Дополнение 2. Статистические оценки

В этом дополнении рассматриваются оценки параметров статистических моделей. Определяются важнейшие свойства оценок: несмещённость, состоятельность, асимптотическая нормальность. Обсуждается метод максимального правдоподобия, позволяющий получать наиболее точные оценки для выборок большого размера. Приводятся примеры использования простейших визуальных методов: диаграмма размахов, гистограмма.

2.1. Статистическая модель

Простейшая *статистическая модель* представляет собой некоторое семейство функций распределения. Предполагается, что среди функций, входящих в это семейство, найдётся такая, которую можно рассматривать как достаточно точное приближение к функции распределения интересующего исследователя признака (рис. 1).

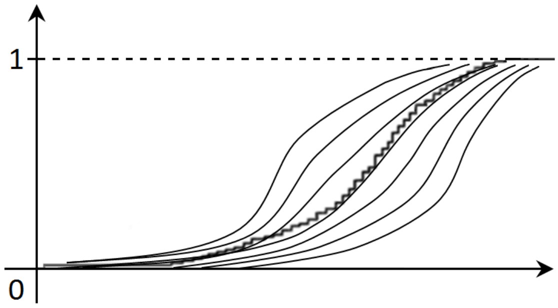

Рис. 1

Модели бывают параметрическими и непараметрическими. В параметрических моделях функции распределения в семействе зависят от одного или нескольких параметров. Например, в экспоненциальной модели функции распределения зависят от параметра σ (см. рис. 5 в разделе 1.4 темы 1), в модели испытаний Бернулли — от вероятности «успеха» p, где $0 < p < 1$.

Одной из самых часто применяемых является *нормальная модель*, зависящая от параметра сдвига μ и параметра масштаба σ. Плотность распределения нормальной случайной величины X задаётся формулой

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (1)$$

График плотности получается из графика, приведённого на рис. 11 в разделе 1.6, сдвигом на величину μ и растяжением в σ раз. Нетрудно вычислить, что математическое ожидание $\mathbf{M}X = \mu$ и дисперсия $\mathbf{D}X = \sigma^2$. Для нормального распределения используется обозначение $N(\mu, \sigma^2)$. Случайная величина $Z = (X - \mu) / \sigma$ имеет распределение $N(0, 1)$ и называется *стандартной нормальной*. Стандартный нормальный закон с функцией распределения $\Phi(x)$ уже встречался в разделе 1.6 при обсуждении центральной предельной теоремы.

В непараметрических моделях на функции распределения накладывается требование принадлежности к некоторому классу, например, классу Ψ_c всех непрерывных распределений или классу Ψ_s распределений с положительной симметричной плотностью. Непараметрические модели менее ограничительны и, следовательно, более применимы для практического применения, чем параметрические. Однако последние в случае адекватности позволяют сделать более точные выводы (обнаружить меньшее систематическое отклонение).

2.2. Оценивание параметров модели

Главная задача статистического анализа, кратко говоря, заключается в подборе модели, наиболее точно описывающей наблюдения. При параметрическом подходе дело сводится к нахождению (оцениванию) параметров модели наилучшим, в некотором смысле, образом.

Пусть $\{F_\theta(x)\}$ — некоторое семейство функций распределения, зависящих от параметра θ, где θ принимает значения из заданного множества Θ. Например, для произвольного θ > 0 рассмотрим *равномерное распределение на отрезке* [0, θ], имеющее функцию распределения

$$F_\theta(x) = \begin{cases} 0, & \text{если } x \le 0; \\ x/\theta, & \text{если } 0 < x < \theta; \\ 1, & \text{если } x \ge \theta. \end{cases} \qquad (2)$$

Равномерную модель с неизвестным параметром масштаба θ образуют функции, заданные формулой (2), где параметр θ принадлежит множеству Θ = $(0, \infty)$. Некоторые функции из этого семейства изображены на рис. 2.

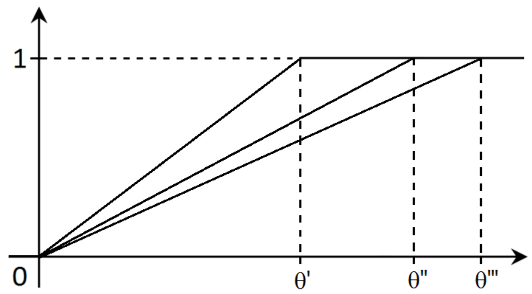

Рис. 2

Представим, что выборка X_1, … , X_n была моделирована методом обратной функции на основе одного из равномерных распределений, скажем, со значением параметра θ_0. Легко убедиться, что обратной функцией к функции $F_\theta(x) = x/\theta$ служит $F_\theta^{-1}(x) = \theta x$. Поэтому формула для моделирования наблюдений имеет следующий вид:

$$X_i = \theta_0 \eta_i, \quad i = 1, \ldots, n, \qquad (3)$$

где η_i взяты наудачу из отрезка [0, 1]. Для примера рассмотрим координаты (с точностью до 0,1) десяти точек, взятых наудачу из отрезка [0, θ_0], где θ_0 — неизвестное целое двузначное число:

3,5 3,2 25,6 8,8 11,6 26,6 18,2 0,4 12,3 30,1.

Попробуйте угадать использованное значение θ_0 с помощью рис. 3. (*Ответ приведён ниже*.)

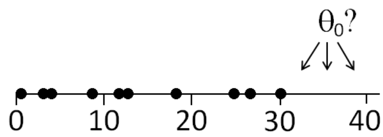

Рис. 3

Вопрос. Может ли θ_0 быть равным а) 30, б) 99?

Ответ: а) нет, так как максимум наблюдений равен 30,1, поэтому θ_0 должно быть не меньше 31; б) значение 99 возможно, но представляется крайне маловероятным: вероятность, что все десять наблюдений при $\theta_0 = 99$ окажутся не больше 30,1, равна $(30{,}1/99)^{10} \approx 6{,}75 \cdot 10^{-6}$.

Каким образом узнать (хотя бы приближённо), какое именно значение параметра θ было использовано. В общем случае, как оценить неизвестный параметр по наблюдениям $X_1, \ldots, X_n$? Будем это делать с помощью некоторых функций $\hat{\theta}(X_1, \ldots, X_n)$, называемых *статистическими оценками.* Эти функции должны принимать (хотя бы для выборок большого размера n) значения близкие к истинному[1] значению параметра.

В равномерной модели на отрезке [0, θ] в качестве оценок неизвестного параметра θ можно взять, например, функции

$$\hat{\theta}_1 = \max\{X_1, \ldots, X_n\} \quad \text{или} \quad \hat{\theta}_2 = 2(X_1 + \ldots + X_n)/n. \tag{4}$$

Действительно, интуитивно понятно, что для выборки большого размера n наибольший элемент среди $X_1, \ldots, X_n$ будет располагаться вблизи правого конца отрезка [0, θ]. В свою очередь, среднее арифметическое наблюдений $\overline{X}$ при увеличении размера выборки будет приближаться к середине отрезка [0, θ], т. е. к θ/2.[2] Поэтому оценка $\hat{\theta}_2 = 2\overline{X}$ будет сходиться к θ.

Какая из оценок $\hat{\theta}_1$ или $\hat{\theta}_2$ является более точной? Как вообще можно сравнивать оценки? Прежде всего, сравнить оценки можно по наличию или отсутствию у них важных свойств: несмещённости, состоятельности, асимптотической нормальности. Давайте познакомимся с этими свойствами.

2.3. Несмещённость и состоятельность

Определение. Если для любого θ из параметрического множества Θ верно равенство

$$\mathbf{M}\hat{\theta}(X_1, \ldots, X_n) = \theta, \tag{5}$$

то оценка $\hat{\theta}$ называется *несмещённой*.

Г. И. Ивченко, Ю. И. Медведев в учебнике «Ведение в математическую статистику» пишут: «Свойство несмещённости интуитивно привлекательно: оно означает, что по крайней мере «в среднем» используемая оценка приводит к желаемому результату».

Например, оценка $\hat{\theta}_1$ из (4) свойством несмещённости не обладает: максимум всегда недооценивает правую границу отрезка. Однако, эту оценку легко подправить: рассмотрим

$$\hat{\theta}_3 = \frac{n+1}{n}\max\{X_1, \ldots, X_n\}. \tag{6}$$

Тогда согласно задаче 2 в конце дополнения для любого $\theta > 0$ справедливо равенство $\mathbf{M}\hat{\theta}_3 = \theta$.

В свою очередь, оценка $\hat{\theta}_2$ из (4) является несмещённой, так как математическое ожидание суммы случайных величин равно сумме математических ожиданий слагаемых:[3]

$$\mathbf{M}\hat{\theta}_2 = 2 \cdot (\mathbf{M}X_1 + \ldots + \mathbf{M}X_n)/n = 2 \cdot \theta/2 = \theta.$$

---

[1] Выбранному «природой» при моделировании наблюдений $X_1, \ldots, X_n$.

[2] Сходимость к θ/2 имеет место в силу закона больших чисел (см. 1.6), поскольку $\mathbf{M}X_i = \theta/2$.

[3] Строго говоря, это свойство верно, если не возникает неопределённости вида $\infty - \infty$ (см. 1.5).

Само по себе свойство несмещённости не достаточно для того, чтобы оценка хорошо приближала неизвестный параметр. Например, в модели испытаний Бернулли первый элемент выборки X_1 является несмещённой оценкой: $\mathbf{M}X_1 = p$ (см. 1.5). Однако его возможные значения 0 и 1 даже не принадлежат параметрическому множеству $\Theta = (0, 1)$. Необходимо, чтобы погрешность оценки стремилась к нулю с увеличением размера выборки. Это свойство в математической статистике называется *состоятельностью.*

Определение. Если для любого θ из параметрического множества Θ и любого $\varepsilon > 0$

$$\mathbf{P}(|\hat{\theta}(X_1,\ldots,X_n) - \theta| > \varepsilon) \to 0 \quad \text{при } n \to \infty, \tag{7}$$

то оценка $\hat{\theta}$ называется *состоятельной*.

Сходимость (7) означает, что любые отклонения оценки $\hat{\theta}$ от параметра θ становятся сколь угодно маловероятными при увеличении размера выборки. Другими словами, состоятельность означает концентрацию вероятностной массы около истинного значения параметра при возрастании n. На рис. 4 изображены графики плотностей распределения оценки $\hat{\theta}$ для двух разных n. Вероятностная масса распределения внутри $[\theta - \varepsilon, \theta + \varepsilon]$ стремится к 1 при $n \to \infty$.

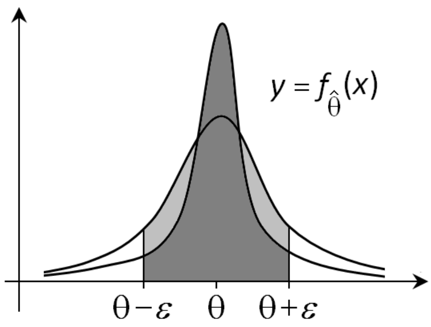

Рис. 4

В задаче 3 в конце дополнения предлагается убедиться, что оценки $\hat{\theta}_1$ и $\hat{\theta}_3$ состоятельны. Оценка $\hat{\theta}_2$, как отмечалось выше, также состоятельна в силу закона больших чисел. Обе оценки $\hat{\theta}_2$ и $\hat{\theta}_3$ являются несмещёнными. Какая из них точнее? Простейшей мерой точности оценки служит её дисперсия, определённая ранее в разделе 1.5. Несложно вычислить, что

$$\mathbf{D}\hat{\theta}_3 / \mathbf{D}\hat{\theta}_2 = 3/(n+2).$$

Правая часть при всех n не превосходит 1 (т. е. $\mathbf{D}\hat{\theta}_3 \le \mathbf{D}\hat{\theta}_2$) и стремится к нулю при $n \to \infty$. Следовательно, $\hat{\theta}_3$ имеет над $\hat{\theta}_2$ преимущество в точности, возрастающее с увеличением n.

Возвращаясь к примеру из 2.2 с равномерной выборкой размера $n = 10$ из отрезка $[0, \theta]$, сообщим, что при моделировании использовалось значение $\theta_0 = 35$. Легко вычислить, что $\hat{\theta}_2 = 28{,}1$ и $\hat{\theta}_3 = 33{,}1$. Значит, в данном случае вторая оценка оказалась точнее, в то время как первая даже приняла недопустимое значение, которое меньше максимума выборки 30,1.

2.4. Асимптотическая нормальность

В общем случае дисперсии оценок могут не существовать или быть бесконечными. Однако если такие оценки обладают свойством асимптотической нормальности, то можно сравнить их асимптотические дисперсии, которые характеризуют точность оценок для выборки большого размера.

Определение. $\hat{\theta}(X_1,\ldots,X_n)$ будем называть *асимптотически нормальной оценкой*, имеющей *асимптотическую дисперсию* $\widehat{\sigma}^2$, если

$$\mathbf{P}(\sqrt{n}(\hat{\theta}-\theta)/\widehat{\sigma} \le x) \to \Phi(x) \quad \text{при } n \to \infty, \tag{8}$$

где $\Phi(x)$ — функция распределения стандартного нормального закона (см. формулу (16) из 1.6).

Например, центральная предельная теорема из 1.6 в свете данного определения является утверждением об асимптотической нормальности выборочного среднего $\overline{X}_n = (X_1 + \ldots + X_n)/n$ как оценки для μ — неизвестного среднего значения признака. Асимптотическая дисперсия $\widehat{\sigma}^2$ при этом совпадает с дисперсией σ^2 признака в генеральной совокупности.

Можно доказать, что асимптотическая нормальность оценки влечёт её состоятельность.

Почему в качестве показателя *относительной эффективности* двух асимптотически нормальных оценок разумно использовать отношение их дисперсий $\widehat{\sigma}_2^2/\widehat{\sigma}_1^2$, а не отношение $\widehat{\sigma}_2/\widehat{\sigma}_1$? Пусть требуется оценить параметр θ с заданной точностью Δ, причём за каждое наблюдение X_i приходится платить цену C. Тогда размеры выборок n_1 и n_2, обеспечивающие заданную точность первой и второй оценок соответственно, определяются из соотношения

$$\Delta = \widehat{\sigma}_1 \Big/ \sqrt{n_1} = \widehat{\sigma}_2 \Big/ \sqrt{n_2}\,.$$

Таким образом, $\widehat{\sigma}_2^2/\widehat{\sigma}_1^2 = n_2/n_1 = (Cn_2)/(Cn_1)$, т. е. относительная эффективность есть отношение затрат при использовании второй оценки к затратам при использовании первой оценки.

Приведём ещё примеры асимптотически нормальных оценок.

Наряду с математическим ожиданием представляют практический интерес такие характеристики распределения как p-квантили (в частности, медиана и квартили). Ради простоты рассмотрим случай непрерывной строго возрастающей функции распределения $F(x)$.

Определение. Пусть $0 < p < 1$. Значение аргумента функции распределения $F(x)$, при котором она достигает уровня p, называется *p-квантилью*[4] и обозначается через x_p (рис. 5).

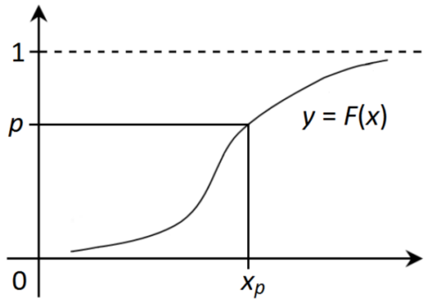

Рис. 5

Короче говоря, $x_p = F^{-1}(p)$. Частные случаи p-квантилей при p = 1/4, 1/2, 3/4 называются, соответственно, *нижней квартилью*, *медианой* и *верхней квартилью* распределения.

Медиана, наряду с математическим ожиданием, является характеристикой «центра» распределения. У перекошенного распределения, имеющего сильно асимметричную плотность, медиана заметно отличается от математического ожидания.

Согласно данным Росстата средняя зарплата в 2016 г. в Москве составила 59 тыс. руб., медианная зарплата — 44 тыс. руб. (74% от средней). Во всей Российской Федерации аналогичные показатели составляли, соответственно, 31 тыс. руб. и 23 тыс. руб. (75% от средней).[5]

[4] Также p-квантиль называют $(100\cdot p)$-й *процентилью* или *персентилью*.

[5] http://www.gks.ru/free_doc/new_site/population/bednost/tabl/tab-bed1-2-6.htm

Для каких признаков медиана представляет интерес? Прежде всего, для неаддитивных (от additio — (лат.) *добавление*), т. е. таких признаков, для которых не имеет смысла суммирование значений. Бессмысленно складывать, скажем, следующие показатели: продолжительность жизни человека, нагрузка разрушения при испытании некоторого материала на прочность, температура больного. В противном случае появляется такая анекдотическая характеристика, как «средняя температура по больнице, включая морг». Напротив, аддитивными признаками являются, например, надой молока от коровы, запас древесины у заготавливаемого дерева, ежемесячный доход гражданина.

В свою очередь, *межквартильный интервал* ($x_{1/4}$, $x_{3/4}$) служит альтернативой области наиболее типичных значений ($\mathbf{M}\xi - \sqrt{\mathbf{D}\xi}$, $\mathbf{M}\xi + \sqrt{\mathbf{D}\xi}$) (см. рис. 9 из раздела 1.5). Действительно, $\mathbf{P}(x_{1/4} \le \xi \le x_{3/4}) = 1/2$, т. е. вероятность попадания ξ в межквартильный интервал равна 50%, поэтому примерно половина элементов выборки оказывается внутри этого интервала.

Децильный коэффициент. Как пример социологического показателя, строящегося на основе квантилей $x_{0,1}$ и $x_{0,9}$ распределения дохода граждан, рассмотрим *децильный коэффициент* — отношение доходов 10% самых богатых граждан страны к доходам 10% самых бедных.[6] Этот коэффициент служит важным индикатором экономической стабильности государства.

По данным Росстата в 2017 году децильный коэффициент в Российской Федерации был равен 15,5.[7] Для сравнения приведём его значения, упорядоченные по возрастанию, в ряде стран за 2012-2015 годы.[8]

Название страны	Доля дохода 10% богатых	Доля дохода 10% бедных	Децильный коэффициент	Год
Казахстан	22,3	4,3	5,2	2015
Белоруссия	22,2	4,0	5,6	2015
Финляндия	22,0	3,9	5,6	2014
Словения	21,1	3,8	5,6	2014
Чехия	22,1	3,9	5,7	2014
Норвегия	21,6	3,5	6,2	2014
Германия	24,9	3,3	7,5	2013
Великобритания	26,2	2,9	9,0	2014
Китай	31,4	2,0	15,7	2012
США	30,2	1,7	17,8	2013
Гондурас	37,8	1,2	31,5	2015
Колумбия	39,6	1,2	33,0	2015
Бразилия	40,5	1,1	36,8	2015
Замбия	45,2	1,0	45,2	2015

«Как только децильный коэффициент достигает 10, в стране создаются условия для социальных беспорядков, — пояснил "Известиям" в 2007 году глава Института экономики РАН Руслан Гринберг. — Это правило не действует разве что в Америке, где коэффициент держится на уровне 10-12.[9] Но там это считается нормальным, поскольку философия американцев отличается от нашей. Там считается: если ты бедный, то сам виноват».[10]

---

[6] Иногда его называют *коэффициентом фондов*, а децильным коэффициентом — отношение $x_{0,9} / x_{0,1}$.
[7] http://ac.gov.ru/events/016271.html
[8] По результатам опросов Всемирного банка, см. http://data.worldbank.org/indicator, раздел «Poverty»
[9] По данным Всемирного банка децильный коэффициент в США составлял 14,8 в 1991 г. и 16,6 в 2000 г.
[10] https://archive.is/20120907172509/www.izvestia.ru/economic/article3107047/

Каким образом можно оценить p-квантиль (в частности, медиану и квартили) по выборке? В R по умолчанию используется оценка Q_p, определяемая следующим образом. Упорядочим элементы выборки $X_1, \dots, X_n$ по возрастанию и получим так называемый *вариационный ряд*:

$$X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}. \tag{9}$$

В частности, $X_{(1)}$ — минимум выборки, $X_{(n)}$ — максимум выборки. Отметим на плоскости точки с координатами $(X_{(i)}, (i-1)/(n-1))$, где $i = 1, \dots, n$. Последовательно соединив их отрезками, получим ломаную, задающую график функции $\widetilde{F}(x)$ (рис. 6). Для произвольного числа p из интервала (0, 1) положим $Q_p = \widetilde{F}^{-1}(p)$.

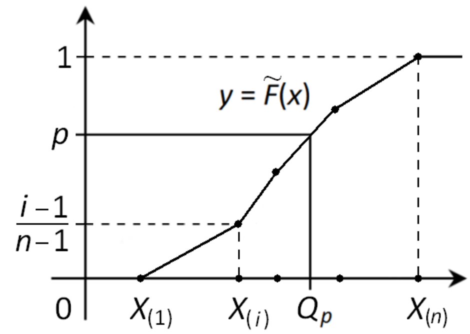

Рис. 6

В частности, оценка $Q_{1/2}$ называется *выборочной медианой*. Её также обозначают через *MED*.[11] Она явно выражается через члены вариационного ряда: *MED* является средним членом вариационного ряда для выборки нечётного размера и равна полусумме двух средних членов для выборки чётного размера, т. е.

$$MED = \begin{cases} X_{((n+1)/2)} & \text{при нечётном } n, \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{при чётном } n. \end{cases}$$

Возвращаясь к случаю произвольного $0 < p < 1$, приведём условия, обеспечивающие асимптотическую нормальность Q_p как оценки для p-квантили x_p.

Теорема. Если существует плотность распределения $f(x) = F'(x)$ и $f(x_p) > 0$, то

$$\mathbf{P}(\sqrt{n}(Q_p - x_p)/\hat{\sigma} \le x) \to \Phi(x) \quad \text{при } n \to \infty, \tag{10}$$

где асимптотическая дисперсия задаётся формулой $\hat{\sigma}^2 = p(1-p)/f^2(x_p)$.

Оценивание параметра сдвига для закона Коши. Рассмотрим задачу оценки неизвестного параметра сдвига θ распределения Коши, имеющего плотность

$$f_\theta(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}. \tag{11}$$

Несмотря на то, что график плотности $f_\theta(x)$ симметричен относительно точки θ, можно доказать, что $\overline{X}$ не сходится к параметру θ, а имеет точно такое же распределение, как и каждое отдельное наблюдение X_i. Причина заключается в том, что для распределения Коши математическое ожидание не определено (см. объяснение в разделе 1.5 темы 1). Поэтому

[11] Median — (англ.) *срединный*.

закон больших чисел для среднего арифметического $\overline{X}$ не выполняется. Напротив, ввиду того, что плотность $f_\theta(x)$ положительна, выполняются условия приведённой выше теоремы. Поэтому выборочная медиана $Q_{1/2}$ (MED) является асимптотически нормальной и, следовательно, состоятельной оценкой параметра сдвига θ.

В задаче 6 в конце дополнения предлагается сравнить точность $\overline{X}$ и MED в моделях сдвига двух других распределений.

Индекс массы тела. В антропологии в качестве показателя степени ожирения (похудения) человека используется *индекс массы тела* (ИМТ):

$$\textbf{ИМТ} = 10^6 \cdot \textbf{ВЕС} \,/\, \textbf{РОСТ}^2.$$

Здесь предполагается, что ВЕС измеряется в килограммах, а РОСТ — в сантиметрах. На рис. 7 приведены оценки квантилей уровней 0,1; 0,25; 0,5; 0,75; 0,9 индекса массы тела мальчиков возраста от 7 до 17 лет, проживавших в 2011 году в Москве (на основе выборки из 500 человек). Чтобы визуально уменьшить влияние случайности, квантили сглажены кубическими многочленами. Норму веса задает межквартильная полоса. Если ИМТ мальчика превышает 0,75-квантиль, то его вес является избыточным. Превышение же 0,9-квантили свидетельствует об ожирении, представляющем опасность для здоровья.

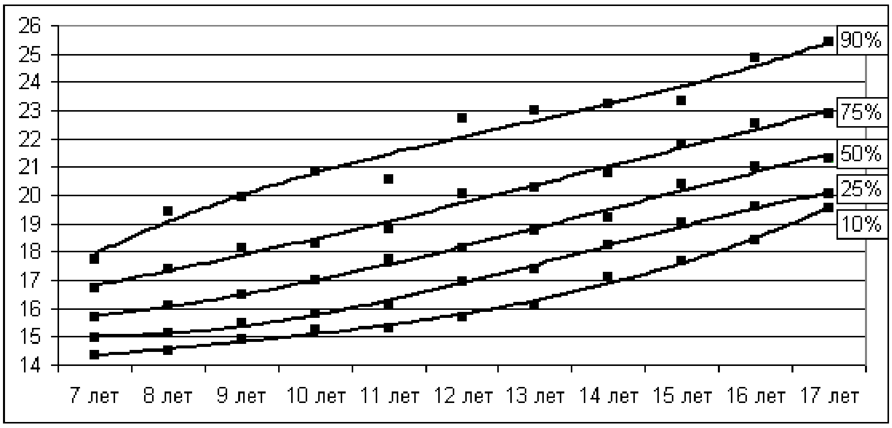

Рис. 7

Приведём ещё один пример асимптотически нормальной оценки. Рассмотрим несмещённую (согласно см. задаче 5 в конце дополнения) оценку неизвестной дисперсии σ^2 элементов выборки

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad (12)$$

которая уже встречалась в разделе 1.6. Для неё выполняется следующая

Теорема. Допустим, что *центральный момент четвёртого порядка* $\mu_4 = \mathbf{M}(X - \mathbf{M}X)^4 < \infty$. Тогда

$$\mathbf{P}(\sqrt{n}(S^2 - \sigma^2)/\widehat{\sigma} \le x) \to \Phi(x) \quad \text{при } n \to \infty,$$

где асимптотическая дисперсия $\widehat{\sigma}^2 = \mu_4 - \sigma^4$.

Отметим, что существуют состоятельные оценки, не обладающие свойством асимптотической нормальности. Примером может служить оценка $\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$ в модели равномерного распределения на отрезке $[0, \theta]$. Действительно, в задаче 4 в конце дополнения предлагается для любого $x > 0$ установить сходимость

$$\mathbf{P}(n(\theta - \hat{\theta}_1) \le x) \to 1 - e^{-x/\theta} \quad \text{при } n \to \infty. \tag{13}$$

Таким образом, погрешность оценки $\hat{\theta}_1$ имеет порядок малости $1/n$, и предельным законом для распределения этой оценки служит не нормальное, а экспоненциальное распределение.

2.5. Метод максимального правдоподобия

Каким образом можно построить оценку неизвестного параметра θ для заданной статистической модели (семейства распределений)? Одним из наиболее полезных общих методов получения оценок является метод максимального правдоподобия.

В 1778 г. швейцарским физиком, механиком и математиком Даниилом Бернулли (племянником Якоба Бернулли, установившего в 1713 году справедливость закона больших чисел для частоты «успехов» в независимых испытаниях) была опубликована в изданиях Петербургской Академии наук работа «Наиболее вероятное определение по нескольким расходящимся между собой наблюдениям и устанавливаемое отсюда наиболее правдоподобное заключение», где впервые был высказан и использован для оценки неизвестного параметра принцип максимального правдоподобия.[12]

Как мы покажем ниже, оценки максимального правдоподобия являются наиболее точными для выборок большого размера при выполнении для семейства распределений определённых условий гладкости и интегрируемости (так называемых *условий регулярности*). Для знакомства с методом максимального правдоподобия нам понадобится формальное определение независимости случайных величин.

Определение. Случайные величины $\xi_1, \ldots, \xi_n$ называются *независимыми*, если для любых чисел $x_1, \ldots, x_n$ справедливо равенство

$$\mathbf{P}(\xi_1 \le x_1, \ldots, \xi_n \le x_n) = \mathbf{P}(\xi_1 \le x_1) \cdot \ldots \cdot \mathbf{P}(\xi_n \le x_n). \tag{14}$$

Можно доказать, что для дискретных случайных величин $\xi_1, \ldots, \xi_n$ независимость равносильна выполнению для произвольных чисел $x_1, \ldots, x_n$ равенства

$$\mathbf{P}(\xi_1 = x_1, \ldots, \xi_n = x_n) = \mathbf{P}(\xi_1 = x_1) \cdot \ldots \cdot \mathbf{P}(\xi_n = x_n). \tag{15}$$

Выбираемые наудачу элементы выборки X_i являются независимыми и одинаково распределёнными случайными величинами. Сначала предположим для простоты, что они имеют дискретное распределение. Введём обозначение $f_\theta(x) = \mathbf{P}(X_i = x)$, где θ — неизвестный параметр модели. Тогда совместная вероятность из (15) запишется в виде

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n) = f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n). \tag{16}$$

В правой части (16) стоит функция от аргументов $x_1, \ldots, x_n$ и θ. Рассматриваемая как функция от единственного аргумента θ при фиксированных значениях аргументов $x_1, \ldots, x_n$, она называется *функцией правдоподобия* и обозначается через $L(\theta)$.[13] Величину $L(\theta_0)$ в фиксированной точке θ_0 можно считать мерой правдоподобия θ_0 для имеющейся реализации выборки $x_1, \ldots, x_n$. Представляется разумным в качестве оценки параметра θ взять наиболее правдоподобное значение $\tilde{\theta} = \tilde{\theta}(x_1, \ldots, x_n)$, при котором функция $L(\theta)$ достигает максимума. Как правило, проще найти точку максимума $\ln L(\theta)$, совпадающую с $\tilde{\theta}$ ввиду монотонности логарифма.

---

[12] Широкое распространение метод получил после публикации статьи Рональда Фишера в 1912 году.

[13] Likelihood — (англ.) *правдоподобие*.

Определение. Оценка $\tilde{\theta}(X_1,\dots,X_n)$ называется *оценкой максимального правдоподобия* (ОМП).[14]

Рассмотрим в качестве примера модель испытаний Бернулли с неизвестной вероятностью «успеха» θ (см. 1.3). Для отдельного испытания X_i при $x = 0$ или 1 верно представление

$$f_\theta(x) = \mathbf{P}(X_i = x) = \theta^x(1-\theta)^{1-x}.$$

Тогда из формулы (16) выводим, что $L(\theta) = \theta^s(1-\theta)^{n-s}$, где $s = x_1 + \ldots + x_n$. Функция правдоподобия представляет собой многочлен n-й степени от переменной θ. График функции $L(\theta)$ для случая $s > 1$ и $n - s > 1$ имеет вид, изображённый на рис. 8.

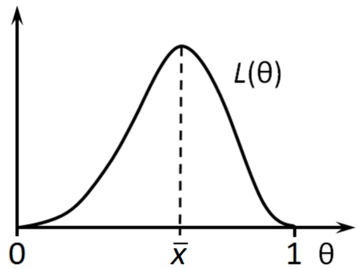

Рис. 8

Найдём точку максимума $\ln L(\theta) = s \ln \theta + (n - s) \ln(1 - \theta)$. Дифференцируя по θ, получим уравнение $s/\theta - (n - s)/(1 - \theta) = 0$. Решив его, найдём корень $\tilde{\theta} = s/n = (x_1 + \ldots + x_n)/n = \bar{x}$. Таким образом, ОМП для вероятности «успеха» θ есть не что иное, как частота «успехов» $\bar{X}$.

В случае непрерывных моделей будем использовать обозначение $f_\theta(x)$ для плотности распределения отдельного элемента выборки. Функцию правдоподобия $L(\theta)$ определим как правую часть формулы (16).

В качестве примера непрерывной модели рассмотрим модель сдвига показательного закона с плотностью $f_\theta(x) = e^{-(x-\theta)}I_{\{x \geq \theta\}}$. В этой модели функция правдоподобия имеет вид

$$L(\theta) = e^{-(x_1+\ldots+x_n)}e^{n\theta}I_{\{x_1 \geq \theta\}} \cdot \ldots \cdot I_{\{x_n \geq \theta\}} = e^{-(x_1+\ldots+x_n)}e^{n\theta}I_{\{x_{(1)} \geq \theta\}}.$$

При фиксированных $x_1,\ldots, x_n$ функция $L(\theta)$ экспоненциально быстро растёт по переменной θ до значения $x_{(1)}$, после которого становится равной нулю из-за индикатора $I_{\{x_{(1)} \geq \theta\}}$ (рис. 9). Отсюда заключаем, что ОМП в данной модели служит минимум выборки $X_{(1)}$. Заметим также, что здесь $L(\theta)$ имеет разрыв в точке $x_{(1)}$, поэтому она не является дифференцируемой по θ функцией на всей прямой, и ОМП нельзя находить путём приравнивания нулю её производной.

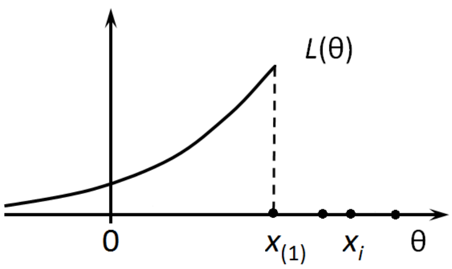

Рис. 9

В задаче 7 предлагается найти ОМП ещё для двух параметрических моделей.

[14] Обратите внимание, что в качестве аргументов в функцию $\tilde{\theta}(x_1,\dots,x_n)$ подставляются элементы выборки $X_1, \ldots, X_n$. Поэтому ОМП является случайной величиной.

Метод максимального правдоподобия понятным образом обобщается на случай векторного параметра $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$. В случае, когда $L(\boldsymbol{\theta})$ гладко зависит от параметра $\boldsymbol{\theta}$, векторная оценка максимального правдоподобия $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_m)$ находится путём решения системы (вообще говоря, нелинейных) уравнений:

$$(\ln L(\boldsymbol{\theta}))'_{\theta_j} = 0, \quad j = 1, \dots, m.$$

Иногда это решение можно найти явно (см. задачу 8 в конце дополнения), но обычно приходится вычислять его приближённо с помощью численных (итерационных) методов.

Какими свойствами обладают оценки максимального правдоподобия? Ради простоты ограничимся случаем скалярного параметра. Оказывается, что ОМП являются асимптотически нормальными при выполнении *условий регулярности* (гладкости) статистической модели.[15]

Теорема. В условиях регулярности для достаточно большого размера выборки n существует с вероятностью 1 решение уравнения правдоподобия $(\ln L(\theta))' = 0$, дающее состоятельную оценку $\tilde{\theta}$, причём

$$\mathbf{P}(\sqrt{n}(\tilde{\theta} - \theta)/\hat{\sigma} \le x) \to \Phi(x) \quad \text{при } n \to \infty,$$

где асимптотическая дисперсия задаётся формулой $\hat{\sigma}^2 = 1/I(\theta)$, в которой функция

$$I(\theta) = \mathbf{D}(\ln f_\theta(X_1))'_\theta$$

называется *информацией Фишера*.

Теорема. Пусть выполнены условия регулярности и $\hat{\theta}$ — асимптотически нормальная оценка с асимптотической дисперсией $\hat{\sigma}^2$, непрерывно зависящей от параметра θ. Тогда $\hat{\sigma}^2 \ge 1/I(\theta)$.

Таким образом, в условиях регулярности ОМП обладает наименьшей возможной асимптотической дисперсией $1/I(\theta)$. Это свойство оценок максимального правдоподобия называется *асимптотической эффективностью*.

2.6. Диаграмма размахов и гистограмма

Прежде чем вычислять статистические оценки полезно применить графические методы анализа наблюдений. Простейшими визуальными методами, позволяющими получить представление о виде распределения выборки, являются диаграмма размахов и гистограмма.[16]

Упрощённая диаграмма размахов демонстрирует положение выборочной медианы и двух размахов: между минимумом и максимумом, между выборочными квартилями $Q_{1/4}$ и $Q_{3/4}$. Интервал $(Q_{1/4}, Q_{3/4})$ оценивает область наиболее типичных значений $(x_{1/4}, x_{3/4})$ (рис. 10). Его длина $H = Q_{3/4} - Q_{1/4}$ называется *межквартильным размахом*.

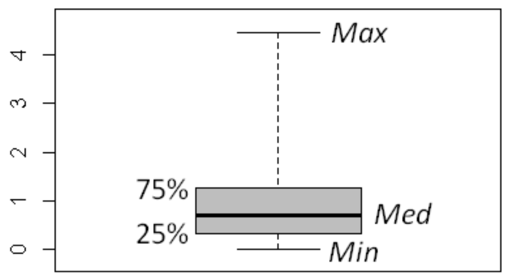

Рис. 10

[15] Один из вариантов условий регулярности приведён в книге Лагутин М. Б. «Наглядная математическая статистика, на с. 114, 119.

[16] Box-and-whisker plot — (англ.) *диаграмма размахов,* histogram — (англ.) *гистограмма.*

Обычно более интересна немного обобщённая диаграмма, называемая *диаграммой размахов с «выбросами»* (рис. 11).[17] В языке R для её построения применяется функция boxplot.

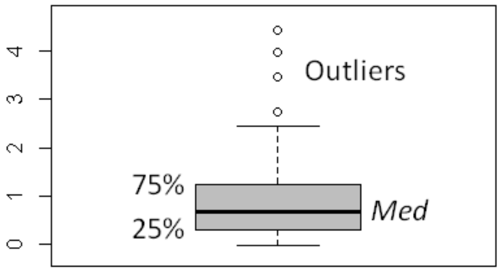

Рис. 11

«Усы» на такой диаграмме соответствуют наименьшему и наибольшему значению, попавшим в интервал ($Q_{1/4}$ – 1,5H, $Q_{3/4}$ + 1,5H), «кружочки» — наблюдениям, оказавшимся вне этого интервала: далёким «хвостовым» значениям или выделяющимся наблюдениям («выбросам»).

Для обнаружения «безусловных выбросов» на практике обычно задают range = 3. В свою очередь, значение 1,5 позволяет выявить «возможные выбросы» и перекос распределения. Почему берут именно 1,5, а не 1 или 2? Если распределение в точности нормальное, то вероятность $p_{1,5}$, что отдельное наблюдение окажется вне интервала ($x_{1/4}$ – 1,5h, $x_{3/4}$ + 1,5h), где $h = x_{3/4} - x_{1/4}$, нетрудно подсчитать с помощью функций pnorm и qnorm языка R, вычисляющих, соответственно, функцию распределения стандартного нормального распределения $\Phi(x)$ и обратную к ней функцию $\Phi^{-1}(x)$. Применив их, получим $p_{1,5} \approx 0{,}007$. Если взять 1 или 2 вместо 1,5, то получим, соответственно, $p_1 \approx 0{,}043$ и $p_2 \approx 0{,}0007$. Распределения реальных признаков обычно имеют более тяжёлые «хвосты», чем нормальный закон. Поэтому при пороге 1 к «выбросам» будут причислены многие «хвостовые» наблюдения. Напротив, при пороге 2 некоторые «возможные выбросы» не будут обнаружены.

Ещё одной диаграммой, демонстрирующей форму распределения, является гистограмма. Она служит оценкой плотности распределения наблюдений. *Гистограмма* представляет собой ряд столбиков разной высоты. Исследователь задаёт целое число N — количество столбиков гистограммы. Диапазон от наименьшего наблюдения $X_{(1)}$ до наибольшего наблюдения $X_{(n)}$ разбивается на N равных по длине промежутков. Затем подсчитываются количества наблюдений, попавших в каждый промежуток. Над каждым промежутком строится столбик, высота которого пропорциональна количеству наблюдений в промежутке. В языке R для её построения применяется функция hist (или функция truehist из пакета MASS).

С помощью гистограммы можно обнаружить такие особенности распределения как:
а) отсутствие симметрии у распределения признака;
б) наличие «выбросов»;
в) неоднородность наблюдений (плотность распределения имеет несколько максимумов).
Гистограммы с указанными особенностями приведены на рис. 12.

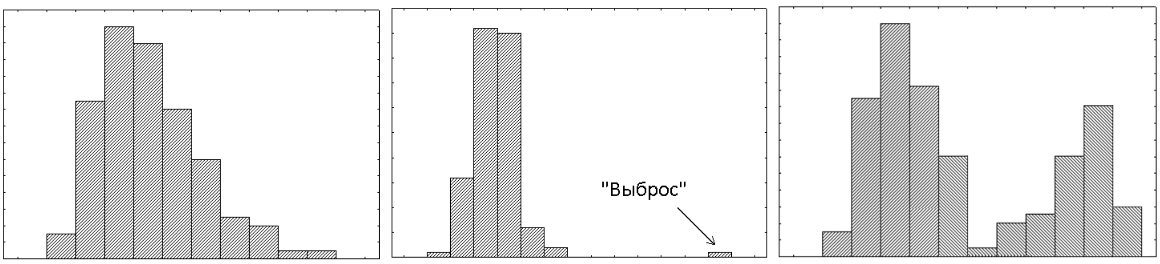

Рис. 12

[17] Outlier — (англ.) *«выброс», т. е. сильно выделяющееся значение.*

В качестве примера приведём на рис. 13 гистограмму, построенную для моделированной методом обратной функции выборки размера 100 из стандартного нормального закона:

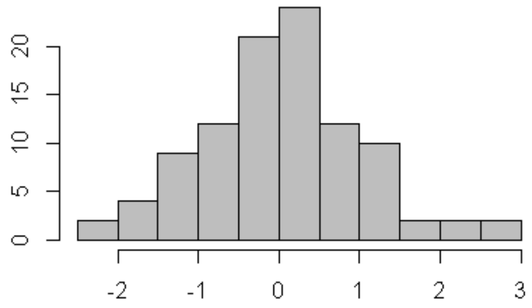

Рис. 13

Сколько столбиков задавать для выборки размера n? Понятно, что слишком малое число столбиков N не позволит гистограмме достаточно точно передать форму распределения. Напротив, если N неоправданно велико, то гистограмма будет слишком «зазубренной» и ненадёжной оценкой для плотности распределения. Иначе говоря, в этом случае дисперсия оценки будет слишком большой, и число максимумов гистограммы превзойдёт истинное число максимумов плотности («ложная» неоднородность).

В статистических пакетах используются разные варианты выбора по умолчанию числа столбиков N. Иногда используется предложенная Стёджесом (Sturges) эмпирическая формула $N = \log_2 n + 1$. Альтернативными вариантами служат формулы для длины промежутка разбиения d, предложенные Скоттом (Scott) в 1979 году или Фридманом и Диаконисом (Freedman, Diaconis) в 1981 году соответственно:

$$d = 3{,}5Sn^{-1/3},$$

$$d = 2Hn^{-1/3}.$$

В первой формуле S обозначает выборочное стандартное отклонение, определённое выше формулой (12). Коэффициент 3,5 подобран из соображения повышения точности оценки для случая, когда распределение похоже на нормальный закон. Во второй формуле H обозначает межквартильный размах $Q_{3/4} - Q_{1/4}$. Формула Фридмана и Диакониса позволяет исключить влияние «выбросов» благодаря устойчивости к ним межквартильного размаха. Гистограмма, построенная на основе этой формулы, обычно имеет больше столбиков, чем гистограмма, построенная на основе формулы Скотта.

Задачи

В задачах 1-4 выборка берётся из равномерного распределения на отрезке $[0, \theta]$.

1. Найти формулу для функции распределения оценки $\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$, используя тождество

$$\mathbf{P}(\max\{X_1, \ldots, X_n\} \le x) = \mathbf{P}(X_1 \le x, \ldots, X_n \le x).$$

2. Вычислить $\mathbf{M}\hat{\theta}_1$.

3. Доказать состоятельность оценок $\hat{\theta}_1$ и $\hat{\theta}_3$ (вторая оценка определяется формулой (6)).

4. Убедиться, что имеет место сходимость (13).

5. Пусть X — произвольный признак с конечной дисперсией σ^2. Доказать, что оценка S^2, заданная формулой (12), служит несмещённой оценкой для σ^2.

6. Для выборки большого размера сравнить по точности оценки $\overline{X}$ и *MED* ($Q_{1/2}$) в следующих двух моделях с неизвестным параметром сдвига θ:

а) равномерное распределение на отрезке $[\theta - 1/2, \theta + 1/2]$;
б) нормальное распределение $N(\theta, 1)$.

7. Найти ОМП:
а) для показательной модели с неизвестным параметром λ, плотность распределения которой задаётся формулой (10) из раздела 1.4 дополнения 1;

б) для параметра сдвига μ в нормальной модели $N(\mu, 1)$ (см. формулу (1)).

8*. Найти ОМП для параметров сдвига и масштаба ($\boldsymbol{\theta} = (\mu, \sigma)$) в нормальной модели $N(\mu, \sigma^2)$.