3. Статистические методы анализа данных

Статистические методы полезны как на этапе разведочного, так и на этапе подтверждающего анализа.

Они являются математической основой методов машинного обучения.

Основные задачи статистического анализа

- Оценка неизвестной вероятности события,
- оценка неизвестной функции распределения,
- оценка параметров распределения, вид которого известен,
- оценка зависимости случайной величины (СВ) от другой СВ,
- проверка статистических гипотез (о виде неизвестного распределения, о параметрах распределения и др.).

3.1 Основные понятия статистического анализа

Пусть требуется изучить совокупность однородных объектов относительно некоторого признака *X*, характеризующего эти объекты.

На практике обследование <u>каждого</u> из объектов совокупности проводится редко.

Причины:

- очень большое (бесконечное) число объектов в изучаемой совокупности;
- обследование объекта требует больших материальных затрат;
- обследование объекта связано с его повреждением или разрушением;
- и т. д.

Обычно из всей совокупности случайно отбирают ограниченное число объектов и их подвергают изучению

Генеральная совокупность

Совокупность значений, соответствующих всем возможным результатам отдельных наблюдений, называется ***генеральной совокупностью***.

Распределение СВ X – изучаемой величины – в такой совокупности задается некоторым законом распределения:

- плотностью распределения $f(x)$ для непрерывной СВ;
- вероятностями $p_1 = \mathrm{P}(X = x_1)$, $p_2 = \mathrm{P}(X = x_2)$, … для дискретной СВ.

На практике этот закон распределения обычно <u>неизвестен</u>

Выборка

Результаты случайно выбранных n наблюдений образуют ***выборку объема n***.

Это набор реализаций СВ X

<u>Предположения относительно наблюдений</u>:

- наблюдения независимы;
- результаты наблюдений одинаково распределены.

Набор реализаций одной, а не нескольких СВ с различными распределениями (отсутствие смеси в выборке)

Повторная и бесповторная выборка

Повторной называется выборка, в которой исследованный (измеренный) объект возвращается в генеральную совокупность перед отбором следующего объекта.

Бесповторной называется выборка, в которой исследованный объект в генеральную совокупность не возвращается.

Репрезентативность выборки

Чтобы по данным выборки можно было с уверенностью судить об исследуемом признаке, выборка должна правильно представлять пропорции генеральной совокупности.

Это требование формулируется так:
выборка должна быть *репрезентативной*.

Выборка будет репрезентативной, если <u>все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку</u>.

Варианты и частоты

Пусть из генеральной совокупности (возможных значений признака X), извлечена выборка объема n, в которой

значение x_1 наблюдалось n_1 раз,
значение x_2 – n_2 раз,
… ,
значение x_k – n_k раз, $\sum_{i=1}^{k} n_i = n$.

Значения x_i называются ***вариантами***, числа n_i – ***частотами***, величины $\frac{n_i}{n}$ – ***относительными частотами***.

Вариационный ряд

Последовательность вариант, записанных в возрастающем порядке (с учетом частот каждой варианты), называется ***вариационным рядом***.

Может быть построен для порядковых и количественных признаков. «Возрастание» определяется заданным отношением порядка

Статистический ряд

Статистическим распределением выборки или ***статистическим рядом*** называется перечень вариант (обычно в порядке возрастания) и соответствующих им частот (или относительных частот):

Варианты	x_1	x_2	…	x_k
Частоты	n_1	n_2	…	n_k

(3.1)

Это статистический аналог ряда распределения СВ (вместо вероятностей p_i – частоты n_i или относительные частоты $\frac{n_i}{n}$).

Статистический и вариационный ряды

Пример.

Распределение частот:

Варианты	3	5	7	8
Частоты	7	10	9	4

$n = 30.$

Распределение относительных частот:

Варианты	3	5	7	8
Относительные частоты	$\frac{7}{30}$	$\frac{10}{30}$	$\frac{9}{30}$	$\frac{4}{30}$

Оценки вероятностей $p_i = P(X = x_i)$

Вариационный ряд:

3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8

Группированный статистический ряд

Если результаты наблюдений представляют собой не дискретную, а непрерывную случайную величину, то статистический закон распределения может быть представлен ***группированным статистическим рядом*** (диапазон наблюдавшихся значений делится на частичные интервалы – *разряды*; для каждого интервала n_i – сумма частот вариант, попавших в этот интервал).

Интервалы	(x_0, x_1)	(x_1, x_2)	…	(x_{k-1}, x_k)
Частоты	n_1	n_2	…	n_k

(3.2)

Статистики

Пусть имеется выборка

$$x_1, x_2, \ldots, x_n. \qquad (3.3)$$

Любая функция от данных (элементов выборки) (3.3) называется ***статистикой***.

Замечание.

X – СВ $\Rightarrow$ статистика также является СВ.

Порядковые статистики

Пусть выборка (3.3) упорядочена по возрастанию (построен вариационный ряд выборки):

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}. \qquad (3.4)$$

Статистики, построенные на основе вариационного ряда (3.4), называются ***порядковыми***.

Примером порядковых статистик являются *выборочные квантили*.

Выборочные квантили

Выборочная квантиль порядка α, $\alpha \in (0, 1)$, – это статистика, равная элементу вариационного ряда (3.4) с номером $[\alpha \cdot n]+1$,

где [z] – целая часть z.

Выборочные квантили позволяют получить оценки квантилей СВ *X* по данным выборки

Выборочные квантили

Пример.

Имеется выборка значений СВ X – количество кликов по названию организации за день (данные 2ГИС):

219, 235, 207, 234, 228, 256, 312, 278, 261, 235, 308, 211, 279, 223, 283, 294, 254, 256, 233, 241, 252, 231.

Требуется оценить снизу количество кликов, которое некоторая организация получит с вероятностью 0,9.

Выборочные квантили

Пример (продолжение).

Требуется найти оценку значения t, для которого $P(X>t) = 0,9$.

$P(X \le t) = 1 - P(X>t) = 0,1$

следует получить оценку 10%-ной квантили, т. е. выборочную квантиль порядка 0,1.

$[0,1 \cdot 22]+1 = [2,2]+1 = 3$,

поэтому нужен третий элемент вариационного ряда.

Выборочные квантили

Пример (продолжение).

Упорядоченная по возрастанию выборка:

207, 211, 219, 223, 228, 231, 233, 234, 235, 235, 241, 252, 254, 256, 256, 261, 278, 279, 283, 294, 308, 312.

$x_{(3)} = 219$ с вероятностью 0,9 организация получит больше 219 кликов в день.

Эмпирическая функция распределения

Пусть известно статистическое распределение частот количественного признака X.

Обозначим:

n_x – число наблюдений, в которых зафиксировано значение признака, меньшее x.

Относительная частота события $X < x$ (оценка вероятности $P(X < x)$) равна $\frac{n_x}{n}$.

Является функцией от x.
Находится эмпирическим (опытным) путем

Эмпирическая функция распределения

Эмпирической функцией распределения выборки объема n называется функция

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число наблюдений, меньших x.

Это статистический аналог «теоретической» функции распределения СВ X $F(x) = \mathrm{P}(X < x)$
(и для дискретных, и для непрерывных признаков).

При больших n функция $F^*(x)$ используется для приближенного представления функции $F(x)$.

Эмпирическая функция распределения

Пример 1.

Построим эмпирическую функцию распределения выборки

Варианты	3	5	7	8
Частоты	7	10	9	4

1) Ясно, что $F^*(x) = 0$ при $x \le 3$.
2) Значения признака, меньшие x, при $3 < x \le 5$, наблюдались $n_x = 7$ раз, поэтому для таких x

$$F^*(x) = \frac{7}{30}.$$

Эмпирическая функция распределения

Пример 1 (продолжение).

Варианты	3	5	7	8
Частоты	7	10	9	4

3) При $5 < x \le 7$ $n_x = 7+10 = 17$, поэтому для таких x

$$F^*(x) = \frac{17}{30}.$$

4) При $7 < x \le 8$ $n_x = 7+10+9 = 26$, поэтому для таких x

$$F^*(x) = \frac{26}{30}.$$

5) При $x > 8$ $n_x = 30$, и $F^*(x) = 1$.

Эмпирическая функция распределения

Пример 1 (продолжение).

Итог:

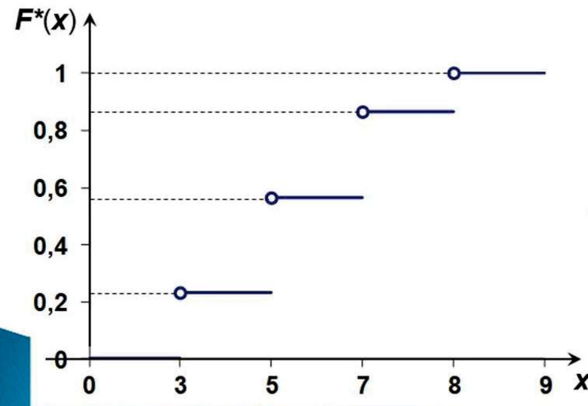
$$F^*(x) = \begin{cases} 0, & x \le 3; \\ \dfrac{7}{30}, & 3 < x \le 5; \\ \dfrac{17}{30}, & 5 < x \le 7; \\ \dfrac{26}{30}, & 7 < x \le 8; \\ 1, & x > 8. \end{cases}$$

Варианты	3	5	7	8
Частоты	7	10	9	4

Эмпирическая функция распределения

Пример 1 (продолжение).

Варианты	3	5	7	8
Частоты	7	10	9	4



$$F^*(x) = \begin{cases} 0, & x \le 3; \\ \dfrac{7}{30}, & 3 < x \le 5; \\ \dfrac{17}{30}, & 5 < x \le 7; \\ \dfrac{26}{30}, & 7 < x \le 8; \\ 1, & x > 8. \end{cases}$$

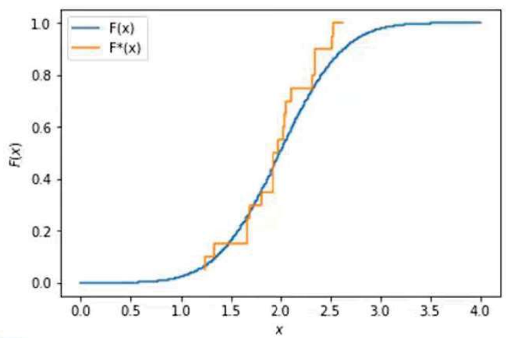
Эмпирическая функция распределения

Пример 2.

Выборка 20 сгенерированных значений СВ, имеющей нормальное распределение с параметрами $m = 2$ и $\sigma = 0.5$;

эмпирическая функция распределения выборки и теоретическая функция распределения нормального закона с указанными параметрами.

2.525, 2.064, 1.661, 1.927,
1.809, 1.976, 1.241, 2.627,
2.117, 2.036, 1.922, 1.336,
2.325, 2.340, 2.345, 1.661,
2.041, 2.518, 1.924, 1.694



Свойства эмпирической функции распределения

1) Для любого x $\quad 0 \le F^*(x) \le 1$.

2) $F^*(x)$ – неубывающая функция.

3) Если x_1 – наименьшая варианта, то
$F^*(x)=0$ при $x \le x_1$;
если x_k – наибольшая варианта, то
$F^*(x)=1$ при $x > x_k$.

$F^*(x)$ – разрывная ступенчатая функция; число скачков равно числу различных значений СВ, полученных в результате наблюдений.

Полигон частот

Полигоном частот называется ломаная, соединяющая точки с координатами (x_1, n_1), (x_2, n_2), … , (x_k, n_k).

Полигоном относительных частот называется ломаная, соединяющая точки с координатами (x_1, p^*_1), (x_2, p^*_2), … , (x_k, p^*_k),

где
$$p_i^* = \frac{n_i}{n} \ .$$

Статистический аналог многоугольника распределения

Очевидно, что для непрерывных признаков при достаточно большом объеме выборки n использование статистического ряда (3.1) становится неудобным (выборка содержит много различных значений, большинство из которых не повторяются).

В таких случаях используют группированный статистический ряд (3.2), который визуализируется с помощью *гистограмм*.

Гистограммы

Позволяют приближенно оценить функцию плотности распределения СВ для непрерывных признаков.

Пусть выборка задана группированным статистическим рядом (3.2).
Обозначим: $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, k$.

Гистограмма частот

Гистограмма частот – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы (x_{i-1}, x_i), $i = 1, 2, \ldots, k$, а высоты равны $\frac{n_i}{h_i}$.

Площадь i–го прямоугольника равна n_i, суммарная площадь всех прямоугольников – объему выборки n.

Гистограмма относительных частот

Гистограмма относительных частот – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы $(x_{i-1},\ x_i)$, $i = 1, 2, \ldots, k$, а высоты равны $\dfrac{n_i}{n \cdot h_i}$.

Статистический аналог кривой распределения

Площадь i–го прямоугольника равна относительной частоте $\dfrac{n_i}{n}$, суммарная площадь всех прямоугольников равна 1.

Гистограммы

Пример.

Гистограммы частот (слева) и относительных частот (справа) двух выборок из 1000 сгенерированных значений СВ, имеющей нормальное распределение с параметрами $m = 2$ и $\sigma = 0.5$