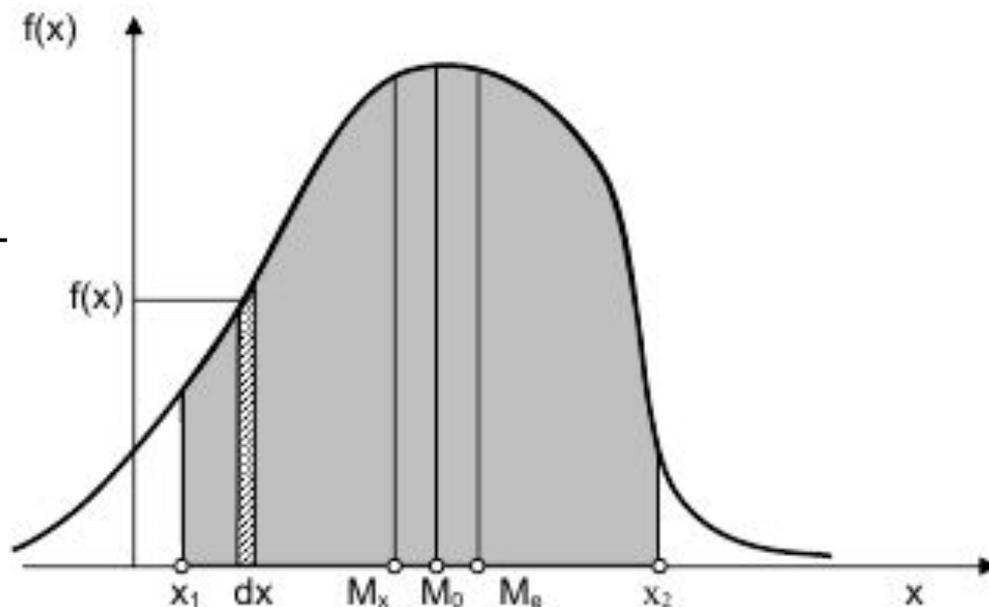
Для непрерывной случайной величины математическое ожидание определяется интегралом

$$M_x = \int_{-\infty}^{+\infty} xf(x)dx,$$

где $f(x)$ – плотность распределения непрерывной случайной величины.

Можно отметить, что геометрический смысл математического ожидания непрерывной случайной величины – это абсцисса центра тяжести фигуры под кривой плотности распределения $f(x)$. Сказанное проиллюстрируем на рис. 2, где видно, что произведение $f(x)dx$ есть площадь элементарного участка под кривой $f(x)$, а x – абсцисса этого участка, т.е. расстояние от начала координат.

Следовательно, интеграл (15) дает абсциссу центра тяжести всей площади фигуры под кривой $f(x)$.



Кроме математического ожидания центр рассеивания случайной величины можно еще охарактеризовать такими параметрами ее распределения, как мода и медиана

Мода M_0 – значение случайной величины, соответствующее локальному максимуму плотности вероятностей для непрерывной случайной величины или локальному максимуму вероятности для дискретной случайной величины.

Для примера 1 (см. табл. 2), при условии, что $p_i \approx W_i$, мода M_0 числа аварий на шахте равна 5, поскольку именно этому значению данной дискретной случайной величины соответствует локальный максимум вероятности, равный 0,33.

Медиана M_e – значение случайной величины, для которого функция распределения принимает значение S , или имеет место «скачок» со значения, меньшего чем S , до значения, большего чем S .

В примере 1, если предположить, что функция распределения от четырех аварий $F(4)$ (вероятность того, что число аварий на шахте в течение месяца будет не более четырех) равна $0,06 + 0,11 + 0,17 = 0,34$, а функция распределения $F(5) = 0,34 + 0,33 = 0,67$, то медианой M_e такой дискретной случайной величины, как количество аварий на шахте в течение месяца, будет значение $M_e = 5$.

Дисперсия случайной величины σ_x^2 – математическое ожидание случайной величины $(X - M_x)^2$.

Для дискретной случайной величины дисперсия определяется следующим математическим выражением:

$$\sigma_x^2 = \sum_{i=1}^n (x_i - M_x)^2 \cdot p(x_i)$$

В примере 1 (опять же, если предположить, что $p_i \approx W_i$) значение дисперсии числа остановок доменной печи равно:

$$\sigma_x^2 = (2 - 4,87)^2 \cdot 0,06 + (3 - 4,87)^2 \cdot 0,11 + (4 - 4,87)^2 \cdot 0,17 + (5 - 4,87)^2 \cdot 0,33 + (6 - 4,87)^2 \cdot 0,22 + (7 - 4,87)^2 \cdot 0,11 = 1,7931.$$

Для непрерывной случайной величины дисперсия определяется выражением

$$\sigma_x^2 = \int_{-\infty}^{+\infty} (x - M_x)^2 \cdot f(x) dx$$

где x – значения непрерывной случайной величины X ;

$f(x)$ – плотность распределения;

M_x – математическое ожидание.

Дисперсия имеет размерность квадрата единицы измерения случайной величины, а положительное значение квадратного корня из дисперсии называется средним квадратичным отклонением.

Среднее квадратичное отклонение σ_x – неотрицательный квадратный корень из дисперсии.

$$\sigma = +\sqrt{\sigma_x^2} \cdot (18)$$

Для примера 1 среднее квадратичное отклонение количества аварий на шахте в течение месяца равно $\sigma_x = + (1,7931)^{(1/2)} = 1,34$.

Дадим определение еще одного параметра распределения случайной величины, который носит название квантиль.

Квантиль порядка P , x_p – значение случайной величины, для которого функция распределения принимает значение P или имеет место «скачок» со значения, меньшего чем P , до значения, большего чем P :

$$F(x_p) = P. \quad (19)$$

Из этого определения квантиля следует, что медиана M_e – это квантиль порядка S , т.е. $M_e = x_{0,5}$.

Вероятность попадания случайной величины X в интервал $[x_{p1}, x_{p2}]$ равна

$$P(x_{p1} < X \leq x_{p2}) = P(X \leq x_{p2}) - P(X \leq x_{p1}) = F(x_{p2}) - F(x_{p1}) = P_2 - P_1 \quad (2.20)$$

В примере 1 квантиль порядка $0,95$ количества аварий на шахте скорее всего равен семи $x_{0,95} = 7$, поскольку $F(6) \approx 0,06 + 0,11 + 0,17 + 0,33 + 0,22 = 0,89$, а $F(7) \approx 0,89 + 0,11 = 1,00$.