УДК 519.22

Муратов А.Н.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань
Бекасов Р.Д.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань
Нурутдинов А.О.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань
Нурисламов Р.Н.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань
Муродов Э.Р.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань
Нишонов А.Б.
Студент, Казанский национальный исследовательский технологический университет, Россия, г.Казань

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Аннотация: Математическая статистика — наука о математических методах систематизации и использования статистических данных для научных и практических выводов. Во многих своих разделах математическая статистика опирается на теорию вероятностей, позволяющую оценить надежность и точность выводов, делаемых на основании ограниченного статистического материала (напр., оценить необходимый объем выборки для получения результатов требуемой точности при выборочном обследовании).

В теории вероятностей рассматриваются случайные величины с заданным распределением или случайные эксперименты, свойства которых целиком известны. Предмет теории вероятностей — свойства и взаимосвязи этих величин (распределений).

Ключевые слова: статистика, теория вероятностей, эксперимент, гистограмма.

Muratov Arslan Narmuratovich
Student, Kazan National Research Technological University, Russia, Kazan

Bekasov Roman Dmitrievich
Student, Kazan National Research Technological University, Russia, Kazan
Murodov Emil Rustamovich
Student, Kazan National Research Technological University, Russia, Kazan
Nurislamov Ramil Nurmukhamadovich
Student, Kazan National Research Technological University, Russia, Kazan
Nurutdinov Akmal Olegovich
Student, Kazan National Research Technological University, Russia, Kazan
Nishonov Abdulaziz Bakhtiyor ugli
Student, Kazan National Research Technological University, Russia, Kazan

MATHEMATICAL STATISTICS

Abstract: Mathematical statistics is the science of mathematical methods of systematizing and using statistical data for scientific and practical conclusions. In many of its sections, mathematical statistics are based on probability theory, which makes it possible to assess the reliability and accuracy of conclusions made on the basis of limited statistical material (e.g., estimate the required sample size to obtain the results of the required accuracy in a sample survey).

Probability theory considers random variables with a given distribution or random experiments whose properties are entirely known. The subject of probability theory is the properties and relationships of these quantities (distributions).

Keywords: statistics, probability theory, experiment, histogram.

Математическая статистика — наука о математических методах анализа данных, полученных при проведении массовых наблюдений (измерений, опытов). В зависимости от математической природы конкретных результатов наблюдений статистика математическая делится на статистику чисел, многомерный статистический анализ, анализ функций (процессов) и

временных рядов, статистику объектов нечисловой природы. Существенная часть статистики математической основана на вероятностных моделях. Выделяют общие задачи описания данных, оценивания и проверки гипотез. Рассматривают и более частные задачи, связанные с проведением выборочных обследований, восстановлением зависимостей, построением и использованием классификаций (типологий) и др.

Для описания данных строят таблицы, диаграммы, иные наглядные представления, например, корреляционные поля. Вероятностные модели обычно не применяются. Некоторые методы описания данных опираются на продвинутую теорию и возможности современных компьютеров. К ним относятся, в частности, кластер-анализ, нацеленный на выделение групп объектов, похожих друг на друга, и многомерное шкалирование, позволяющее наглядно представить объекты на плоскости, в наименьшей степени исказив расстояния между ними.

Методы оценивания и проверки гипотез опираются на вероятностные модели порождения данных. Эти модели делятся на параметрические и непараметрические. В параметрических моделях предполагается, что изучаемые объекты описываются функциями распределения, зависящими от небольшого числа (1-4) числовых параметров. В непараметрических моделях функции распределения предполагаются произвольными непрерывными. В статистике математической оценивают параметры и характеристики распределения (математическое ожидание, медиану, дисперсию, квантили и др.), плотности и функции распределения, зависимости между переменными (на основе линейных и непараметрических коэффициентов корреляции, а также параметрических или непараметрических оценок функций, выражающих зависимости) и др. Используют точечные и интервальные (дающие границы для истинных значений) оценки.

В математической статистике есть общая теория проверки гипотез и большое число методов, посвященных проверке конкретных гипотез. Рассматривают гипотезы о значениях параметров и характеристик, о

проверке однородности (то есть о совпадении характеристик или функций распределения в двух выборках), о согласии эмпирической функции распределения с заданной функцией распределения или с параметрическим семейством таких функций, о симметрии распределения и др.

Большое значение имеет раздел математической статистики, связанный с проведением выборочных обследований, со свойствами различных схем организации выборок и построением адекватных методов оценивания и проверки гипотез.

Задачи восстановления зависимостей активно изучаются более 200 лет, с момента разработки К. Гауссом в 1794 г. метода наименьших квадратов. В настоящее время наиболее актуальны методы поиска информативного подмножества переменных и непараметрические методы.

Разработка методов аппроксимации данных и сокращения размерности описания была начата более 100 лет назад, когда К. Пирсон создал метод главных компонент. Позднее были разработаны факторный анализ [1] и многочисленные нелинейные обобщения.

Различные методы построения (кластер-анализ), анализа и использования (дискриминантный анализ) классификаций (типологий) именуют также методами распознавания образов (с учителем и без), автоматической классификации и др.

Математические методы в статистике основаны либо на использовании сумм (на основе Центральной Предельной Теоремы теории вероятностей) или показателей различия (расстояний, метрик), как в статистике объектов нечисловой природы. Строго обоснованы обычно лишь асимптотические результаты. В настоящее время компьютеры играют большую роль в математической статистике. Они используются как для расчетов, так и для имитационного моделирования (в частности, в методах размножения выборок и при изучении пригодности асимптотических результатов).

Основные понятия выборочного метода

Пусть $\xi : \Omega \to \mathbb{R}$ — случайная величина, наблюдаемая в случайном эксперименте. Предполагается, что вероятностное пространство задано (и не будет нас интересовать).

Будем считать, что, проведя n раз этот эксперимент в одинаковых условиях, мы получили числа X_1, X_2, ..., X_n — значения этой случайной величины в первом, втором, и т.д. экспериментах. Случайная величина ξ имеет некоторое распределение $\mathcal{F}$, которое нам частично или полностью неизвестно.

Рассмотрим подробнее набор $\mathbf{X} = (X_1, \ldots, X_n)$, называемый выборкой.

В серии уже произведенных экспериментов выборка — это набор чисел. Но если эту серию экспериментов повторить еще раз, то вместо этого набора мы получим новый набор чисел. Вместо числа X_1 появится другое число — одно из значений случайной величины ξ. То есть X_1 (и X_2, и X_3, и т.д.) — переменная величина, которая может принимать те же значения, что и случайная величина ξ, и так же часто (с теми же вероятностями). Поэтому до опыта X_1 — случайная величина, одинаково распределенная с ξ, а после опыта — число, которое мы наблюдаем в данном первом эксперименте, т.е. одно из возможных значений случайной величины X_1.

Выборка $\mathbf{X} = (X_1, \ldots, X_n)$ объема n — это набор из n независимых и одинаково распределенных случайных величин («копий ξ»), имеющих, как и ξ, распределение $\mathcal{F}$.

Что значит «по выборке сделать вывод о распределении»? Распределение характеризуется функцией распределения, плотностью или таблицей, набором числовых характеристик — $\mathsf{E}\xi$, $\mathsf{D}\xi$, $\mathsf{E}\xi^k$ и т.д. По выборке нужно уметь строить приближения для всех этих характеристик.

Выборочное распределение.

Рассмотрим реализацию выборки на одном элементарном исходе ω_0 — набор чисел $X_1 = X_1(\omega_0)$, ..., $X_n = X_n(\omega_0)$. На подходящем вероятностном пространстве введем случайную величину ξ^*, принимающую значения X_1,

$\dots$, X_n с вероятностями по $1/n$ (если какие-то из значений совпали, сложим вероятности соответствующее число раз). Таблица распределения вероятностей и функция распределения случайной величины ξ^* выглядят так:

ξ	X_1	$\dots$	X_n
P	$1/n$	$\dots$	$1/n$

$$F_n^*(y) = \sum_{X_i<y} \frac{1}{n} = \frac{\text{количество } X_i \in (-\infty, y)}{n}.$$

Распределение величины ξ^* называют эмпирическим или выборочным распределением. Вычислим математическое ожидание и дисперсию величины ξ^* и введем обозначения для этих величин:

$$E\xi^* = \sum_{i=1}^{n} \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}, \quad D\xi^* = \sum_{i=1}^{n} \frac{1}{n} (X_i - E\xi^*)^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = S^2.$$

Точно так же вычислим и момент порядка k

$$E\,(\xi^*)^k = \sum_{i=1}^{n} \frac{1}{n} X_i^k = \frac{1}{n} \sum_{i=1}^{n} X_i^k = \overline{X^k}.$$

В общем случае обозначим через $\overline{g(X)}$ величину

$$E\,g(\xi^*) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \overline{g(X)}.$$

Если при построении всех введенных нами характеристик считать выборку X_1 , $\dots$, X_n набором случайных величин, то и сами эти характеристики — $F_n^*(y)$, $\overline{X}$, S^2, $\overline{X^k}$, $\overline{g(X)}$ — станут величинами случайными. Эти характеристики выборочного распределения используют для оценки (приближения) соответствующих неизвестных характеристик истинного распределения.

Причина использования характеристик распределения ξ^* для оценки характеристик истинного распределения ξ (или X_1) — в близости этих распределений при больших n.

Рассмотрим, для примера, n подбрасываний правильного кубика. Пусть $X_i \in \{1, \ldots, 6\}$ — количество очков, выпавших при i-м броске, $i = 1, \ldots, n$. Предположим, что единица в выборке встретится n_1 раз, двойка — n_2 раз и т.д. Тогда случайная величина ξ^* будет принимать значения **1**, … , **6** с вероятностями $\frac{n_1}{n}$, …, $\frac{n_6}{n}$ соответственно. Но эти пропорции с ростом n приближаются к $1/6$ согласно закону больших чисел. То есть распределение величины ξ^* в некотором смысле сближается с истинным распределением числа очков, выпадающих при подбрасывании правильного кубика.

Мы не станем уточнять, что имеется в виду под близостью выборочного и истинного распределений. В следующих параграфах мы подробнее познакомимся с каждой из введенных выше характеристик и исследуем ее свойства, в том числе ее поведение с ростом объема выборки.

Эмпирическая функция распределения, гистограмма

Поскольку неизвестное распределение $\mathcal{F}$ можно описать, например, его функцией распределения $F(y) = P(X_1 < y)$, построим по выборке «оценку» для этой функции.

Определение 1.

Эмпирической функцией распределения, построенной по выборке $\mathbf{X} = (X_1, \ldots, X_n)$ объема n, называется случайная функция $F_n^* : \mathbb{R} \times \Omega \to [0, 1]$, при каждом $y \in \mathbb{R}$ равная

$$F_n^*(y) = \frac{\text{количество } X_i \in (-\infty, y)}{n} = \frac{1}{n} \sum_{i=1}^{n} I(X_i < y).$$

Напоминание: Случайная функция

$$I(X_i < y) = \begin{cases} 1, & \text{если } X_i < y, \\ 0 & \text{иначе} \end{cases}$$

называется индикатором события $\{X_i < y\}$. При каждом y это — случайная величина, имеющая распределение Бернулли с параметром $p = P(X_i < y) = F(y)$. почему?

Иначе говоря, при любом y значение $F(y)$, равное истинной вероятности случайной величине X_1 быть меньше y, оценивается долей элементов выборки, меньших y.

Если элементы выборки X_1, ..., X_n упорядочить по возрастанию (на каждом элементарном исходе), получится новый набор случайных величин, называемый вариационным рядом:

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n-1)} \leqslant X_{(n)}.$$

Здесь

$$X_{(1)} = \min\{X_1, \ldots, X_n\}, \quad X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

Элемент $X_{(k)}$, $k = 1, \ldots, n$, называется k-м членом вариационного ряда или k-й порядковой статистикой.

Пример 1.

Выборка: $\mathbf{X} = (0;2;1;2,6;3,1;4,6;1;4,6;6;2,6;6;7;9;\ 9;2,6)$.

Вариационный ряд: $(0;1;1;2;2,6;2,6;2,6;3,1;4,6;4,6;6;\ 6;7;9;9)$.

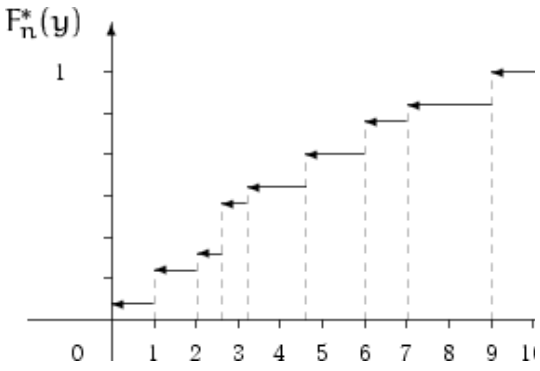


Рисунок 1 - Пример 1

Эмпирическая функция распределения имеет скачки в точках выборки, величина скачка в точке X_i равна m/n, где m — количество элементов выборки, совпадающих с X_i.

Можно построить эмпирическую функцию распределения по вариационному ряду:

$$F_n^*(y) = \begin{cases} 0, & \text{если } y \leqslant X_{(1)}, \\ \dfrac{k}{n}, & \text{если } X_{(k)} < y \leqslant X_{(k+1)}, \\ 1 & \text{при } y > X_{(n)}. \end{cases}$$

Другой характеристикой распределения является таблица (для дискретных распределений) или плотность (для абсолютно непрерывных). Эмпирическим, или выборочным аналогом таблицы или плотности является так называемая гистограмма.

Гистограмма строится по группированным данным. Предполагаемую область значений случайной величины ξ(или область выборочных данных) делят независимо от выборки на некоторое количество интервалов (не обязательно одинаковых). Пусть A_1, ..., A_k — интервалы на прямой, называемые интервалами группировки. Обозначим для $j = 1, \ldots, k$ через ν_j число элементов выборки, попавших в интервал A_j:

$$\nu_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^{n} I(X_i \in A_j), \quad \text{здесь} \quad \sum_{j=1}^{k} \nu_j = n. \qquad (1)$$

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна ν_j. Общая площадь всех прямоугольников должна равняться единице. Пусть l_j — длина интервала A_j. Высота f_j прямоугольника над A_j равна

$$f_j = \frac{\nu_j}{n l_j}.$$

Полученная фигура называется гистограммой.

Пример 2.

Имеется вариационный ряд (см. пример 1):

$(0;1;1;2;2,6;2,6;2,6;3,1;4,6;\ 4,6;6;6;7;9;9).$

Разобьем отрезок $[0,10]$ на 4 равных отрезка. В отрезок $A_1 = [0;2,5)$ попали 4 элемента выборки, в $A_2 = [2,5;5)$ — 6, в $A_3 = [5;7,5)$ — 3, и в отрезок $A_4 = [7,5;10]$ попали 2 элемента выборки. Строим гистограмму (рис. 2). На рис. 3 — тоже гистограмма для той же выборки, но при разбиении области на 5 равных отрезков.

Рис. 2. Пример 2

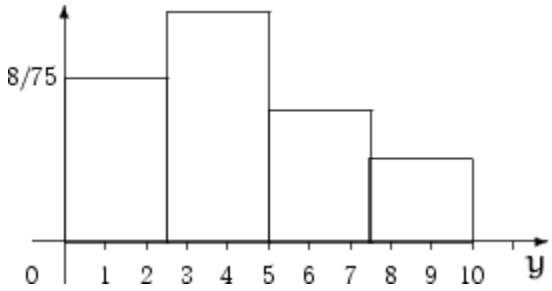
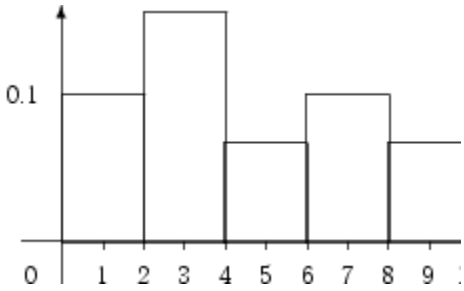


Рис. 3. Пример 2



Замечание 1.

В курсе «Эконометрика» утверждается, что наилучшим числом интервалов группировки («формула Стерджесса») является $k = k(n) = 1 + [3.322 \lg n]$.

Здесь $\lg n$ — десятичный логарифм, поэтому $k = 1 + [\log_2 10 \log_{10} n] = 1 + [\log_2 n]$, т.е. при увеличении выборки вдвое число интервалов группировки увеличивается на 1. Заметим, что чем больше интервалов группировки, тем лучше. Но, если брать число интервалов, скажем, порядка n, то с ростом n гистограмма не будет приближаться к плотности.

Справедливо следующее утверждение:

Если плотность распределения элементов выборки является непрерывной функцией, то при $k(n) \to \infty$ так, что $k(n)/n \to 0$, имеет место поточечная сходимость по вероятности гистограммы к плотности.

Так что выбор логарифма разумен, но не является единственно возможным.

Математическая (или теоретическая) статистика опирается на методы и понятия теории вероятностей, но решает в каком-то смысле обратные задачи.

Если мы наблюдаем одновременно проявление двух (или более) признаков, т.е. имеем набор значений нескольких случайных величин — что можно сказать об их зависимости? Есть она или нет? А если есть, то какова эта зависимость?

Часто бывает возможно высказать некие предположения о распределении, спрятанном в «черном ящике», или о его свойствах. В этом случае по опытным данным требуется подтвердить или опровергнуть эти предположения («гипотезы»). При этом надо помнить, что ответ «да» или «нет» может быть дан лишь с определенной степенью достоверности, и чем дольше мы можем продолжать эксперимент, тем точнее могут быть выводы. Наиболее благоприятной для исследования оказывается ситуация, когда можно уверенно утверждать о некоторых свойствах наблюдаемого эксперимента — например, о наличии функциональной зависимости между наблюдаемыми величинами, о нормальности распределения, о его симметричности, о наличии у распределения плотности или о его дискретном характере, и т.д.

Итак, о (математической) статистике имеет смысл вспоминать, если

- имеется случайный эксперимент, свойства которого частично или полностью неизвестны,
- мы умеем воспроизводить этот эксперимент в одних и тех же условиях некоторое (а лучше — какое угодно) число раз.

СПИСОК ЛИТЕРАТУРЫ

1. Баумоль У. Экономическая теория и исследование операций. – М.; Наука, 1999.
2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1995.
3. Боровков А.А. Математическая статистика. М.: Наука, 1994.
4. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. - СПБ: Издательство «Лань», 2003.
5. Коршунов Д.А., Чернова Н.И. Сборник задач и упражнений по математической статистике. Новосибирск: Изд-во Института математики им. С.Л.Соболева СО РАН, 2001.
6. Пехелецкий И.Д. Математика: учебник для студентов. - М.: Академия, 2003.
7. Суходольский В.Г. Лекции по высшей математике для гуманитариев. - СПБ Издательство Санкт-петербургского государственного университета. 2003
8. Феллер В. Введение в теорию вероятностей и ее приложения. - М.: Мир, Т.2, 1984.
9. Харман Г., Современный факторный анализ. — М.: Статистика, 1972.