МИСИ

МГСУ

СТРОИТЕЛЬНОЕ МАТЕРИАЛОВЕДЕНИЕ

СТАТИСТИЧЕСКИЕ РЕШЕНИЯ ТЕХНОЛОГИЧЕСКИХ ЗАДАЧ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ»

СТАТИСТИЧЕСКИЕ МЕТОДЫ РЕШЕНИЯ ТЕХНОЛОГИЧЕСКИХ ЗАДАЧ

*Рекомендовано Учебно-методическим объединением вузов РФ
по образованию в области строительства
в качестве учебного пособия для подготовки магистров
по направлению 08.04.01 (270800) Строительство
(магистерская программа «Строительное материаловедение»)
(24.02.2015 г., № 102-15/875)*

Москва 2015

УДК 691.3:311
ББК 38.3
С78

Рецензенты:
доктор технических наук *Л.А. Алимов*, профессор кафедры технологии вяжущих веществ и бетонов ФГБОУ ВПО «МГСУ»;
профессор, доктор технических наук *А.Ф. Бурьянов*, исполнительный директор Российской гипсовой ассоциации;
кандидат технических наук И.В. *Бессонов*, ведущий научный сотрудник НИИСФ РААСН

Авторы:
О.В. Александрова, Т.А. Мацеевич, Л.В. Кирьянова, В.Г. Соловьев

С78 **Статистические** методы решения технологических задач : учебное пособие / О.В. Александрова, Т.А. Мацеевич, Л.В. Кирьянова [и др.] ; М-во образования и науки Рос. Федерации, Моск. гос. строит. ун-т. Москва : МГСУ, 2015. 160 с.
ISBN 978-5-7264-1076-0

Освещены вопросы планирования и обработки результатов эксперимента в области строительных материалов. Изложены процедуры математической обработки для оценки результатов эксперимента, проверки статистических гипотез. Рассмотрены вопросы планирования эксперимента с целью математического описания и выявления важнейших факторов, воздействующих на объект исследований с области строительных материалов. Приведены результаты исследований, связанные с определением свойств строительных материалов.

Для студентов, обучающихся по направлению 08.04.01 (270800) Строительство (магистерская программа «Строительное материаловедение»).

УДК 691.3:311
ББК 38.3

ISBN 978-5-7264-1076-0

Редактор *А.К. Смирнова*
Корректор *В.К. Чупрова*
Компьютерная правка *О.В. Суховой*
Верстка макета *О.Г. Горюновой*
Дизайн обложки *Д.Л. Разумного*

Подписано в печать 20.07.2015 г. И-30. Формат 60×84/16.
Усл.-печ. л. 9,3. Уч.-изд. 9,00. Тираж 100 экз. Заказ 252

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Московский государственный строительный университет».
Издательство МИСИ – МГСУ.
129337, Москва, Ярославское ш., 26.
Тел. (495) 287-49-14, вн. 13-71, (499) 188-29-75, (499) 183-97-95.
E-mail: ric@mgsu.ru, rio@mgsu.ru.
Отпечатано в типографии Издательства МИСИ – МГСУ.
Тел. (499) 183-91-90, (499) 183-67-92, (499) 183-91-44

Глава 1. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Случайная величина и ее закон распределения

Анализ значений величин, полученных в результате эксперимента, производится исходя из основных понятий, теорем и методов теории вероятностей и математической статистики.

Теория вероятностей — раздел математики, в котором изучаются закономерности случайных явлений.

Случайная величина — переменная величина, численные значения которой зависят от результата опыта. Обычно случайные величины обозначают большими латинскими буквами X, Y, Z и т.п., а возможные значения случайных величин — x_i, y_i, z_i.

На практике, как правило, используют случайные величины двух типов — дискретные и непрерывные.

Случайная величина называется дискретной, если множество ее значений конечно или счетно (т.е. множество бесконечное, но элементы его можно пронумеровать). Дискретность (от латин. *discretus* — разделенный, прерывистый) — прерывность.

Пример: дискретная случайная величина X — число отказавших элементов в приборе.

Законом распределения дискретной случайной величины называется правило, по которому каждому возможному значению случайной величины ставится в соответствие вероятность, с которой случайная величина может принять это значение. Закон распределения дискретной случайной величины может быть задан:

- *таблицей* (*рядом распределения*). В первой строке таблицы находятся значения дискретной случайной величины, во второй — соответствующие вероятности. Данный ряд распределения может быть обозначен формулой

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

$$\sum_{i=1}^{n} P_i = 1;$$

- *графически* (*многоугольником распределения*). По оси OX откладываются возможные значения дискретной случайной величины, по оси OY — вероятности этих значений, и для наглядности полученные точки соединяются отрезками (рис. 1.1);

• *аналитически* (*функцией распределения*). Функция распределения случайной величины X — функция $F(x)$, равная для любого значения x вероятности того, что случайная величина X примет значение меньшее, чем x:

$$F(x) = P(X < x).$$

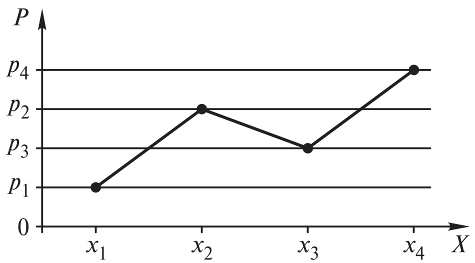

Рис. 1.1. Графическое обозначение дискретной случайной величины

$F(x)$ иногда называют интегральной функцией распределения, или интегральным законом распределения. Графиком $F(x)$ дискретной случайной величины является ступенчатая функция (рис. 1.2).

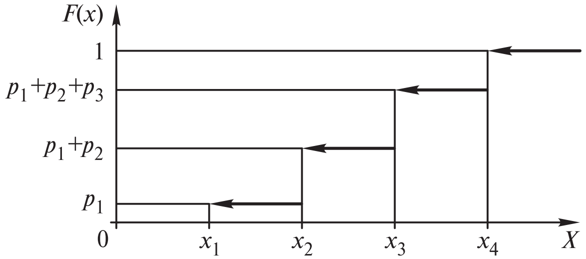

Рис. 1.2. Ступенчатая функция дискретной случайной величины

Случайная величина называется *непрерывной*, если множество ее значений целиком заполняет некоторый интервал. Например, непрерывная случайная величина Y — время безотказной работы прибора.

Непрерывная случайная величина может быть задана функцией распределения $F(x)$ или плотностью распределения $f(x)$.

Плотностью распределения непрерывной случайной величины X называется производная от функции распределения

$$f(x) = F'(x).$$

$f(x)$ иногда называется дифференциальной функцией распределения, или дифференциальным законом распределения. Кривая, являющаяся графиком плотности распределения, называется кривой распределения этой случайной величины (рис. 1.3).

При описании непрерывной случайной величины часто используют так называемые квантили. *Квантилем*, отвечающим заданной вероятности p, называют такое значение $x = x_p$, при котором функция распределения принимает значение, равное p (рис. 1.4), т.е.

$$F(x_p) = p.$$

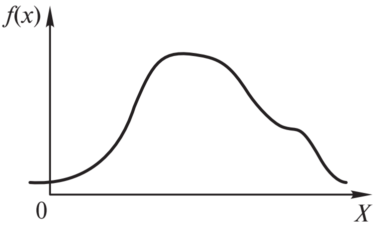

Рис. 1.3. Кривая распределения случайной величины

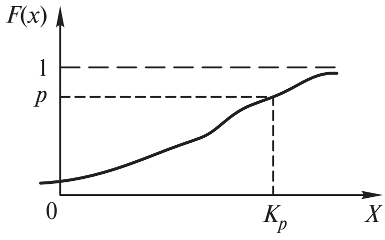

Рис. 1.4. Квантиль x с заданной вероятностью p

Некоторые квантили имеют особые названия. Например, квантиль, отвечающий значению $p = 0{,}5$, называют *медианой* распределения M_e. Медиана используется в качестве характеристики центра распределения. Квантили, соответствующие значениям $p = 0{,}25$ и $p = 0{,}75$, называют нижним и верхним квартилями (от латин. *quarta* — четверть). Зная значения достаточного числа квантилей, можно представить себе ход возрастания функции распределения.

Числовые характеристики *случайных величин* бывают двух видов:

• характеристики положения — математическое ожидание $M(X)$, мода M_0;

• характеристики рассеивания — дисперсия $D(X)$, среднеквадратическое отклонение $\sigma(X)$.

Математическое ожидание $M(X)$ для дискретной случайной величины X — это сумма произведений всех возможных значений дискретной случайной величины на вероятности этих значений:

$$M(X) = \sum_{i=1}^{n} x_i p_i.$$

Математическое ожидание $M(X)$ для непрерывной случайной величины X — это значение интеграла следующего вида:

$$M(X) = \int_{-\infty}^{+\infty} x f(x) dx.$$

Математическое ожидание можно воспринимать как некоторое значение, около которого группируются все возможные значения случайной величины.

Модой случайной величины M_0 называется наиболее вероятное значение этой случайной величины. Для дискретной случайной величины модой является то ее значение, у которого самая большая вероятность. Для непрерывной случайной величины модой является то ее значение, в котором плотность вероятности максимальная (рис. 1.5).

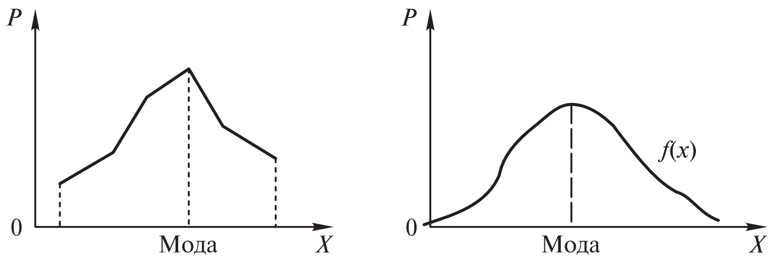

Рис. 1.5. Мода случайных величин

Дисперсией случайной величины $D(X)$ называют математическое ожидание квадрата отклонения этой случайной величины от ее математического ожидания:

$$D(X) = M\{[X - M(X)^2]\}.$$

Дисперсия (от латин. *dispersion* — рассеивание) — мера рассеивания, отклонения от среднего. На практике для вычисления дисперсии используют следующее ее свойство:

$$D(X) = M(X^2) - M^2(X),$$

т.е. для дискретной случайной величины

$$D(X) = \sum_{i=1}^{n} x_i^2 p_i - \left(\sum_{i=1}^{n} x_i p_i \right)^2,$$

а для непрерывной

$$D(X) = \int_{-\infty}^{+\infty} x^2 f(x) dx - \left(\int_{-\infty}^{+\infty} x f(x) dx \right)^2.$$

Среднеквадратическим отклонением случайной величины $\sigma(X)$ называют корень квадратный из дисперсии

$$\sigma(X) = \sqrt{D(X)}.$$

Коэффициент асимметрии определяется отношением

$$\gamma = \frac{M(X - M(X))^3}{(\sigma(X))^3}.$$

Если коэффициент асимметрии положителен, более «длинная» часть кривой плотности распределения лежит правее моды, что видно на рис. 1.6.

Если коэффициент асимметрии отрицателен, более «длинная» часть кривой плотности распределения лежит левее моды, что также видно на рис. 1.6.

Остановимся подробнее на наиболее часто встречающемся на практике законе распределения непрерывной случайной величины, который является предельным законом (т.е. к нему приближаются другие законы распределения при часто встречающихся типичных условиях), — нормальном законе распределения.

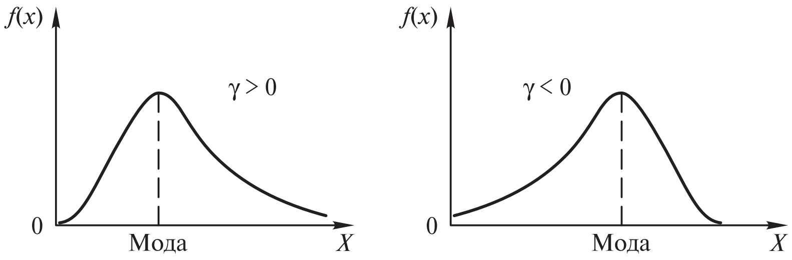

Рис. 1.6. Коэффициенты асимметрии для различных распределений

Непрерывная случайная величина называется распределенной по *нормальному закону* (или закону Гаусса), если ее плотность вероятности имеет следующий вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Кривая плотности распределения нормально распределенной случайной величины также называется *кривой Гаусса* (рис. 1.7).

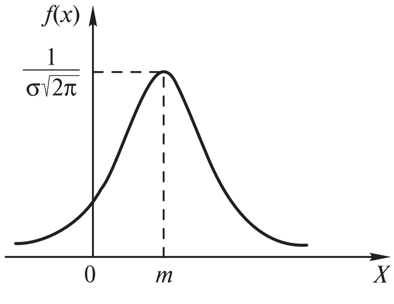

Рис. 1.7. Кривая Гаусса

Прямая $x = m$ является осью симметрии для кривой Гаусса. Параметр m не влияет на форму кривой, он определяет сдвиг по оси OX. Параметр σ определяет растяжение (или сжатие) кривой: чем больше значение σ, тем более пологая кривая.

Если $m = 0$, $\sigma = 1$, то кривая, как и соответствующее распределение, называется *нормированной*.

Для нормально распределенной случайной величины X математическое ожидание $M(X) = m$, а дисперсия $D(X) = \sigma^2$. Мода и медиана совпадают с математическим ожиданием, коэффициент асимметрии равен нулю.

Вероятность попадания нормально распределенной случайной величины X в интервал от α до β

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - m}{\sigma}\right) - \Phi\left(\frac{\alpha - m}{\sigma}\right),$$

где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} d$$

— функция Лапласа. Значения функции Лапласа для положительных значений аргумента x табулированы, а для отрицательных x значение $\Phi(x)$ находят их условия ее нечетности

$$\Phi(-x) = -\Phi(x).$$

При $x \geq 5$ полагают, что $\Phi(-x) \approx 0{,}5$.

На практике часто используют «*правило трех сигм*», которое позволяет указать интервал практически возможных значений нормально распределенной случайной величины: если случайная величина распределена нормально, то абсолютная величина ее отклонения от математического ожидания не превосходит утроенного среднеквадратического отклонения с вероятностью, близкой к единице:

$$P(|X - M(X)| < 3\sigma) = 0{,}997 \approx 1.$$

Более подробно с введенными понятиями можно ознакомиться в [1—3].

1.2. Простейшие приемы статистического описания

Математическая статистика — раздел математики, в котором занимаются разработкой методов получения, описания, обработки опытных данных с целью изучения закономерностей случайных массовых явлений.

Множество значений случайной величины, полученных в результате эксперимента, представляет собой статистическую совокупность. *Генеральной статистической совокупностью* называется совокупность, содержащая в себе все возможные значения случайной величины. *Выборочной статистической совокупностью* (*выборкой*) называется совокупность, содержащая в себе только некоторую часть элементов генеральной совокупности. Число опытов n, содержащихся в выборке, называется *объемом выборки*.

Выборка должна быть *репрезентативной*, т.е. объекты выборки должны правильно представлять признаки генеральной совокупности. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществлять случайно.

Пусть имеется набор экспериментальных данных (выборка) объемом n. *Вариационным рядом* называют упорядоченные по возрастанию числовых значений элементы выборки $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$. Величина $x_{(k)}$ называется *k-й порядковой статистикой*. Крайние члены $x_{(1)} = x_{\min}$ и $x_{(n)} = x_{\max}$ называются *экстремальными* (соответственно минимальным и максимальным) *значениями выборки*. Промежуток между экстремальными значениями называют *интервалом варьирования*. Разность $x_{(n)} - x_{(1)}$ называется *размахом выборки*.

Выборочной медианой (серединой) выборки является величина

$$\widetilde{Me} = \begin{cases} x_{(m+1)} \text{ при } n = 2m+1 \text{ (} n \text{ — нечетная величина);} \\ \dfrac{x_{(m)} + x_{(m+1)}}{2} \text{ при } n = 2m \text{ (} n \text{ — четная величина).} \end{cases}$$

Таким образом, выборочная медиана — это либо средний член вариационного ряда (если в выборке нечетное число данных), либо среднее арифметическое двух средних членов вариационного ряда (если в выборке четное число данных).

При статистическом анализе дискретной случайной величины используется простая таблица частот. Пусть выборка содержит $k (k \le n)$ различных значений и значение x_i встречается n_i раз, тогда величину n_i называют *частотой*, а значение x_i — *вариантой*. Сумма всех частот равна объему выборки

$$\sum_{i=1}^{k} n_i = n.$$

Множество пар (x_i, n_i), где для каждой варианты указана ее частота, называют *статистическим рядом*, который записывают в виде простой таблицы частот (табл. 1.1).

Таблица 1.1

Таблица частот

№	Значение	Частота
1	$x_{\min}$	n_1
...	...	...
m	$x_{\max}$	n_m

Если генеральная случайная величина — непрерывная, то таблица частот будет интервальной. Интервал варьирования разбивают на несколько, проводя группировку выборочных данных.

Число интервалов группировки (табл. 1.2) находится по эмпирической формуле (в которой округление идет до ближайшего целого):

$$k \approx 1 + 3{,}2 \cdot \lg n,$$

где n — объем выборки.

Длина интервала

$$h = \frac{x_{\max} - x_{\min}}{k},$$

где $x_{\max} - x_{\min}$ — размах выборки;

$x_{\max}$ — максимальное значение выборки;

$x_{\min}$ — минимальное значение выборки.

Таблица 1.2

Интервалы таблицы частот

№ интервала	Интервал	Частота
1	$(x_{\min}; x_{\max} + h)$	n_1
2	$(x_{\min} + h; x_{\max} + 2h)$	n_2
...	...	...
k	$(x_{\min} - h; x_{\max})$	n_k

Примечание. Частота n_k — количество данных, попавших в k-й интервал.

Полигоном частот называется ломаная, концы отрезков которой имеют координаты $\left(x_i; \dfrac{n_i}{n}\right)$. Для непрерывных величин полигон представляет собой ломаную, которая проходит через точки, абсциссы которых являются серединами интервалов группировки (рис. 1.8).

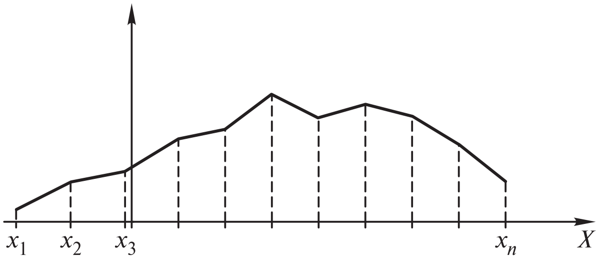

Рис. 1.8. Полигон частот

Эмпирическая функция распределения (рис. 1.9)

$$F_n(x) = \frac{v_n(x)}{n},$$

где $v_n(x)$ — число элементов выборки, оказавшихся меньше x. Функция $F_n(x)$ при большом числе наблюдений n близка в каждой точке x к теоретической функции распределения $F(x)$ наблюдаемой случайной величины X [о функции $F_n(x)$ говорят как о статистическом аналоге для $F(x)$].

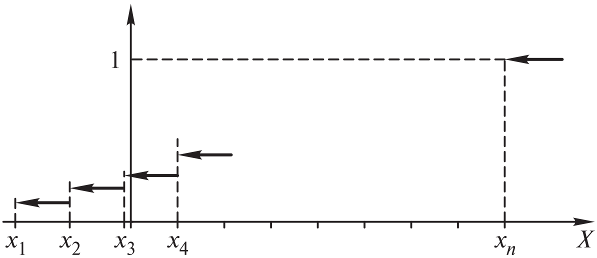

Рис. 1.9. Эмпирическая функция распределения

Гистограммой называется объединение прямоугольников, полученных следующим образом: по оси OX от $x_{\min}$ откладываются интервалы длиной d до $x_{\max}$, на каждом интервале строится прямоугольник с высотой $h_i = \dfrac{p_i}{d}$, где p_i — относительная частота попадания в i-й интервал группировки, $p_i = \frac{n_i}{n}$. При увеличении объема выборки и уменьшении длины интервала гистограмма будет приближаться к кривой плотности распределения (рис. 1.10).

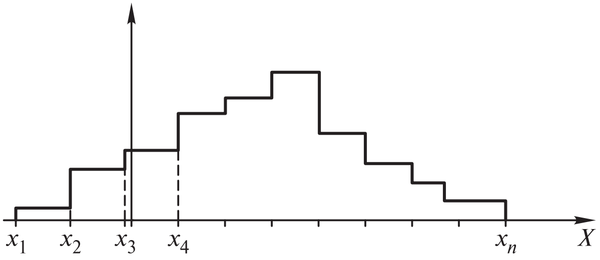

Рис. 1.10. Гистограмма функции распределения

Пример 1.1. В результате измерений активности цемента получено $n = 400$ значений, МПа. Значения активности цемента измеряются в диапазоне 28,5—55,5 МПа.

Требуется построить гистограмму активности цемента, соответствующую данному диапазону.

Решение

Рассчитываем значение числа интервалов по формуле

$$k = 1 + 3{,}2 \lg 400 \approx 9.$$

Определяем длину интервала:

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{55{,}5 - 28{,}5}{9} = 30.$$

Далее определяем границы интервалов и заносим их во второй столбец табл. 1.3.

Таблица 1.3

Таблица распределения

№ интервала	Граница интервала	Середина интервала	Число наблюдений в интервале n_i
1	2	3	4
1	28,5—31,5	30,0	10
2	31,5—34,5	33,0	22
3	34,5—37,5	36,0	42

Оглавление