ГЛАВА 2. ТЕОРИЯ ОЦЕНОК

§2.1. Точечные оценки и их свойства

Статистическое оценивание – совокупность методов математической статистики, которые используются для приближенного определения неизвестных распределений вероятностей по результатам наблюдений.

Обычно предполагают, что вид функции распределения $F(x,\theta_1.....\vartheta_m)$ известен, и определению подлежат лишь значения неизвестных параметров θ_j. Поэтому на практике обычно задается распределение вероятностей $F(x,\theta_1.....\vartheta_m)$ и по значениям $(x_1,x_2....x_n)$ случайной выборки $(\xi_1,\xi_2,.....\xi_n)$ оценивают лишь неизвестные параметры θ_j. При этом в соответствие с классической вероятностной моделью предполагается, что $(\xi_1,\xi_2,.....\xi_n)$- последовательность независимых случайных величин, имеющих одно и то же распределение вероятностей $F(x,\theta_1.....\vartheta_m)$.

Определение 1. Точечной оценкой $\overline{\theta}_n$ параметра θ называется такая статистика $\overline{\theta}_n = f(\xi_1,\xi_2,...\xi_n)$, значение которой $f(x_1,x_2,.....x_n)\approx\theta$.

Определение 2. Если $\overline{\theta}_n \underset{n\to\infty}{\overset{вер}{\to}} \theta$, то оценка $\overline{\theta}_m$ называется состоятельной.

Определение 3. Если $M(\overline{\theta}_n)=\theta$, то оценка $\overline{\theta}_m$ называется несмещенной.

Определение 4. Если $D(\overline{\theta}_n)$ минимальна, то $\overline{\theta}_m$ называется эффективной оценкой.

§2.2. Выборочное среднее и выборочная дисперсия

Пусть X - физическая величина. Пусть $(x_1,....x_n)$- выборка объема n. В соответствии с классической вероятностной моделью сопоставим:

1) Физической величине X случайную величину ξ, распределенную по закону $f_\xi(x)$; $M(\xi)\equiv a, D(\xi)\equiv\sigma^2$.

2) Каждому выборочному значению x_i сопоставим случайную величину ξ_i, где ξ_i- взаимно независимые одинаково распределенные случайные величины, т.е. $f_{\xi_1}(x)=.....=f_{\xi_n}=f_\xi(x)$ и соответственно

$$M(\xi_1)=.....=M(\xi_n)\equiv a,\ D(\xi_1)=.....=D(\xi_n)\equiv\sigma^2\ . \qquad (1)$$

Тогда по теореме Чебышева

$$\begin{cases}\dfrac{1}{n}\sum\limits_{i=1}^{n}\xi_i \underset{n\to\infty}{\overset{вер}{\to}} a\\ \dfrac{1}{n}\sum\limits_{i=1}^{n}(\xi_i-a)^2 \underset{n\to\infty}{\overset{вер}{\to}} \sigma^2\end{cases} \qquad (2)$$

Поэтому величину $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ выбирают в качестве состоятельной оценки математического ожидания $M(\xi)\equiv a$, а величину $\frac{1}{n}\sum_{i=1}^{n}(\xi_i-a)^2$ - в качестве состоятельной оценки дисперсии $D(\xi)$.

Введем следующие обозначения:

$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$ - случайная величина; $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ - значение $\overline{\xi}$; $\overline{S}^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-a)^2$ - случайная величина; $\overline{S}_1^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2$ - значение $\overline{S}^2$; $\widetilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2$ - случайная величина; $\widetilde{S}_1^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ -значение $\widetilde{S}^2$.

Исследуем смещение введенных точечных оценок. Для этого рассмотрим математическое ожидание

$$M(\overline{\xi})=\frac{1}{n}M(\xi_1+\xi_2+\ldots+\xi_n)=\frac{1}{n}\sum_{i=1}^{n}M(\xi_i)=M(\xi),$$

так что $\overline{\xi}$ является несмещенной точечной оценкой для математического ожидания. Аналогично, $M(\overline{S}^2)=D(\xi)$. Поэтому $\overline{\xi}$ и $\overline{S}^2$ являются несмещенными точечными оценками для математического ожидания и дисперсии. Так как на практике $M(\xi)=a$ заранее неизвестно, то приходится пользоваться оценкой $\widetilde{S}^2$, которая является смещенной. Однако из оценки $\widetilde{S}^2$ можно получить несмещенную оценку следующим образом. Для этого преобразуем сумму квадратов

$$\sum_{i=1}^{n}(\xi_i-a)^2=\sum_{i=1}^{n}\left[(\xi_i-\overline{\xi})+(\overline{\xi}-a)\right]^2=$$
$$=\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2+\sum_{i=1}^{n}(\overline{\xi}-a)^2+2(\overline{\xi}-a)\sum_{i=1}^{n}(\xi_i-\overline{\xi})=$$
$$=\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2+n(\overline{\xi}-a)^2$$

Откуда находим

$$\widetilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-a)^2-(\overline{\xi}-a)^2 \qquad (3)$$

Из выражения (3) следует

$$M\left[\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi})^2\right]=\frac{1}{n}M\left[\sum_{i=1}^{n}(\xi_i-a)^2\right]-M\left[(\overline{\xi}-a)^2\right]=$$
$$=\frac{1}{n}\sum_{i=1}^{n}M(\xi_i-a)^2-D(\overline{\xi})=\frac{1}{n}\sum_{i=1}^{n}D(\xi_i)-D(\overline{\xi})= \qquad (4)$$
$$=D(\xi)-D(\overline{\xi})=D(\xi)-\frac{1}{n^2}D\left(\sum_{i=1}^{n}\xi_i\right)=D(\xi)-\frac{1}{n}D(\xi)=\frac{n-1}{n}D(\xi)$$

т.е. $\widetilde{S}^2$ является смещенной оценкой. Для получения несмещенной оценки, согласно формуле (4), достаточно умножить оценку $\widetilde{S}^2$ на $\frac{n}{n-1}$

При этом получим несмещенную оценку для дисперсии $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i - \overline{\xi}\right)^2$. Нетрудно видеть, что

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

где $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}x_i^2$

Исследуем эффективность выборочного среднего. Рассмотрим величину

$$\widetilde{\xi} = a_1\xi_1 + a_2\xi_2 + ... + a_n\xi_n .$$

По свойству математического ожидания

$$M\left(\widetilde{\xi}\right) = M(\xi)\sum_i a_i .$$

Условие несмещенности требует, чтобы

$$\sum_i a_i = 1 . \qquad (5)$$

Докажем минимальность оценки $\overline{\xi}$. Для этого найдем условный экстремум функции

$$u = \sum_{i=1}^{n} a_i^2 ,$$

при условии (5). Построим функцию Лагранжа

$$L = u + \lambda(a_1 + a_2 + ... + a_n - 1).$$

Необходимое условие экстремума имеет вид

$$\frac{\partial L}{\partial a_i} = 2a_i + \lambda = 0 ,$$

так что

$$a_i = \frac{\lambda}{2} . \qquad (6)$$

Подставляя (6) в (5), найдем множитель Лагранжа

$$\lambda = \frac{2}{n} . \qquad (7)$$

Тогда из (5) с учетом (7) получим

$$a_i = \frac{1}{n} . \qquad (8)$$

Из (8) следует, что второй дифференциал

$$d^2u = 2\sum_{i=1}^{n} a_i^2 > 0 ,$$

так что значения (8) соответствуют минимуму функции $u = \sum_{i=1}^{n} a_i^2$. При этом $\tilde{\xi} = \bar{\xi} = \frac{1}{n}\sum_i \xi_i$ и дисперсия $D(\bar{\xi})$ минимальна. Следовательно, выборочное среднее является эффективной оценкой математического ожидания.

§2.3. Мода и медиана, четвертые значения и прямоугольная диаграмма, размах варьирования и среднее абсолютное отклонение, коэффициент вариации

Кроме выборочного среднего и выборочной дисперсии применяются и другие числовые характеристики вариационного ряда. Укажем главные из них.

Определение 1. Выборочной модой $\overline{M}_0$ называют вариационное значение, которое имеет наибольшую частоту. Например, для ряда

x_i	1	4	7	9
m_i	5	1	20	6

мода равна $\overline{M}_0 = 7$.

Определение 2. Выборочной медианой $\overline{M}_e$ называют число, которое делит вариационный ряд на части, равные по числу вариант:

$$M_e = \begin{cases} x_{k+1}, & n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & n = 2k \end{cases}$$

где $n-$ число вариационных значений.

Например, для ряда 2,3,5,6,7 $\overline{M}_e = 5$; для ряда 2,3,5,6,7,9 $\overline{M}_e = 5.5$.

Четвертое значение. Если разделить пополам нижнюю и верхнюю половины упорядоченного ряда, то полученные значения называют нижним и верхним ***четвертым значением*** $x_{1/4}$ или $x_{3/4}$.

Например, для ряда 14, 15, 20, 22, 25, 33, 39, 40. 45, 47, 56, 57, 65, 69. 71, 72, 77, 77, 123, 144

$$\overline{M}_e = \frac{47+56}{2} = 51.5, \;\; x_{1/4} = \frac{25+33}{2} = 29, \;\; x_{3/4} = \frac{71+72}{2} = 71.5$$

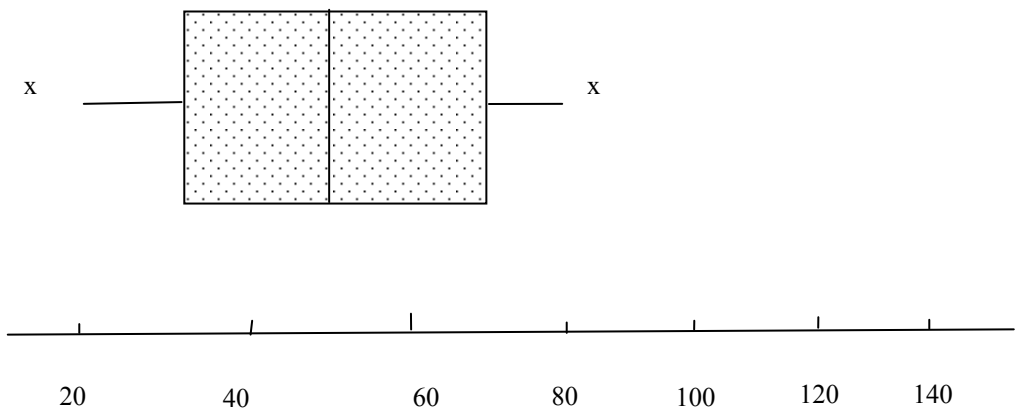
Из способа образования средних ясно, что, по меньшей мере, половина всех выборочных значений лежит между обоими четвертыми значениями. Четвертые значения используют при построении ***«Прямоугольной диаграммы»,*** которую строят следующим образом.

На числовую прямую переносят выборочные значения. Над ними рисуют прямоугольник («ящик»), который начинается у нижнего четвертого значения $x_{1/4}$ и заканчивается у верхнего четвертого значения $x_{3/4}$. Отрезок, перпендикулярный числовой прямой, отмечает выборочную медиану $\overline{M}_e$, Крестиками обозначают крайние выборочные значения. Линии, выходящие из ящика вправо и влево, показывают распределение вне ящика. Линии продолжают так далеко, что за линиями каждый раз лежит не более 10% данных.

В предыдущем примере

$$x_{\min} = 14, x_{1/4} = 29,\ \overline{M}_e = 51.5,\ x_{3/4} = 71.5,\ x_{\max} = 144\,.$$

В этом примере ***«Прямоугольная диаграмма»*** выглядит следующим образом:



Определение 3. Размахом варьирования R называют разность между максимальными и минимальными вариационными значениями:

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

Например, для ряда 1,3,4,5,6,10 размах варьирования R=10-1=9 (Размах варьирования является простейшей характеристикой рассеяния вариационного ряда).

Определение 4. Средним абсолютным отклонением Θ называют среднее арифметическое абсолютных отклонений:

$$\Theta = \frac{1}{N}\sum_{j=1}^{k} m_j \left| x_j - \overline{x} \right|$$

Например, для ряда

x_i	1	3	6	16
m_i	4	10	5	1

имеем $\bar{x} = 4;\ \Theta = 2.2$ (Θ служит для характеристики рассеяния вариационного ряда)

Определение 5. Коэффициентом вариации V называют выраженное в процентах отношение среднего квадратичного отклонения S к выборочному среднему $\bar{x}$

$$V = \frac{S}{\bar{x}} 100\%$$

(V служит для сравнения величин рассеяния двух вариационных рядов: тот из рядов имеет большее рассеяние, у которого V больше).

Пример. Используя коэффициент вариации, проанализируем данные годовых уровней прибыли трех компаний. Сравним результаты их деятельности за 10 лет, и выясним деятельность какой компании более успешна.

Год	Татнефть	Сибнефть	Томскнефть
1983	14.2	-6.2	37.5
1984	12.3	13.3	-10.6
1985	-16.2	-8.4	40.3
1986	15.4	27.3	5.4
1987	17.2	28.2	6.2
1988	10.3	14.5	10.2
1989	-6.3	-2.4	13.8
1990	-7.8	-3.1	11.5
1991	3.4	15.6	-6.2
1992	12.2	18.2	27.5

Решение. По данным выборки средняя прибыль компании «Татнефть» равна 5.47 с выборочным среднеквадратичным отклонением 11.63, для «Сибнефть» соответствующие величины равны 9.7 и 13.69, «Томскнефть» – 13.56 и 16.97. Тогда коэффициент вариации для «Татнефть» принимает значение $V_1 = \frac{S_1}{\bar{x}_1} 100\% = 213\%$, для «Сибнефть» – $V_2 = \frac{S_2}{\bar{x}_2} 100\% = 141\%$, а для «Томскнефть» – $V_3 = \frac{S_3}{\bar{x}_3} 100\% = 125\%$. Таким образом, деятельность компании «Томскнефть» является наиболее успешной.

§2.4. Оценка качества продукции

Рассмотрим задачу, связанную с выборкой из конечной совокупности. Пусть каждая партия готовой продукции содержит N изделий, причем $M_1, M_2, \ldots$ – количества дефектных изделий в этих партиях, а $m_1, m_2, \ldots$ – соответствующие количества дефектных изделий, обнаруженных в выборках объема n. Согласно условию бездефектной приемки, партия с номером i передается потребителю, если $m_i = 0$, в противном случае она бракуется. Предположим, что контроль изделий сопряжен с их уничтожением, и поэтому потребитель либо получает партию объема $R_i = 0$ при $m_i > 0$, либо партию объема $R_i = N - n$ с количеством дефектных изделий $D_i = M_i$ при $m_i = 0$, причем значения R_i известны, а значения D_i неизвестны. Отношение $Q = \frac{\sum\limits_i D_i}{\sum\limits_i R_i}$ называют долей пропущенного брака, а его математическое ожидание $q = M(Q)$ – средней долей пропущенного брака. Задача математической статистики заключается в оценке q по значениям R_i, зафиксированным в результате применения выборочного метода. Если значения M_i трактовать как реализации независимых одинаково распределенных случайных величин с известным законом распределения $P(M_i = r) \equiv p_r$, то согласно формуле Байеса, статистическая оценка среднего числа пропущенных дефектных изделий в принятых партиях выражается формулой

$$\overline{D} = M[M / m = 0] = \frac{\sum\limits_{r=1}^{N-n} r \frac{C_{N-r}^n}{C_N^n} p_r}{P(m = 0)},$$

причем

$$\overline{D} \le \frac{(N-n)P(m=1)}{P(m=0)},$$

где

$$P(m = k) = \sum_{r=0}^{N-n} \frac{C_r^k C_{N-r}^n}{C_N^n} p_r,\ k = 0,1,2....$$

Поэтому оценка

$$\overline{q} = \frac{\overline{D}}{N-n}$$

средней доли пропущенного брака в принятых партиях удовлетворяет неравенству

$$\overline{q} = \frac{\overline{D}}{N-n} \le \frac{P(m=1)}{nP(m=0)} \approx \frac{\nu_1}{n\nu_0},$$

где ν_0 – число принятых партий, а ν_1 – количество тех забракованных партий, в которых обнаружено ровно 1 дефектное изделие.

Основой количественного описания изменчивости качественного признака дефектности внутри отдельной совокупности является среднее этой совокупности (генеральное среднее)

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

где $X_1 = X_2 = ... = X_M = 1,\ X_{M+1} = X_{M+2} = ... = X_{N-M} = 0,$ так что в случае качественного признака дефектности $\overline{X} = \frac{M}{N}$, N – объем всей совокупности, M – количество дефектных деталей в данной отдельной совокупности. Оценкой для генерального среднего $\overline{X} = \frac{M}{N}$ является выборочное среднее $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{m}{n}$, где m – число дефектных объектов в выборке объема n. За характеристику изменчивости дефектности принимают дисперсию

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \frac{M(N-M)}{N^2}.$$

О точности оценоки $\frac{m}{n}$ судят по их дисперсиям $\frac{\sigma^2}{n}$ (в случае повторной выборки) и $\frac{\sigma^2(N-n)}{n(N-1)}$ (в случае бесповторной выборки). Случайные величины $\frac{m}{n}$ при $n \ge 30$ приближенно подчиняются нормальному распределению, так что отклонение $\frac{m}{n}$ от $\frac{M}{N}$, превышающие по абсолютной величине $2\frac{\sigma}{\sqrt{n}}$, может при $n \ge 30$ осуществиться в среднем приблизительно в одном случае из 20.

§2.5. Методы получения точечных оценок

Метод моментов

В тех случаях, когда в качестве параметров распределения выступают его моменты, то полученные оценки позволяют непосредственно получить оценки этих параметров.

В общем же случае точечные оценки $\overline{\theta}_n$ получают по методу моментов следующим образом. Пусть, например, плотность вероятности (СВ) ξ есть $f_\xi(x, \theta_1, \theta_2)$. Тогда моменты

$$\begin{cases} M(\xi) = \int\limits_{-\infty}^{\infty} x\, f_{\xi}(x, \theta_1, \theta_2)\, dx \\ D(\xi) = \int\limits_{-\infty}^{\infty} (x - M)^2 f(x, \theta_1, \theta_2)\, dx \end{cases} \qquad (1)$$

представляют собой некоторые функции от параметров θ_1, θ_2. Написанные выше соотношения (1) рассматривают как уравнения относительно этих параметров, после решения которых в полученных выражениях заменяют теоретические моменты $M(\xi), D(\xi)$ их оценками $\bar{x}, S^2$. Это и дает оценки θ_1, θ_2 по методу моментов.

Пример1. Оценить параметры нормального распределения

$$f(x, a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2} \right].$$

Решение. Из теории вероятностей известно, что

$$M(\xi) = a$$

$$D(\xi) = \sigma^2$$

Так как $M(\xi) \approx \bar{x}, D(\xi) \approx S^2$, то из полученных выше уравнений находим точечные оценки параметров нормального распределения $a = \bar{x}, \sigma^2 = S^2$.

Пример2. Оценим неизвестный параметр λ распределения Пуассона $P(m, \lambda) = \dfrac{\lambda^m e^{-\lambda}}{m!}$.

Решение. Так как распределение является дискретным, то

$$M(\xi) = \sum_{m=0}^{\infty} m\, P(m, \lambda) = \sum m \frac{\lambda^m e^{-\lambda}}{m!} = \lambda e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^{m-1}}{(m-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Заменяя в левой части полученного уравнения теоретическое математическое ожидание $M(\xi)$ его точечной оценкой $\bar{x}$, получим

$$\lambda = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} m_i \ ,$$

где N –число серий испытаний, n-число испытаний в серии, m_i-число появлений события A в i-й серии испытаний

Пример 3. По методу моментов найдем точечную оценку параметра p биномиального распределения $P(m, p) = C_m^n\, p^m (1-p)^{n-m}$.

Решение. Для этого найдем математическое ожидание

$$M(\xi) = \sum_{m=0}^{n} m\, P[m, p] = \sum_{m=0}^{n} m\, C_m^n\, p^m (1-p)^{n-m} = np \sum_{m=1}^{n} C_{m-1}^{n-1} p^{m-1} (1-p)^{n-m} = np$$

Откуда находим

$$p = \frac{1}{Nn} \sum_{i=1}^{N} m_i \ .$$

Замечание. Достоинством метода моментов является относительная простота решения уравнений. Недостатки: 1) метод не применим, когда моменты нужного порядка не существуют; 2) кроме того, оценки,

получаемые по методу моментов, часто неэффективны, поэтому обычно их используют в качестве первых приближений для нахождения последующих приближений с большей эффективностью.

Метод наибольшего правдоподобия

Пусть $(x_1, x_2, ...x_n)$- простая выборка; $(\xi_1, \xi_2, ...\xi_n)$- случайная выборка, соответствующая простой выборке; ξ_i- последовательность взаимно независимых случайных величин, имеющих одинаковую плотность вероятности $f_\xi(x, \theta_1, ...\theta_m)$ с подлежащими оценке параметрами θ_j. По методу наибольшего правдоподобия приближенные значения параметров θ_j находятся из условия экстремума функции

$$l(x_1, ...x_n, \theta_1, ...\theta_m) = \ln(L),$$

где функция правдоподобия L равна выражению

$$L(x_1, ...x_n, \theta_1, ...\theta_m) = \prod_{i=1}^{n} f_\xi(x_i, \theta_1, ...\theta_m)$$

Решая уравнения правдоподобия $\frac{\partial l}{\partial \theta_j} = 0 \ (j = 1,2...m)$, находят значения оценок параметров θ_j.

Замечание1. Для получения оценок параметров дискретного распределения в предыдущие формулы вместо $f_\xi(x_i, \theta_1, ...\theta_m)$ надо подставить соответствующие вероятности $P(\xi = x_i, \vartheta_1, ...\vartheta_m)$.

Замечание2. Достоинства метода: 1) оценки, полученные по методу наибольшего правдоподобия, являются состоятельными; 2) оценки, полученные по методу наибольшего правдоподобия, являются эффективными. Недостатки метода: 1) оценки могут оказаться смещенными; 2) нахождение оценок по методу наибольшего правдоподобия приводят к сложным уравнениям.

Пример1. Оценим истинное среднее значение измеряемой величины в случае неравноточных измерений, результаты которых распределены по нормальному закону.

Пусть проведена серия независимых измерений одной и той же величины, осуществленных с различной точностью, т.е.

$$M(\xi_i) = a, D(\xi_i) \equiv \sigma_i^2 \text{ (дисперсии различны).}$$

Необходимо оценить неизвестное математическое ожидание a. **Решение.** Составим функцию правдоподобия для данного случая

$$L(x_1, ...x_n, a) = \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_i - a)^2}{2\sigma_i^2} \right)$$

Тогда $$l(x_1, ...x_n, a) = \ln(L) = \sum_{i=1}^{n} \left[-\ln(\sigma_i) - \frac{1}{2}\ln(2\pi) - \frac{(x_i - a)^2}{2\sigma_i^2} \right]$$

При этом из уравнения правдоподобия

$$\frac{\partial l}{\partial a}=\sum_{i=1}^{n}\frac{(x_i-a)}{\sigma_i^2}=0$$

окончательно найдем оценку $a=\sum_{i=1}^{n}g_i x_i, g_i=\frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$. Из g_i видно, что чем больше дисперсия σ_i^2 i-го измерения, тем меньший вклад это измерение вносит в оценку измеряемой величины.

Покажем, что полученная оценка является несмещенной. В соответствии с классической вероятностной моделью сопоставим выборочным значениям x_i случайные величины ξ_i. При этом получим оценку $\overline{a}_n=\sum_{i=1}^{n}g_i\xi_i$. Тогда $M(\overline{a}_n)=\sum_{i=1}^{n}g_i M(\xi_i)=\sum_{i=1}^{n}g_i a=a$, так что оценка $\overline{a}_n$ является несмещенной.

Докажем состоятельность оценки. Для этого найдем дисперсию оценки.

$$D(\overline{a}_n)=\sum_{i=1}^{n}D(g_i\xi_i)=\sum_{i=1}^{n}g_i^2\sigma_i^2=\frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\underset{n\to\infty}{\to}0$$

Следовательно, при $n\to\infty$ $\overline{a}_n\to a$ и оценка состоятельная.

Пример2. По методу наибольшего правдоподобия найдем точечные оценки параметров a,σ нормального распределения

$$f_\xi(x,a,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]$$

в случае равноточных измерений.

Решение. Найдем логарифм функции правдоподобия

$$l=\ln(L)=\ln\left[\prod_{i=1}^{n}f_\xi(x_x,a,\sigma)\right]=\sum_{i=1}^{n}\left[-\ln(\sigma)-\frac{1}{2}\ln(2\pi)-\frac{(x_i-a)^2}{2\sigma^2}\right]=$$

$$=-n\ln(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2-\frac{n}{2}\ln(2\pi)$$

Отсюда находим уравнения правдоподобия

$$\begin{cases}\frac{\partial l}{\partial a}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a)=0,\\ \frac{\partial l}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-a)^2=0.\end{cases}$$

Решая эти уравнения, получаем нужные точечные оценки параметров a, σ, $a = \frac{1}{n}\sum_{i=1}^{n} x_i, \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$. В этом примере проявился недостаток метода: точечная оценка дисперсии является смещенной.

Пример 3. Оценим параметр λ распределения Лапласа

$$f(x;\lambda) = \frac{\lambda}{2} e^{-\lambda|x|}.$$

Решение. Найдем логарифм функции правдоподобия

$$l = \ln(L) = \ln\left(\prod_{i=1}^{n} f(x_i;\lambda)\right) = \sum_{i=1}^{n}\left[\ln\left(\frac{\lambda}{2}\right) - \lambda|x_i|\right].$$

Тогда из уравнения правдоподобия находим

$$\lambda = \frac{n}{\sum_{i=1}^{n}|x_i|}.$$

Пример 4. Оценим параметр λ распределения Пуассона

$$P(m,\lambda) = \frac{\lambda^{m} e^{-\lambda}}{m!}.$$

Решение. Найдем функцию

$l = \ln(L) = \sum_{i=1}^{N} \ln(P(m_i,\lambda)) = \sum_{i=1}^{N} m_i \ln(\lambda) - \sum_{i=1}^{N}(m_i!) - \lambda N$. При этом уравнение правдоподобия примет вид $\frac{dl}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{N} m_i - N = 0$. Откуда следует, что $\lambda = \frac{1}{N}\sum_{i=1}^{N} m_i$, где N – число серий испытаний, m_i-число появлений события в i-й серии.

Пример 5. Оценим параметр p биномиального распределения

$$P(m,p) = C_n^k p^m (1-p)^{n-m}$$

Решение. В этом случае функция l равна

$$l = \ln(L) = \sum_{i=1}^{N} \ln(P(m_i,p)) = \sum_{i=1}^{N} m_i \ln(p) + \sum_{i=1}^{N}(n-m_i)\ln(1-p) + \sum_{i=1}^{N} \ln\left(C_{m_i}^{n}\right)$$

Откуда следует уравнение правдоподобия

$$\frac{dl}{dp} = \frac{1}{p}\sum_{i=1}^{N} m_i - \frac{1}{1-p}\sum_{i=1}^{N}(n-m_i) = 0,$$

решение которого дает $p = \frac{1}{nN}\sum_{i=1}^{N} m_i$.

Пример 6. Требуется узнать общее число N студентов ИГНД**.** Для этого случайным образом выберем 20 студентов, пометим их и отпустим. Спустя некоторое время выберем 50 студентов, среди которых, допустим, оказался 1 меченый студент.

Решение. Пусть ξ – случайная величина числа меченых студентов. Тогда вероятность того, что среди 50 отобранных студентов окажется 1 меченый студент, равна

$$P(\xi=1)=\frac{C_1^{20}C_{N-20}^{50-1}}{C_N^{50}}=20\frac{(N-50)....(N-68)}{N(N-1)....(N-19)}.$$

Так как выборка состоит только из одного выборочного значения $x_1=1$, то функция правдоподобия равна

$$L(N)=P(\xi=1).$$

Функция правдоподобия недиффиренцируема, так как ее аргумент N принимает только дискретные значения 69, 70… Поэтому значение N, при котором $L(N)$ достигает наибольшей величины, найдем перебором. Для этого получим таблицу значений $L(N)$ при $N=69...2000$. Из таблицы следует, что $L(N)$ достигает наибольшего значения при $N=1000$.

§2.6. Групповая, внутригрупповая и межгрупповая дисперсии

Групповая, внутригрупповая и межгрупповая дисперсии используются в корреляционном анализе и однофакторном дисперсионном анализе. Введем эти понятия.

Пусть $(x_1,x_2,...x_n)$-выборка объема n. Допустим, что все выборочные значения x_l разбиты на k групп $(l=1,2,...n)$. Введем следующие обозначения: j-номер группы; m_j- число вариационных значений в j группе; x_{ij}- i вариационное значение j группы $(i=1,2,..m_j)$; n_{ij}-частота x_{ij} элемента; $\sum_{i=1}^{m_j}\sum_{j=1}^{k}n_{ij}=n$, $\sum_{i=1}^{m_j}n_{ij}=N_j$, N_j- число элементов в j группе; $\bar{x}=\frac{1}{n}\sum_{l=1}^{n}x_l$ -общее среднее; $\bar{x}_j=\frac{1}{N_j}\sum_{i=1}^{m_j}n_{ij}x_{ij}$ -групповые средние.

Рассматривая каждую группу как самостоятельную совокупность, можно ввести оценку дисперсии для выборочных значений x_{ij}, принадлежащих j группе, относительно группового среднего $\bar{x}_j$ следующим образом

$$D_j=\frac{1}{N_j}\sum_{i=1}^{m_j}n_{ij}(x_{ij}-\bar{x}_j)^2 \qquad (1)$$

Пример 1. Пусть дан статистический ряд:

x_l	2	3	4	5	8
ν	1	2	7	2	3

Разобьем выборку на две группы

x_{i1}	2	4	5
n_{i1}	1	7	2

x_{i2}	3	8
n_{i2}	2	3

$$N_1 = 1+7+2 = 10,\ N_2 = 2+3 = 5$$

Найдем групповые средние

$$\bar{x}_1 = \frac{1\cdot 2 + 7\cdot 4 + 2\cdot 5}{10} = 4,\ \bar{x}_2 = \frac{6+24}{5} = 6.$$

Найдем искомые групповые дисперсии

$$D_1 = \frac{1\cdot(2-4)^2 + 7(4-4)^2 + 2(5-4)^2}{10} = 0.6,$$

$$D_2 = \frac{2\cdot(3-6)^2 + 3(8-6)^2}{5} = 6.$$

Зная дисперсию каждой группы, можно найти их среднее арифметическое. Внутригрупповой дисперсией называют среднее арифметическое групповых дисперсий

$$D_{внгр} = \frac{1}{n}\sum_{j=1}^{k} N_j D_j\ . \qquad (2)$$

Пример 2. Данные те же. Найдем внутригрупповую дисперсию.
Искомая внутригрупповая дисперсия равна

$$D_{внгр} = \frac{10\cdot 0.6 + 5\cdot 6}{15} = \frac{12}{5}.$$

Зная групповые средние и общее среднее, можно найти дисперсию групповых средних относительно общего среднего. Межгрупповой дисперсией называют дисперсию $\bar{x}_j$ относительно $\bar{x}$

$$D_{межгр} = \frac{1}{n}\sum_{j=1}^{k} N_j\left(\bar{x}_j - \bar{x}\right)^2\ . \qquad (3)$$

Пример 3. Данные те же. Найдем сначала общее среднее.

$$\bar{x} = \frac{1}{n}\sum_{l=1}^{n} n_l x_l = \frac{2+28+10+6+24}{15} = \frac{14}{3}.$$

Используя вычисленные выше величины $\bar{x}_1 = 4,\ \bar{x}_2 = 6$, найдем искомую межгрупповую дисперсию

$$D_{межгр} = \frac{10\left(4-\frac{14}{3}\right)^2 + 5\left(6-\frac{14}{3}\right)^2}{15} = \frac{8}{9}.$$

Общей дисперсией называют дисперсию x_l относительно $\bar{x}$

$$D_{общ} = \frac{1}{n}\sum_{l=1}^{n}(x_l - \bar{x})^2 . \qquad (4)$$

Пример 4. Данные те же. Вычислим общую дисперсию.

$$D_{общ} = \frac{\left(2-\frac{14}{3}\right)^2 + 7\left(4-\frac{14}{3}\right)^2 + 2\left(5-\frac{14}{3}\right)^2 + 3\left(8-\frac{14}{3}\right)^2 + 2\left(8-\frac{14}{3}\right)^2}{15} = \frac{148}{45}.$$

Интересно отметить, что $\frac{148}{45} = \frac{12}{5} + \frac{8}{9}$, т.е. $D_{общ} = D_{внгр} + D_{межгр}$, Оказывается, что это свойство является общим. Докажем это.
Действительно, по определению

$$D_{общ} = \frac{1}{n}\sum_{l=1}^{n}(x_l - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{m_j} n_{ij}(x_{ij} - \bar{x})^2 =$$

$$= \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{m_j} n_{ij}\left[(x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})\right]^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{m_j} n_{ij}(x_{ij} - \bar{x}_j)^2 + \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{m_j} n_{ij}(\bar{x}_j - \bar{x})^2 +$$

$$+ \frac{2}{n}\sum_{j=1}^{k}(\bar{x}_j - \bar{x})\sum_{i=1}^{m_j} n_{ij}(x_{ij} - \bar{x}_j) = \frac{1}{n}\sum_{j=1}^{k} N_j D_j + \frac{1}{n}\sum N_j(\bar{x}_j - \bar{x})^2 = D_{внгр} + D_{межгр}$$

Что и требовалось показать.

§2.7. Доверительный интервал для оценки математического ожидания нормального распределения в случае, когда среднее квадратичное отклонение известно

Все оценки, рассмотренные выше – точечные. При выборке малого объема точечные оценки могут приводить к грубым ошибкам. Поэтому при небольшом объеме выборки следует пользоваться интервальными оценками.

Интервальная оценка-оценка, которая определяется двумя числами-концами интервала. Пусть: θ - неизвестный параметр распределения; $\overline{\theta}$ - точечная оценка этого параметра; $\gamma = P\left(\overline{\theta} - \delta < \theta < \overline{\theta} + \delta\right)$ - вероятность того, что интервал $\left(\overline{\theta} - \delta < \theta < \overline{\theta} + \delta\right)$ покрывает значение неизвестного параметра θ (γ называется доверительной вероятностью).

Интервал $\left(\overline{\theta} - \delta < \theta < \overline{\theta} + \delta\right)$, который покрывает значение параметра θ с заданной доверительной вероятностью γ, называется доверительным интервалом.

Пусть X – физическая величина. В соответствии с классической вероятностной моделью:

1)сопоставим физической величине X случайную величину ξ, распре деленную по нормальному закону

$$f_{\xi}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right],$$

причем среднее квадратичное отклонение $\sigma=\sqrt{D(\xi)}$ этого распределение считается известным; $a=M(\xi)$-математическое ожидание. Пусть $(x_1,x_2,...x_n)$ -выборка из генеральной совокупности ξ;

2) каждому выборочному значению x_i сопоставим случайную величину ξ_i; ξ_i-взаимно независимые и одинаково распределенные случайные величины, т.е. $f_{\xi_i}(x)=f_{\xi}(x),\ M(\xi_i)=a,\ \sigma(\xi_i)=\sigma$. Пусть $\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$. Докажем утверждение.

Утверждение 1. Если случайная величина ξ распределена нормально с параметрами a,σ, то случайная величина $\overline{\xi}$ также распределена нормально, но с параметрами $a,\frac{\sigma}{\sqrt{n}}$.

Доказательство

Для случайных величин ξ_i, распределенных нормально

$$f_{\xi_i}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right],$$

характеристической функцией является

$$f_{\xi_i}(t)=\int_{-\infty}^{\infty}\exp(itx)\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]dx=\exp(ita)\exp\left(-\frac{t^2\sigma^2}{2}\right). \quad (1)$$

Тогда по первому свойству характеристических функций и с учетом выражения (1) получим

$$f_{\frac{\xi_i}{n}}(t)=f_{\xi_i}\left(\frac{t}{n}\right)=\exp\left(\frac{ita}{n}\right)\exp\left(-\frac{t^2\sigma^2}{2n^2}\right). \quad (2)$$

По второму свойству характеристических функций из формулы (2) следует

$$f_{\overline{\xi}}(t)=\prod_{i=1}^{n}f_{\xi_i}\left(\frac{t}{n}\right)=\exp(ita)\exp\left(-\frac{t^2\sigma^2}{2n}\right). \quad (3)$$

Из сравнения выражений (3) и (1) заключаем, что $\overline{\xi}$ распределена нормально, но с параметрами $a,\frac{\sigma}{\sqrt{n}}$, что и требовалось доказать.

Потребуем, чтобы выполнялось соотношение

$$P\left\lfloor \left|a-\overline{\xi}\right|<\delta \right\rfloor =\gamma \qquad (4)$$

Из теории вероятностей известно, что

$$P\left(\left|a-\xi_i\right|<\delta\right)=\int_{a-\delta}^{a+\delta} f_{\xi_i}(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{a-\delta}^{a+\delta}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)dx=$$

$$=\left|\begin{array}{l}\dfrac{x-a}{\sigma}=t\\ dx=\sigma dt\end{array}\right|=\frac{2}{\sqrt{2\pi}}\int_0^{\frac{\delta}{\sigma}}\exp\left(-\frac{t^2}{2}\right)dt\equiv 2\Phi_0\left(\frac{\delta}{\sigma}\right), \qquad (5)$$

где функция Лапласа $\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int_0^x\exp\left(-\frac{t^2}{2}\right)dt$, значения которой приведены в соответствующих таблицах. Из формулы (5) и доказанной выше теоремы следует, что

$$P\left(\left|a-\overline{\xi}\right|<\delta\right)=2\Phi_0\left(\frac{\delta\sqrt{n}}{\sigma}\right) \qquad (6)$$

Введем обозначение $\lambda_\gamma=\frac{\delta\sqrt{n}}{\sigma}$. Тогда получим, что $\delta=\lambda_\gamma\frac{\sigma}{\sqrt{n}}$, При этом выражение (6) перепишется следующим образом

$$P\left(\left|a-\overline{\xi}\right|<\lambda_\gamma\frac{\sigma}{\sqrt{n}}\right)=2\Phi_0\left(\lambda_\gamma\right) \qquad (7)$$

Подставляя формулу (7) в равенство (4), находим

$$P\left(\overline{\xi}-\lambda_\gamma\frac{\sigma}{\sqrt{n}}<a<\overline{\xi}+\lambda_\gamma\frac{\sigma}{\sqrt{n}}\right)=2\Phi_0\left(\lambda_\gamma\right)=\gamma \qquad (8)$$

Заменяя $\overline{\xi}$ на $\overline{x}$, из уравнения (8) получим доверительный интервал

$$\overline{x}-\lambda_\gamma\frac{\sigma}{\sqrt{n}}<a<\overline{x}+\lambda_\gamma\frac{\sigma}{\sqrt{n}} \qquad (9)$$

где λ_γ определяется из равенства

$$\Phi_0\left(\lambda_\gamma\right)=\frac{\gamma}{2}$$

с помощью таблицы функции Лапласа.

Пример. На станции технического обследования 36 автомобилей исследовались затраты времени на ремонт двигателя. Было зафиксировано выборочное среднее $\overline{x}=4.1$час. Среднее квадратичное отклонение задано и равно $\sigma=3$. Построить 95% доверительный интервал для средних затрат времени на ремонт двигателя.

Решение. Из уравнения $\Phi(\lambda_\gamma)=\frac{\gamma}{2}=0.475$ по таблице находим $\lambda_\gamma=1.96$. Тогда ошибка среднего равна $\lambda_\gamma\frac{\sigma}{\sqrt{n}}=0.98$. Окончательно получим, что затраты времени в среднем удовлетворяют неравенству $3.12<a<5.08$.

§2.8. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном среднем квадратичном отклонении

Пусть ξ - случайная величина, распределенная нормально. Среднее квадратичное отклонение σ считается неизвестным. В этом случае нахождение доверительного интервала основывается на доказанном в теории вероятностей утверждении.

Утверждение 2. Если случайная величина ξ имеет нормальное распределение, то случайная величина

$$T=\frac{\overline{\xi}-a}{S}\sqrt{n} \qquad (1)$$

имеет распределение Стьюдента с k=n-1 степенями свободы, где

$$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i,\ S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i-\overline{\xi}\right)^2}\ .$$

Плотность вероятности распределения Стьюдента имеет вид

$$f_T(x)=B_n\left(1+\frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

здесь $B_n=\dfrac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi(n-1)}}$, $\Gamma(x)$- гамма-функция.

Из приведенного утверждения следует

$$P(|T|<t_\gamma)=2\int_0^{t_\gamma}f_T(x)dx \qquad (2)$$

Потребуем, чтобы выполнялось равенство

$$P\left(\overline{x}-t_\gamma\frac{S}{\sqrt{n}}<a<\overline{x}+t_\gamma\frac{S}{\sqrt{n}}\right)=\gamma \qquad (3)$$

Значения интеграла $2\int_0^{t_\gamma}f_T(x)dx$ приводятся в соответствующей таблице (Стьюдента). Из равенств (2-3) получим, что для оценки математического

ожидания a случайной величины ξ, распределенной по нормальному закону, при неизвестной дисперсии служит доверительный интервал

$$\bar{x} - t_\gamma \frac{S}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{S}{\sqrt{n}}, \qquad (4)$$

где коэффициент Стьюдента t_γ находится из уравнения $2\int_0^{t_\gamma} f_T(x)dx = \gamma$ по таблице Стьюдента по входным данным k и γ.

Пример. Пусть по выборке объема n=9 вычислено выборочное среднее $\bar{x} = 42.319$ и выборочное среднее квадратичное отклонение $S = 5.0$. Пусть доверительная вероятность равна $\gamma = 0.95$.

Решение. Используя входные данные n=9, $\gamma = 0.95$ по таблице находим коэффициент Стьюдента $t_\gamma = 2.31$. Найдем точность оценки $t_\gamma \frac{S}{\sqrt{n}} = 3.85$. Найдем границы доверительного интервала

$$\begin{cases} \bar{x} - t_\gamma \frac{S}{\sqrt{n}} = 36.469 \\ \bar{x} + t_\gamma \frac{S}{\sqrt{n}} = 46.169 \end{cases}$$

Итак, с надежностью 0.95 истинное значение измеряемой величины заключено в доверительном интервале $38.469 < a < 46.169$.

§2.9. Доверительный интервал для оценки среднего квадратичного отклонения нормального распределения

Пусть ξ - случайная величина, распределенная нормально. Возникает задача: по выборочному среднему квадратичному отклонению S оценить неизвестное теоретическое среднее квадратичное отклонение σ. Для решения этой задачи потребуем, чтобы выполнялось соотношение

$$P(S - \delta < \sigma < S + \delta) = \gamma \qquad (1)$$

Преобразуем неравенство $(S - \delta < \sigma < S + \delta)$ следующим образом

$$(S - \delta < \sigma < S + \delta) \Rightarrow S(1 - q) < \sigma < S(1 + q), \qquad (2)$$

где $q = \frac{\delta}{S}$. Неравенство (2) эквивалентно двум неравенствам

$$\begin{cases} \sigma < S(1+q) \\ \sigma > S(1-q) \end{cases} \qquad (3)$$

Возможны два случая. В первом случае $q < 1$. В этом случае из неравенств (3) находим

$$\begin{cases} \dfrac{S\sqrt{n-1}}{\sigma} > \dfrac{\sqrt{n-1}}{1+q} \\ \dfrac{S\sqrt{n-1}}{\sigma} < \dfrac{\sqrt{n-1}}{1-q} \end{cases} \tag{4}$$

Введем обозначение $\chi = \dfrac{S\sqrt{n-1}}{\sigma}$. Тогда соотношения (4) перепишутся в следующем виде

$$\frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q} \tag{5}$$

Во втором случае $q > 1$. В этом случае из неравенств (3) находим

$$\begin{cases} \chi > \dfrac{\sqrt{n-1}}{1+q} \\ \chi > \dfrac{-\sqrt{n-1}}{q-1} \end{cases} \tag{6}$$

Так как при $q > 1$ $\dfrac{\sqrt{n-1}}{1+q} > \dfrac{-\sqrt{n-1}}{q-1}$, то из выражений (6) находим

$$\frac{\sqrt{n-1}}{1+q} < \chi < \infty \tag{7}$$

Объединяя неравенства (5) и (7), окончательно имеем

$$\begin{cases} \dfrac{\sqrt{n-1}}{1+q} < \chi < \dfrac{\sqrt{n-1}}{1-q}, & q < 1 \\ \dfrac{\sqrt{n-1}}{1+q} < \chi < \infty, & q > 1 \end{cases} \tag{8}$$

Построение доверительного интервала (2) основывается на следующем доказанном в теории вероятностей утверждении.

Утверждение 3. Если случайная величина ξ распределена нормально, то случайная величина $\chi = \dfrac{S}{\sigma}\sqrt{n-1}$ имеет плотность вероятности

$$f_\chi(x) = C_n x^{n-2} \exp\left(-\frac{x^2}{2}\right), \tag{9}$$

где $C_n = \dfrac{1}{2^{\frac{n-3}{2}}\Gamma\left(\dfrac{n-1}{2}\right)}$.

Используя функцию (9) число q находят из следующих интегральных уравнений

$$\begin{cases} \int\limits_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} f_\chi(x)dx = \gamma \ , \quad q<1 \\ \int\limits_{\frac{\sqrt{n-1}}{1+q}}^{\infty} f_\chi(x)dx = \gamma \ , \quad q>1 \end{cases} \tag{10}$$

Вычислив по выборке S и найдя величину q, удовлетворяющую уравнениям (10), по таблице, получим доверительные интервалы

$$\begin{cases} S(1-q) < \sigma < S(1+q), \ q<1 \\ 0 < \sigma < S(1+q), \ q>1 \end{cases} \tag{11}$$

Пример. Пусть по выборке объема n=25 найдено выборочное среднее квадратичное отклонение S=0.8. Пусть доверительная вероятность равна $\gamma = 0.95$. По таблице по данным $\gamma = 0.95$ и n=25 найдем $q = 0.32$. Тогда доверительный интервал таков:

$$0.544 < \sigma < 1.056$$

Пример.
Дано: $n = 10, S = 0.16, \gamma = 0.999$. По таблице $q = 1.80$.
Тогда $0 < \sigma < 0.448$.

§2.10. Доверительный интервал для оценки параметра показательного распределения

Пусть ξ - случайная величина, распределенная по показательному закону

$$f(x,\lambda) = \lambda e^{-\lambda x}, \ x \geq 0,$$

где λ – неизвестный параметр. Найдем интервальную оценку для парамера λ. Построение доверительного интервала основывается на следующем утверждении.

Утверждение 4. Если случайные величины ξ_i одинаково распределены по показательному закону, то статистика

$$\Gamma = \lambda \sum_{i=1}^{n} \xi_i \tag{1}$$

имеет гамма-распределение

$$f_\Gamma(x) = \frac{1}{\Gamma(n)} x^{n-1} e^{-x}, \ x \geq 0, \tag{2}$$

а функция распределения статистики (1) равна

$$F_{\Gamma}(x)=\int_{-\infty}^{x} f_{\Gamma}(x)dx=1-e^{-x}\sum_{k=0}^{n}\frac{x^{k}}{k!}, \qquad (3)$$

причем предполагается, что $P(\Gamma \le q_1)=P(\Gamma \ge q_2)=\frac{1-\gamma}{2}$.

Доказательство

По определению функции распределения

$$P(\Gamma \le q_1)=F_{\Gamma}(q_1)$$

и

$$P(\Gamma \ge q_2)=1-P(\Gamma < q_2)=1-F_{\Gamma}(q_2).$$

Поэтому квантили q_1 и q_2 определяются уравнениями

$$F_{\Gamma}(q_1)=\frac{1-\gamma}{2},\ F(q_2)=\frac{1+\gamma}{2}. \qquad (4)$$

Используя формулы (3) и (4), можно найти квантили q_1, q_2. Тогда решая неравенство

$$q_1 < \Gamma < q_2$$

относительно параметра λ, получим случайный интервал

$$\frac{q_1}{\sum_{i=1}^{n} x_i} < \lambda < \frac{q_2}{\sum_{i=1}^{n} x_i},$$

который содержит параметр λ с вероятностью γ,

§2.11. Интервальная оценка вероятности биномиального распределения

Найдем доверительный интервал для оценки вероятности по относительной частоте. Если n достаточно велико и вероятность p не очень близка к нулю или единице, то по интегральной формуле Муавра-Лапласа

$$P[|W-p|<\delta]=2\Phi_0\left(\frac{\delta}{\sigma}\sqrt{n}\right), \qquad (1)$$

где $\sigma=\sqrt{pq}$. Для построения доверительного интервала потребуем, чтобы с надежностью γ выполнялось соотношение

$$P\left[|W-p|<\delta\right]=2\Phi_0\left(\frac{\delta}{\sigma}\sqrt{n}\right)=2\Phi_0(t)=\gamma, \qquad (2)$$

где $t=\delta\sqrt{\frac{n}{pq}}$, так что $\delta=t\sqrt{\frac{pq}{n}}$ и, следовательно,

$$P\left[|W-p|<t\sqrt{\frac{pq}{n}}\right]=2\Phi_0(t)=\gamma.$$

Таким образом, с надежностью γ выполняется неравенство

$$|W-p|<t\sqrt{\frac{p(1-p)}{n}}. \qquad (3)$$

Возведя обе части неравенства (3) в квадрат, получим равносильное квадратное неравенство

$$\left[\left(\frac{t^2}{n}\right)+1\right]p^2-2\left[W+\left(\frac{t^2}{n}\right)\right]p+W^2<0. \qquad (4)$$

Дискриминант трехчлена положительный, так что его корни действительные и различные:

меньший корень

$$p_1=\frac{n}{n+t^2}\left[W+\frac{t^2}{2n}-t\sqrt{\frac{W(1-W)}{n}+\left(\frac{t}{2n}\right)^2}\right], \qquad (5)$$

больший корень

$$p_2=\frac{n}{n+t^2}\left[W+\frac{t^2}{2n}+t\sqrt{\frac{W(1-W)}{n}+\left(\frac{t}{2n}\right)^2}\right]. \qquad (6)$$

Таким образом, искомый доверительный интервал имеет вид

$$p_1<p<p_2,$$

где p_1 и p_2 находятся по формулам (5,6).

Пример. Для изучения различных демографических характеристик населения выборочно обследовано 300 семей Томской области. Оказалось,

что среди обследованных семей 15% состоят из 2 человек. В каких пределах находится в генеральной совокупности доля семей, состоящих из 2 человек, если принять доверительную вероятность равной $\gamma = 0.97$?

Решение. Так как обьем выборки велик, то интервальная оценка вероятности принимает простой вид

$$p = W \pm \lambda\sqrt{\frac{W(1-W)}{n}}, \quad \Phi_0(\lambda) = \frac{\gamma}{2} = \frac{0.97}{2} = 0.485,\ W = 0.15.$$

Из интегрального уравнения $\Phi_0(\lambda) = \dfrac{0.97}{2}$ следует, что $\lambda = 2.17$, так что в генеральной совокупности доля семей $p \cdot 100\%$, состоящих из 2 человек, заключена в пределах

$$10.5\% < 100p < 19.5\%.$$

§2.12. Методы Монте-Карло

Простейший метод Монте-Карло

Пусть требуется вычислить определенный интеграл

$$I \equiv \int_a^b f(x)dx. \qquad (1)$$

Выберем определенную на отрезке $[a,b]$ произвольную плотность распределения $\varphi_\xi(x) > 0$ случайной величины ξ, которая нормирована условием

$$\int_a^b \varphi_\xi(x)dx = 1.$$

Введем случайную величину

$$\eta = \frac{f(\xi)}{\varphi_\xi(\xi)}.$$

Вычислим математическое ожидание случайной величины η

$$M(\eta) = \int_a^b \left[\frac{f(x)}{\varphi_\xi(x)}\right] \cdot \varphi_\xi(x)dx = \int_a^b f(x)dx = I. \qquad (2)$$

Формула (2) означает, что можно вычислить интеграл (1), вычислив математическое ожидание случайной величины η. Для вычисления $M(\eta)$ воспользуемся методами математической статистики. Выберем N значений $x_1, x_2, \ldots x_N$ случайной величины ξ, тогда по теореме Чебышева при достаточно большом N имеем

$$I = M(\eta) \approx \frac{1}{N}\sum_{j=1}^{N} \frac{f(x_j)}{\varphi(x_j)}, \qquad (3)$$

где $\frac{1}{N}\sum_{j=1}^{N}\frac{f(x_j)}{\varphi(x_j)}$ – точечная оценка математического ожидания $M(\eta)$.

На основании правила трех сигм для нормального распределения справедлива следующая формула

$$P\left[\left|\frac{1}{N}\sum_{j=1}^{N}\frac{f(x_j)}{\varphi(x_j)}-I\right|<3\frac{S(\eta)}{\sqrt{N}}\right]=0.997. \qquad (4)$$

Здесь P - вероятность события $\left[\left|\frac{1}{N}\sum_{j=1}^{N}\frac{f(x_j)}{\varphi(x_j)}-I\right|<3\frac{S(\eta)}{\sqrt{N}}\right]$, $S^2(\eta)$- смещенная выборочная дисперсия случайной величины η и

$$S^2(\eta)=\frac{1}{N}\sum_{j=1}^{N}y_j^2-\left[\frac{1}{N}\sum_{j=1}^{N}y_j\right]^2,\ y_j=\frac{f(x_j)}{\varphi_\xi(x_j)}. \qquad (5)$$

Формула (4) означает, что с вероятностью близкой к 1, абсолютная погрешность вычисления интеграла (1) не превосходит величины $3\frac{S(\eta)}{\sqrt{N}}$, так что погрешность Δ_M метода равна

$$\Delta_M=3\frac{S(\eta)}{\sqrt{N}}. \qquad (6)$$

При реализации метода Монте-Карло обычно в качестве $\varphi_\xi(x)$ используют равномерное распределение

$$\varphi_\xi(x)=\begin{cases}\frac{1}{b-a},\ x\in[a,b]\\ 0,\ x\notin[a,b]\end{cases}. \qquad (7)$$

Тогда с учетом (7) формулы (3) и (5) примут вид

$$I=\frac{b-a}{N}\sum_{j=1}^{N}f(x_j), \qquad (8)$$

$$S^2(\eta)=\frac{(b-a)^2}{N}\sum_{j=1}^{N}f^2(x_j)-I^2, \qquad (9)$$

где выборочные значения x_j из равномерно распределенной генеральной совокупности ξ реализуются с помощью системной функции $\mathrm{rnd}(t)$

$$x_j=a+\mathrm{rnd}(b-a),\ j=1,2,...,N.$$

Геометрический метод Монте-Карло

Вычислим однократный интеграл

$$\int_0^1 f(x)dx. \qquad (1)$$

Дополнительно потребуем, чтобы для функции $f(x)$ было справедливо неравенство

$$0 \le f(x) \le 1. \qquad (2)$$

Для реализации метода генерируется N независимых случайных точек

$$P_j(\xi_j, \eta_j),$$

где ξ_j, η_j – независимые случайные величины равномерно распределенные на отрезке $[0,1]$. Пусть x_j, y_j – выборочные значения этих величин и

$$x_j = a + \mathrm{rnd}(b-a),\ y_j = f(a) + \mathrm{rnd}(f(b) - f(a)).$$

Если для двух случайных величин ξ_j и η_j окажется, что

$$f(x_j) < y_j, \qquad (3)$$

то событие считается неблагоприятным. Если же

$$f(x_j) \ge y_j, \qquad (4)$$

то событие считается благоприятным, так как в этом случае точка с координатами (x_j, y_j) попадает в область криволинейной трапеции, ограниченной осями координат и кривой подынтегральной функции.

Пусть оказалось, что после реализации N испытаний число благоприятных событий равно ν, которое можно найти с помощью системной единичной функции $\Phi(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0 \end{cases}$ компьютерной программы «Mathcad. 2001» по формуле

$$\nu = \sum_j \Phi(f(x_j) - y_j).$$

По теореме Бернулли, являющейся частным случаем теоремы Чебышева, и свойству плотности вероятности относительная частота благоприятных событий $\frac{\nu}{N}$ приблизительно равна площади криволинейной трапеции, т.е.

$$\int_0^1 f(x)dx \approx \frac{\nu}{N}, \qquad (5)$$

и при $N \to \infty$ справедливо равенство

$$\int_0^1 f(x)dx = \lim_{N\to\infty}\left(\frac{\nu}{N}\right).$$

Рассмотрим теперь общий случай. Пусть нам необходимо вычислить интеграл

$$\int_a^b f(x)dx, \qquad (6)$$

где

$$c \le f(x) \le d.$$

Введем новую функцию

$$\varphi(x) = \frac{f(x) - c}{d - c},$$

которая изменяется от 0 до 1. Тогда

$$\int_a^b f(x)dx = (d-c)\int_a^b \varphi(x)dx + c(b-a). \qquad (7)$$

При вычислении интеграла $\int_a^b \varphi(x)dx$ выполним замену переменной $t = \frac{x-a}{b-a}$, тогда интеграл (7) будет равен

$$\int_a^b f(x)dx = (d-c)(b-a)\int_0^1 \varphi((b-a)t + a)dt + c(b-a), \qquad (8)$$

где интеграл, стоящий в правой части (8) вычисляется с помощью геометрического метода Монте-Карло по формуле (5).

Погрешность метода можно определить, воспользовавшись интегральной формулой Муавра-Лапласа

$$P\left[\left|\frac{\nu}{N} - I\right| < \varepsilon\right] = 2\Phi\left(\varepsilon\sqrt{\frac{N}{p(1-p)}}\right) = \gamma, \qquad (9)$$

где обычно доверительная вероятность $\gamma = 0.99$, p -вероятность попадания отдельной точки в область криволинейной трапеции, $\sigma = \sqrt{p(1-p)}$, $\sigma_{\max} = \frac{1}{2}$, откуда по таблице функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ находим $\varepsilon = \frac{2.58}{2\sqrt{N}}$.

Формула (9) означает, что с вероятностью $\gamma = 0.99$ близкой к единице абсолютная погрешность вычисления интеграла I не превосходит величины $\frac{2.58}{2\sqrt{N}}$, так что погрешность геометрического метода Монте-Карло равна

$$\Delta_M \approx \frac{2.58}{2\sqrt{N}}. \qquad (10)$$