Александр Орлов. Математика случая

Книга написан с целью преодоления разрыва между курсами по теории вероятностей и математической статистике и практическими потребностями специалистов широкого профиля, использующих статистические методы. Рассмотрены все основные понятия, используемые при применении современных статистических методов. Особое внимание уделено непараметрическим подходам и статистике нечисловых данных. Книга адресована всем, кому необходимо в сжатые сроки овладеть понятийной базой статистических методов.

Александр Иванович Орлов. Математика случая. Вероятность и статистика – основные факты. – М.: МЗ-Пресс, 2004. – 110 с.

Глава 1. Вероятность и статистика нужны всем

По типу решаемых задач математическая статистика обычно делится на три раздела: описание данных, оценивание и проверка гипотез.

По виду обрабатываемых статистических данных математическая статистика делится на четыре направления:

- одномерная статистика (статистика случайных величин), в которой результат наблюдения описывается действительным числом;
- многомерный статистический анализ, где результат наблюдения над объектом описывается несколькими числами (вектором);
- статистика случайных процессов и временных рядов, где результат наблюдения – функция;
- статистика объектов нечисловой природы, в которой результат наблюдения имеет нечисловую природу, например, является множеством (геометрической фигурой), упорядочением или получен в результате измерения по качественному признаку.

Глава 2. Основы теории вероятностей

Теория вероятностей	Теория множеств
Пространство элементарных событий	Множество
Элементарное событие	Элемент этого множества
Событие	Подмножество
Достоверное событие	Подмножество, совпадающее с множеством
Невозможное событие	Пустое подмножество $\varnothing$
Сумма $A+B$ событий A и B	Объединение $A \cup B$
Произведение AB событий A и B	Пересечение $A \cap B$
Событие, противоположное A	Дополнение A
События A и B несовместны	$A \cap B$ пусто
События A и B совместны	$A \cap B$ не пусто

Рис. 1. Соответствие терминов теории вероятностей и теории множеств

Определение в рамках аксиоматического подхода на базе математической модели, предложенной А.Н. Колмогоровым. Пусть конечное множество Ω = {ω} является пространством элементарных событий, соответствующим некоторому опыту. Пусть каждому ω ∈ Ω поставлено в соответствие неотрицательное число P(ω), называемое вероятностью элементарного события ω, причем сумма вероятностей всех элементарных событий равна 1, т.е.

$$(1)\ \sum_{\omega\in\Omega} P(\omega) = 1$$

Тогда пара {Ω,P}, состоящая из конечного множества Ω и неотрицательной функции P, определенной на Ω и удовлетворяющей условию (1), называется *вероятностным пространством*. Вероятность события A равна сумме вероятностей элементарных событий, входящих в A, т.е. определяется равенством

$$(2)\ P(A) = \sum_{\omega\in\Omega} P(\omega)$$

Случайная величина – это функция, определенная на пространстве элементарных событий ω. Всегда наблюдается лишь *реализация случайной величины*, т.е. ее значение, соответствующее именно тому элементарному исходу опыта (элементарному событию), которое осуществилось в конкретной реальной ситуации. А функция от элементарного события – это теоретическое понятие, основа вероятностной модели реального явления или процесса.

Математическим ожиданием случайной величины X называется число

$$(3)\ M(X) = \sum_{\omega\in\Omega} X(\omega)\cdot P(\omega)$$

Т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям соответствующих элементарных событий.

Математическое ожидание показывает, вокруг какой точки группируются значения случайной величины. Необходимо также уметь измерить изменчивость случайной величины относительно математического ожидания. $M[(X - a)^2]$ достигает минимума по a при $a = M(X)$. Поэтому за показатель изменчивости случайной величины естественно взять именно $M[(X - M(X))^2]$.

Дисперсией случайной величины X называется число $\sigma^2 = D(X) = M[(X - M(X))^2]$.

Если случайные величины X и У независимы, то дисперсия их суммы X+У равна сумме дисперсий: D(X+Y) = D(X) + D(Y).

Пусть $X_1, X_2, ..., X_k$ – попарно независимые случайные величины. Пусть Y_k – их сумма, $Y_k = X_1+X_2+...+X_k$. Тогда математическое ожидание суммы равно сумме математических ожиданий слагаемых, $M(Y_k) = M(X_1)+ M(X_2)+...+M(X_k)$, дисперсия суммы равна сумме дисперсий слагаемых, $D(Y_k) = D(X_1)+D(X_2)+… +D(X_k)$.

Эти утверждения играют важную роль при изучении выборочных характеристик, поскольку результаты наблюдений, включенные в выборку, рассматриваются в математической статистике как реализации независимых случайных величин.

Теорема Бернулли. Пусть m – число наступлений события A в k независимых (попарно) испытаниях, и p – вероятность наступления события A в каждом из испытаний. Тогда при любом ε > 0 справедливо неравенство

$$(4)\ P\left\{\left|\frac{m}{k} - p\right| \geq \varepsilon\right\} \leq \frac{p(1-p)}{k\varepsilon^2}$$

Проверка статистических гипотез. Одна из основных характеристик метода проверки гипотезы – уровень значимости, т.е. вероятность отвергнуть проверяемую (нулевую) гипотезу H_0, когда она верна. Если для выбранного уровня значимости α вероятность отвергнуть ее меньше α, то гипотезу отвергают, как говорят, на уровне значимости α. Если эта вероятность больше или равна α, то гипотезу не отвергают. Обычно в вероятностно-статистических методах принятия решений выбирают α = 0,05, реже α = 0,01 или α = 0,1.

Теорема Бернулли позволяет оценить неизвестную вероятность p числом m/k, поскольку доказано, что при возрастании k вероятность того, что m/k отличается от p более чем на какое-либо фиксированное число, приближается к 0.

Глава 3. Суть вероятностно-статистических методов

Для принятия решений опираются на вероятностную модель реального явления или процесса, т.е. математическую модель, в которой объективные соотношения выражены в терминах теории вероятностей. Математическая статистика решает обратную задачу по отношению к теории вероятностей. Ее цель – на основе результатов наблюдений получить выводы о вероятностях, лежащих в основе вероятностной модели. Например, на основе частоты появления дефектных изделий в контрольной выборке можно сделать выводы о вероятности дефектности всей партии.

Используются два параллельных ряда понятий – относящиеся к теории (вероятностной модели) и относящиеся к практике (выборке результатов наблюдений). Например, теоретической вероятности соответствует частота, найденная по выборке. Математическому ожиданию (теоретический ряд) соответствует выборочное среднее арифметическое (практический ряд). Как правило, выборочные характеристики являются оценками теоретических. При этом величины, относящиеся к теоретическому ряду, «находятся в головах исследователей», недоступны для непосредственного измерения.

Чтобы перенести выводы с выборки на более обширную совокупность, необходимы те или иные предположения о связи выборочных характеристик с характеристиками этой более обширной совокупности. Эти предположения основаны на соответствующей вероятностной модели.

Итак, использование вероятностных моделей на основе оценивания и проверки гипотез с помощью выборочных характеристик – вот суть вероятностно-статистических методов принятия решений.

Глава 4. Случайные величины и их распределения

Распределение числовой случайной величины – это функция, которая однозначно определяет вероятность того, что случайная величина принимает заданное значение или принадлежит к некоторому заданному интервалу.

Распределение может быть задано с помощью функции распределения $F(x) = P(X<x)$, определяющей для всех действительных x вероятность того, что случайная величина X принимает значения, меньшие x.

Характеристики положения указывают на «центр» распределения. Большое значение в статистике имеет квантиль порядка p = ½. Он называется медианой случайной величины X или ее функции распределения F(x) и обозначается Me(X). Медиана характеризует случайную величину лучше, чем математическое ожидание.

Характеристики разброса: дисперсия $D(X) = \sigma^2$, среднее квадратическое отклонение σ и коэффициенту вариации ν. Коэффициент вариации – это отношение среднего квадратического отклонения к математическому ожиданию:

$$(5)\ \nu = \frac{\sigma}{M(X)}$$

Коэффициент вариации измеряет разброс в относительных единицах, в то время как среднее квадратическое отклонение – в абсолютных.

Преобразования случайных величин

По каждой случайной величине X определяют еще три величины. Центрированная случайная величина Y – это разность между данной случайной величиной X и ее математическим ожиданием M(X), т.е. Y = X – M(X). Нормированная случайная величина V – это отношение данной случайной величины X к ее среднему квадратическому отклонению σ, т.е. V = X/σ. Приведенная случайная величина U – это центрированная и нормированная случайная величина:

$$(6)\ U = \frac{X - M(X)}{\sigma}$$

Для приведенной случайной величины M(U) = 0, D(U) = 1.

С каждой случайной величиной X можно связать множество случайных величин Y, заданных формулой $Y = aX + b$. Вместо $Y = aX + b$ часто используют запись

$$(7)\ Y = \frac{X - c}{d}$$

Число c называют параметром сдвига, а число d – параметром масштаба. Формула (7) показывает, что X – результат измерения некоторой величины – переходит в Y – результат измерения той же величины, если начало измерения перенести в точку c, а затем использовать новую единицу измерения, в d раз большую старой.

Для масштабно-сдвигового семейства (7) распределение Y называют стандартным.

Моменты случайных величин

При обработке данных используют такие характеристики случайной величины X как моменты порядка q, т.е. математические ожидания случайной величины X^q, q = 1, 2, ... Так, само математическое ожидание – это момент порядка 1. Для дискретной случайной величины момент порядка q может быть рассчитан как

$$(8)\ m_q = M(X^q) = \sum_i x_i^q P(X = x_i)$$

Для непрерывной случайной величины

$$(9)\ m_q = M(X^q) = \int_{-\infty}^{\infty} x_i^q f(x)dx$$

Моменты порядка q называют также начальными моментами порядка q, в отличие от родственных характеристик – центральных моментов порядка q, задаваемых формулой

$$(10)\ \mu_q = M[(X - M(X))^q)], q = 2, 3, \dots$$

Так, дисперсия – это центральный момент порядка 2.

Центральная предельная теорема

Пусть X_1, X_2, ..., X_n – независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, i = 1, 2, ..., n. Тогда $M(X_1 + X_2 + ... + X_n) = nm$, $D(X_1 + X_2 + ... + X_n) = n\sigma^2$.

Рассмотрим приведенную случайную величину U_n для суммы $X_1 + X_2 + ... + X_n$, а именно,

$$(11)\ U_n = \frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}}$$

По аналогии с (6), $M(U_n) = 0$, $D(U_n) = 1$.

Центральная предельная теорема утверждает, что для любого x существует предел

$$(12)\ \lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \dots + X_n - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x)$$

где Φ(x) – интегральная функция стандартного нормального распределения.

Центральная предельная теорема (ЦПТ) носит свое название по той причине, что она является центральным, наиболее часто применяющимся математическим результатом теории вероятностей и математической статистики. Суть ЦПТ – независимо от вида распределения X, средние значения достаточно больших выборок из X (X_1, X_2, ..., X_n в пределе при $n\to\infty$) распределены нормально.

Семейство нормальных распределений

Для центральных моментов третьего и четвертого порядка нормального распределения справедливы равенства $\mu_3 = 0$, $\mu_4 = \sigma^4$. Эти равенства лежат в основе классических методов проверки того, что результаты наблюдений подчиняются нормальному распределению. В настоящее время нормальность обычно рекомендуется проверять по критерию W Шапиро – Уилка.

Распределения Пирсона (хи-квадрат), Стьюдента и Фишера

С помощью нормального распределения определяются еще три весьма распространенных распределения.

Распределение Пирсона χ^2 (хи - квадрат) – распределение случайной величины X

$$(13)\ X = X_1^2 + X_2^2 + \cdots + X_n^2$$

где случайные величины X_1, X_2, ..., X_n независимы и имеют одно и тоже нормальное распределение N(0,1) с μ = 0 и σ = 1. При этом число слагаемых n, называется *числом степеней свободы*.

Распределение хи-квадрат широко используют при оценивании дисперсии (с помощью доверительного интервала), при проверке гипотез согласия, однородности, независимости и др.

Распределение t Стьюдента – это распределение случайной величины

$$(14)\ T = \frac{U\sqrt{n}}{\sqrt{X}}$$

где случайные величины U и X независимы, U имеет распределение стандартное нормальное распределение $N(0,1)$, а X – распределение хи-квадрат с n степенями свободы.

Распределение Стьюдента – одно из наиболее известных распределений среди используемых при анализе реальных данных. Его применяют при оценивании математического ожидания, прогнозного значения и других характеристик с помощью доверительных интервалов, по проверке гипотез о значениях математических ожиданий, коэффициентов регрессионной зависимости, гипотез однородности выборок и т.д.

Распределение Фишера – это распределение случайной величины

$$(15)\ F = \frac{\frac{1}{k_1}X_1}{\frac{1}{k_2}X_2}$$

где случайные величины X_1 и X_2 независимы и имеют распределения хи-квадрат с числом степеней свободы k_1 и k_2 соответственно.

Распределение Фишера используют при проверке гипотез об адекватности модели в регрессионном анализе, о равенстве дисперсий и в других задачах прикладной статистики.

Кроме семейства нормальных распределений, широко используют логарифмически нормальные распределения, экспоненциальные, Вейбулла-Гнеденко, гамма-распределения.

Логарифмически нормальные распределения

Случайная величина X имеет логарифмически нормальное распределение, если случайная величина Y = lgX имеет нормальное распределение.

Из центральной предельной теоремы следует, что произведение $X = X_1X_2...X_n$ независимых положительных случайных величин X_i при больших n можно аппроксимировать логарифмически нормальным распределением.

Экспоненциальные распределения

Рассмотрим *поток событий*, т.е. последовательность событий, происходящих одно за другим в какие-то моменты времени. Например, поток вызовов на телефонной станции. В теории потоков событий справедлива теорема, аналогичная центральной предельной теореме, но в ней речь идет не о суммировании случайных величин, а о суммировании потоков событий. Рассматривается суммарный поток, составленный из большого числа независимых потоков. Например, вызовов, поступающих на телефонную станцию от отдельных абонентов. Доказано, что в случае, когда характеристики потоков не зависят от времени, суммарный поток полностью описывается одним числом λ – интенсивностью потока. Для суммарного потока рассмотрим случайную величину X – длину промежутка времени между последовательными событиями. Ее функция распределения:

$$(16)\ F(x;\lambda) = P(X \le x) = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Это распределение называется экспоненциальным распределением, т.к. в формуле (16) участвует экспоненциальная функция. Величина 1/λ – масштабный параметр. Функция экспоненциального распределения F(x, λ) и его плотность f(x, λ) связаны соотношением

$$(17)\ f(x;\lambda) = \lambda \cdot (1 - F(x;\lambda))$$

Распределения Вейбулла – Гнеденко

Пусть X – случайная величина, характеризующая длительность функционирования изделия, предприятия или жизни живого существа. Важную роль играет интенсивность отказа

$$(18)\ \lambda(x) = \frac{f(x)}{1 - F(x)}$$

где $F(x)$ и $f(x)$ – функция распределения и плотность случайной величины X.

Распределение Вейбулла – Гнеденко

$$(19)\ F(x) = \begin{cases} 1 - e^{-\lambda_0 x^b}, x \geq 0, \\ 0, \qquad\quad x < 0. \end{cases}$$

Экспоненциальное распределение – частный случай распределения Вейбулла – Гнеденко при b = 1.

Распределение Вейбулла – Гнеденко применяется также при построении вероятностных моделей ситуаций, в которых поведение объекта определяется наиболее слабым звеном. Пусть $X_1, X_2, \ldots, X_n$ – независимые одинаково распределенные случайные величины, $X(1)$ = min ($X_1, X_2, \ldots, X_n$), а $X(n)$ = max ($X_1, X_2, \ldots, X_n$).

В ряде прикладных задач большую роль играют $X(1)$ и $X(n)$, в частности, при исследовании максимально возможных значений (рекордов) тех или иных значений, например, страховых выплат или потерь из-за коммерческих рисков, при изучении пределов упругости и выносливости стали, ряда характеристик надежности и т.п. Показано, что при больших n распределения $X(1)$ и $X(n)$ хорошо описываются распределениями Вейбулла – Гнеденко.

Наиболее часто используют три семейства дискретных распределений: биномиальные, гипергеометрические и Пуассона.

Биномиальное распределение

Пусть общее число независимых испытаний n. В каждом из них с вероятностью p появляется событие A. Тогда число испытаний Y, в которых появилось событие A, имеет биномиальное распределение. Для него вероятность принятия случайной величиной Y значения y определяется формулой

$$(20)\ P(Y = y|p,n) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0, 1, 2, \ldots, n$$

где

$$(21)\ \binom{n}{y} = \frac{n!}{y!\,(n-y)!} = C_n^y -$$

– число сочетаний из n элементов по y, известное из комбинаторики. Для всех y, кроме $0, 1, 2, \ldots, n$, имеем $P(Y=y)=0$. Биномиальное распределение при фиксированном объеме выборки n задается параметром p, т.е. биномиальные распределения образуют однопараметрическое семейство. Они применяются при изучении предпочтений потребителей, выборочном контроле качества продукции, в демографии, социологии, медицине, биологии и др.

Характеристики биномиального распределения: $M(Y) = np$, $D(Y) = np(1 - p)$.

Гипергеометрическое распределение

Используется при выборочном контроле конечной совокупности объектов объема N по альтернативному признаку. Каждый контролируемый объект классифицируется либо как обладающий признаком A, либо как не обладающий этим признаком. Гипергеометрическое распределение имеет случайная величина Y, равная числу объектов, обладающих признаком A в случайной выборке объема n, где n < N. Для гипергеометрического распределения вероятность принятия случайной величиной Y значения y имеет вид

$$(22)\ P(Y = y|N, d, n) = \frac{\binom{n}{y}\binom{N-n}{D-y}}{\binom{N}{D}}$$

где D – число объектов, обладающих признаком A, в рассматриваемой совокупности объема N. При этом y принимает значения от $max\{0,\ n - (N - D)\}$ до $min\{n, D\}$, при прочих y вероятность в формуле (22) равна 0. Таким образом, гипергеометрическое распределение определяется тремя параметрами – объемом генеральной совокупности N, числом объектов D в ней, обладающих рассматриваемым признаком A, и объемом выборки n.

Для рассматриваемого типа выборок используются термин *случайная выборка без возвращения*.

При $N > 10n$ гипергеометрическое распределение аппроксимируется биномиальным.

Распределение Пуассона

Случайная величина Y имеет распределение Пуассона, если

$$(23)\ P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, y = 0, 1, 2, \ldots$$

где λ – параметр распределения Пуассона, и P(Y=y)=0 для всех прочих y (0! =1). Для распределения Пуассона M(Y) = λ, D(Y) = λ.

Распределение Пуассона является предельным случаем биномиального распределения, когда вероятность p осуществления события мала, но число испытаний n велико, причем np = λ. Поэтому распределение Пуассона также называют *законом редких событий*.

Глава 5. Основные проблемы прикладной статистики – описание данных, оценивание и проверка гипотез

Статистические данные – это результаты наблюдений (измерений, испытаний, опытов, анализов). Функции результатов наблюдений, используемые, в частности, для оценки параметров распределений и (или) для проверки статистических гипотез, называют *статистиками*. Статистики, являющиеся выборочными аналогами характеристик случайных величин (математического ожидания, медианы, дисперсии, моментов и др.) и используемые для оценивания этих характеристик, называют *статистическими характеристиками*.

Эмпирической функцией распределения $F_n(x)$ называется доля элементов выборки, меньших x. Чтобы записать выражение для эмпирической функции распределения в виде формулы, введем функцию $c(x, y)$ двух переменных:

$$(24)\ c(x, y) = \begin{cases} 0, x \le 0, \\ 1, x > 0. \end{cases}$$

Случайные величины, моделирующие результаты наблюдений, обозначим $X_1(\omega)$, $X_2(\omega)$, ..., $X_n(\omega)$, $\omega \in \Omega$. Тогда эмпирическая функция распределения $F_n(x)$ имеет вид

$$(25)\ F_n(x) = F_n(x, \omega) = \frac{1}{n} \sum_{1 \le i \le n} c(x, X_i(\omega))$$

Из закона больших чисел следует, что для каждого действительного числа x эмпирическая функция распределения $F_n(x)$ сходится к функции распределения $F(x)$ результатов наблюдений при n→∞. В.И. Гливенко доказал, что сходимость равномерна по x, т.е.

$$(26)\ \sup_x |F_n(x) - F(x)| \to 0$$

при n→∞ (сходимость по вероятности).

Для функции $g(x)$ под $supg(x)$ понимают наименьшее x из чисел a таких, что $g(x) \le a$ при всех x. Если функция $g(x)$ достигает максимума в точке x_0, то

$$(27)\ \sup_x g(x) = g(x_0)$$

В таком случае вместо sup пишут max. А.Н. Колмогоров усилил результат В.И. Гливенко для непрерывных функций распределения $F(x)$. Рассмотрим случайную величину

(28) $D_n = \sqrt{n} \sup_x |F_n(x) - F_0(x)|$

и ее функцию распределения

(29) $K_n(x) = P\{D_n \leq x\}$

По теореме А.Н. Колмогорова

(30) $\lim_{n\to\infty} K_n(x) = K(x)$

при каждом x, где $K(x)$ – функция распределения Колмогорова.

Работа А.Н. Колмогорова породила одно из основных направлений математической статистики – т.н. *непараметрическую статистику*. В настоящее время непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат широко используются. Они были разработаны для проверки согласия с полностью известным теоретическим распределением, т.е. предназначены для проверки гипотезы H_0: $F(x) \equiv F_0(x)$. Основная идея критериев Колмогорова, омега-квадрат и аналогичных им состоит в измерении расстояния между функцией эмпирического распределения и функцией теоретического распределения. Различаются эти критерии видом расстояний в пространстве функций распределения.

Выборочные характеристики распределения

По аналогии с генеральной совокупностью для описания данных выборки используют выборочное среднее $\bar{x}$, выборочную медиану $\tilde{x}$, выборочную дисперсию s^2, выборочное среднее квадратическое отклонение s, размах R.

Основные понятия, используемые при оценивании

Оценивание – это определение приближенного значения неизвестного *параметра* генеральной совокупности по результатам наблюдений. Оценивание проводят с помощью оценок – *статистик* (функций от результатов наблюдений). Оценивание бывает двух видов – точечное оценивание и оценивание с помощью доверительной области.

Точечное оценивание – способ оценивания, заключающийся в том, что значение оценки принимается как неизвестное значение параметра распределения. Например, математическое ожидание m можно оценить с помощью выборочного среднего арифметического $\bar{x}$, выборочной медианы $\tilde{x}$ или полусуммы минимального и максимального членов вариационного ряда $m^{**} = [x(1)+x(n)]/2$. Наличие нескольких методов оценивания одних и тех же параметров приводит к необходимости выбора между этими методами.

Сравнение проводят на основе таких показателей качества методов оценивания, как состоятельность, несмещенность, эффективность и др. Рассмотрим оценку Θ_n числового параметра Θ, определенную при n = 1, 2, ... Оценка Θ_n называется *состоятельной*, если она сходится по вероятности к значению оцениваемого параметра Θ при безграничном возрастании объема выборки. Все (за редчайшими исключениями) оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются состоятельными.

Несмещенная оценка Θ_n – это оценка параметра Θ, математическое ожидание которой равно значению оцениваемого параметра: $M(\Theta_n) = \Theta$. Оценки, для которых соотношение $M(\Theta_n) = \Theta$ неверно, называются *смещенными*. При этом разность между математическим ожиданием оценки Θ_n и оцениваемым параметром Θ, т.е. $M(\Theta_n) - \Theta$, называется смещением оценки.

Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра. Доказано, что $\bar{x}$ и s_0 являются эффективными оценками параметров m и σ нормального распределения. В то же время для выборочной медианы $\tilde{x}$ справедливо предельное соотношение

(31) $\lim_{n\to\infty} \frac{D(\bar{x})}{D(\tilde{x})} = \frac{2}{\pi} = 0{,}637$

Другими словами, эффективность выборочной медианы, т.е. отношение дисперсии эффективной оценки $\bar{x}$ параметра m к дисперсии несмещенной оценки $\tilde{x}$ этого параметра при больших n близка к 0,637. Именно из-за сравнительно низкой эффективности выборочной медианы в качестве оценки

математического ожидания нормального распределения обычно используют выборочное среднее арифметическое.

Подавляющее большинство оценок Θ_n, используемых в вероятностно-статистических методах, являются асимптотически нормальными, т.е. для них справедливы предельные соотношения:

$$(32)\ \lim_{n\to\infty} P\left\{\frac{\Theta_n - M(\Theta_n)}{\sqrt{D(\Theta_n)}}\right\} < x = \Phi(x)$$

для любого x, где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Это означает, что для больших объемов выборок (несколько десятков или сотен наблюдений) распределения оценок полностью описываются их математическими ожиданиями и дисперсиями, а качество оценок – значениями средних квадратов ошибок $d_n(\Theta_n)$.

Доверительное оценивание

Доверительная область – это область в пространстве параметров, в которую с заданной вероятностью входит неизвестное значение оцениваемого параметра распределения. «Заданная вероятность» называется *доверительной вероятностью* и обычно обозначается γ. Пусть Θ – пространство параметров. Рассмотрим статистику $\Theta_1 = \Theta_1(x_1, x_2, \ldots, x_n)$ – функцию от результатов наблюдений $x_1, x_2, \ldots, x_n$, значениями которой являются подмножества пространства параметров Θ. Так как результаты наблюдений – случайные величины, то Θ_1 – также случайная величина, значения которой – подмножества множества Θ, т.е. Θ_1 – случайное множество.

Статистика Θ_1 называется *доверительной областью*, соответствующей доверительной вероятности γ, если

$$(33)\ P\{\theta \in \Theta_1(x_1, x_2, \ldots, x_n)\} = \gamma$$

При оценке одного числового параметра в качестве доверительных областей обычно применяют доверительные интервалы.

Доверительный интервал – это интервал, который с заданной вероятностью γ накроет неизвестное значение оцениваемого параметра распределения. Пусть $x_1, x_2, \ldots, x_n$ – выборка из нормального закона $N(m,\sigma)$, параметры m и σ неизвестны. Укажем доверительные границы для m. Известно, что случайная величина

$$(34)\ Y = \sqrt{n}\frac{\bar{x} - m}{s_0}$$

имеет распределение Стьюдента с $(n-1)$ степенью свободы, где $\bar{x}$ – выборочное среднее арифметическое и s_0 – выборочное среднее квадратическое отклонение. Тогда в качестве нижней и верхней доверительных границ следует взять

$$(35)\ \bar{x} \mp t_\gamma(n-1)\frac{s_0}{\sqrt{n}}$$

При обработке экономических, управленческих или технических статистических данных обычно используют значение доверительной вероятности γ = 0,95. Применяют также значения γ = 0,99 или γ = 0,90. Иногда встречаются иные значения γ.

Доверительное оценивание для дискретных распределений

Для дискретных распределений, таких, как биномиальное, гипергеометрическое или распределение Пуассона (а также распределения статистики Колмогорова (28) и других непараметрических статистик), функции распределения имеют скачки.

Основные понятия, используемые при проверке гипотез

Статистическая гипотеза – любое предположение, касающееся неизвестного распределения случайных величин. Различают нулевую и альтернативную гипотезы. Нулевая гипотеза – гипотеза, подлежащая проверке. Альтернативная гипотеза – каждая допустимая гипотеза, отличная от нулевой. Нулевую гипотезу обозначают H_0, альтернативную – H_1 (от англ. Hypothesis).

Например, в качестве показателей стабильности технологического, экономического, управленческого или иного процесса используют ряд характеристик распределений контролируемых показателей, в

частности, коэффициент вариации ν = σ/M(X). Требуется проверить нулевую гипотезу H_0: $\nu \le \nu_0$ при альтернативной гипотезе H_1: $\nu > \nu_0$, где ν_0 – некоторое заранее заданное граничное значение.

Конкретная задача проверки статистической гипотезы полностью описана, если заданы нулевая и альтернативная гипотезы. Выбор метода проверки статистической гипотезы, свойства и характеристики методов определяются как нулевой, так и альтернативной гипотезами. Для проверки одной и той же нулевой гипотезы при различных альтернативных гипотезах, возможно, следует использовать различные методы.

Параметрические и непараметрические гипотезы

Предположение, которое касается неизвестного значения параметра распределения, входящего в некоторое параметрическое семейство распределений, называется *параметрической гипотезой*. Предположение, при котором вид распределения неизвестен (не предполагается, что оно входит в некоторое параметрическое семейство распределений), называется *непараметрической гипотезой*. Таким образом, если распределение $F(x)$ результатов наблюдений в выборке, согласно принятой вероятностной модели, входит в некоторое параметрическое семейство $\{F(x;\theta), \theta \in \Theta\}$, т.е. $F(x) = F(x;\theta_0)$ при некотором $\theta_0 \in \Theta$, то рассматриваемая гипотеза – параметрическая, в противном случае – непараметрическая.

Непараметрические задачи делятся на два класса: в одном из них речь идет о проверке утверждений, касающихся функций распределения, во втором – о проверке утверждений, касающихся характеристик распределений. Статистическая гипотеза называется *простой*, если она однозначно задает распределение результатов наблюдений, вошедших в выборку. В противном случае статистическая гипотеза называется *сложной*.

Однозначно определенный способ проверки статистических гипотез называется *статистическим критерием*. Статистический критерий строится с помощью статистики $U(x_1, x_2, \ldots, x_n)$ – функции от результатов наблюдений $x_1, x_2, \ldots, x_n$. В пространстве значений статистики U выделяют критическую область Ψ, т.е., область со следующим свойством: если значения применяемой статистики принадлежат данной области, то отклоняют (иногда говорят отвергают) нулевую гипотезу, в противном случае – не отвергают.

Статистику U, используемую при построении определенного статистического критерия, называют статистикой этого критерия. Например, критерий Колмогорова, основанный на статистике (28). При этом D_n называют статистикой критерия Колмогорова.

Статистические критерии делятся на параметрические и непараметрические. Параметрические критерии используются в параметрических задачах проверки статистических гипотез, а непараметрические – в непараметрических задачах.

Уровень значимости и мощность

При проверке статистической гипотезы возможны ошибки. Есть два рода ошибок. Ошибка первого рода заключается в том, что отвергают нулевую гипотезу, в то время как в действительности эта гипотеза верна. Ошибка второго рода состоит в том, что не отвергают нулевую гипотезу, в то время как в действительности эта гипотеза неверна.

Вероятность ошибки первого рода называется *уровнем значимости* и обозначается α. Вероятность ошибки второго рода β=$P\{U \notin \Psi \,|H_1\}$. Обычно используют не эту вероятность, а ее дополнение до 1, т.е. $1 - P\{U \notin \Psi \,|H_1\}$. Эта величина носит название *мощности критерия* – вероятность того, что нулевая гипотеза будет отвергнута, когда альтернативная гипотеза верна.

В статистическом приемочном контроле α – риск изготовителя, β – риск потребителя. При статистическом регулировании технологического процесса α – риск излишней наладки, β – риск незамеченной разладки.

Глава 6. Некоторые типовые задачи прикладной статистики и методы их решения

В современной математической статистике разработан ряд общих методов определения оценок и доверительных границ – метод моментов, метод максимального правдоподобия, метод одношаговых оценок, метод устойчивых (робастных) оценок, метод несмещенных оценок и др.

Метод моментов основан на использовании выражений для моментов рассматриваемых случайных величин через параметры их функций распределения. Оценки метода моментов

получают, подставляя выборочные моменты вместо теоретических в функции, выражающие параметры через моменты.

В методе максимального правдоподобия, разработанном Р.А.Фишером, в качестве оценки параметра θ берут значение θ^*, для которого максимальна так называемая функция правдоподобия $f(x_1, \theta)\, f(x_2, \theta) \ldots f(x_n, \theta)$ где $x_1, x_2, \ldots, x_n$ – результаты наблюдений; $f(x, \theta)$ – их плотность распределения, зависящая от параметра θ, который необходимо оценить.

В непараметрических задачах оценивания принимают вероятностную модель, в которой результаты наблюдений $x_1, x_2, \ldots, x_n$ рассматривают как реализации n независимых случайных величин с функцией распределения $F(x)$ общего вида. От $F(x)$ требуют лишь выполнения некоторых условий типа непрерывности, существования математического ожидания и дисперсии и т.п. Подобные условия не являются столь жесткими, как условие принадлежности к определенному параметрическому семейству.

Непараметрическое оценивание функции распределения

Правила определения оценок и доверительных границ в параметрическом случае строятся на основе параметрического семейства распределений $F(x;\theta)$. При обработке реальных данных возникает вопрос – соответствуют ли эти данные принятой вероятностной модели? Т.е. статистической гипотезе о том, что результаты наблюдений имеют функцию распределения из семейства $\{F(x;\theta), \theta \in \Theta\}$, при некотором $\theta = \theta_0$? Такие гипотезы называют *гипотезами согласия*, а критерии их проверки – критериями согласия.

Если истинное значение параметра $\theta = \theta_0$ известно, функция распределения $F(x;\theta_0)$ непрерывна, то для проверки гипотезы согласия часто применяют критерий Колмогорова, основанный на статистике

$$(36)\quad D_n = \sqrt{n} \sup_x |F_n(x) - F_0(x, \theta_0)|$$

где $F_n(x)$ – эмпирическая функция распределения.

Если истинное значение параметра θ_0 неизвестно, например, при проверке гипотезы о нормальности распределения результатов наблюдения (т.е. при проверке принадлежности этого распределения к семейству нормальных распределений), то иногда используют статистику

$$(37)\quad D_n(\theta^*) = \sqrt{n} \sup_x |F_n(x) - F_0(x, \theta^*)|$$

Она отличается от статистики Колмогорова D_n тем, что вместо истинного значения параметра θ_0 подставлена его оценка θ^*.

Распределение статистики $D_n(\theta^*)$ сильно отличается от распределения статистики D_n. В качестве примера рассмотрим проверку нормальности, когда $\theta = (m, \sigma^2)$, а $\theta^* = (\bar{x}, s^2)$. Для этого случая квантили распределений статистик D_n и $D_n(\theta^*)$ приведены на рис. 2. Видно, что квантили отличаются примерно в 1,5 раза.

p	0,85	0,90	0,95	0,975	0,99
Квантили порядка p для D_n	1,138	1,224	1,358	1,480	1,626
Квантили порядка p для $D_n(\theta^*)$	0,775	0,819	0,895	0,955	1,035

Корреляция и регрессия

Целями исследования зависимости между признаками являются доказательство наличия связи между признаками и изучение этой связи. Для доказательства наличия связи между двумя случайными величинами X и Y применяют *корреляционный анализ*. Если совместное распределение X и Y является нормальным, то статистические выводы основывают на выборочном коэффициенте линейной корреляции, в остальных случаях используют коэффициенты ранговой корреляции Кендалла и Спирмена, а для качественных признаков – критерий хи-квадрат.

Основная задача *регрессионного анализа* состоит в оценке неизвестных параметров a и b, задающих линейную зависимость y от x. Для решения этой задачи применяют разработанный еще К.Гауссом в 1794 г. метод наименьших квадратов, т.е. находят оценки неизвестных параметров модели a и b из условия минимизации суммы квадратов

$$(38)\quad \sum_{1 \le i \le n} (y_i - a x_i - b)^2$$

по переменным a и b.

Выбор планов эксперимента, т.е. точек x_i, в которых будут проводиться эксперименты по наблюдению y_i – предмет теории планирования эксперимента.