Элементы математической статистики. Первичная обработка результатов измерений

Математическая статистика - раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для установления существующих закономерностей.

Предметом математической статистики является изучение СВ по результатам наблюдений (опыта, эксперимента).

Пусть рассматривается некоторый наблюдаемый признак X.

Определение 44.1. Генеральной совокупностью наблюдаемого признака X называется множество значений, которые может принимать наблюдаемый признак.

Определение 44.2. Часть наблюдений, отобранных случайным образом для изучения из генеральной совокупности, называется выборкой.

Определение 44.3. Число наблюдений n в выборке называется ее объемом.

Полученные в результате наблюдений данные надо представить в удобном для обозрения и анализа виде. Для этого проводится первичная обработка результатов измерений, которая осуществляется по-разному в зависимости от типа наблюдаемого признака.

Ранжирование. Если наблюдаемый признак X является ДСВ, то первичная обработка результатов наблюдений заключается в ранжировании, т.е. в расположении упомянутых в выборке значений в порядке возрастания с указанием количества повторов каждого значения.

Пример. Пусть X – количество сбоев в работе станка в течении смены. В результате проведенных наблюдений получены следующие данные: 5,1,3,2,4,1,2,3,4,5,3,2,2,1,2,5,4,4,4,3,1,2,3,4,5. Требуется произвести первичную обработку результатов измерений.

◁ Наблюдаемый признак является ДСВ, поэтому первичная обработка есть ранжирование:

X	1	2	3	4	5
n_i	4	6	5	6	4

Значения 1,2,3,4,5, встречающиеся в приведенной таблице, называют вариантами, числа 4,6,5,6,4 – их частотами, а саму таблицу - вариационным рядом частот. Отметим, что сумма частот n_i совпадает с объемом выборки: 4+6+5+6+4=25. ▷

Иногда вместо вариационного ряда частот используют вариационный ряд относительных частот, в котором вместо частот n_i используются относительные частоты $w_i = \frac{n_i}{n}$, причем $\sum_{i=1}^{k} w_i = 1.$

В примере 1 вариационный ряд относительных частот имеет вид:

X	1	2	3	4	5
w_i	0,16	0,24	0,20	0,24	0,16

Интервальная обработка выборки. Если генеральная совокупность не является дискретным множеством, то построение вариационного ряда производится при помощи интервальной обработки выборки.

При этом число интервалов рассчитывается по эмпирической формуле Стерджеса: $k \approx 1 + 3{,}322 \lg n$, шаг h определяется по формуле $h = \frac{R}{k}$, где $R = x_{\max} - x_{\min}$ - размах варьирования, $x_{\max}$ – наибольшая, а $x_{\min}$ – наименьшая варианта ряда.

В качестве начала первого интервала x_0 выбирают $x_0 = x_{\min} - \frac{h}{2}$.

Вариационный ряд частот получают из интервального ряда частот, заменяя каждый из интервалов его серединой.

Пример. Построить вариационный ряд частот и относительных частот по результатам измерений:
2,3; 2,5; 2,7; 2,35; 2,71; 2,32; 2,36; 2,44; 2,61; 2,67; 2,83; 2,86; 3,01; 3,12; 3,14; 2,61;2,49; 2,57; 2,52; 2,54; 3,03; 3,05.

◁ Очевидно, в рассматриваемом случае n =25; $x_{\min} = 2{,}3$; $x_{\max} = 3{,}3$, поэтому

$$h = \frac{3{,}3 - 2{,}3}{1 + 3{,}322 \lg 25} \approx 0{,}2 .$$

Руководствуясь изложенными выше соображениями, строим интервальный ряд частот:

X	[2,2; 2,4)	[2,4; 2,6)	[2,6; 2,8)	[2,8; 3,0)	[3,0; 3,2)	[3,2; 3,4]
n_i	4	6	5	2	6	2

Заменяя интервалы их серединами, получаем вариационный ряд частот:

X	2,3	2,5	2,7	2,9	3,1	3,3
n_i	4	6	5	2	6	2

а затем вариационный ряд относительных частот:

X	2,3	2,5	2,7	2,9	3,1	3,3
w_i	0,16	0,24	0,2	0,08	0,24	0,08

Графическое изображение результатов измерений.

Полигон и гистограмма распределения

Полигон служит для изображения вариационного ряда и представляет собой ломаную линию, отрезки которой последовательно соединяют точки $(x_1;n_1)$, $(x_2;n_2)$,...,$(x_k;n_k)$, где x_i – варианты, а n_i – соответствующие им частоты.

Полигоном относительных частот называют ломаную, отрезки которой последовательно соединяют точки $(x_1;w_1)$, $(x_2;w_2)$,...,$(x_k;w_k)$, где w_i – относительные частоты.

Гистограмма служит для графического изображения интервальных рядов и представляет собой ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$. Площадь i-го прямоугольника равна $\frac{hn_i}{h}=n_i$ – сумме частот, попавших в i-й интервал, а площадь гистограммы частот равна сумме всех частот, т.е. объему выборки *n.*

В случае гистограммы относительных частот высоты прямоугольников равны $\frac{w_i}{h}$, площадь i-го прямоугольника равна w_i – относительной частоте варианты i-го интервала, а площадь всей гистограммы равна сумме всех относительных частот, т. е. единице.

Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Эмпирическая функция распределения

Одним из способов обработки вариационного ряда является построение эмпирической функции распределения.

Определение 44.4. Эмпирической функцией распределения (функцией распределения выборки) называется относительная частота того, что СВ X примет значение меньше заданного x: $F_n(x)=\frac{n_x}{n}$, где n_x – число вариант, меньших x ($x \in \mathbb{R}$), n – объем выборки.

Эмпирическая функция распределения при неограниченном увеличении объема выборки приближается к функции распределения наблюдаемого признака *X*. Точнее, имеет место следующее утверждение.

Теорема Гливенко-Кантелли. Эмпирическая функция распределения при неограниченном увеличении объема выборки сходится к функции распределения наблюдаемого признака по вероятности равномерно по *x,* т.е.

$$P(\lim_{n\to\infty}(\sup_{x\in R}(|F(x)-F_n(x)|)=0)=1$$

Точечное оценивание параметров

Пусть X – наблюдаемый признак с известным видом функции (плотности, закона) распределения. Будем предполагать, что функция распределения зависит от параметров: $\theta_1,\theta_2,...,\theta_m$: $F(x,\theta_1,\theta_2,...,\theta_m)$.

Определение 44.5. Точечной оценкой параметра θ_i называют всякую формулу, которая по результатам выборки позволяет рассчитывать приближенное значение параметра: $\tilde{\theta}_i = \tilde{\theta}_i(X_1, X_2,...,X_k)$.

К оценке любого параметра предъявляется ряд требований, которым она должна удовлетворять, чтобы быть близкой к истинному значению параметра. В связи с этим, среди оценок выбирают наилучшие, используя для этого следующие критерии:

1. ***Несмещенность.*** Точечная оценка параметра называется несмещенной, если математическое ожидание оценки совпадает с истинным значением этого параметра: $M[\tilde{\theta}_i(X_1, X_2,...,X_k)] = \theta_i$
2. ***Состоятельность.*** Точечная оценка параметра называется состоятельной, если при неограниченном увеличении объема выборки СВ $\tilde{\theta}_i(X_1, X_2,...,X_k)$ сходится по вероятности к истинному значению этого параметра, т.е. если $\lim\limits_{n\to\infty} P(\left|\tilde{\theta}_i(X_1, X_2,...,X_k) - \theta_i\right| < \varepsilon) = 1$ каково бы ни было $\varepsilon > 0$.
3. ***Эффективность.*** Несмещенная точечная оценка параметра называется эффективной, если она имеет наименьшую дисперсию среди всех несмещенных оценок рассматриваемого параметра.

Точечная оценка математического ожидания

Пусть задан вариационный ряд частот наблюдаемого признака X. Назовем выборочным средним и будем обозначать $\bar{x}_в$ среднее арифметическое значений, наблюденных в выборке:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \quad \text{или, если все } n_i = 1, \text{ то } \bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Выборочное среднее является точечной оценкой математического ожидания наблюдаемого признака, и эта оценка наилучшая в силу следующей теоремы.

Теорема 44.1. Выборочное среднее есть несмещенная, состоятельная, эффективная оценка математического ожидания.

Точечные оценки дисперсии

По аналогии с математическим ожиданием точечной оценкой дисперсии будем считать выборочную дисперсию, которая вычисляется по одной из формул:

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot n_i \qquad \text{или} \quad D_в = \frac{1}{n}\sum_{i=1}^{k} x_i^{\,2} n_i - \bar{x}_в^{\,2}$$

Теорема 44.2. Оценка $D_в$ является смещенной, а именно $M[D_в] = \frac{n-1}{n}\sigma^2$, где σ^2 - дисперсия наблюдаемого признака.

Следствие. Несмещенной оценкой дисперсии является

$$s^2 = \frac{n}{n-1} D_в = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2 \cdot n_i \ .$$

Эту оценку называют исправленной выборочной дисперсией.

Дробь $\frac{n}{n-1}$ называют поправкой Бесселя. Очевидно, эта поправка стремится к 1 при увеличении объема выборки и при $n > 50$ разница между исправленной дисперсией и дисперсией выборки практически неощутима. Пользуясь законом больших чисел, можно показать, что обе рассмотренные оценки являются состоятельными. Однако, исправленная выборочная дисперсия не является эффективной оценкой дисперсии. Несмещенной, состоятельной, эффективной оценкой дисперсии является следующая оценка $s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$. Но эта оценка практически неприменима, ибо для ее построения необходимо знание точного значения математического ожидания a.

Выборочное среднее квадратическое отклонение определяется по формуле: $\sigma_в = \sqrt{D_в}$.

Исправленное среднее квадратическое отклонение: $s = \sqrt{s^2}$.

Определение 44.6. Медианой Me вариационного ряда называется значение признака, приходящееся на середину ряда. Если $n = 2k$ (ряд имеет четное число членов), то $Me = \frac{x_k + x_{k+1}}{2}$, и если $n = 2k+1$ (число членов ряда нечетное), то $Me = x_{k+1}$.

Определение 44.7. Модой Mo распределения вариационного ряда называется варианта, имеющая наибольшую частоту.

Интервальное оценивание параметров

Пусть наблюдаемый признак X зависит от некоторых параметров и Θ - один из этих параметров.

Определение 44.8. Оценка неизвестного параметра называется интервальной, если она определяется двумя числами – концами интервала.

Определение 44.9. Интервал $\left(\underline{\Theta},\overline{\Theta}\right)$ называется доверительным интервалом, соответствующим доверительной вероятности γ, если вероятность того, что истинное значение параметра Θ находится на этом интервале, равна γ:

$$P\left(\Theta \in \left(\underline{\Theta},\overline{\Theta}\right)\right)=\gamma$$

В качестве доверительной вероятности выбирают достаточно близкие к единице значения, с целью получения информации об изучаемом параметре с вероятностью, близкой к единице. В связи с этим, иногда доверительную вероятность называют надежностью.

Стандартными являются следующие значения доверительной вероятности:

$$\gamma = 0{,}95; \qquad \gamma = 0{,}99; \qquad \gamma = 0{,}995;$$

Замечание. Иногда вместо доверительной вероятности использую величину $\alpha = 1-\gamma$, называемую уровнем значимости.

Построение доверительного интервала для математического ожидания нормального распределения при известном σ

Пусть наблюдаемый признак X распределен по нормальному закону и известно его среднее квадратическое отклонение σ.

Интервал $\left(\overline{x}_в - \frac{z_\gamma}{\sqrt{n}}\sigma, \overline{x}_в + \frac{z_\gamma}{\sqrt{n}}\sigma\right)$

является доверительным для математического ожидания с доверительной вероятностью γ. Величина z_γ определяется из уравнения

$$2\Phi\left(z_\gamma\right)=\gamma, \quad \text{где} \quad \Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt.$$

Функция Лапласа $\Phi(x)$ задается таблично и, соответственно, решение уравнения также производится при помощи таблиц. При этом следует учитывать, что функция $\Phi(x)$ нечетна, т.е., $\Phi(-x)=-\Phi(x)$.

Пример. СВ X имеет нормальное распределение со средним квадратическим отклонением $\sigma = 3$. Найти доверительный интервал по выборке 2,1; 2,3; 2,4; 2,6; 2,8; 2,7; 2,5;2,.4; 2,7; с уровнем значимости 0,05.

◁ Доверительная вероятность (надежность) в рассматриваемом случае

γ = 1 – 0,05 = 0,95.·

Найдем $z_{0,95}$, решив уравнение $2\Phi(z_{0,95}) = 0{,}95$ или $\Phi(z_{0,95}) = 0{,}475$.

Пользуясь таблицами для функции $\Phi(x)$, получаем $z_{0,95}=1{,}96$.

Очевидно, объем выборки в рассматриваемом случае n=9.

Найдем $\bar{x}_в$:

$$\bar{x}_в=\frac{1}{9}\left(2{,}1+2{,}3+2{,}4+2{,}6+2{,}8+2{,}7+2{,}5+2{,}4+2{,}7\right)=\frac{1}{9}\cdot 22{,}5=2{,}5$$

Следовательно, доверительный интервал имеет вид:

$$\left(2{,}5-\frac{3\cdot 1{,}96}{3};\ \ 2{,}5+\frac{3\cdot 1{,}96}{3}\right)\ \text{или}\ \ (0{,}54;\,4{,}46)\qquad \triangleright$$

Построение доверительного интервала для математического ожидания нормального распределения при неизвестном σ

Пусть наблюдаемый признак X распределен по нормальному закону и его среднее квадратическое отклонение σ не известно.

Интервал $\left(\bar{x}_в-\frac{t_{k,\gamma}\bar{s}_в}{\sqrt{n}},\bar{x}_в+\frac{t_{k,\gamma}\bar{s}_в}{\sqrt{n}}\right)$ является доверительным интервалом для неизвестного математического ожидания СВ X, соответствующим надежности γ. Величина $t_{k,\gamma}$, зависящая от надежности γ и числа степеней свободы $k=n-1$, определяется по таблицам t – распределения Стьюдента.

В таблице $t_{k,\gamma}$ находится на пересечении столбца, соответствующего γ = 0,95 (0,99; 0,995) и строки, указывающей число степеней свободы,

$\bar{s}_в^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-a\right)^2$.

Пример. СВ X имеет нормальное распределение. Найти доверительный интервал для математического ожидания по выборке: 5, 6, 4, 6, 7, 4, 8, 7, 9, 4 с уровнем значимости 0,05.

$\triangleleft$ По выборке находим $\bar{x}_в$ и $\bar{s}_в$:

$$\bar{x}_в=\frac{1}{10}\left(5+6+4+6+7+4+8+7+9+4\right)=6$$

$$\bar{s}_в^2=\frac{1}{9}\left((4-6)^2\cdot 3+(5-6)^2+(7-6)^2\cdot 2+(8-6)^2+(9-6)^2\right)=3{,}11$$

Найдем $t_{k,\gamma}$ по таблице:

$\gamma=1-0.05=0{,}95$

$t_{k,\gamma}=2{,}26$

Следовательно, доверительный интервал имеет вид

$$\left(6-\frac{2{,}26\cdot 1{,}76}{\sqrt{10}};\ 6+\frac{2{,}26\cdot 1{,}76}{\sqrt{10}}\right),\quad \text{или}\quad (4{,}75;\ 7{,}25).\qquad \triangleright$$