7. Основные понятия математической статистики

Генеральная совокупность и выборка. Эмпирическая функция распределения и гистограмма. Точечные оценки параметров распределения: несмещенность, состоятельность, эффективность. Точечные оценки математического ожидания и дисперсии. Интервальные оценки. Доверительный интервал.

7.1. Предмет математической статистики

Математическая статистика — это наука, которая методами теории вероятностей на основании результатов наблюдений изучает закономерности в массовых случайных явлениях. Математическая статистика (не путать со статистикой — разделом экономической теории) указывает способы сбора и группировки статистических данных (результатов наблюдений), разрабатывает методы их обработки для оценки характеристик распределения, для установления зависимости случайной величины от других, для проверки статистических гипотез о виде распределения или значениях его параметров. Математическая статистика возникла и развивалась параллельно с теорией вероятностей.

7.2. Генеральная совокупность и выборка

Значительная часть математической статистики связана с необходимостью описать большую совокупность объектов. Ее называют генеральной. Если *генеральная совокупность* слишком многочисленна, или ее объекты труднодоступны, или имеются другие причины, не позволяющие изучить все объекты, прибегают к изучению какой-то части объектов. Эта выбранная для полного изучения часть называется *выборкой*. Необходимо, чтобы выборка наилучшим образом представляла генеральную совокупность, т.е. была *репрезентативной* (*представительной*). Если генеральная совокупность мала или совсем неизвестна, не удается предложить ничего лучшего, чем чисто случайный выбор.

Пример 7.1. *Пусть необходимо оценить качество изделий, выпускаемых определенным цехом машиностроительного предприятия. Для этого выбирают партию изделий и подвергают их контролю с целью дефектирования. Доля бракованных изделий для выбранной партии распространяется затем на всю продукцию цеха.*

Здесь генеральная совокупность — все изделия, выпускаемые цехом, выборка — отобранные для проверки изделия.

Пример 7.2. *Пусть необходимо оценить будущий урожай пшеницы. Для этого выбирают небольшой участок поля, например один квадратный метр, и подсчитывают число зерен во всех колосках и их массу. Приближенно весь урожай равен площади поля в метрах, умноженной на массу зерен, собранную с данного участка. Здесь генеральная совокупность — весь ожидаемый урожай, а выборка — урожай, собранный с одного квадратного метра. Если выбрать «плохой» участок (например, близко к краю поля), то оценка урожая будет заниженной. Если же участок имеет преимущества перед другими (например, лучше освещается солнцем), то оценка урожая будет завышенной.*

Пример 7.3. *Производится социологическое исследование с целью прогноза результатов предстоящих выборов мэра города.*

Здесь генеральная совокупность — все избиратели города, а выборка — число опрошенных респондентов.

Большое значение имеет способ, которым получена выборка. Ошибки при выборе способа отбора приводят к тому, что выборка становится нерепрезентативной.

Если в качестве респондентов взять, например, сто первых встречных с 10 до 12 часов дня, то социологи узнают мнение не всех слоев населения, а только домохозяек, направляющихся в это время за покупками.

Будем проводить испытания и в каждом из них фиксировать значения, которые приняла случайная величина ξ. В результате n испытаний получим выборку n значений, образующих простую статистическую совокупность наблюдений.

Определение 7.1. *Количество наблюдений n называется* ***объемом выборки***.

При большом числе наблюдений (сотни, тысячи) простая статистическая совокупность перестает быть удобной формой записи статистического материала — она становится слишком громадной. Для более экономичной записи наблюдаемые значения группируют.

Пусть в выборке значение x_1 наблюдалось m_1 раз, x_2 — m_2 раз, …, x_k — m_k раз и $\sum\limits_{i=1}^{k} m_i = n$ — объем выборки.

Определение 7.2. *Наблюдаемые значения x_i называют* ***вариантами****, а их последовательность, записанную в возрастающем порядке —* ***вариационным рядом****. Числа наблюдений $m_1, m_2, \ldots, m_k$ называют* ***частотами****.*

Разность $max(x_i) - min(x_i)$ называется ***размахом вариационного ряда****.*

Статистическим распределением выборки называют перечень вариант и соответствующих им частот — табл. 7.1.

Таблица 7.1

Статистическое распределение			
варианты x_i	x_1	$\ldots$	x_k
частоты m_i	m_1	$\ldots$	m_k

7.3. Эмпирическая функция распределения и гистограмма

С каждым испытанием, в котором наблюдается некоторая случайная величина ξ, можно связать случайное событие $\xi = x_i$, но иногда удобнее рассматривать событие $\xi < x_i$.

Определение 7.3. ***Эмпирической*** *(статистической)* ***функцией распределения*** *случайной величины ξ называется функция $F^*(x)$, которая при каждом x равна относительной частоте события $\xi < x$, т.е. отношению m_x — числа наблюдений меньших x к объему выборки n:*

$$F^*(x) = P^*(\xi < x) = \frac{m_x}{n}\ .$$

Пример 7.4. *Построить эмпирическую функцию распределения для данной выборки:*

Варианты x_i	*1*	*4*	*6*	*7*	*8*	*10*
Частоты m_i	*5*	*10*	*15*	*5*	*10*	*5*

◀Объем выборки n равен 5+10+15+5+10+5=50. Наименьшая варианта равна 1, следовательно $F^*(x) = 0$ при $x \leqslant 1$. Значение $x < 4$, а именно $x = 1$ наблюдалось 5 раз, следовательно $F^*(x) = \frac{5}{50} = 0{,}1$ при $1 < x \leqslant 4$. Значения $x < 6$, а именно $x = 1$ и $x = 4$, наблюдались 5+10=15 раз, следовательно $F^*(x) = \frac{15}{50} = 0{,}3$ при $4 < x \leqslant 6$. Аналогично получаем $F^*(x) = \frac{30}{50} = 0{,}6$ при

$6 < x \leqslant 7$ и т.д. Так как 10 — наибольшая варианта, $F^*(x) = 1$ при $x > 10$.

$$F^*(x) = \begin{cases} 0 & \text{при} \quad x \leqslant 1, \\ 0{,}1 & \text{при} \quad 1 < x \leqslant 4, \\ 0{,}3 & \text{при} \quad 4 < x \leqslant 6, \\ 0{,}6 & \text{при} \quad 6 < x \leqslant 7, \\ 0{,}7 & \text{при} \quad 7 < x \leqslant 8, \\ 0{,}9 & \text{при} \quad 8 < x \leqslant 10, \\ 1 & \text{при} \quad x > 10. \end{cases}$$

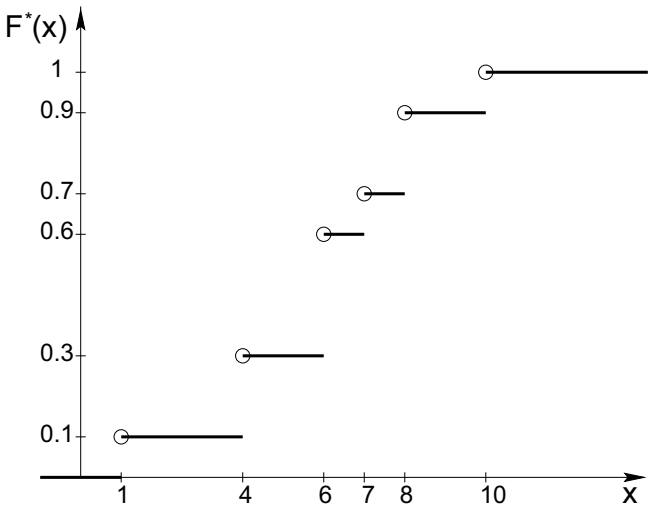

Рисунок 29. *Эмпирическая функция распределения*

График найденной функции представлен на рис. 29.
Из определения $F^*(x)$ вытекают ее свойства:

1) $0 \leqslant F^*(x) \leqslant 1$;

2) $F^*(x)$ — ступенчатая неубывающая функция;

3) если x_1 — наименьшая, а x_k — наибольшая варианты, то $F^*(x) = 0$ при $x \leqslant x_1$ и $F^*(x) = 1$ при $x > x_k$.

Гистограмма представляет выборку более наглядно. Для построения гистограммы разделим весь диапазон наблюдений на s интервалов вида $(a_{j-1}; a_j]$ и определим количество наблюдений m_j, попавших в j-й интервал. Относительная частота наблюдений, попавших в j-й интервал, равна $P_j^* = \dfrac{m_j}{n}$ $(m_1 + \ldots + m_s = n)$, сумма всех относительных частот, очевидно, равна единице. Для построения гистограммы по оси ординат откладываются значения

$\frac{P_j^*}{\Delta a_j} = \frac{m_j}{n \cdot (a_j - a_{j-1})}$. Полученная фигура, состоящая из прямоугольников, называется гистограммой относительных частот. Площадь каждого прямоугольника равна относительной частоте наблюдений, попавших в данный интервал. Для данных примера 7.4 получаются следующие значения:

N п/п	a_{j-1}	a_j	m_j	$P_j^* = \frac{m_j}{n}$	$\frac{P_j^*}{\Delta a_j}$
1	0	3	5	0.1	1/30
2	3	6	25	0.5	5/30
3	6	9	15	0.3	3/30
4	9	12	5	0.1	1/30

Получившаяся гистограмма представлена на рис. 30.

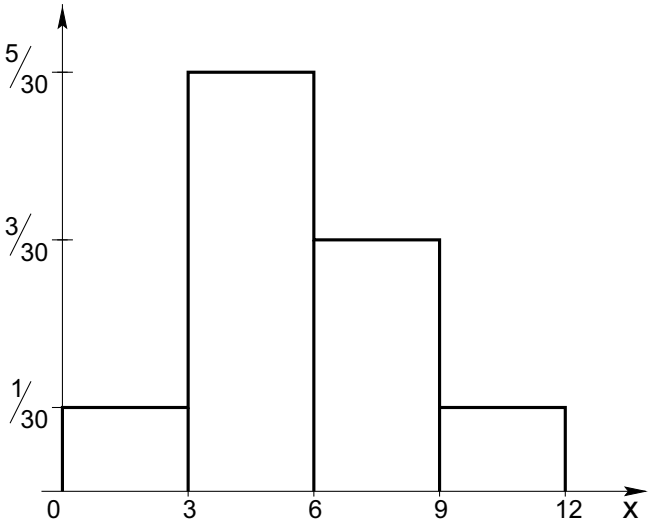

Рисунок 30. *Гистограмма относительных частот*

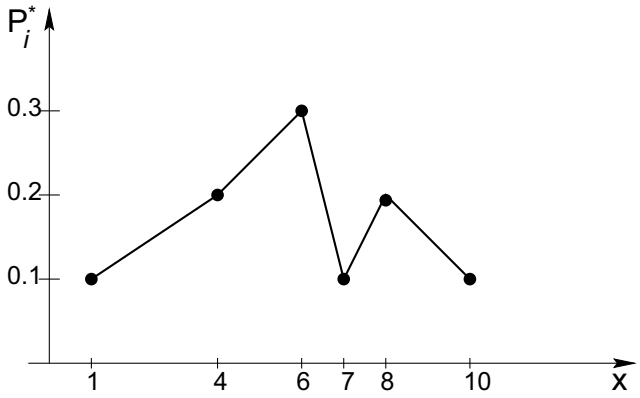

Рисунок 31. *Полигон относительных частот*

Другим наглядным способом представления распределения является полигон относительных частот. Для его построения по оси абсцисс откладываются варианты, а по оси ординат — относительные частоты (рис. 31), и полученные точки соединяются ломаной линией.

Для выборки из генеральной совокупности значений непрерывной случайной величины гистограмма является статистическим аналогом плотности распределения, а для дискретной случайной величины полигон относительных частот является статистическим аналогом многоугольника вероятностей. При увеличении объема выборки эти статистические характеристики в определенном смысле приближаются к своим теоретическим аналогам.

Замечание 7.1. *Наряду с гистограммой и полигоном относительных частот иногда рассматривают соответственно гистограмму и полигон частот, отличающиеся масштабом по оси ординат — все значения по оси ординат умножаются на n — объем выборки. Понятно, что форму получаемых фигур это не изменяет.*

7.4. Числовые характеристики статистического распределения

Статистическая функция распределения и гистограмма являются полными характеристиками результатов наблюдения случайной величины в данной серии испытаний. Однако иногда целесообразно ограничиться более простой, хотя и неполной характеристикой распределения.

Простейшей характеристикой распределения является *выборочное среднее*, которое для простой статистической совокупности вычисляется по формуле:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{7.1}$$

Если данные сгруппированы, то:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot x_i. \tag{7.2}$$

Иными словами, выборочное среднее представляет собой среднее взвешенное значение, причем веса равны соответствующим частотам.

Для характеристики разброса значений случайной величины относительно ее среднего значения используется *выборочная дисперсия*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{(x - \bar{x})^2} \tag{7.3}$$

для простой совокупности и

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} m_i(x_i - \bar{x})^2 \tag{7.4}$$

для сгруппированного распределения (распределения представленного в виде Таблица 7.1).

Очевидно, выборочная дисперсия имеет ту же размерность, что и квадрат случайной величины. Практически удобно пользоваться величиной, имеющей ту же размерность, что и данная случайная величина.

Для этого достаточно из дисперсии извлечь квадратный корень.

Эта величина

$$S = \sqrt{S^2} \tag{7.5}$$

называется *выборочным средним квадратическим отклонением* (СКО).

На практике вместо формулы (7.3) бывает удобнее применять другую:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \overline{x^2} - \overline{x}^2 \tag{7.6}$$

для простой совокупности и

$$S^2 = \frac{1}{n}\sum_{i=1}^{k} m_i x_i^2 - (\bar{x})^2 \tag{7.7}$$

для сгруппированного распределения.

Докажем формулу (7.6):

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 = \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \overline{x}^2.$$

Определение 7.4. ***Модой*** *статистического распределения (обозначается M_O) называется значение, которое наиболее часто встречается в исследуемой выборке.*

Для простой статистической совокупности мода вычисляется простым подсчетом. Например, для вариационного ряда $\{2, 2, 4, 5, 5, 5, 5, 6, 6, 7\}$, $M_O = 5$, т.к. значение 5 встречается чаще других.

Для сгруппированной выборки значение для моды необходимо аппроксимировать. Формула будет приведена ниже (7.8).

Статистические распределения, которые имеют несколько наиболее часто встречающихся значений, называются ***мультимодальными*** или ***полимодальными***.

Например, для вариационного ряда: $\{1, 2, 2, 3, 4, 5, 5, 6, 6, 7\}$, модами будут три значения $M_O = \{2, 5, 6\}$.

Определение 7.5. ***Медианой*** *(M_e) называется значение, которое разбивает выборку на две равные части. Половина наблюдений лежит ниже (левее) медианы, а другая половина выше (правее) медианы.*

Для простой статистической совокупности медиана вычисляется следующим образом. Исследуемая выборка $\{x_i\}$ сортируется в порядке не убывания значений элементов. Далее, если объем выборки – нечетное число, то $M_e = x_{(n+1)/2}$, иначе $M_e = \big(x_{n/2} + x_{n/2+1}\big)/2$.

Например, для вариационного ряда $\{1, 3, 5, 7, 9, 9, 12\}$ медиана равна четвертому элементу $M_e = 7$, а для вариационного ряда $\{1, 3, 5, 7, 9, 12\}$ медиана равна среднему арифметическому третьего и четвертого элементов $M_e = (5 + 7)/2 = 6$.

Пример 7.5. *Выборка задана в виде распределения частот:*

x_i	*1*	*5*	*9*	*12*	*16*	*18*
m_i	*5*	*10*	*30*	*25*	*22*	*8*

Найти моду, медиану, выборочное среднее и выборочную дисперсию.

◀Здесь объем выборки

$$n = 5 + 10 + 30 + 25 + 22 + 8 = 100.$$

Мода равна 9, так значение $x_3 = 9$ встречается максимальное число раз – $m_3 = 30$. Медиана равна 12, так как, если расположить все элементы выборки в виде неубывающий последовательности ее значений, то 50 и 51 значение будут равны 12.

Найдем выборочное среднее $\bar{x}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot x_i = 0{,}01 \cdot \big(5 \cdot 1 + 10 \cdot 5 + 30 \cdot 9 + 25 \cdot 12 + 22 \cdot 16 + 8 \cdot 18\big) = 11{,}21.$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot x_i^2 = 0{,}01 \cdot \big(5 \cdot 1^2 + 10 \cdot 5^2 + 30 \cdot 9^2 + 25 \cdot 12^2 + 22 \cdot 16^2 + 8 \cdot 18^2\big) = 145{,}09.$$

Выборочная дисперсия:

$$S^2 = \overline{x^2} - (\bar{x})^2 = 145{,}09 - 11{,}21^2 = 19{,}426.$$

►

Для выборки, представленной в виде сгруппированного по интервалам распределения, значение моды аппроксимируется в некоторую точку модального интервала (внутри которого находится максимальное значение):

$$M_0 = X_0 - h\frac{f_{mo} - f_{mo-1}}{f_{m0-1} - 2f_{mo} + f_{mo+1}}, \tag{7.8}$$

где X_0 — нижнее значение модального интервала; $f_{mo}, f_{mo-1}, f_{mo+1}$ — значение частот в модальном, предыдущем и следующем интервалах, соответственно; h — размах модального интервала.

Для выборки, представленной в виде сгруппированного по интервалам распределения, значение медианы аппроксимируется в некоторую точку M_e медианного интервала по формуле:

$$M_e = X_0 + h\frac{0{,}5\sum\limits_{k=1}^{s} f_k - \sum\limits_{k=1}^{m_e-1} f_k}{f_{me}}, \tag{7.9}$$

где X_0 — нижняя граница интервала, в котором находится медиана (медианный интервал); f_{me} — значение частоты в медианном интервале; h — размах медианного интервала.

Пример 7.6. *Проведена выборка из* $n = 100$ *значений* $\{x_1, x_2, \ldots, x_n\}$ *некоторой непрерывной случайной величины* ξ. *Для удобства обработки результатов, полученные значения, заключенные на отрезке* $[0; 30]$, *разбили на 6 интервалов постоянной длины и поместили в таблицу.*

i	1	2	3	4	5	16
$(a_{i-1}; a_i]$	[0; 5]	(5; 10]	(10; 15]	(15; 20]	(20; 25]	(25; 30]
m_i	5	10	30	25	22	8

По полученной выборке, заключенной в таблицу, аппроксимировать моду, медиану, выборочное среднее, выборочное среднеквадратическое отклонение (СКО) и функцию распределения $F(x)$ *непрерывной случайной величины* ξ.

◄ Аппроксимируем значение моды случайной величины ξ. Здесь третий интервал является модальным $X_0 = 10, f_{mo} = 30, f_{mo-1} = 10, f_{mo+1} = 25$, $h = 5$.

$$M_0 = 10 - 5 \cdot \frac{30 - 10}{10 - 2 \cdot 30 + 25} = 14.$$

Аппроксимируем значение медианы случайной величины ξ.
Здесь четвертый интервал – медианный $X_0 = 15, f_{me} = 25, h = 5$.

$$M_e = 15 + 5 \cdot \frac{50 - (5 + 10 + 30)}{25} = 16.$$

Для вычисления выборочных характеристик значения частот группируются в центре интервалов, значение которых равны:
$D = \{2{,}5; 7{,}5; 12{,}5; 17{,}5; 22{,}5; 27{,}5\}$.

Находим выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i d_i =$$

$$= \frac{1}{100}(5 \cdot 2{,}5 + 10 \cdot 7{,}5 + 30 \cdot 12{,}5 + 25 \cdot 17{,}5 + 22 \cdot 22{,}5 + 8 \cdot 27{,}5) = 16{,}15.$$

Выборочную дисперсию находим по формуле: $S^2 = \overline{x^2} - (\bar{x})^2$: $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} m_i d_i^2 =$

$$= \frac{1}{100}(5 \cdot 2{,}5^2 + 10 \cdot 7{,}5^2 + 30 \cdot 12{,}5^2 + 25 \cdot 17{,}5^2 + 22 \cdot 22{,}5^2 + 8 \cdot 27{,}5^2) = 301{,}25.$$

Выборочное СКО:

$$S^2 = \overline{x^2} - (\bar{x})^2 = 301{,}25 - 16{,}15^2 = 40{,}4275.$$

$$S = \sqrt{S^2} = \sqrt{40{,}4275} \approx 6{,}358.$$

Для аппроксимации функции распределения будем считать, что случайная величина на каждом i-том интервале распределена равномерно. Следовательно, на каждом интервале функция распределения является линейной. Находим массив относительных частот:
$P_i^* = m_i/n = \{0{,}05; 0{,}1; 0{,}3; 0{,}25; 0{,}22; 0{,}08\}$ и массив накопленных относительных частот: $Q_i = \sum_{k=1}^{i} P_k^* = \{0; 0{,}05; 0{,}15; 0{,}45; 0{,}7; 0{,}92; 1\}, i = 0, 1, \ldots, 6$.

По точкам с координатами: $M(a_i; Q_i), i = 0, 1, \ldots, 6$, проводим кусочно-линейную кривую. Уравнение эмпирической функции распределения имеет вид

$$F^*(x) = \begin{cases} 0 & \text{при} \quad x \leqslant 0, \\ 0{,}01x & \text{при} \quad x \in (0, 5], \\ 0{,}05 + 0,02(x-5) & \text{при} \quad x \in (5, 10], \\ 0{,}15 + 0,06(x-10) & \text{при} \quad x \in (10, 15], \\ 0{,}45 + 0,05(x-15) & \text{при} \quad x \in (15, 20], \\ 0{,}7 + 0,044(x-20) & \text{при} \quad x \in (20, 25], \\ 0{,}92 + 0,016(x-25) & \text{при} \quad x \in (25, 30], \\ 1 & \text{при} \quad x > 30. \blacktriangleright \end{cases}$$

7.5. Точечные оценки параметров распределения

Выборочное среднее, выборочная дисперсия и СКО являются примерами точечных оценок параметров распределения.

Определение 7.6. ***Точечной оценкой*** $\widetilde{a}_n$ *неизвестного параметра* a *распределения случайной величины* ξ *называется функция от наблюдений:*

$$\widetilde{a}_n = \widetilde{a}(x_1, \ \ldots, x_n).$$

Для изучения свойств этой оценки ее рассматривают как функцию от n независимых случайных величин $\xi_1, \ldots, \xi_n$, имеющих такое же распределение, что и ξ; $x_1, \ldots, x_n$ в этом случае рассматриваются как наблюдения над этими случайными величинами: x_i — полученное значение ξ_i, $i = 1, 2, \ldots, n$. Сама оценка $\widetilde{a}_n$ в этом случае является случайной величиной.

Перечислим свойства точечной оценки $\widetilde{a}_n$, которые могут считаться «хорошими».

Определение 7.7. *Оценка* $\widetilde{a}_n$ *называется* ***состоятельной****, если при* $n \to \infty$ *она сходится по вероятности к оцениваемому параметру* a*:*

$$\lim_{n\to\infty} P(|\widetilde{a}_n - a| < \varepsilon) = 1 \ \textit{для} \ \forall \ \varepsilon > 0.$$

Определение 7.8. *Оценка* $\widetilde{a}_n$ *называется* ***несмещенной****, если ее математическое ожидание равно оцениваемому параметру* a*:*

$$M(\widetilde{a}_n) = a.$$

Иногда точечные оценки обладают более слабым свойством: их смещение $M(\widetilde{a}_n) - a$ стремится к нулю при $n \to \infty$. Такие оценки называются асимптотически несмещенными.

Определение 7.9. *Несмещенная оценка $\widetilde{a}_n$ называется* ***эффективной****, если ее дисперсия – наименьшая, по сравнению с другими несмещенными оценками.*

На практике оценка не всегда удовлетворяет всем этим требованиям одновременно.

Пример 7.7. *Доказать, что выборочное среднее $\bar{x}$ является несмещенной и состоятельной оценкой для математического ожидания (генерального среднего) случайной величины.*

◀Обозначим $M(\xi) = a,\ D(\xi) = \sigma^2$. Рассматривая $\bar{x}$ как случайную величину, найдем ее математическое ожидание. При этом, как было отмечено ранее, считаем

$M(\xi_1) = \ldots = M(\xi_n) = a, \quad D(\xi_1) = \ldots = D(\xi_n) = \sigma^2.$

$$M(\xi) = M\left(\frac{\sum\limits_{i=1}^{n}\xi_i}{n}\right) = \frac{\sum\limits_{i=1}^{n}M(\xi_i)}{n} = \frac{na}{n} = a.$$

Несмещенность выборочного среднего доказана. Оценим теперь дисперсию выборочного среднего:

$$D(\xi) = D\left(\frac{\sum\limits_{i=1}^{n}\xi_i}{n}\right) = \frac{\sum\limits_{i=1}^{n}D(\xi_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

В соответствии с неравенством Чебышева (теорема 5.3), получаем $\forall\,\varepsilon > 0$:

$$1 \geqslant P\big(\big|\xi - M(\xi)\big| < \varepsilon\big) \geqslant 1 - \frac{\sigma^2/n}{\varepsilon^2}.$$

Заменяя $M(\xi) = a$ и переходя к пределу при $n \to \infty$, получаем

$$1 \geqslant \lim_{n\to\infty} P\big(\big|\xi - a\big| < \varepsilon\big) \geqslant 1,$$

откуда получаем:

$$\lim_{n\to\infty} P\big(\big|\xi - a\big| < \varepsilon\big) = 1.$$

Это равенство и означает состоятельность оценки $\bar{x}$.

Замечание 7.2. *Можно доказать, что выборочное среднее будет эффективной оценкой математического ожидания в случае, когда случайная величина имеет нормальное распределение.*

Аналогично доказывается, что выборочная дисперсия S^2 является состоятельной и смещенной оценкой дисперсии σ^2:

$$M(S^2) = \frac{n-1}{n}\,\sigma^2. \tag{7.10}$$

Примем это без доказательства.

При малых объемах выборки n для оценки дисперсии σ^2 используют исправленную выборочную дисперсию S^{*^2}:

$$S^{*^2} = \frac{n}{n-1}\,S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{7.11}$$

Оценка S^{*^2} является несмещенной, состоятельной оценкой дисперсии σ^2.

Формула (7.11) позволяет вычислять S^{*^2} для простой совокупности. Для сгруппированных данных используют аналогичную формулу (7.12):

$$S^{*^2} = \frac{1}{n-1}\sum_{i=1}^{k} m_i(x_i - \bar{x})^2. \tag{7.12}$$

Замечание 7.3. *Исправленное СКО S^* является смещенной оценкой СКО S.*

7.6. Распределения, используемые в статистике

Познакомимся с некоторыми непрерывными распределениями, которые применяются в математической статистике.

Распределение χ^2 (хи-квадрат).

Пусть имеется n независимых стандартных нормальных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, $\xi_i \sim N(0;1)$, $i = 1, \ldots, n$.

Определение 7.10. *Распределение случайной величины $\chi^2_n = \sum_{i=1}^{n} \xi_i^2$ называется χ^2—**распределением** с n степенями свободы.*

Очевидно, что случайная величина $\chi^2_n \geqslant 0$.

Плотность этого распределения имеет вид:

$$f(x) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}}e^{-\frac{x}{2}} & \text{при} \quad x \geqslant 0, \\ 0 & \text{при} \quad x < 0. \end{cases}$$

Здесь $\Gamma(x) = \int\limits_0^\infty t^{x-1} \cdot e^{-t}dt$ — гамма функция, являющаяся обобщением понятия факториала: $\Gamma(x) = (x-1)!$ при $x \geqslant 1$.

Замечание 7.4. *Если случайные величины* $\xi_1, \ \ldots, \xi_n$ *связаны какой-нибудь зависимостью, например* $\xi_1 + \ldots + \xi_n = n \cdot \overline{x}$, *то число степеней свободы уменьшается, случайная величина* $\sum\limits_{i=1}^{n} \xi_i^2$ *будет иметь распределение* χ^2_{n-1}.

Распределение Стьюдента.

Пусть имеется $n+1$ независимая нормальная стандартная случайная величина $\zeta, \xi_1, \ \ldots, \xi_n$.

Определение 7.11. *Распределение случайной величины*

$$t = \frac{\zeta}{\sqrt{\dfrac{\chi^2_n}{n}}}$$

называется ***распределением Стьюдента*** *с* n *степенями свободы.*

Плотность этого распределения имеет вид:

$$f_{st}(x) = \frac{\left(1 + \frac{x^2}{2}\right)^{\frac{n+1}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}.$$

Поскольку распределение симметрично относительно нуля (плотность — четная функция), математическое ожидание равно нулю.

Стьюдент — псевдоним английского статистика Госсета.

F—Распределение Фишера—Снедекора.

Пусть имеется $n+k$ независимых нормальных стандартных величин: $\xi_1, \ldots, \xi_n$; $\zeta_1, \ldots, \zeta_k$; $\xi_i \sim N(0;1), \ i = 1, \ \ldots, n$; $\zeta_i \sim N(0;1), \ j = 1, \ldots, k$.

Определение 7.12. *Распределение случайной величины*

$$F_{n,k} = \frac{\dfrac{\chi_n^2}{n}}{\dfrac{\chi_k^2}{k}}$$

называется ***F—распределением Фишера—Снедекора*** *(распределением Фишера или F—распределением) с* n, k *степенями свободы.*

Очевидно, что случайная величина $F_{n,k} \geqslant 0$.

Плотность этого распределения имеет вид:

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{m+k}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right) \cdot \Gamma\left(\dfrac{k}{2}\right)} \left(\dfrac{n}{k}\right)^{\frac{n}{2}} \cdot x^{\frac{n}{2}-1} \left(1 + \dfrac{n}{k}x\right)^{-\frac{n+k}{2}} & \text{при} \quad x \geqslant 0, \\ 0 & \text{при} \quad x < 0. \end{cases}$$

Для всех этих распределений имеются таблицы значений плотности и функции распределения; их можно также вычислить с помощью прикладных программ на ЭВМ (таких, как Excel, Mathcad, Maxima и проч.).

7.7. Интервальные оценки параметров распределения

Наряду с рассмотренными точечными оценками, определяемыми одним числом, используют интервальные оценки неизвестных параметров, определяемые двумя числами — концами интервала, дающими вероятностную оценку сверху и снизу неизвестного параметра распределения.

Интервальные оценки целесообразно применять при малом объеме выборки, когда дисперсия точечной оценки велика и она может сильно отличаться от оцениваемого параметра.

Определение 7.13. ***Доверительным интервалом*** *для несмещенного параметра* a *называют интервал* $(a_1; a_2)$ *со случайными границами, зависящими от наблюдений:* $a_1 = a_1(x_1, \ldots, x_n)$, $a_2 = a_2(x_1, \ldots, x_n)$, *накрывающий неизвестный параметр с заданной вероятностью* γ*:* $P(a \in (a_1; a_2)) = \gamma$. ***Вероятность*** γ *называется* ***доверительной вероятностью*** *или* ***надежностью доверительного интервала****.*

Обычно γ задают равным 0,95; 0,99 и более.

Доверительный интервал $\mathbf{I}_\gamma$ *для неизвестного математического ожидания нормального распределения при известной дисперсии имеет вид:*

$$\mathbf{I}_\gamma = \Big(\bar{x} - \tau_{\gamma/2}\frac{\sigma}{\sqrt{n}};\ \bar{x} + \tau_{\gamma/2}\frac{\sigma}{\sqrt{n}}\Big), \tag{7.13}$$

где величина $\tau_{\gamma/2}$ определяется из уравнения:

$$\Phi(\tau_{\gamma/2}) = \frac{\gamma}{2} \tag{7.14}$$

по таблицам функции Лапласа или с помощью компьютера, а $\bar{x}$ — выборочное среднее.

Замечание 7.5. *При возрастании объема выборки* n*, как видно из (7.13), доверительный интервал уменьшается. При увеличении надежности* γ *увеличивается величина* $\tau_{\gamma/}$*, т.к. функция Лапласа в (7.14) возрастающая; следовательно, увеличивается и доверительный интервал (7.13).*

Для получения доверительного интервала (7.13) заметим, что если независимые случайные величины $\xi_i \sim N(a;\sigma)$, $i = 1, \ldots, n$, то среднее арифметическое $\bar{\xi} = (\xi_1 + \ldots \xi_n)/n$ тоже распределено нормально с параметрами:

$$M(\bar{\xi}) = a, \quad \sigma(\bar{\xi}) = \frac{\sigma}{\sqrt{n}}. \tag{7.15}$$

Формулы (7.15) были получены в примере 7.7. Будем искать доверительный интервал для a в виде:

$$P(|\bar{\xi} - a| < \varepsilon) = \gamma, \tag{7.16}$$

где γ — заданная доверительная вероятность. Для определения ε воспользуемся формулой (5.8), которая в данном случае с учетом (7.15) принимает вид:

$$P(|\bar{\xi} - a| < \varepsilon) = 2\Phi\Big(\frac{\varepsilon}{\sigma/\sqrt{n}}\Big).$$

Найдем ε из уравнения:

$$2\Phi\Big(\frac{\varepsilon}{\sigma/\sqrt{n}}\Big) = \gamma \implies \Phi\Big(\frac{\varepsilon}{\sigma/\sqrt{n}}\Big) = \frac{\gamma}{2} \implies \frac{\varepsilon}{\sigma/\sqrt{n}} = \tau_{\frac{\gamma}{2}} \implies$$
$$\varepsilon = \tau_{\frac{\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

С учетом полученной величины ε доверительный интервал (7.16) принимает вид (7.13).

Пример 7.8. *Найти доверительный интервал* $\mathbf{I}_\gamma$ *для неизвестного математического ожидания* a *нормально распределенной случайной величины со средним квадратическим отклонением* $\sigma = 2$ *по выборке объема* $n = 64$ *с выборочным средним* $\bar{x} = 5{,}2$. *Надежность доверительного интервала* $\gamma = 0{,}95$.

◀ Из уравнения (7.14) по таблице приложения 2 находим для $\frac{\gamma}{2} = 0{,}475$ $\tau_{\frac{\gamma}{2}} = 1{,}96$. Подставляя найденное значение в (7.13), получаем $\mathbf{I}_\gamma = (4{,}71; 5{,}69)$.

Ответ: $\mathbf{I}_\gamma = (4{,}71; 5{,}69)$.

Доверительный интервал $\mathbf{I}_\gamma$ *для неизвестного математического ожидания нормального распределения при неизвестной дисперсии имеет вид:*

$$\mathbf{I}_\gamma = \left(\bar{x} - t_\gamma \frac{S^*}{\sqrt{n}};\ \bar{x} + t_\gamma \frac{S^*}{\sqrt{n}}\right), \tag{7.17}$$

где величина t_γ определяется по таблице приложения 3 критических точек распределения Стьюдента для $\alpha = 1 - \gamma$ и $k = n - 1$ или с помощью компьютера из уравнения для функции распределения Стьюдента $F_{st}(x)$ с $n - 1$ степенью свободы:

$$F_{st}(t_\gamma) = \frac{1 + \gamma}{2}, \tag{7.18}$$

где $\bar{x}$ и S^* — соответственно выборочное среднее и исправленное СКО.

Для получения доверительного интервала (7.17) примем без доказательства, что если независимые случайные величины $\xi_i \sim N(a; \sigma)$, $i = 1, \ldots, n$, то случайная величина

$$t = \frac{\bar{\xi} - a}{S^*/\sqrt{n}} \tag{7.19}$$

имеет распределение Стьюдента с $n - 1$ степенью свободы (см п. 7.6).

Обозначим t_γ значение, при котором с вероятностью γ выполняется следующее неравенство:

$$P(|t| < t_\gamma) = \gamma. \tag{7.20}$$

С учетом четности плотности распределения Стьюдента $f_{st}(t)$ значение t_γ определяется из условия:

$$P(|t| < t_\gamma) = \gamma \iff P(|t| > t_\gamma) = 1 - \gamma \implies P(t > t_\gamma) = \frac{1 - \gamma}{2} \iff$$
$$\iff 1 - F_{st}(t_\gamma) = \frac{1 - \gamma}{2} \iff F_{st}(t_\gamma) = \frac{1 + \gamma}{2}.$$

Подставляя в (7.20) выражение (7.19), получаем:

$$P\left\{\left|\frac{\bar{\xi}-a}{S^*/\sqrt{n}}\right|<t_\gamma\right\}=\gamma \iff P\left\{-t_\gamma<\frac{\bar{\xi}-a}{S^*/\sqrt{n}}<t_\gamma\right\}=\gamma,$$

откуда получаем для a доверительный интервал в виде (7.17).

Замечание 7.6. *В некоторых пакетах прикладных программ для ЭВМ, например в Excel, под распределением Стьюдента понимается* $1-F_{st}(x)$. *Поэтому, задавая значение* $1-\gamma$ *и число свободы, с помощью обратной функции можно сразу получить значение* t_γ *для двустороннего интервала (без использования (7.18)). Указанные особенности можно узнать из инструкций к программам.*

Пример 7.9. *Найти доверительный интервал* $\mathbf{I}_\gamma$ *для неизвестного математического ожидания* a *нормально распределенной случайной величины с выборочным средним* $\bar{x}=10{,}5$ *и исправленным СКО* $S^*=1{,}6$ *по выборке объема* $n=16$. *Надежность доверительного интервала* $\gamma=0{,}99$.

◀ По таблице приложения 3 для числа степеней свободы $k=n-1=15$ и $\alpha=1-\gamma=0{,}01$ находим $t_\gamma=2{,}95$. Подставляя полученное значение в (7.17), получаем значение для радиуса доверительного интервала ε:

$$\varepsilon=t_\gamma\frac{S^*}{\sqrt{n}}=2{,}95\frac{1{,}6}{\sqrt{16}}=2{,}95\cdot 0{,}4=1{,}18.$$

Находим доверительный интервал

$\mathbf{I}_\gamma=(10{,}5-1{,}18;10{,}5+1{,}18)=(9{,}32;11{,}68)$.

Ответ: $\mathbf{I}_\gamma=(9{,}32;11{,}68)$.

7.8. Выборочный коэффициент корреляции

Рассмотрим выборку объема n из генеральной совокупности значений двумерной случайной величины $(\xi;\eta)$, т.е. n пар наблюдений $(x_i;y_i)$. Поскольку многие значения в этой выборке могут повторяться, их заносят в так называемую корреляционную таблицу (табл. 7.4).

Таблица 7.4

Корреляционная таблица					
$\xi\backslash\eta$	y_1	y_2	$\dots$	y_s	n_{i*}
x_1	n_{11}	n_{12}	$\dots$	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	$\dots$	n_{2s}	n_{2*}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{ks}	n_{k*}
n_{*j}	n_{*1}	n_{*2}	$\dots$	n_{*s}	n

В первом столбце этой таблицы перечислены значения x_i, во второй строке — y_i. Данные представлены в виде вариационных рядов. На пересечении i-й строки и j-го столбца — соответствующая частота n_{ij}, т.е. количество раз, которое наблюдение $(x_i;y_j)$ встретилось в выборке. При обработке корреляционной таблицы в последнем столбце указывают сумму частот по строкам $n_{i*}=\sum\limits_{j=1}^{s} n_{ij}$, а в последней строке — сумму частот по столбцам $n_{*j}=\sum\limits_{i=1}^{k} n_{ij}$. Сумма всех элементов последнего столбца или строки даст объем выборки

$$n=\sum_{i=1}^{k}\sum_{j=1}^{s} n_{ij}=\sum_{i=1}^{k} n_{i*}=\sum_{j=1}^{s} n_{*j}.$$

Первый и последний столбцы корреляционной таблицы образуют статистическое распределение выборки случайной величины ξ, а первая и последняя строки образуют выборку случайной величины η. Обработав их, как описано в предыдущей лекции, получим числовые характеристики

$$\overline{x}=\frac{\sum\limits_{i=1}^{k} n_{i*}x_i}{n},\quad \overline{x^2}=\frac{\sum\limits_{i=1}^{k} n_{i*}x_i^2}{n},\quad S_x^2=\overline{x^2}-\overline{x}^2,$$

$$\overline{y} = \frac{\sum\limits_{j=1}^{s} n_{*j} y_j}{n}, \quad \overline{y^2} = \frac{\sum\limits_{j=1}^{s} n_{*j} y_j^2}{n}, \quad S_y^2 = \overline{y^2} - \overline{y}^2.$$

Определение 7.14. ***Выборочным коэффициентом корреляции*** r^*_{xy} *называется:*

$$r^*_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y}, \quad \text{где} \tag{7.21}$$

$$\overline{xy} = \frac{\sum\limits_{i=1}^{k} \sum\limits_{j=1}^{s} n_{ij} x_i y_j}{n}. \tag{7.22}$$

Выборочный коэффициент корреляции является статистической оценкой коэффициента корреляции, рассмотренного в п.6.5, и он обладает следующими свойствами, которые мы приведем без доказательства:

1) $r^*_{xy} = r^*_{yx}$;

2) Выборочный коэффициент корреляции находится в пределах от -1 до 1: $-1 \leqslant r^*_{xy} \leqslant 1$;

3) $|r^*_{xy}| = 1$ тогда и только тогда, когда между значениями x_i и y_i имеется линейная зависимость. Чем ближе r^*_{xy} к нулю, тем хуже эта зависимость аппроксимируется линейной.

Определение 7.15. *Условным средним* $\overline{y_x}$ *называют среднее арифметическое значений* η *при фиксированном значении* $\xi = x$.

Для корреляционной таблицы 7.4 условное среднее $\overline{y_x}$ *получается усреднением значений* η *по строке, соответствующей* $\xi = x$. *Так, например,* $\overline{y_{x_1}} = \frac{\sum\limits_{j=1}^{S} y_j n_{1j}}{n_1 * .}$. *Аналогично определяется условное среднее* $\overline{x_y}$.