

Лекция 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. Её называют генеральной. Если *генеральная совокупность* слишком многочисленна, или её 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 удаётся предложить ничего лучшего, чем чисто случайный выбор.

ПРИМЕР 7.1. Пусть необходимо оценить качество изделий, выпускаемых определённым 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. Приблизённо 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 m испытаний получим выборку n значений, образующих простую статистическую совокупность наблюдений.

ОПРЕДЕЛЕНИЕ 7.1. Количество наблюдений n называется объёмом выборки.

При большом числе наблюдений (сотни, тысячи) простая статистическая совокупность перестаёт быть удобной формой записи статистического материала — она становится слишком громоздкой. 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 $\sum_{i=1}^k m_i = n$ — объём 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, \dots, m_k называют частотами.

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

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

Таблица 7.1

Статистическое распределение			
варианты x_i	x_1	\dots	x_k
частоты m_i	m_1	\dots	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 объёму 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

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

5+10=15 раз, следовательно $F^*(x) = \frac{15}{50} = 0,3$ при $4 < x \leq 6$. Аналогично получаем $F^*(x) = \frac{30}{50} = 0,6$ при $6 < x \leq 7$ и т.д. Так как 10 — наибольшая варианта, $F^*(x) = 1$ при $x > 10$.

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

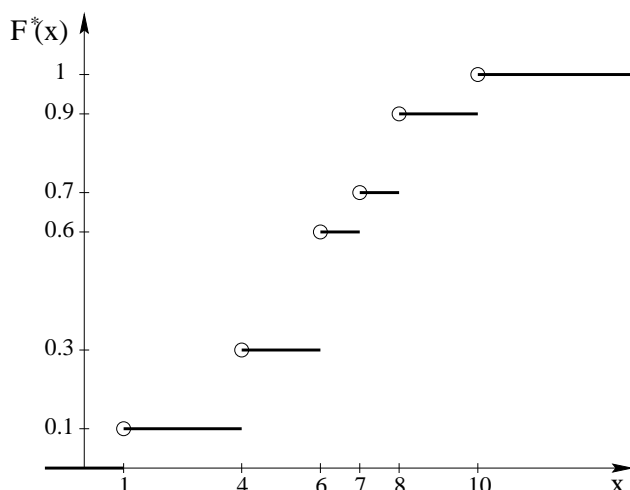


Рис. 25. Эмпирическая функция распределения

График найденной функции представлен на рис. 25.

Из определения $F^*(x)$ вытекают её свойства:

- 1) $0 \leq F^*(x) \leq 1$;
- 2) $F^*(x)$ — ступенчатая неубывающая функция;
- 3) если x_1 — наименьшая, а x_k — наибольшая варианты, то $F^*(x) = 0$ при $x \leq x_1$ 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 $P_j^* = \frac{m_j}{n}$ ($m_1 + \dots + m_s = n$), сумма всех 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. 26.

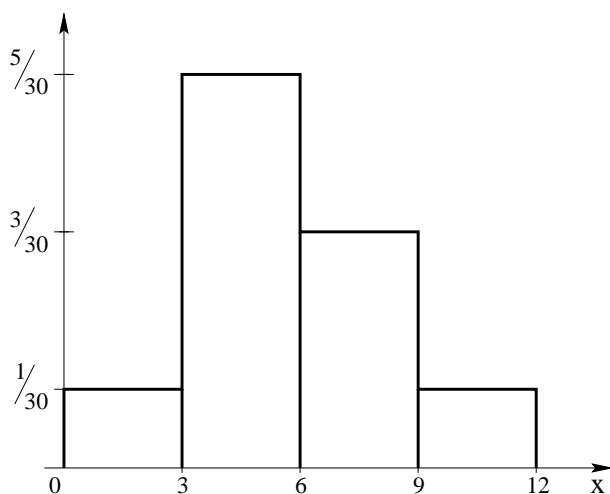


Рис. 26. Гистограмма относительных частот

Другим наглядным способом представления распределения является полигон относительных частот. Для его построения по оси абсцисс откладываются варианты, а по оси ординат — относительные частоты (рис. 27), и полученные точки соединяются ломаной линией.

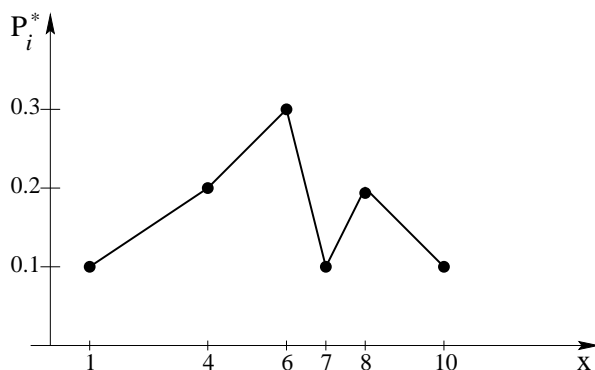


Рис. 27. Полигон относительных частот

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 — объём 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:

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

Если данные сгруппированы, то:

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

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

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

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

для простой совокупности и

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

для сгруппированного распределения.

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

$$S = \sqrt{S^2} \quad (7.5)$$

называется *выборочным средним квадратическим отклонением* (СКО).

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

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

для простой совокупности и

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

для сгруппированного распределения.

Докажем формулу (7.6):

$$\begin{aligned} S^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{1}{n} \sum_{i=1}^n x_i + \bar{x}^2 = \overline{x^2} - 2\bar{x} \cdot \bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2. \end{aligned}$$

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 подсчётом. 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: $\{1, 2, 2, 3, 4, 5, 5, 6, 6, 7\}$, моды 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 мода вычисляется следующим образом. Исследуемая выборка $\{x_i\}$ сортируется в порядке не убывания значений элементов. Далее, если объём выборки нечётное число, то $M_e = x_{(n+1)/2}$, иначе $M_e = (x_{n/2} + x_{n/2+1})/2$.

Например, для вариационного ряда $\{1, 3, 5, 7, 9, 9, 12\}$ медиана равна четвёртому элементу $M_e = 7$, а для вариационного ряда $\{1, 3, 5, 7, 9, 12\}$ медиана равна третьему и четвёртому элементам $M_e = (5 + 7)/2 = 6$.

Для сгруппированной выборки числовое значение для медианы необходимо аппроксимировать. Формула будет приведена ниже (7.9).

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

$$M_0 = X_0 - h \frac{f_{m0} - f_{m0-1}}{f_{m0-1} - 2f_{m0} + f_{m0+1}}, \quad (7.8)$$

где X_0 — нижнее значение модального интервала; $f_{m0}, f_{m0-1}, f_{m0+1}$ — значение частот в модальном, предыдущем и следующем интервалах, соответственно; h — размах модального интервала.

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

$$M_e = X_0 + h \frac{0,5 \sum_{k=1}^s f_k - \sum_{k=1}^{m_1-1} f_k}{f_{me}}, \quad (7.9)$$

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

ПРИМЕР 7.5. Покажем, как построить гистограмму и эмпирическую функцию распределения, вычислить числовые характеристики статистического распределения с помощью программы на *Maxima*. Сформируем массив x выборки $n = 500$ с помощью датчика случайных чисел.

Maxima-программа:

```
(%i0) kill(all)$  fpprintprec:4$  numer:true$  n:500$
(%i4) load(distrib)$

/*Генерируем выборку объёма n псевдослучайными числами.*/

(%i5) x:random_normal(145, 15, n)$

/*Изменяем значение выборки добавлением случайных чисел в диапазоне от 0 до 10.*/

(%i6)x:makelist(x[i]+random(10),i,1,n)$

/*Загружаем библиотеку descriptive.*/

(%i7) load (descriptive)$

/*Строим гистограмму частот.*/
```

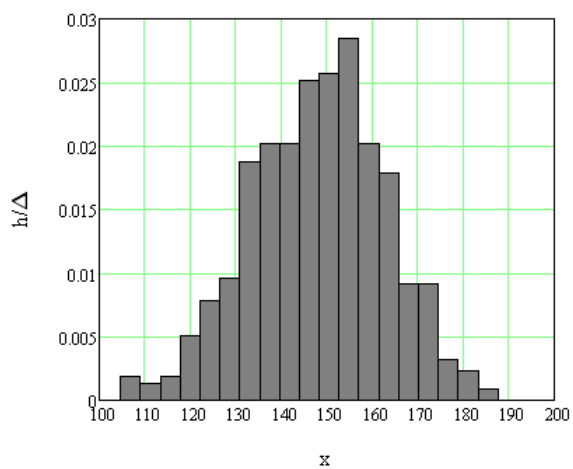


Рис. 28. Гистограмма относительных частот для примера 7.5

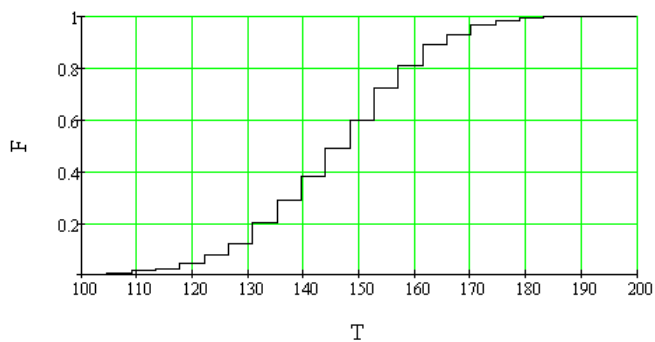


Рис. 29. Эмпирическая функция распределения для примера 7.5

```
(%i8) histogram(x, nclasses=20, title="закон распределения",
    xlabel="x", ylabel="частоты",
    fill_color=black, fill_density=0.05);

/* Сортируем список в порядке возрастания значений. */
(%i9) x:sort(x)$

/* Разбиваем выборку объёма n на s интервалов длиной delta. */
(%i10) s:20$ delta:(x[n]-x[1])/s;

/* T – массив узловых координат разбиения. */
(%i12) T:makelist(x[1]+delta*k,k,-1,s+1);

/* Координаты средних точек отрезков. */
(%i13) t:makelist((T[m]+T[m+1])/2,m,1,s+2);

/* Частоты попадания в соответствующие отрезки. */
(%i14) h:makelist(0, i, 1, s+2); for j:1 while j<=n do(
    k:fix((x[j] - x[1])/delta)+1,h[k]:h[k]+1);

/* Контрольная сумма объёма выборки. */
(%i15) sum(h[i],i,1,s+2);
(%o15) 500

/* Эмпирическая функция распределения. */
(%i16) F[1]:h[1]; for j:2 while j<=s+2 do(F[j]:F[j-1]+h[j]);
(%i17) listarray(F);

/* График эмпирическая функция распределения. */
(%i18) wxplot2d([[ 'discrete,makelist([t[j],F[j]/n],j,1,s+2)]],
    [style,[lines,3,5]], [gnuplot_preamble,"set grid",
    [ylabel,""]])$

/* Найдём также некоторые числовые характеристики данной вы-
борки: Выборочное среднее. */
(%i19) mean(x);
```

/* Выборочная дисперсия.

(%i20) var(x);

/* Выборочное среднее квадратическое отклонение.* /

(%i21) std(x);

/* Несмещенная выборочная дисперсия.* /

(%i22) var1(x);

/* Несмещенное среднее квадратическое отклонение.* /

(%i23) std1(x);

/* Медиана.* /

(%i24) median(x);

7.5. Точечные оценки параметров распределения

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

ОПРЕДЕЛЕНИЕ 7.6. Точечной оценкой \tilde{a}_n неизвестного параметра a распределения случайной величины ξ называется функция от наблюдений:

$$\tilde{a}_n = \tilde{a}(x_1, \dots, x_n).$$

Для изучения свойств этой оценки её рассматривают как функцию от n независимых случайных величин ξ_1, \dots, ξ_n , имеющих такое же распределение, что и ξ ; x_1, \dots, x_n в этом случае рассматриваются как наблюдения над этими случайными величинами: x_1 — полученное значение ξ_1 , x_2 — наблюдаемое значение ξ_2 и т.д. Сама оценка \tilde{a}_n в этом случае является случайной величиной.

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

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

$$\lim_{n \rightarrow \infty} P\{|\tilde{a}_n - a| < \varepsilon\} = 1 \text{ для } \forall \varepsilon > 0.$$

ОПРЕДЕЛЕНИЕ 7.8. Оценка \tilde{a}_n называется **несмещенной**, если её математическое ожидание равно оцениваемому параметру a :

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

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

ОПРЕДЕЛЕНИЕ 7.9. Несмещённая оценка \tilde{a}_n называется **эффективной**, если её дисперсия наименьшая по сравнению с другими несмещёнными оценками.

На практике оценка не всегда удовлетворяет всем этим требованиям одновременно.

ПРИМЕР 7.6. Доказать, что выборочное среднее \bar{x} является несмещённой 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 её математическое ожидание. При этом, как было отмечено ранее, считаем

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

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

Несмещённость выборочного среднего доказана. Оценим теперь дисперсию выборочного среднего:

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

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

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

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

$$1 \geq \lim_{n \rightarrow \infty} P\{|\xi - a| < \varepsilon\} \geq 1,$$

откуда получаем:

$$\lim_{n \rightarrow \infty} P\{|\xi - a| < \varepsilon\} = 1.$$

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 смещённой оценкой дисперсии σ^2 :

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

Примем это без доказательства.

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

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

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

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

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

ЗАМЕЧАНИЕ 7.3. Исправленное СКО S^* является смещённой 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 $\xi_1, \xi_2, \dots, \xi_n$, $\xi_i \sim N(0; 1)$, $i = 1, \dots, n$.

ОПРЕДЕЛЕНИЕ 7.10. Распределение случайной величины $\chi_n^2 = \sum_{i=1}^n \xi_i^2$ называется χ^2 -распределением с n степенями свободы.

Очевидно, что случайная величина $\chi_n^2 \geq 0$.

Плотность этого распределения имеет вид:

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

Здесь $\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t} dt$ — гамма функция, являющаяся обобщением понятия факториала: $\Gamma(x) = (x-1)!$ при $x \geq 1$.

ЗАМЕЧАНИЕ 7.4. Если случайные величины ξ_1, \dots, ξ_n связаны какой-нибудь зависимостью, например $\xi_1 + \dots + \xi_n = n \cdot \bar{x}$, то число степеней свободы уменьшается, случайная величина $\sum_{i=1}^n \xi_i^2$ будет иметь распределение χ_{n-1}^2 .

Распределение Стьюдента.

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

ОПРЕДЕЛЕНИЕ 7.11. Распределение случайной величины

$$t = \frac{\zeta}{\sqrt{\frac{\chi_n^2}{n}}}$$

называется распределением Стьюдента с n степенями свободы.

Плотность этого распределения имеет вид:

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

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

Стьюдент — псевдоним английского статистика Госсета.

F—Распределение Фишера—Снедекора.

Пусть имеется $n + k$ независимых стандартных величин: ξ_1, \dots, ξ_n ; ζ_1, \dots, ζ_k ; $\xi_i \sim N(0; 1)$, $i = 1, \dots, n$; $\zeta_j \sim N(0; 1)$, $j = 1, \dots, k$.

ОПРЕДЕЛЕНИЕ 7.12. Распределение случайной величины

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

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

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

Плотность этого распределения имеет вид:

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

Для всех этих распределений имеются таблицы 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 объёме выборки, когда дисперсия точечной оценки велика и она может сильно отличаться от оцениваемого параметра.

ОПРЕДЕЛЕНИЕ 7.13. Доверительным интервалом для несмещённого параметра a называют интервал $(a_1; a_2)$ со случайными границами, зависящими от наблюдений: $a_1 = a_1(x_1, \dots, x_n)$, $a_2 = a_2(x_1, \dots, x_n)$, накрывающий неизвестный параметр с заданной вероятностью γ : $P\{a \in (a_1; a_2)\} = \gamma$. Вероятность γ называется

доверительной вероятностью или надёжностью доверительного интервала.

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

Доверительный интервал \mathbf{I}_γ для неизвестного математического ожидания нормального распределения при известной дисперсии имеет вид:

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

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

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

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

ЗАМЕЧАНИЕ 7.5. При возрастании объёма выборки n , как видно из (7.13), доверительный интервал уменьшается. При увеличении надёжности γ увеличивается величина $\tau_{\gamma/2}$, 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 $\xi_i \sim N(a; \sigma)$, $i = 1, \dots, n$, то среднее арифметическое $\bar{\xi} = (\xi_1 + \dots + \xi_n)/n$ тоже распределено нормально с параметрами:

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

Формулы (7.15) были получены в примере 7.1. Будем искать доверительный интервал для a в виде:

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

где γ — заданная доверительная вероятность. Для определения ε воспользуемся формулой (5.7), которая в данном случае с учётом (7.15) принимает вид:

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

Найдём ε из уравнения:

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

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

ПРИМЕР 7.7. Найти доверительный интервал I_γ для неизвестного математического ожидания a нормально распределённой случайной величины со средним квадратическим отклонением $\sigma = 2$ по выборке объёма $n = 64$ 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 $I_\gamma = (4,71; 5,69)$.

Ответ: $I_\gamma = (4,71; 5,69)$.

Доверительный интервал I_γ для неизвестного математического ожидания нормального распределения при неизвестной дисперсии имеет вид:

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

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:

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

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 $\xi_i \sim N(a; \sigma)$, $i = 1, \dots, n$, то случайная величина

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

имеет распределение Стьюдента с $n - 1$ степенью свободы (см п. 93.4).

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

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

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

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

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 7.8. Найти доверительный интервал \mathbf{I}_γ для неизвестного математического ожидания a нормально распределённой случайной величины с выборочным средним $\bar{x} = 10,5$ и исправленным СКО $S^* = 1,6$ по выборке объёма $n = 16$. 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 объёма n из генеральной совокупности значений двумерной случайной величины $(\xi; \zeta)$, 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). В первом столбце этой таблицы перечислены значения x_i , во втором — y_i в виде вариационных рядов. На пересечении i -й строки

Таблица 7.4

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

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 $n_{i.} = \sum_{j=1}^s n_{ij}$, а в последней строке — сумму частот по

столбцам $n_{.j} = \sum_{i=1}^k n_{ij}$. Сумма всех элементов последнего столбца или строки даст объём 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 ζ . Обработав их, как описано в п. 96.4 предыдущей лекции, получим числовые характеристики

$$\bar{x} = \frac{\sum_{i=1}^k n_{i.} x_i}{n}, \quad \overline{x^2} = \frac{\sum_{i=1}^k n_{i.} x_i^2}{n}, \quad S_x^2 = \overline{x^2} - \bar{x}^2,$$

$$\bar{y} = \frac{\sum_{j=1}^s n_{\cdot j} y_j}{n}, \quad \overline{y^2} = \frac{\sum_{j=1}^s n_{\cdot j} y_j^2}{n}, \quad S_y^2 = \overline{y^2} - \bar{y}^2.$$

ОПРЕДЕЛЕНИЕ 7.14. Выборочным коэффициентом корреляции r_{xy}^* называется:

$$r_{xy}^* = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}, \quad \text{где} \quad (7.21)$$

$$\overline{xy} = \frac{\sum_{i=1}^k \sum_{j=1}^s n_{ij} x_i y_j}{n}. \quad (7.22)$$

Выборочный коэффициент корреляции является статистической оценкой коэффициента корреляции, рассмотренного в лекции 95, 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 : $-1 \leq r_{xy}^* \leq 1$;
- 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 ζ при фиксированном значении $\xi = x$.

Для корреляционной таблицы 7.4 условное среднее $\overline{y_x}$ получается усреднением значений ζ по строке, соответствующей $\xi = x$.

Так, например, $\overline{y_{x_1}} = \frac{\sum_{j=1}^s y_j n_{1j}}{n_1}$. Аналогично определяется условное среднее $\overline{x_y}$.