И. ВЫСОЦКИЙ,
г. Москва

Задача 3

ЗАДАЧИ С УЛИЦЫ

Задача о такси в Майкопе

Сколько Яндекс-такси в Майкопе?

Замечательный город Курск состоит из нескольких удаленных друг от друга жилых массивов и промзон. Волгоград вытянут вдоль реки почти на 100 км. Санкт-Петербург — огромный мегаполис, к тому же разрезанный на части Невой с разводными мостами. Поэтому машины такси, работающие в таких городах, тоже «разбиты на кластеры». Закончив поездку в удаленный район, таксист старается взять «домашний» заказ.

Но если город не очень велик и расположен на местности компактно, то можно считать, что работающие машины такси случайно перемешаны и на протяжении нескольких дней их состав мало меняется. Например, такова столица Адыгеи город Майкоп. Его улицы пересекают друг друга под прямыми углами, а река Белая отсекает лишь небольшой участок (рис. 1).

Рис. 1

Среди достопримечательностей Майкопа — единственный в России Математический парк (рис. 2), дорожки в котором повторяют знаменитый граф Кенигсбергских мостов[1].

Рис. 2

[1] Подробно о Математическом парке и его создателях можно прочесть на странице math-park.ru.

Есть дополнительные материалы на сайте raum.math.ru.

Приехав в Майкоп на семинар учителей математики, автор поселился довольно далеко от университета, поэтому на протяжении трех дней активно пользовался Яндекс-такси, разъезжая между гостиницей, университетом, Математическим парком и центром города. В девятый раз приехала та же машина, которая приезжала на третий вызов, а до этого повторов не было.

Будем считать, что на выезд приезжает случайная машина и что на протяжении рассматриваемого периода времени совокупность работающих в городе такси неизменна. Можно ли оценить количество машин Яндекс-такси, работающих в городе, имея лишь эту информацию? Влияет ли на оценку то, что в девятый раз приехала именно третья машина, а не первая или пятая? При этих условиях можно сформулировать задачу.

Задача. Сколько машин Яндекс-такси работает в городе, если в последовательности случайных машин первый раз машина повторилась при девятой поездке?

Задачи оценивания отличаются от большинства математических задач тем, что, помимо информации, данной в условии, нужно иметь в виду метод, которым сделана оценка. Разные методы часто дают разные оценки. Кратко можно сказать, что оценку нельзя отрывать от метода. Мы используем три разных метода и получим три разные точечные оценки.

Оценка наибольшего правдоподобия

Принцип наибольшего правдоподобия можно сформулировать так: поскольку нет причин считать осуществившееся событие маловероятным, давайте считать его наиболее вероятным. Это соображение и дает метод, а полученную оценку называют *оценкой наибольшего или максимального правдоподобия* (ОМП).

Нужно предположить, что всего в городе n машин, и составить выражение для вероятности p_n события A «первое повторение случилось при девятой поездке». Индекс n подчеркивает, что вероятность ищется в предположении, что машин ровно n. Затем нужно выяснить, при каком n эта вероятность оказывается наибольшей. Полученное значение и будет ОМП.

Чтобы найти p_n, воспользуемся обычным правилом умножения вероятностей. В первый раз приехала какая-то машина. Вероятность этого 1 или $\frac{n}{n}$. Вероятность того, что на второй вызов приехала какая-то другая, равна $\frac{n-1}{n}$. При этом условии вероятность того, что на третий вызов приехала машина, которой не было раньше, равна $\frac{n-2}{n}$ и т.д. Таким образом, вероятность того, что восемь раз такси не повторялись, равна

$$\frac{n-1}{n}\cdot\frac{n-2}{n}\cdot\ldots\cdot\frac{n-7}{n}=\frac{(n-1)!}{(n-8)!\cdot n^7}.$$

Зато в девятый раз приехала одна из первых восьми машин, значит,

$$p_n=\frac{(n-1)!}{(n-8)!\cdot n^7}\cdot\frac{8}{n}=\frac{8(n-1)!}{(n-8)!\cdot n^8}$$

или

$$p_n=\frac{8n!}{(n-8)!\cdot n^9}, \qquad (1)$$

записываем кому как удобнее. Полученную вероятность называют *функцией правдоподобия*. В данном случае — это функция от аргумента n.

Нужно узнать, при каком значении n эта последовательность принимает наибольшее значение. Ясно, что постоянный множитель 8 в числителе можно отбросить, а это означает, что информация о том, была ли эта повторившаяся машина просто одна какая-то из восьми предыдущих или конкретно третья или пятая, не играет роли.

Чтобы найти наибольшее значение, воспользуемся прямым вычислением в MS Excel (рис. 3).

Рис. 3

Для вычисления факториалов можно использовать функцию ФАКТР(), но электронным таблицам тяжело вычислять факториалы больших чисел[2]. Поэтому сделаем преобразование, выразив p_n через p_{n-1}:

$$p_n = \frac{8n!}{(n-8)!\cdot n^9} = \frac{8(n-1)!}{((n-1)-8)!\cdot(n-1)^9}\cdot\frac{n}{n-8}\cdot\left(\frac{n-1}{n}\right)^9 =$$

$$= p_{n-1}\cdot\left(1-\frac{1}{n}\right)^9 : \left(1-\frac{8}{n}\right). \qquad (2)$$

Нужно честно вычислить лишь первую вероятность:

$$p_8 = \frac{8\cdot 8!}{0!\cdot 8^9} = \frac{7!}{8^7},$$

поскольку в городе не может быть меньше, чем восемь такси. Все последующие вероятности мы найдем с помощью рекурсии (2).

Ответ: ОМП равна 33.

Если вы относите себя к математическим туристам и считаете, что прямой подсчет не является удовлетворительным способом решения, можете прочитать следующую часть, где рассказано, как эту же оценку можно получить без MS Excel, оставаясь в рамках математики девятого класса.

Оценка максимального правдоподобия без MS Excel

Возвращаемся на шаг назад к формуле (2) и видим, что нужно лишь узнать, при каких n верно неравенство $p_n \ge p_{n-1}$, то есть множитель $\left(1-\frac{1}{n}\right)^9 : \left(1-\frac{8}{n}\right)$ больше или равен единице. Иными словами, нужно решить неравенство

$$\left(1-\frac{1}{n}\right)^9 \ge 1-\frac{8}{n}.$$

Первое, что приходит в голову, — это замена $x=\frac{1}{n}$. Получается неравенство

$$(1-x)^9 \ge 1-8x.$$

Уже понимая, что все плохо, раскроем скобки с помощью формулы бинома Ньютона, перенесем все члены в одну сторону и разделим обе части на x (он положителен):

$$1-9x+36x^2-84x^3+126x^4-\ldots-x^9 \ge 1-8x,$$

$$1-36x+84x^2-126x^3+\ldots+x^8 \le 0.$$

Мы (и не только мы) не умеем алгебраически решать неравенства восьмой степени, если только они не специально составлены хитрым автором задачника. Конкретно это неравенство не решается ни заменой, ни группировкой и никаким другим нашим любимым способом. Разумеется, можно было бы попробовать графическое решение исходного неравенства, но оно будет очень грубым.

Что делать — неясно. Но тут на помощь приходит нематематическое соображение. Майкоп — многотысячный город, где *много* такси. То есть число n не может быть малым. А значит, число x намного меньше единицы. Поэтому все слагаемые, начиная с $126x^3$, можно отбросить, считая их пренебрежимо малыми по сравнению с первыми тремя членами левой части. Нас не удивляют приближенные равенства. Значит, нас не должно удивить *приближенное неравенство*

$$84x^2 - 36x + 1 \le 0,$$

которое мы умеем решать обычными средствами:

$$\frac{9-2\sqrt{15}}{42} \le x \le \frac{9+2\sqrt{15}}{42},$$

то есть

$$\frac{42}{9+2\sqrt{15}} \le n \le \frac{42}{9-2\sqrt{15}}.$$

Учитывая, что n заведомо не меньше 8, видим, что последовательность p_n растет до тех пор, пока n остается меньше

$$\frac{42}{9-2\sqrt{15}} = 33{,}49\ldots,$$

а далее — убывает:

$$\ldots < p_{31} < p_{32} < p_{33} > p_{34} > p_{35} > \ldots,$$

и наибольшая вероятность равна p_{33} при $n = 33$.

Оценка наименьших квадратов

Метод наименьших квадратов основан на статистической устойчивости: наблюдаемые значения не могут сильно отличаться от ожидаемых. Например, частоты событий не могут слишком сильно отличаться от их вероятностей. Отсюда метод: *нужно минимизировать меру различия между наблюдаемым и ожидаемым*. Как устроить меру различия для нескольких наблюдений одновременно? Например, в качестве такой меры можно взять сумму квадратов разностей частот и вероятностей и найти, при каком значении n она минимальна. Полученную оценку называют *оценкой наименьших квадратов* (ОНК).

В нашем случае имеется последовательность случайных испытаний: повторилась ли машина при первой поездке (да/нет)? Очевидно, нет. Повторилась ли машина при второй поездке (да/нет)? Мы знаем, что не повторилась, хотя уже могла. И так далее: первым утвердительным станет ответ на вопрос, повторилось ли такси на девятой поездке.

[2] Этой трудности можно избежать, вычисляя вместо самой функции правдоподобия ее логарифм, но автор предпочитает обойтись минимумом средств, оставив простор для эксперимента читателю.

Поскольку в каждом из этих опытов наблюдение было единственным, частота события «машина повторилась» равна 1 или 0. Результат наблюдений можно записать в виде последовательности

$$0\,0\,0\,0\,0\,0\,0\,0\,1.$$

Теперь посмотрим, каковы вероятности этих событий. Разумеется, условные. Вероятность повторения такси при первой поездке 0, вероятность повторения при второй поездке равна $\frac{1}{n}$, и так далее до девятой поездки, где вероятность повторения равна $\frac{8}{n}$ (все это при условии, что раньше повторений не было). Получаем функцию

$$f(n)=(0-0)^2+\left(0-\frac{1}{n}\right)^2+\left(0-\frac{2}{n}\right)^2+\ldots$$

$$\ldots+\left(0-\frac{7}{n}\right)^2+\left(1-\frac{8}{n}\right)^2, \qquad (3)$$

наименьшее значение которой нужно найти.

Преобразуем ее:

$$f(n)=\left(\frac{1}{n}\right)^2+\left(\frac{2}{n}\right)^2+\ldots+\left(\frac{7}{n}\right)^2+\left(\frac{n-8}{n}\right)^2=$$

$$=\frac{1+4+\ldots+49+64+n^2-16n}{n^2}.$$

Квадраты сложим с помощью формулы Архимеда[3]:

$$f(n)=\frac{n^2-16n+204}{n^2}=1-\frac{16}{n}+\frac{204}{n^2}.$$

Замена $x=\frac{1}{n}$ дает квадратный трехчлен

$$204x^2-16x+1,$$

который принимает наименьшее значение при $x=\frac{8}{204}$. Значит, $n=\frac{204}{8}=25{,}5$.

Ответ: ОНК равна 25,5.

Оценка моментов

И здесь тоже не будем много теоретизировать. Известен закон больших чисел: *выборочные характеристики, скорее всего, не слишком сильно отклоняются от своих теоретических аналогов*. На этом основан *метод моментов*: будем считать, что наблюдаемое среднее значение случайной величины совпадает с ее математическим ожиданием, наблюдаемая дисперсия — с теоретической дисперсией и т.п. Оценку, полученную таким способом, называют *оценкой моментов* (ОМ или ОММ).

[3] Формула Архимеда для суммы первых натуральных квадратов: $1+4+\ldots+m^2=\frac{m(m+1)(2m+1)}{6}$.

Будем наблюдать случайную величину X «номер последней поездки без повторения». Выборка состоит из единственного наблюдения 8 (если бы по городу ездил еще кто-то, у этого кого-то тоже случилось повторение и это стало нам известно, то было бы уже два наблюдения, но, увы, таких сведений у нас нет).

Делать нечего, будем пользоваться тем, что есть:

$$\mathrm{E}\,X=8. \qquad (4)$$

Нужно выразить $\mathrm{E}\,X$. Для каждой поездки придумаем *индикатор* того, что ни в этой поездке, ни в предыдущих такси не повторялись. Достаточно взять n индикаторов, поскольку поездок без повторения не может быть больше n, то есть больше общего числа машин такси в городе. Индикаторы введем по следующему правилу:

$$I_k=\begin{cases}1, \text{ если до } k\text{-й поездки включительно повторов нет,}\\ 0, \text{ если есть.}\end{cases}$$

Пока повторов нет, индикаторы равны 1, а как только в какой-то поездке машина повторилась, индикатор для этой поездки и всех последующих равен 0. Поэтому

$$X=I_1+I_2+I_3+\ldots+I_n. \qquad (5)$$

Вероятность того, что $I_k=1$, найти несложно:

$$\mathrm{P}(I_k=1)=\frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-(k-1)}{n}=\frac{n!}{(n-k)!\cdot n^k}.$$

Значит,

$$\mathrm{E}\,I_k=1\cdot\frac{n!}{(n-k)!\cdot n^k}+0\cdot\left(1-\frac{n!}{(n-k)!\cdot n^k}\right)=\frac{n!}{(n-k)!\cdot n^k}.$$

Перейдем в равенстве (5) к математическим ожиданиям:

$$\mathrm{E}\,X=\frac{n!}{(n-1)!\cdot n}+\frac{n!}{(n-2)!\cdot n^2}+\ldots+\frac{n!}{1!\cdot n^{n-1}}+\frac{n!}{0!\cdot n^n}=$$

$$=\frac{n!}{n^n}\cdot\left(\frac{n^0}{0!}+\frac{n^1}{1!}+\frac{n^2}{2!}+\ldots+\frac{n^n}{n!}\right)$$

(в последнем выражении изменен порядок слагаемых в скобках).

Подставим результат в равенство (4):

$$\frac{n!}{n^n}\cdot\left(\frac{n^0}{0!}+\frac{n^1}{1!}+\frac{n^2}{2!}+\ldots+\frac{n^n}{n!}\right)=8.$$

Это уравнение выглядит намного страшнее, чем неравенство восьмой степени. Тем не менее его можно решить приближенно, причем довольно остроумным способом. Умножим и разделим левую часть на e^n:

$$\frac{n!e^n}{n^n}\cdot\left(\frac{n^0}{0!}e^{-n}+\frac{n^1}{1!}e^{-n}+\frac{n^2}{2!}e^{-n}+\ldots+\frac{n^n}{n!}e^{-n}\right).$$

Выражение в скобках равно вероятности события «некоторая величина Y, имеющая распределение Пуассона со средним значением $\lambda = n$, не превосходит своего среднего, то есть n»:

$$\frac{n^0}{0!}e^{-n}+\frac{n^1}{1!}e^{-n}+\frac{n^2}{2!}e^{-n}+\ldots+\frac{n^n}{n!}e^{-n}=\mathrm{P}(Y\le n).$$

При целом среднем λ медиана распределения Пуассона тоже равна λ. Это можно пояснить с помощью рисунка 4: при больших λ распределение Пуассона очень близко к нормальному, а математическое ожидание нормального распределения совпадает с его медианой.

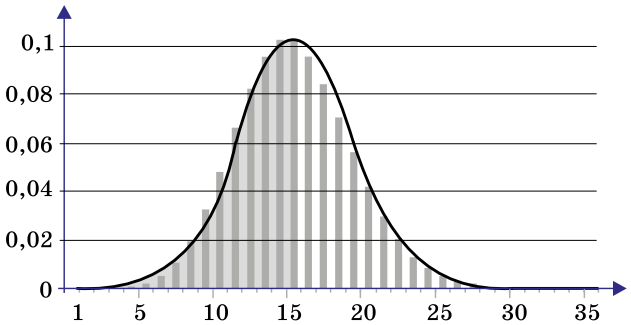

Рис. 4

На рисунке 4 изображена диаграмма распределения Пуассона с $\lambda = 15$ (для примера) и функция плотности нормального распределения со средним 15 и дисперсией 15. Площадь закрашенной фигуры (левая половина) равна в точности $\frac{1}{2}$. Поэтому и сумма длин соответствующих столбиков приблизительно равна $\frac{1}{2}$. Значит,

$$\mathrm{P}(Y\le n)\approx\frac{1}{2}.$$

Остался первый множитель $\frac{n!e^n}{n^n}$, который найдем приближенно с помощью формулы Стирлинга $n!\approx\sqrt{2\pi n}n^ne^{-n}$:

$$\sqrt{2\pi n}\cdot\frac{n^n}{e^n}\cdot\frac{e^n}{n^n}\cdot\frac{1}{2}\approx 8,$$

откуда

$$n\approx\frac{128}{\pi}=40{,}7436\ldots$$

Ответ: ОММ равна 40,74...

Какая из трех полученных оценок точнее? На этот вопрос ответа нет, по крайней мере при нашей скудной информации. Дело в том, что *любая точечная оценка сама по себе является случайной величиной*. Свойства этих оценок, например, их математические ожидания и дисперсии, известны в отдельных важных случаях, да и то лишь до определенной степени. К сожалению, наш случай в список изученных не попадает. Интуитивное желание полагаться на что-то промежуточное и избегать экстремальных значений обращает наше пристальное внимание на ОМП, которая меньше, чем ОММ, но больше, чем ОНК. Но интуиция — плохой советчик в подобных делах.

Лучший способ разобраться — моделирование. В приложении (файл MS Excel) на странице «Моделирование» проводится 27 одинаковых экспериментов. Нужно указать количество такси (по умолчанию 41), в шестой строке указано, на какой поездке машина первый раз повторилась в каждом эксперименте. Справа — описательные характеристики полученной выборки. Таблицу можно расширять, увеличивая число экспериментов, а можно много раз ее пересчитывать, поставив курсор в любую пустую ячейку и нажимая клавишу Delete. Каждый раз будут получаться новые выборки, которые можно объединять с предыдущими, получая растущую выборку.

Задачи для самостоятельного решения

1. Пусть, как и раньше, мы располагаем информацией только о девяти поездках, но известно, что машина повторилась не на девятой, а на шестой поездке, но больше повторов не было.

а) Как изменится ОМП?

б) Как изменится ОНК?

2. Найдите ОМП числа машин такси, обладая следующей информацией: всего поездок было 10 и машины повторялись два раза — при третьей поездке и при последней. Есть ли в этом условии лишняя информация (не влияющая на оценку)?

3. Мы получили дополнительную информацию от друга о его поездках в том же городе, при тех же условиях и о том, когда у него случился первый повтор машины. Как модифицировать метод наибольшего правдоподобия для оценки числа машин?

4. Как модифицировать метод моментов в условии задачи 3?

5. Как модифицировать метод наименьших квадратов в условии задачи 3?

Ответы к задачам из статьи «Задача о варенье»

2. $\left(\frac{1}{500}+1\right)^{249}$, то есть приблизительно 1,645 таза.

3. По формуле $\mathrm{E}J_5=e^5-4e^4+\frac{9}{2}e^3-\frac{4}{3}e^2+\frac{1}{24}$ получается 10,6666620686224... Приближенная формула $2\cdot5+\frac{2}{3}$ дает значение 10,(6).