Тема 7. Статистические оценки

7.1. Задачи математической статистики

Как правило, методы математической статистики применяются к данным, полученным в результате случайного выбора n объектов из некоторой «необъятной» генеральной совокупности $\Omega = \{\omega\}$. Обычно предполагается, что объекты $\omega_1, \ldots, \omega_n$ выбираются из Ω наудачу с возвращением (см. раздел 1.2). Из-за «необъятности» множества Ω крайне маловероятно выбрать объект повторно, поэтому практически выбор с возвращением не отличается от выбора без возвращения.

Каждый объект ω обладает рядом характеристик (*признаков*): $X(\omega), Y(\omega), Z(\omega), \ldots$. Например, объектами могут быть слова некоторого языка, а признаками — некоторые лингвистические характеристики: часть речи, длина слова, частота встречаемости, количество синонимов и т. п. Обычно данные представляются в виде таблицы «объекты — признаки», которая может содержать незаполненные ячейки (*пропуски*).

	Признак 1	Признак 2	Признак 3	Признак 4
Объект 1	7	1,35	глагол	235
Объект 2	15	2,44	существительное	
Объект 3	9	0,67	наречие	554

Исследователя могут интересовать разные вопросы. Например, какое распределение имеет определённый признак, связаны или нет некоторые признаки между собой, однородны ли объекты по совокупности признаков или они разбиваются на компактные изолированные группы (*кластеры*), как построить модель прогнозирования одного из признаков на основе некоторых других и т. п.

Применение моделей и методов математической статистики часто позволяет получить наиболее полные и полезные с практической точки зрения ответы на подобные вопросы.

Рассмотрим пример, относящийся к демографии. Выясним, какое распределение имеет продолжительность жизни граждан России.[1] На рис. 1 приведена столбиковая диаграмма (также называемая *гистограммой*)[2] продолжительности жизни T в России в 2010 году.

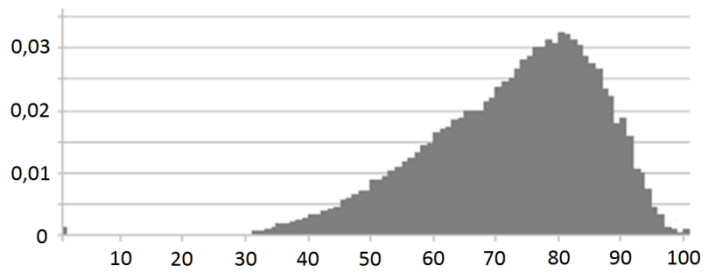

Рис. 1

Высота столбика гистограммы равна доле p_t тех, кто прожил t лет, среди всех умерших в России в течение 2010 года. Отметим, что «горка» заметно перекошена — правый склон намного круче. Наиболее высокие столбики, соответствующие наиболее вероятной продолжительности жизни,

[1] На основе сведений из статьи Горшкова В. А. «Распределение продолжительности жизни в России. Альтерна-
[2] Термин, впервые использованный Карлом Пирсоном в 1895 году, происходит от ἱστός — (др.-греч.) столб.

располагаются вблизи 80 лет. Небольшой «горбик» вблизи 65 лет соответствует наиболее вероятной продолжительности жизни мужчин.[3] Подобные диаграммы за разные годы представляют несомненный интерес для страховщиков и социальных служб.

Вероятность $F_T(t) = \mathbf{P}(T \le t)$, что выбранный наудачу россиянин проживёт не более t лет, равна сумме высот первых t столбиков гистограммы:

$$F_T(t) = p_1 + \ldots + p_t,$$

Иначе говоря, функция распределения $F_T(t)$ выражает «накопленную» долю людей, проживших не менее t лет.[4] Для произвольных возрастов $a < b$ доля людей, проживших больше a лет, но не более b лет, равна $F_T(b) - F_T(a)$ (рис. 2).

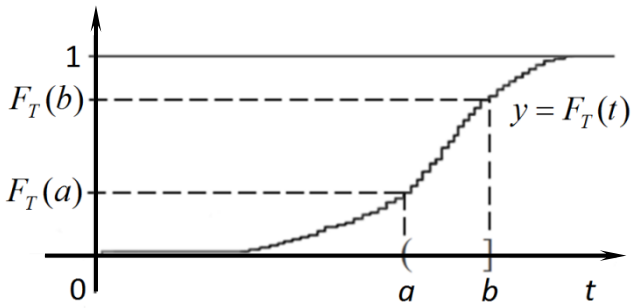

Рис. 2

Понятно, что полную информацию о продолжительности жизни во всей России получить затруднительно. Для оценки функции распределения $F_T(t)$ можно использовать выборочный метод. Скажем, изучить данные о смертности граждан, хранящиеся в нескольких случайно отобранных отделениях ЗАГС[5] из разных регионов страны. Предположим, что в результате запросов исследователь получил сведения о продолжительности жизни n человек. На основе выборки можно подсчитать частоты $\hat{p}_1, \hat{p}_2, \ldots,$ по которым вычисляется *выборочная функция распределения*

$$\hat{F}_T(t) = \hat{p}_1 + \ldots + \hat{p}_t.$$

Тогда для заданных возрастов $a < b$ оценкой неизвестной доли $F_T(b) - F_T(a)$ будет служить частота $\hat{F}_T(b) - \hat{F}_T(a)$. Принципиальный вопрос — насколько большим должен быть размер выборки n для обеспечения заданной точности оценки.

7.2. Статистические модели

Давайте пока ограничимся задачей выяснения вида распределения некоторого признака X, т. е. задачей получения достаточно точной оценки для неизвестной функции распределения $F_X(x)$. Простейшая *статистическая модель*, используемая для данной задачи, представляет собой некоторое семейство Ψ функций распределения $F(x, \theta)$, зависящих помимо аргумента x ещё и от

[3] По данным Росстата за период с 2010 по 2017 годы средняя продолжительность жизни в России постоянно увеличивалась и к июлю 2017 года достигла 72,5 лет (67,5 лет у мужчин, 77,4 лет у женщин).

[4] Раньше $F_T(t)$ называли *кумулятивной функцией распределения* (cumulative — (англ.) *накопленный*).

[5] Запись актов гражданского состояния.

параметра θ. Предполагается, что среди функций, входящих в Ψ, найдётся такая, которую может послужить достаточно хорошим приближением к функции распределения признака $F_X(x)$ (рис. 3).

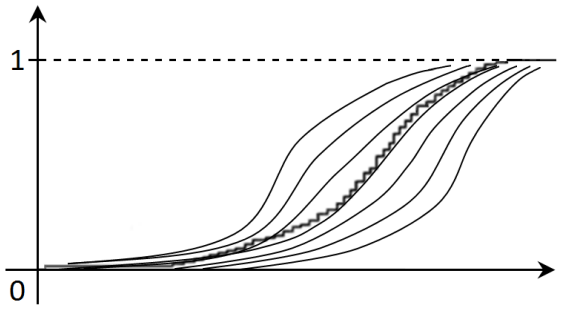

Рис. 3

Например, в модели испытаний Бернулли параметром является вероятность «успеха» p. В экспоненциальной (показательной) модели функции распределения

$$F_T(x) = \begin{cases} 0, & \text{если } x \le 0, \\ 1 - e^{-\lambda x}, & \text{если } x > 0 \end{cases}$$

зависят от параметра $\lambda > 0$ (рис. 4).

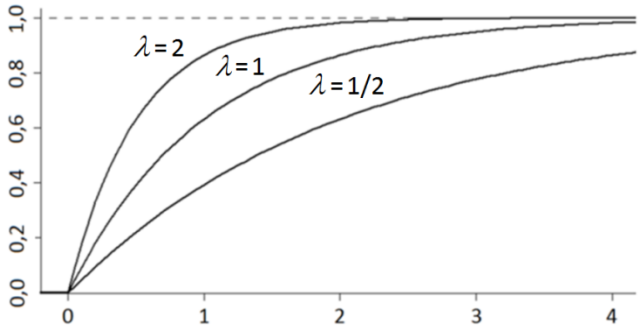

Рис. 4

Одной из (неоправданно) часто применяемых на практике является *нормальная модель*, зависящая от двух параметров: параметра сдвига μ и параметра масштаба $\sigma > 0$. Плотность распределения нормальной случайной величины X задаётся формулой

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

Нетрудно вычислить, что математическое ожидание $\mathsf{M}X = \mu$ и дисперсия $\mathsf{D}X = \sigma^2$. Для нормального распределения используется обозначение $N(\mu, \sigma^2)$. Случайная величина $Z = (X - \mu)/\sigma$ имеет распределение $N(0, 1)$ и называется *стандартной нормальной*.

Вопрос 1. Как доказать, что $\mathsf{M}Z = 0$ и $\mathsf{D}Z = 1$? (*Указание*. Используйте свойства математического ожидания и дисперсии.)

Стандартный нормальный закон с функцией распределения $\Phi(x)$ уже встречался в разделе 6.4 при обсуждении центральной предельной теоремы. График плотности $f_X(x)$, заданной формулой (1), получается из графика стандартной нормальной плотности $\varphi(x)$, изображённого на рис. 1 в теме 6, сдвигом на величину μ и растяжением в σ раз.

7.3. Оценки параметров

Определение. *Выборкой* называется набор из n независимых (см. раздел 6.1) и одинаково распределённых случайных величин $X_1, \ldots, X_n$.

Согласно приведённой ниже задаче 3 выборкой является набор $X(\omega_1), \ldots, X(\omega_n)$, где $\omega_1, \ldots, \omega_n$ — объекты конечной генеральной совокупности, выбранные наудачу с возвращением, X — произвольный признак отдельного объекта.

Пусть $\Psi = \{F(x, \theta)\}$ — некоторое семейство функций распределения, зависящих от параметра θ, который принимает значения из заданного множества Θ.

В математической статистике имеющиеся у исследователя данные — наблюдения $x_1, \ldots, x_n$ некоторого признака X — рассматриваются как реализация выборки $X_1, \ldots, X_n$, элементы которой распределены согласно одной из функций, входящих в параметрическое семейство $\Psi = \{F(x, \theta)\}$. Какой именно — неизвестно. Требуется выбрать в семействе Ψ функцию распределения, наиболее соответствующую (в некотором смысле) наблюдениям. Иначе говоря, требуется на основе наблюдений $x_1, \ldots, x_n$ оценить неизвестный параметр θ.

В качестве примера приведём модель *равномерного распределения на отрезке* $[0, \theta]$, в которой функции распределения из семейства Ψ задаются формулой

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

Параметр θ принадлежит множеству $\Theta = (0, \infty)$. Графики некоторых $F(x, \theta)$ изображены на рис. 5.

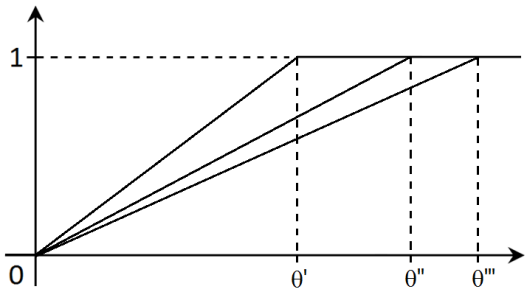

Рис. 5

Положим $X = \theta Y$, где случайная величина Y равномерно распределена на отрезке $[0, 1]$.

Вопрос 2. Какую функцию распределения имеет случайная величина X?

Пусть $Y_1, \ldots, Y_n$ — независимые и равномерно распределённые на $[0, 1]$ случайные величины. Тогда для заданного числа $\theta_0 > 0$ случайные величины $X_i = \theta_0 Y_i$, $i = 1, \ldots, n$, образуют выборку из равномерного распределения на отрезке $[0, \theta_0]$.

Следующие 10 наблюдений (округлённые до одной цифры после запятой)

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

были смоделированы при некотором θ_0 по формуле $X_i = \theta_0 Y_i$ с помощью компьютерной программы, имитирующей случайный выбор точек Y_i из отрезка $[0, 1]$. Попробуйте угадать (оценить) неизвестный параметр θ_0 на основе рис. 6, на котором изображены наблюдения.

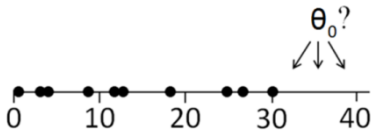

Рис. 6

Вопрос 3. Может ли параметр θ_0 быть равным: а) 30; б) 100?

Ответ: а) Нет: максимум наблюдений равен 30,1, поэтому θ_0 должен быть не меньше этого числа. б) Значение 100 допустимо, однако оно представляется крайне маловероятным. Действительно, вероятность, что при $\theta_0 = 100$ все 10 наблюдений окажутся меньше 30,1, равна $(30{,}1/100)^{10} \approx 6{,}1 \cdot 10^{-6}$.

Как выбрать наиболее подходящее к наблюдениям значение θ_0? Рассмотрим этот вопрос сразу в случае произвольной параметрической модели: как оценить неизвестный параметр θ по наблюдениям $x_1, \ldots, x_n$? Будем это делать с помощью некоторых функций n переменных $\hat{\theta}(x_1, \ldots, x_n)$.

Определение. *Статистическими оценками* называются функции $\hat{\theta}(X_1, \ldots, X_n)$.

В модели равномерного распределения на отрезке $[0, \theta]$ в качестве оценок неизвестного параметра θ разумно взять, например, функции

$$\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$$

или

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

Действительно, интуитивно понятно, что для выборки большого размера n наибольшее среди значений случайных величин $X_1, \ldots, X_n$ с большой вероятностью будет располагаться вблизи правого конца отрезка $[0, \theta]$.

В свою очередь, при увеличении размера выборки n среднее арифметическое

$$\overline{X} \equiv (X_1 + \ldots + X_n)/n,$$

также называемое *выборочным средним*, в силу закона больших чисел (см. раздел 6.3 из темы 6) будет сходиться по вероятности к $\mathrm{M}X_1 = \theta/2$ (середине отрезка $[0, \theta]$). Поэтому оценка $\hat{\theta}_2 = 2\overline{X}$ будет сходится по вероятности к самому параметру θ.

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

7.4. Несмещённость и состоятельность

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

$$\mathrm{M}\hat{\theta}(X_1, \ldots, X_n) = \theta, \tag{3}$$

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».

Оценка $\hat{\theta}_1 = \max\{X_1, \ldots, X_n\}$ для модели равномерного распределения на отрезке $[0, \theta]$ свойством несмещённости не обладает: максимум всегда недооценивает правую границу отрезка. Однако оценку $\hat{\theta}_1$ нетрудно подправить. Положим

$$\hat{\theta}_3 = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}. \tag{4}$$

Ниже в задаче 1 предлагается проверить, что $\hat{\theta}_3$ является несмещённой оценкой.

В свою очередь, оценка $\hat{\theta}_2 = 2\overline{X}$ не имеет смещения. Действительно, используя свойства математического ожидания (см. раздел 2.3), имеем:

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

Само по себе свойство несмещённости недостаточно для того, чтобы оценка хорошо приближала неизвестный параметр. Например, в модели испытаний Бернулли первый элемент выборки X_1 является несмещённой оценкой неизвестной вероятности «успеха» p: $\mathrm{M}X_1 = p$. Однако его возможные значения 0 и 1 даже не принадлежат параметрическому множеству $\Theta = (0, 1)$. Помимо несмещённости оценки необходимо, чтобы её погрешность стремилась к 0 с увеличением размера выборки n. Это свойство в математической статистике называется *состоятельностью.*

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

$$\mathbf{P}(|\hat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0 \quad \text{при} \quad n \to \infty, \tag{5}$$

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

Сходимость (5) означает, что любые отклонения оценки $\hat{\theta}$ от параметра θ становятся сколь угодно маловероятными при увеличении размера выборки n. Другими словами, состоятельность означает концентрацию с ростом n вероятностной массы около истинного значения параметра. На рис. 7 изображены графики плотностей распределения оценки $\hat{\theta}$ для двух разных значений n. Для состоятельной оценки $\hat{\theta}$ вероятностная масса её распределения, сконцентрированная внутри отрезка $[\theta - \varepsilon, \theta + \varepsilon]$ при сколь угодно малом $\varepsilon > 0$, стремится к 1 при $n \to \infty$.

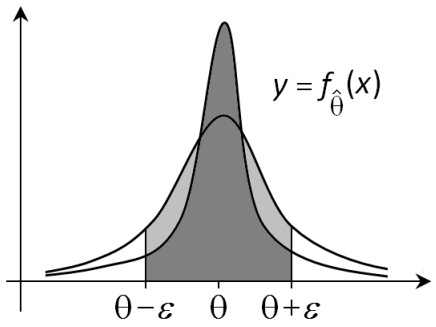

Рис. 7

Ниже в задаче 2 предлагается убедиться, что для модели равномерного распределения на отрезке $[0, \theta]$ оценки $\hat{\theta}_1$ и $\hat{\theta}_3$ являются состоятельными. В силу закона больших чисел оценка $\hat{\theta}_2 = 2\overline{X}$ также состоятельна. Таким образом, оценки $\hat{\theta}_2$ и $\hat{\theta}_3$ обе являются несмещёнными и состоятельными. Какая из них точнее?

Простейшей мерой точности несмещённой оценки служит её дисперсия (см. раздел 2.4). Ниже в задаче 7.1 требуется доказать, что

$$\frac{\mathrm{D}\hat{\theta}_3}{\mathrm{D}\hat{\theta}_2} = \frac{3}{n+2}. \qquad (6)$$

Заметим, что правая часть (6) при всех n не превосходит 1 (т. е. $\mathrm{D}\hat{\theta}_3 \le \mathrm{D}\hat{\theta}_2$) и стремится к 0 при $n \to \infty$. Следовательно, оценка $\hat{\theta}_3$ имеет преимущество в точности над $\hat{\theta}_2$, возрастающее с увеличением n.

Возвращаясь к приведённому выше числовому примеру с выборкой из $[0, \theta_0]$, 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 оценка $\hat{\theta}_3$ оказалась точнее, в то время как оценка $\hat{\theta}_2$ даже приняла недопустимое значение, которое меньше максимума выборки (30,1).

Задачи для решения на занятии

1) Для модели равномерного распределения на отрезке $[0, \theta]$ доказать, что оценка $\hat{\theta}_3$, определённая формулой (4), является несмещённой.

2) Для модели равномерного распределения на отрезке $[0, \theta]$ проверить, что оценка: а) $\hat{\theta}_1$; б) $\hat{\theta}_3$ является состоятельной.

3) Из генеральной совокупности Ω, содержащей m объектов, наудачу с возвращением выбираются объекты $\omega_1, \ldots, \omega_n$. Положим $\tilde{\omega} = (\omega_1, \ldots, \omega_n)$, $\tilde{\Omega} = \{\tilde{\omega}\}$. Тогда $|\tilde{\Omega}| = m^n$. Пусть X — некоторый признак отдельного объекта. Определим на $\tilde{\Omega}$ случайные величины $X_i(\tilde{\omega}) = X(\omega_i)$, $i = 1, \ldots, n$. Доказать, что случайные величины $X_1, \ldots, X_n$: а) одинаково распределены; б) независимы. (*Указание*. Используйте обозначение $\mathrm{A}_x = \{\omega: X(\omega) = x\}$.)

Домашнее задание

Если пятая буква вашей фамилии находится в диапазоне:

«**А – Е**», то «своими» являются задачи 7.1 и 7.5;
«**Ж – М**», то «своими» являются задачи 7.2 и 7.6;
«**Н – Р**», то «своими» являются задачи 7.3 и 7.7;
«**С – Я**», то «своими» являются задачи 7.4 и 7.8.

7.1) Вывести соотношение (6). (*Указание*. Используйте представление $X_i = \theta Y_i$, $i = 1, \ldots, n$, где $Y_1, \ldots, Y_n$ — независимые и равномерно распределённые на отрезке $[0, 1]$ случайные величины.)

7.2) Является ли эта оценка $\hat{\theta}_4 = (n+1)\min\{X_1, \ldots, X_n\}$ в модели равномерного распределения на отрезке $[0, \theta]$ несмещённой?

7.3) Является ли эта оценка $\hat{\theta}_4 = (n+1)\min\{X_1, \ldots, X_n\}$ в модели равномерного распределения на отрезке $[0, \theta]$ состоятельной?

7.4) Является ли состоятельной оценка $\hat{\mu} = \min\{X_1, \ldots, X_n\}$ в модели сдвига показательного закона с функцией распределения $F(x, \mu) = (1 - e^{-(x-\mu)})I_{\{x > \mu\}}$?

7.5) Является ли несмещённой оценка $\hat{\mu} = \min\{X_1, \ldots, X_n\}$ в модели сдвига показательного закона с функцией распределения $F(x, \mu) = (1 - e^{-(x-\mu)})I_{\{x > \mu\}}$?

7.6) Является ли состоятельной оценка $\hat{\sigma} = \overline{X}$ в модели масштаба показательного закона с функцией распределения $F(x, \sigma) = (1 - e^{-x/\sigma})I_{\{x>0\}}$?

7.7) Является ли несмещённой оценка $\hat{\sigma} = \overline{X}$ в модели масштаба показательного закона с функцией распределения $F(x, \sigma) = (1 - e^{-x/\sigma})I_{\{x>0\}}$?

7.8) Доказать, что

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

служит несмещённой оценкой для неизвестной дисперсии $\mathsf{D}X$. (*Указание*. Возведите в квадрат и воспользуйтесь тождеством $\mathsf{M}\overline{X}^2 = \mathsf{D}\overline{X} + (\mathsf{M}\overline{X})^2$, которое следует из формулы (7) темы 2).