ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
Физический факультет
Кафедра высшей математики

А. А. БЫСТРОВ, А. П. КОВАЛЕВСКИЙ,
В. И. ЛОТОВ

ПРАКТИКУМ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ

Учебное пособие

Новосибирск
2009

УДК 519.2 (075)
ББК В171я73-1
Б955

Быстров А. А., Ковалевский А. П., Лотов В. И. Практикум по теории вероятностей: Учеб. пособие / Новосиб. гос. ун-т. Новосибирск, 2009. 120 с.

ISBN 978–5–94356–669–1

Настоящее учебное пособие подготовлено для студентов I курса Физического факультета, изучающих теорию вероятностей во втором семестре. Пособие содержит теоретический материал, задачи для решения в классе и для самостоятельного решения.

Все замечания по содержанию пособия просим передавать авторам. Они будут с благодарностью приняты и учтены в следующих изданиях.

Рецензент
канд. физ.-мат. наук Н. И. Чернова

ISBN 978–5–94356–669–1

Оглавление

Глава 1. Элементарная вероятность

§1. Классическая вероятностная модель

Пусть множество элементарных исходов Ω непусто и содержит конечное число элементов, а класс событий состоит из всех подмножеств Ω. Предположим, что все элементарные исходы равновозможны. Тогда вероятность любого события A вычисляется по формуле:

$$\mathbf{P}(A) = \frac{N(A)}{N(\Omega)}, \tag{1}$$

где $N(A)$ обозначает число элементов множества A.

Напомним несколько комбинаторных формул, полезных при решении задач в рамках классической вероятностной модели.

Количество различных выборок объема k из совокупности объема n *без возвращения и с учетом порядка* равно

$$A_n^k = \frac{n!}{(n-k)!}.$$

Это число называется *числом размещений* *из n элементов по k.*

Количество различных выборок объема k из совокупности объема n *без возвращения и без учета порядка* равно

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

Это число называется *числом сочетаний* *из n элементов по k.*

Количество различных выборок объема k из совокупности объема n *с возвращением и с учетом порядка* равно n^k.

Количество различных выборок объема k из совокупности объема n *с возвращением и без учета порядка* равно C^k_{n+k-1}.

Пример 1.1. Последовательно брошены две монеты. Найти вероятность события A ={Появится хотя бы одна решка}.
Решение. В данном эксперименте четыре равновозможных исхода: {РР}, {ГР}, {РГ}, {ГГ}. Событие состоит из трех исходов {РР}, {ГР}, {РГ}. Таким образом, его вероятность равна 3/4.

Пример 1.2. Какова вероятность того, что в наугад выбранном семизначном телефонном номере нет повторяющихся цифр?
Решение. Элементарные исходы данного эксперимента равновероятны и представляют из себя *упорядоченные* выборки объема 7 из набора цифр $\{0, 1, \ldots, 9\}$. Таким образом, для подсчета общего количества элементарных исходов здесь пригодна схема с возвращением и с учетом порядка:

$$N(\Omega) = 10^7.$$

Подсчитаем теперь количество «благоприятных» элементарных исходов, то есть таких выборок, в которых нет совпадающих элементов. Очевидно, это количество равно числу размещений из 10 по 7 так как здесь реализуется схема с учетом порядка, но без возвращения. Искомая вероятность вычисляется следующим образом:

$$\mathbf{P}(A) = \frac{N(A)}{N(\Omega)} = \frac{A^7_{10}}{10^7} = 0,06048.$$

Пример 1.3. Из колоды, насчитывающей 32 карты, наугад извлекается без возвращения 10 карт. Какова вероятность того, что среди выбранных карт найдется хотя бы один туз?
Решение. Пусть A ={найдется хотя бы один туз}. Вычислим вероятность события A по формуле

$$\mathbf{P}(A) = 1 - \mathbf{P}(\overline{A}),$$

где *дополнительное* событие $\overline{A}$ состоит в том, что среди выбранных карт нет ни одного туза.

Вероятность события $\overline{A}$ найдем по формуле (1). Элементарные исходы данного эксперимента равновозможны и представляют собой неупорядоченные выборки без возвращения объема 10 из совокупности, состоящей из 32 различных карт. Общее число элементарных исходов составляет C_{32}^{10}. Подсчитаем теперь число исходов, благоприятствующих событию $\overline{A}$. Это число совпадает с числом выборок без возвращения объема 10 из совокупности, состоящей из 28 карт (все карты, кроме тузов) и равно C_{28}^{10}. Таким образом, искомая вероятность равна

$$\mathbf{P}(A) = 1 - \mathbf{P}(\overline{A}) = 1 - \frac{C_{28}^{10}}{C_{32}^{10}} = \frac{5729}{7192}.$$

Домашнее задание 1

1.1. Буквы, составляющие фамилию студента, написали на карточках, затем карточки перетасовали и стали выкладывать в ряд в случайном порядке. Какова вероятность того, что в результате получится фамилия студента?

1.2. n книг произвольным образом расставляются на книжной полке. Какова вероятность того, что две фиксированные книги окажутся стоящими рядом?

1.3. У человека в кармане n ключей, из которых только один подходит к его двери. Ключи последовательно извлекаются (без возвращения) до тех пор, пока не появится нужный ключ. Найти вероятность того, что нужный ключ появится при k-м извлечении.

1.4. Из колоды, насчитывающей 36 карт, наугад извлекаются 6 карт. Какова вероятность того, что:

а) среди них окажется туз пик;

б) среди них окажется ровно один туз;

в) среди них окажутся ровно две бубновые карты;

г) среди них окажется хотя бы одна бубновая карта?

1.5. В лотерее n билетов, из которых m выигрышных. Некто приобретает k билетов. Найти вероятность того, что хотя бы один билет окажется выигрышным.

1.6. В лифт восьмиэтажного дома на первом этаже входят 5 человек. Независимо от других каждый может выйти с равными шансами на любом этаже, начиная со второго. Какова вероятность того, что:

а) все выйдут на четвертом этаже;

б) все пятеро выйдут на одном и том же этаже;

в) все пятеро выйдут на разных этажах?

1.7. Числа 1; 2; ...; n расставлены случайным образом. Предполагая, что различные расположения чисел равновероятны, найти вероятность того, что числа 1, 2, 3 расположены в порядке возрастания, но не обязательно рядом.

1.8. Найти вероятность того, что в наугад выбранном трехзначном автомобильном номере:

а) все цифры одинаковы;

б) все цифры различны;

в) только две одинаковые цифры.

Задачи для решения в классе

1.9. Однократно бросается игральная кость. Найти вероятность того, что:

а) выпадет число 3;

б) выпадет число, отличное от трех;

в) выпадет число, не меньшее трех.

1.10. Однократно бросается пара игральных костей. Найти вероятность того, что:

а) сумма выпавших очков окажется равна трем;

б) выпадут одинаковые грани;

в) сумма выпавших очков окажется не меньше шести.

1.11. Ребенок играет с десятью буквами разрезной азбуки: А, А, А, Е, И, К, М, М, Т, Т. Какова вероятность того, что при

случайном расположении букв в ряд он получит слово «МАТЕМАТИКА»?

1.12. Найти вероятность того, что в наугад выбранном четырехзначном автомобильном номере:
а) все цифры одинаковы;
б) все цифры различны;
в) ровно три одинаковые цифры;
г) только две одинаковые цифры;
д) две пары одинаковых цифр.

1.13. На шахматную доску из 64 клеток ставятся наудачу две ладьи разного цвета. С какой вероятностью они не будут «бить» друг друга?

§2. Комбинаторика. Геометрическая вероятность

Рассмотрим следующую типичную задачу.

Пример 2.1. В урне находится n белых и m черных шаров. Наугад выбираем $k \leq n+m$ шаров (без возвращения). Какова вероятность того, что среди извлеченных шаров окажется ровно $l \leq n$ белых и $k-l \leq m$ черных?

Решение. Воспользуемся формулой (1). Общее количество элементарных исходов равно C_{n+m}^k (выборка без возвращения и без учета порядка). Подсчитаем количество благоприятных исходов. Ровно l белых шаров можно выбрать C_n^l способами, а черные шары можно выбрать C_m^{k-l} способами. Общее количество благоприятных исходов равно произведению этих чисел. Таким образом, искомая вероятность равна

$$\frac{C_n^l C_m^{k-l}}{C_{n+m}^k}.$$

Эта формула называется формулой *гипергеометрического* распределения.

Рассмотрим теперь непрерывный аналог классической вероятностной модели. Пусть множество исходов эксперимента Ω представляет из себя ограниченное подмножество в $\mathbb{R}^n$. Через $\boldsymbol{\lambda}(A)$ будем обозначать $n-$мерный объем множества A. Предположим, что $\boldsymbol{\lambda}(\Omega) > 0$ и положим для любого $A \subseteq \Omega$

$$\mathbf{P}(A) = \frac{\boldsymbol{\lambda}(A)}{\boldsymbol{\lambda}(\Omega)}. \tag{2}$$

Это определение вероятности называется *геометрическим*.

Пример 2.2 (задача о встрече). Два деловых человека Ф. и Ч. договорились встретиться в условленном месте между полуночью и часом ночи, причем каждый ждет другого в течение

десяти минут, после чего уходит. Какова вероятность того, что эти два человека встретятся, если каждый может прийти в любое время в указанный промежуток независимо от другого?
Решение. Элементарным исходом этого эксперимента является точка (X, Y) в единичном квадрате, абсцисса которой — время прихода Ф., а ордината — время прихода Ч.:

$$\Omega = \{(X, Y) : 0 \le X \le 1, 0 \le Y \le 1\}.$$

Тогда событие $A = \{$Ф. и Ч. встретятся$\}$ можно записать в виде

$$A = \{(X, Y) : |X - Y| < 1/6\}$$

(см. рис.1). Искомая вероятность вычисляется по формуле (2) и равняется отношению площадей множеств A и Ω:

$$\mathbf{P}(A) = \frac{\lambda(A)}{\lambda(\Omega)} = \frac{11}{36}.$$

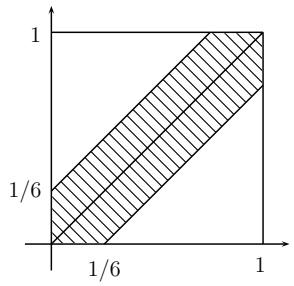

Рис. 1: Задача о встрече

Пример 2.3. На отрезок единичной длины наудачу независимо друг от друга брошены две точки, которые делят отрезок на три части. Пусть A — событие, состоящее в том, что из этих частей можно составить треугольник. Какова вероятность A?
Решение. Обозначим через X координату первой точки, а через Y — координату второй. Как и в предыдущем примере,

множество элементарных исходов представляет из себя единичный квадрат. Длины полученных частей равны $\min(X,Y), |Y-X|, 1-\max(X,Y)$. Чтобы из этих отрезков можно было составить треугольник, должны выполняться следующие соотношения:

$$\begin{cases} \min(X,Y) + |Y-X| > 1 - \max(X,Y); \\ |Y-X| + 1 - \max(X,Y) > \min(X,Y); \\ \min(X,Y) + 1 - \max(X,Y) > |Y-X|. \end{cases} .$$

Площадь множества A (см. рис. 2) равна $1/4$. Таким образом

$$\mathbf{P}(A) = \frac{\boldsymbol{\lambda}(A)}{\boldsymbol{\lambda}(\Omega)} = \frac{1}{4}.$$

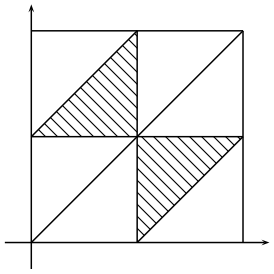

Рис. 2: Пример 2.3

Домашнее задание 2

2.1. Из чисел 1, 2, ..., 49 наугад выбираются и фиксируются 6 чисел, считающиеся выигрышными. Некто, желающий выиграть, наугад называет свои 6 чисел из 49. Какова вероятность, что среди названных им чисел окажется не менее трех выигрышных?

2.2. Группа, состоящая из $3n$ юношей и 3 девушек, делится произвольным образом на три равные по количеству подгруппы. Какова вероятность, что все девушки окажутся в разных подгруппах?

2.3. n студентов произвольным образом расходятся по k аудиториям. Какова вероятность, что в первой аудитории окажется n_1 студентов, во второй — n_2 студентов, ..., в k-й аудитории — n_k студентов, $n_1 + \ldots + n_k = n$?

2.4. В купейный вагон (9 купе по 4 места) продано №+4 билетов. Найти вероятность того, что занятыми оказались ровно пять купе (через № обозначен номер студента по списку группы).

2.5. Из отрезка [0, 1] наугад выбирается число. Какова вероятность, что в десятичной записи этого числа вторая цифра после запятой будет двойкой?

2.6. В квадрат с вершинами (0, 0), (0, 1), (1, 0), (1, 1) наудачу брошена точка. Обозначим ее координаты через X; Y. Предполагается, что вероятность попадания в область, лежащую целиком внутри квадрата, зависит лишь от площади этой области и пропорциональна ей:

а) доказать, что для $0 < u < 1,\ \ 0 < v < 1$ выполнено

$$P\{X < u, Y < v\} = P\{X < u\}P\{Y < v\} = uv;$$

б) найти для $0 < t < 1$ вероятности

1) $P\{|X - Y| < t\}$; 2) $P\{XY < t\}$;

3) $P\{\max(X, Y) < t\}$; 4) $P\{\min(X, Y) < t\}$;

в) найти $P\{X + Y < t\}$ для $0 < t < 2$.

2.7. На отрезок длины l произвольным образом брошены три точки. Пусть X, Y, Z — расстояния до этих точек от левого конца отрезка. Какова вероятность, что из отрезков с длинами X, Y и Z можно составить треугольник?

2.8. На отрезок единичной длины произвольным образом брошены две точки, которые делят отрезок на три части. Какова вероятность, что из этих частей можно составить треугольник?

Задачи для решения в классе

2.9. Недобросовестный казначей заменил 3 из 50 золотых монет в казне на фальшивые. Султан взвешивает 3 наугад выбранных монеты. Какова вероятность того, что казначей будет уличен?

2.10. Группа, состоящая из $2n$ девушек и $2n$ юношей, делится произвольным образом на две равные по количеству подгруппы. Найти вероятность того, что в каждой подгруппе окажется поровну юношей и девушек.

2.11. В ящике имеется 4 зеленых, 5 синих и 6 красных шаров. Наугад выбирается два шара. Какова вероятность того, что:

а) это будут синий и зеленый шары;

б) шары окажутся одного цвета;

в) шары окажутся различных цветов.

2.12. n различных шаров произвольным образом раскладываются по n ящикам. Какова вероятность, что при этом ровно один ящик окажется пустым?

2.13. Из колоды, насчитывающей 52 карты, наугад извлекают 6 карт. Какова вероятность, что среди них будут представители всех четырех мастей?

2.14. Отрезок длины l ломается в произвольной точке. Какова вероятность, что длина наибольшего обломка превосходит $2l/3$?

2.15. Точка бросается наудачу в квадрат. Найти вероятность того, что точка попадет в круг, вписанный в этот квадрат.

2.16. Точка бросается наудачу в треугольник с вершинами в точках (0, 0),(2, 0) и (0, 1). Найти вероятность того, что:

а) абсцисса точки окажется больше 1/2;

б) ордината точки окажется больше 1/2.

§3. Независимые события. Схема Бернулли

События A и B называются *независимыми*, если вероятность пересечения этих событий равна произведению их вероятностей:

$$\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B).$$

События $A_1, \ldots, A_n$ называются *независимыми в совокупности*, если для любого $k \leq n$ и для любого набора индексов $1 \leq i_1 < \ldots < i_k \leq n$ имеет место равенство

$$\mathbf{P}(A_{i_1} \ldots A_{i_k}) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}(A_{i_k}).$$

События $A_1, \ldots, A_n$ называются *независимыми попарно*, если предыдущее равенство имеет место при $k = 2$.

Пример 3.1. Последовательно брошены две монеты. Определить, зависимы ли события $A = \{$выпал герб на первой монете$\}$ и $B = \{$выпала хотя бы одна решка$\}$.
Решение. Вычислим вероятности каждого из событий A, B, AB и проверим равенство $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$. Имеем классическую вероятностную модель. Общее число равновозможных исходов эксперимента равно 4. Число исходов, благоприятствующих событию A равно двум, событию B — трем, и, наконец, событию AB — единице. Таким образом

$$\mathbf{P}(A)\mathbf{P}(B) = \frac{2}{4} \cdot \frac{3}{4} = \frac{3}{8},$$

тогда как

$$\mathbf{P}(AB) = \frac{1}{4}.$$

События зависимы.

Пример 3.2. Одна деталь изготовлена на станке A, а вторая — на станке B. Вероятность изготовления бракованной детали для станка A составляет $0,1$, а для станка B она равна $0,2$.

Найти вероятность того, что ровно одна деталь из двух является бракованной.
Решение. Представим событие $C = \{$ровно одна деталь бракована$\}$ в виде объединения двух несовместных событий $C_1 = \{$только первая деталь бракована$\}$ и $C_2 = \{$только вторая деталь бракована$\}$. В силу независимости событий {станок А выдал брак} и {станок В выдал брак}

$$\mathbf{P}(C_1) = 0,1 \cdot (1 - 0,2) = 0,08,\ \mathbf{P}(C_2) = (1 - 0,1) \cdot 0,2 = 0,18.$$

Таким образом, искомая вероятность

$$\mathbf{P}(C) = \mathbf{P}(C_1) + \mathbf{P}(C_2) = 0,26.$$

Пусть у нас есть последовательность независимых в совокупности испытаний, причем каждое испытание имеет два возможных исхода: «успех» и «неудача». Пусть, кроме того, вероятность успеха не меняется от испытания к испытанию и равна p. Такая модель называется *схемой Бернулли*. Вероятность появления k успехов в n испытаниях вычисляется по *формуле Бернулли*:

$$\mathbf{P}(S_n = k) = C_n^k p^k (1 - p)^{n-k},\quad k = 0, 1, \ldots, n.$$

Пример 3.3. Вася выигрывает у Пети в «Quake» в среднем три партии из четырех. Найти вероятность того, что он выиграет ровно три из четырех сыгранных партий.
Решение. Вероятность успеха равна $0,75$. По формуле Бернулли искомая вероятность равна

$$\mathbf{P}(S_4 = 3) = C_4^3 \,(0,75)^3 \cdot 0,25 = \frac{27}{64}.$$

Домашнее задание 3

3.1. Производят $n > 1$ независимых случайных перестановок букв фамилии студента. Найти вероятность того, что:

а) хотя бы раз получилась фамилия студента;

б) каждый раз получалась фамилия студента;

в) в последний раз получилась фамилия студента.

Сравнить вероятности, найденные в пунктах (а), (б), (в).

3.2. Пусть событие A не зависит от самого себя. Доказать, что тогда $P(A)$ равна 0 или 1.

3.3. Стрелок A поражает мишень с вероятностью 0,6, стрелок B — с вероятностью 0,5, стрелок C — с вероятностью 0,4. Стрелки дали залп по мишени. Какова вероятность, что ровно две пули попали в цель?

3.4. Двое играют в игру, поочередно бросая монету. Выигравшим считается тот, кто первым получит герб. Найти вероятность того, что игра закончится на k-м бросании. Какова вероятность выигрыша для игрока, начинающего игру?

3.5. 10 любителей подледного лова рыбы независимо друг от друга произвольным образом размещаются на льду озера, имеющего форму круга радиуса 1 км. Какова вероятность того, что не менее 5 рыбаков расположатся на расстоянии более 200 м от берега?

3.6. В шар радиуса R наудачу бросаются n точек. Найти вероятность того, что расстояние от центра шара до ближайшей точки будет не меньше a, $0 < a < R$.

3.7. Найти вероятность того, что в n испытаниях схемы Бернулли с вероятностью успеха p появятся $m + l$ успехов, причем l успехов появятся в последних l испытаниях.

3.8. В круг вписан квадрат. Найти вероятность того, что из 10 точек, брошенных наудачу в круг, четыре попадут в квадрат, три — в нижний сегмент, и по одной — в оставшиеся три сегмента.

Задачи для решения в классе

3.9. События $A_1;\ \ldots;\ A_n$ независимы, известны вероятности $p_i = P(A_i);\ \ i = 1;\ \ldots;\ n$. Найти вероятность того, что:

а) произойдет ровно одно из A_i;

б) не произойдет ни одно из A_i;

в) произойдет хотя бы одно из A_i.

3.10. Что вероятнее, выиграть у равносильного противника 3 партии из 4 или 5 партий из 8?

3.11. Шахматисты A и B решили сыграть между собой матч. Известно, что A выигрывает каждую партию у B с вероятностью $2/3$, и с вероятностью $1/3$ проигрывает. В связи с этим для победы в матче игроку A нужно набрать 4 очка, а игроку B для победы достаточно набрать 2 очка (за выигрыш в партии дается очко, за проигрыш — 0 очков, ничьих нет). Равны ли шансы на успех?

3.12. Найти вероятность того, что k-й по порядку успех в серии последовательных испытаний Бернулли произойдет на l-м испытании.

3.13. На отрезок [0, 10] наудачу брошено 5 точек. Найти вероятность того, что две точки попадут в [0, 2], одна — в [2, 3] и две — в [3, 10].

§4. Условные вероятности

Условной вероятностью $\mathbf{P}(A|B)$ события A при условии, что произошло событие B ненулевой вероятности, называется число

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}.$$

Пример 4.1. Известно, что при бросании трех монет выпало не менее двух гербов. Найти вероятность того, что на одной из монет выпала решка.
Решение. Пусть $A = \{$выпала решка$\}$, $B = \{$выпало не менее двух гербов$\}$. Имеем классическую вероятностную модель. Число равновозможных исходов эксперимента равно 8. Событие B состоит из четырех исходов: {ГГГ, ГГР, ГРГ, РГГ}. Событие AB состоит из трех исходов: {ГГР, ГРГ, РГГ}. Таким образом

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} = \frac{3/8}{1/2} = \frac{3}{4}.$$

Пусть теперь имеется событие A и попарно несовместные события положительной вероятности $B_1, \ldots, B_n$ такие, что $A \subseteq (B_1 \cup \ldots \cup B_n)$. Тогда вероятность события A можно вычислить по *формуле полной вероятности:*

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A|B_i)\mathbf{P}(B_i).$$

В тех же условиях, если произошло событие A ненулевой вероятности, то условные (или *апостериорные*) вероятности гипотез B_i могут быть вычислены по *формуле Байеса*:

$$\mathbf{P}(B_i|A) = \frac{\mathbf{P}(A|B_i)\mathbf{P}(B_i)}{\sum\limits_{j=1}^{n} \mathbf{P}(A|B_j)\mathbf{P}(B_j)}.$$

Пример 4.2. Вероятность изготовления бракованной детали для станка A составляет $0,1$, а для станка B она равна $0,2$. Найти вероятность того, что взятая наугад деталь является бракованной, если на первом станке изготавливается в три раза больше деталей, чем на втором.
Решение. Обозначим $H_1 = \{$взятая наугад деталь изготовлена на станке A$\}$, $H_2 = \{$взятая наугад деталь изготовлена на станке B$\}$, $C = \{$взятая наугад деталь бракована$\}$. По условию задачи

$$\mathbf{P}(H_1) = 0,75,\ \mathbf{P}(H_2) = 0,25,\ H_1 \cap H_2 = \emptyset,\ A \subseteq (H_1 \cup H_2) = \Omega,$$

$$\mathbf{P}(C|H_1) = 0,1, \mathbf{P}(C|H_2) = 0,2.$$

Тогда по формуле полной вероятности

$$\mathbf{P}(C) = \mathbf{P}(C|H_1)\mathbf{P}(H_1) + \mathbf{P}(C|H_2)\mathbf{P}(H_2) = 0,075 + 0,05 = 0,125.$$

Пример 4.3. Пусть теперь в условиях предыдущей задачи взятая наугад деталь оказалась бракованной. Найти вероятность того, что она была изготовлена на станке A.
Решение. По формуле Байеса

$$\mathbf{P}(H_1|C) = \frac{\mathbf{P}(C|H_1)\mathbf{P}(H_1)}{\mathbf{P}(C|H_1)\mathbf{P}(H_1) + \mathbf{P}(C|H_2)\mathbf{P}(H_2)} = \frac{0,075}{0,125} = 0,6.$$

Домашнее задание 4

4.1. Наудачу выбирают число первых букв от 2 до m из фамилии студента (здесь m — общее число букв в фамилии) и осуществляют их случайную перестановку. Найти вероятность того, что в результате получится фамилия студента. Найти вероятность того, что выбрали две первых буквы, если известно, что фамилия студента получилась.

4.2. Пусть в условиях задачи 3.3 известно, что две пули из трех попали в цель. Какова вероятность того, что промахнулся C?

4.3. Чтобы найти нужную книгу, студент решил обойти 3 библиотеки. Для каждой библиотеки одинаково вероятно, есть в фондах эта книга или нет, и если книга есть в фондах, то с вероятностью 0,5 она не занята другим читателем. Какова вероятность того, что студент найдет книгу, если известно, что библиотеки комплектуются независимо одна от другой?

4.4. Из n экзаменационных билетов студент знает m, поэтому, если он зайдет первым на экзамен, то с вероятностью m/n он вытащит «хороший» билет. Какова вероятность вытащить «хороший» билет, если студент зайдет на экзамен вторым?

4.5. Допустим, что вероятность попадания в цель при одном выстреле равна p, а вероятность поражения цели при k попаданиях равна $1-q^k$. Какова вероятность того, что цель поражена, если было произведено n выстрелов?

4.6. В продажу поступают телевизоры трех заводов. Продукция первого завода содержит 5% телевизоров со скрытым дефектом, второго — 3%, и третьего — 1%. Какова вероятность приобрести исправный телевизор, если в магазин поступило 20% телевизоров с первого завода, 30% — со второго и 50% — с третьего?

4.7. Известно, что 34% людей имеют первую группу крови, 37% — вторую, 21% — третью и 8% — четвертую. Больному с первой группой можно переливать только кровь первой группы, со второй — кровь первой и второй групп, с третьей — кровь первой и третьей групп, и человеку с четвертой группой можно переливать кровь любой группы. Какова вероятность того, что произвольно взятому больному можно перелить кровь произвольно выбранного донора?

4.8. По каналу связи может быть передана одна из трех последовательностей букв: $AAAA$, $BBBB$, $CCCC$, причем делается это с вероятностями 0,3, 0,4 и 0,3 соответственно. Известно, что действие шумов на приемное устройство уменьшает вероятность правильного приема каждой из переданных букв до 0,6, а вероятность приема каждой переданной буквы за две другие

равны 0,2 и 0,2. Предполагается, что буквы искажаются независимо друг от друга. Найти вероятность того, что была передана последовательность $AAAA$, если на приемном устройстве получено $ABCA$.

Задачи для решения в классе

4.9. Пусть в условиях задачи 2.11 вытянули шары разного цвета. Какова вероятность того, что это синий и зеленый шары?

4.10. Предположим, что 5% всех мужчин и 0,25% всех женщин — дальтоники. Наугад выбранное лицо страдает дальтонизмом. Какова вероятность того, что это мужчина? Считать, что мужчин и женщин одинаковое число.

4.11. Из урны, содержащей 3 белых и 2 черных шара, переложены 2 вытянутых наудачу шара в урну, содержащую 4 белых и 4 черных шара. Затем из второй урны вынут шар. Найти вероятность того, что он белый.

4.12. Некоторое насекомое с вероятностью $\frac{\lambda^k}{k!}e^{-\lambda}$ откладывает k яиц, где $k = 0,\ 1,\ 2, \ldots$, а число λ положительно. Вероятность развития потомка из яйца равна p. Какова вероятность того, что у насекомого будет ровно m потомков?

4.13. В условиях предыдущей задачи у насекомого развилось 10 потомков. Какова вероятность того, что при этом было отложено 20 яиц?

Глава 2. Случайные величины и их распределения

§5. Функции и плотности распределения

Определение. *Функцией распределения* случайной величины X называется

$$F_X(y) = \mathbf{P}(\omega :\ X(\omega) < y) = \mathbf{P}(X < y), \quad -\infty < y < \infty.$$

Случайная величина X называется *дискретной*, если существует конечная или счетная последовательность чисел y_1, y_2, y_3, ... такая, что

$$\sum_{k=1}^{\infty} \mathbf{P}(X = y_k) = 1.$$

Функция распределения дискретной случайной величины называется дискретной.

Примеры дискретных распределений

Случайная величина X имеет *вырожденное распределение*, если $\mathbf{P}(X = a) = 1$.

Случайная величина X имеет *распределение Бернулли*, если

$$\mathbf{P}(X = 1) = p, \quad \mathbf{P}(X = 0) = 1 - p, \quad 0 < p < 1.$$

Случайная величина X имеет *биномиальное распределение*, если

$$\mathbf{P}(X = k) = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n, \quad n \geq 1, \quad 0 < p < 1.$$

При $n = 1$ биномиальное распределение совпадает с распределением Бернулли.

Случайная величина X имеет *распределение Пуассона*, если

$$\mathbf{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0, 1, 2, \ldots;\ \lambda > 0.$$

Случайная величина X имеет *геометрическое распределение*, если

$$\mathbf{P}(X = k) = (1 - p)\, p^{k-1},\ k = 1,\, 2,\, 3,\, \ldots,\quad 0 < p < 1.$$

Данное распределение может встретиться и в другом варианте:

$$\mathbf{P}(X = k) = (1 - p)\, p^{k},\quad k = 0, 1,\, 2,\, 3,\, \ldots.$$

Функция распределения $F_X(y)$ называется *абсолютно непрерывной*, если для любого значения y

$$F_X(y) = \int\limits_{-\infty}^{y} f(t)\, dt;$$

стоящая под знаком интеграла функция $f(t)$ называется *плотностью* распределения.

Примеры абсолютно непрерывных распределений

Здесь мы используем заглавные буквы для обозначения функций распределения, а соответствующие малые буквы — для обозначения плотностей.

1. *Равномерное распределение* на отрезке $[a;\, b]$. Его плотность равна

$$u_{a,b}(t) = \begin{cases} \dfrac{1}{b-a}, & t \in [a;\, b], \\ \\ 0, & \text{иначе.} \end{cases}$$

Для функции распределения имеем формулу

$$U_{a,b}(y) = \begin{cases} 0, & y \le a, \\ \dfrac{y-a}{b-a}, & y \in [a;\, b], \\ 1, & y > b. \end{cases}$$

2. *Нормальное (гауссовское) распределение* Φ_{α,σ^2}. Плотность задается формулой

$$\varphi_{\alpha,\sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\alpha)^2/2\sigma^2}, \ -\infty < t < \infty, \ -\infty < \alpha < \infty, \ \sigma^2 > 0.$$

Функция распределения задается формулой (к сожалению, интеграл не берется в элементарных функциях)

$$\Phi_{\alpha,\sigma^2}(y) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-\frac{(t-\alpha)^2}{2\sigma^2}} \, dt.$$

Если $\alpha = 0$, $\sigma^2 = 1$, то мы получаем *стандартное нормальное распределение* $\Phi_{0,1}$ с плотностью

$$\varphi_{0,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

и с функцией распределения

$$\Phi_{0,1}(y) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-t^2/2} \, dt.$$

График этой функции имеет центр симметрии — точку с координатами $(0,\ 1/2)$, $\Phi_{0,1}(y) = 1 - \Phi_{0,1}(-y)$. Функция $\Phi_{0,1}(y)$ очень быстро стремится к нулю при $y \to -\infty$ (и соответственно так же быстро к единице при $y \to \infty$):

$$\Phi_{0,1}(-3) = 0,00135; \quad \Phi_{0,1}(-1,96) = 0,025; \quad \Phi_{0,1}(-1,64) = 0,05.$$

3. *Показательное (экспоненциальное) распределение* E_α. Плотность показательного распределения задается формулой

$$e_\alpha(t) = \begin{cases} \alpha\, e^{-\alpha t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

Здесь $\alpha > 0$ — параметр распределения.

Функция распределения легко получается интегрированием:

$$E_\alpha(y) = \begin{cases} 0, & y \le 0, \\ 1 - e^{-\alpha y}, & y > 0. \end{cases}$$

4. *Гамма-распределение* $\Gamma_{\alpha,\lambda}$. Плотность гамма-распределения равна

$$\gamma_{\alpha,\lambda}(t) = \begin{cases} \dfrac{\alpha^\lambda}{\Gamma(\lambda)}\, t^{\lambda-1}\, e^{-\alpha t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

Здесь участвуют два параметра $\alpha > 0$, $\lambda > 0$. Напомним, что

$$\Gamma(\lambda) = \int\limits_0^\infty t^{\lambda-1}\, e^{-t}\, dt \ \ —$$

это известная гамма-функция Эйлера; она обладает свойством $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$. Для целых значений $\lambda = n$ имеет место по этой причине $\Gamma(n+1) = n!$.

5. *Стандартное распределение Коши* C. Плотность задается формулой

$$c(t) = \frac{1}{\pi}\, \frac{1}{1+t^2}, \quad -\infty < t < \infty.$$

Интегрируя плотность, находим функцию распределения:

$$C(y) = \frac{1}{\pi} \int\limits_{-\infty}^{y} \frac{1}{1+t^2}\, dt = \frac{1}{2} + \frac{1}{\pi}\, \operatorname{arctg} y.$$

Определение. Функция распределения F относится к *смешанному типу*, если при всех значениях y

$$F(y) = \alpha\, F_1(y) + \beta\, F_2(y),$$

где $F_1(y)$ — абсолютно непрерывная, а $F_2(y)$ — дискретная функции распределения, $\alpha \geq 0,\ \beta \geq 0,\ \alpha + \beta = 1$.

Определение. Функцией распределения случайного вектора X (многомерной функцией распределения, совместной функцией распределения) называется

$$F_{X_1,X_2,\ldots,X_n}(y_1,\, y_2,\, \ldots,\, y_n) = \mathbf{P}(X_1 < y_1,\, X_2 < y_2,\, \ldots,\, X_n < y_n),$$

где перечисление событий через запятую означает одновременное их осуществление, то есть пересечение.

Определение. Случайные величины $X_1,\, X_2,\, \ldots,\, X_n$ называются *независимыми*, если для любых $B_1 \subset \mathbf{R},\ \ldots, B_n \subset \mathbf{R}$ выполняется соотношение

$$\mathbf{P}(X_1 \in B_1,\, X_2 \in B_2,\, \ldots,\, X_n \in B_n) =$$

$$= \mathbf{P}(X_1 \in B_1)\mathbf{P}(X_2 \in B_2)\ldots\mathbf{P}(X_n \in B_n).$$

Домашнее задание 5

5.1. Построить график функции распределения числа испытаний Бернулли, производимых до появления первого успеха включительно.

5.2. Выразить через функцию распределения случайной величины X вероятности следующих событий: $\mathbf{P}\{a < X < b\}$, $\mathbf{P}\{a \leq X < b\},\ \mathbf{P}\{a < X \leq b\},\ \mathbf{P}\{a \leq X \leq b\}$.

5.3. Могут ли функции

а) $f(y) = \frac{1}{2}e^{-|y|}$; б) $f(y) = e^{-y}$; в) $f(y) = \cos y$; г) $f(y) \equiv 1$

быть плотностями распределения?

5.4. Плотность распределения случайной величины X задается формулой

$$f(y) = \begin{cases} Cy^2, & y \in [0;1], \\ 0, & y \notin [0;1]. \end{cases}$$

Найти C и функцию распределения случайной величины X.

5.5. Вычислить функцию гамма-распределения $\Gamma_{\alpha,\lambda}$ в случае, когда $\lambda = n$ — целое число.

5.6. На отрезок длины l произвольным образом бросают две точки. Найти функцию распределения расстояния между ними.

5.7. Точку бросают наудачу в треугольник с вершинами, координаты которых равны $(0;\ 0)$, $(№-5;\ 15-2№)$, $(9-№;\ 11-2№)$. Здесь №— номер студента по списку группы. Найти функции распределения и плотности декартовых координат точки.

5.8. В круг радиуса R наугад бросают точку. Найти:

а) функцию распределения и плотность распределения расстояния этой точки до центра круга;

б) совместную функцию распределения полярных координат точки.

Задачи для решения в классе

5.9. Игрок выигрывает очко, если при подбрасывании монеты выпадает герб, и проигрывает очко в противном случае. Построить график функции распределения суммарного выигрыша игрока после двух бросаний монеты.

5.10. Дискретное совместное распределение случайного вектора $(X,\ Y)$ задается таблицей:

$X \setminus Y$	-1	0	1
-1	0,2	0,1	0,0
1	0,4	0,0	0,3

Найти:

а) одномерные распределения X и Y;

б) закон распределения $X + Y$;

в) закон распределения $Z = Y^2$.

5.11. Какова вероятность того, что значение случайной величины окажется целым, если известно, что она имеет нормальное распределение?

5.12. n точек независимо друг от друга бросаются на отрезок $[0;\ a]$. Найти функции распределения и плотности распределения случайных величин:

а) Y_1 (крайняя слева точка);

б) Y_n (крайняя справа точка);

в) Y_k (k-я по счету слева точка, $k = 1, \ldots, n$).

§6. Преобразования случайных величин

Большая часть задач этого раздела связана с нахождением плотности распределения случайной величины вида $Y = g(X)$, где g — неслучайная функция, а распределение случайной величины X известно. Напомним, что для обоснования существования плотности распределения случайной величины Y и для нахождения этой плотности необходимо представить функцию распределения в виде

$$F_Y(y) = \int\limits_{-\infty}^{y} f(t)\,dt;$$

подынтегральная функция и будет искомой плотностью.

Если же из условия задачи известно, что плотность случайной величины Y существует, то проще найти ее, вычислив сначала функцию расределения, а потом взяв от нее производную:

$$f_Y(t) = \frac{dF_Y(t)}{dt}.$$

Если случайные величины X и Y независимы и имеют плотности распределения $f_X(t)$ и $f_Y(t)$ соответственно, то случайная величина $X + Y$ также будет иметь плотность, равную

$$f_{X+Y}(t) = \int\limits_{-\infty}^{\infty} f_X(u) f_Y(t-u)du = \int\limits_{-\infty}^{\infty} f_Y(v) f_X(t-v)dv.$$

Эти интегралы называются *свертками* плотностей f_X и f_Y.

Пример 6.1. Случайная величина X имеет равномерное распределение на $[0; 1]$. Найти функцию распределения и плотность случайной величины $Y = -\ln X$.

Решение. Для любого t

$$\begin{aligned} F_Y(t) &= \mathbf{P}(-\ln X < t) = \mathbf{P}(\ln X > -t) = \mathbf{P}(X > e^{-t}) = \\ &= 1 - \mathbf{P}(X \le e^{-t}) = 1 - U_{0,1}(e^{-t}) = \begin{cases} 0, & t \le 0, \\ 1 - e^{-t}, & t > 0. \end{cases} \end{aligned}$$

Таким образом, Y распределена по показательному закону с параметром 1, то есть

$$f_Y(t) = e_1(t) = \begin{cases} e^{-t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

Пример 6.2. Случайная величина X имеет стандартное нормальное распределение. Найти функцию распределения и плотность случайной величины $Y = |X|$.

Решение. Для $y \le 0$, очевидно, $F_Y(y) = \mathbf{P}(|X| < y) = 0$, поэтому для нахождения плотности $f_Y(t)$ достаточно представить $F_Y(y)$ в виде интеграла в пределах от 0 до y. Имеем

$$\mathbf{P}(|X| < y) = \frac{1}{\sqrt{2\pi}} \int\limits_{-y}^{y} e^{-\frac{t^2}{2}}\, dt = \frac{2}{\sqrt{2\pi}} \int\limits_{0}^{y} e^{-\frac{t^2}{2}}\, dt,$$

то есть

$$f_Y(t) = \begin{cases} 0, & t \le 0, \\ \frac{2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, & t > 0. \end{cases}$$

Пример 6.3. Найти плотность распределения суммы двух независимых случайных величин, имеющих равномерное распределение на [0; 1].

Решение. Первый способ. Воспользуемся формулой свертки:

$$f_{X+Y}(t) = \int\limits_{-\infty}^{\infty} u_{0,1}(v) u_{0,1}(t - v) dv.$$

Подынтегральные функции обе отличны от нуля только если одновременно $0 \le v \le 1$ и $0 \le t - v \le 1$. При $t < 0$ и для $t > 2$ эти неравенства несовместны, то есть $f_{X+Y}(t) = 0$ для таких t. Если $0 \le t \le 1$, то подынтегральное выражение отлично от нуля только при $0 \le v \le t$ и равно единице, то есть для таких t имеем

$$f_{X+Y}(t) = \int_0^t 1 \cdot dv = t.$$

При $1 \le t \le 2$ подынтегральное выражение отлично от нуля, если $t - 1 \le v \le 1$, поэтому для таких t имеем

$$f_{X+Y}(t) = \int_{t-1}^1 1 \cdot dv = 2 - t.$$

Второй способ. Пара (X, Y) имеет равномерное распределение в единичном квадрате, поэтому вычислять вероятность события $\mathbf{P}(X + Y < t)$ можно геометрическим способом как отношение площадей. Результат будет тем же.

Домашнее задание 6

6.1. Случайная величина X имеет равномерное распределение на $[0;\ \pi]$. Найти функцию распределения и плотность случайной величины $Y = \sin X$.

6.2. Случайная величина X имеет равномерное распределение на $[-\pi/2;\ \pi/2]$. Найти функцию распределения и плотность случайной величины $Y = \operatorname{tg} X$.

6.3. Плотность распределения случайной величины X задается формулой

$$f(t) = \begin{cases} \theta t^{\theta-1}, & t \in [0;1], \\ 0, & t \notin [0;1]. \end{cases}$$

Найти плотность распределения для $Y = -\ln X$.

6.4. Случайные величины X и Y независимы и распределены равномерно на $[0; 1]$. Найти плотность распределения случайной величины $X - Y$.

6.5. Случайная величина X имеет стандартное нормальное распределение. Найти функции распределения и плотности случайных величин: а) $Y_1 = X^2$; б) $Y_2 = \sin X$.

6.6. В условиях предыдущей задачи найти функцию распределения случайной величины $\max(0, X)$. Найти дискретную и абсолютно непрерывную компоненты полученной функции распределения.

6.7. Случайные величины X и Y независимы и имеют одно и то же дискретное распределение $\mathbf{P}\{X = y_k\} = \mathbf{P}\{Y = y_k\} = p_k$, $k \geq 1$. Найти $\mathbf{P}\{X = Y\}$.

6.8. X и Y независимы, причем $\mathbf{P}\{X = 0\} = \mathbf{P}\{X = 1\} = 1/2$, а $\mathbf{P}\{Y < t\} = t$, $0 < t < 1$. Найти функции распределения случайных величин $X + Y$ и XY.

Задачи для решения в классе

6.9. Случайная величина X имеет равномерное распределение на $[0; 1]$. Найти функции распределения и плотности случайных величин: а) $Y_1 = 2X + 1$; б) $Y_2 = X^{-1}$.

6.10. Случайная величина X имеет показательное распределение с параметром α. Найти распределения случайных величин:

а) $Y_1 = [X]$ (целая часть X);

б) $Y_2 = X - [X]$;

в) $Y_3 = X^2$;

г) $Y_4 = \alpha^{-1} \ln X$;

д) $Y_5 = \sqrt{X}$.

6.11. Случайная величина X имеет стандартное распределение Коши. Найти функцию и плотность распределения случайной величины $Y = \operatorname{tg} X$.

6.12. Пусть X и Y — независимые случайные величины, имеющие показательные распределения. Найти $\mathbf{P}\{X = 2Y\}$.

6.13. Случайные величины X и Y независимы и распределены равномерно на $[0;\ 1]$. Найти плотность распределения случайной величины $\max(X,\ 2Y)$.

§7. Математическое ожидание и дисперсия

Пусть случайная величина X дискретна, т. е. для некоторого набора чисел y_1, y_2, ...

$$\sum_{k=1}^{\infty} \mathbf{P}(X = y_k) = 1.$$

Определение. Математическим ожиданием дискретной случайной величины называется

$$\mathbf{E}\, X = \sum_{k=1}^{\infty} y_k \, \mathbf{P}(X = y_k),$$

если этот ряд абсолютно сходится, т. е. если

$$\sum_{k=1}^{\infty} |y_k| \, \mathbf{P}(X = y_k) < \infty.$$

В противном случае мы говорим, что математическое ожидание случайной величины X не существует.

Пример 7.1. Пусть X имеет распределение Бернулли. Тогда

$$\mathbf{E}X = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Пример 7.2. Если X имеет биномиальное распределение, то

$$\begin{aligned} \mathbf{E}X &= \sum_{k=o}^{n} k C_n^k \, p^k (1-p)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} \, p^k (1-p)^{n-k} \\ &= np \sum_{m=0}^{n-1} C_{n-1}^m \, p^m (1-p)^{n-1-m} = np. \end{aligned}$$

Определение. Математическим ожиданием случайной величины X, имеющей абсолютно непрерывное распределение с плотностью $f_X(t)$, называется

$$\mathbf{E}X = \int_{-\infty}^{\infty} t f_X(t) dt,$$

если только

$$\int_{-\infty}^{\infty} |t| f_X(t) dt < \infty.$$

В противном случае считаем, что $\mathbf{E}X$ не существует.

Пример 7.3. Пусть X имеет нормальное распределение с параметрами α, σ^2. Тогда

$$\begin{aligned}
\mathbf{E}X &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} t \exp\left\{-\frac{(t-\alpha)^2}{2\sigma^2}\right\} dt = \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (t-\alpha) \exp\left\{-\frac{(t-\alpha)^2}{2\sigma^2}\right\} dt + \\
&+ \frac{\alpha}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{(t-\alpha)^2}{2\sigma^2}\right\} dt = \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y \exp\left\{-\frac{y^2}{2\sigma^2}\right\} dy + \alpha \int_{-\infty}^{\infty} \varphi_{\alpha,\sigma^2}(t) dt = \alpha.
\end{aligned}$$

Здесь интеграл от плотности нормального распределения равен единице, а предпоследний интеграл равен нулю, так как в нем интегрируется нечетная функция.

Пусть случайная величина X имеет функцию распределения смешанного типа

$$F_X(y) = \alpha F_1(y) + \beta F_2(y),$$

где $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$, $F_1(y)$ — абсолютно непрерывная функция распределения, имеющая плотность $f(t)$, а $F_2(y)$ — дискретная функция распределения, имеющая скачки величиной p_1, p_2, ... в точках y_1, y_2, ...

Тогда, по определению,

$$\mathbf{E}X = \alpha \int_{-\infty}^{\infty} t f(t) dt + \beta \sum_{k=1}^{\infty} y_k p_k,$$

если только абсолютно сходятся участвующие здесь интеграл и сумма ряда.

Математическое ожидание функции от нескольких случайных величин вычисляется следующим образом:

$$\mathbf{E}g(X_1,\ldots,X_n) = \int\limits_{-\infty}^{\infty}\ldots\int\limits_{-\infty}^{\infty} g(t_1,\ldots,t_n)f(t_1,\ldots,t_n)dt_1\ldots dt_n,$$

где $f(t_1,\ldots,t_n)$ — плотность совместного распределения случайного вектора $(X_1,\ldots,X_n)$.

Определение. *Дисперсией* случайной величины X называется

$$\mathbf{D}X = \mathbf{E}(X-\mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

Для дискретных распределений дисперсия вычисляется по формулам

$$\mathbf{D}X = \sum_{k=1}^{\infty}(y_k-\mathbf{E}X)^2\mathbf{P}(X=y_k) = \sum_{k=1}^{\infty} y_k^2\mathbf{P}(X=y_k) - (\mathbf{E}X)^2,$$

для распределений абсолютно непрерывного типа имеем

$$\mathbf{D}X = \int_{-\infty}^{\infty}(t-\mathbf{E}X)^2 f_X(t)dt = \int_{-\infty}^{\infty} t^2 f_X(t)dt - (\mathbf{E}X)^2.$$

Пример 7.4. Пусть X имеет распределение Бернулли. Тогда

$$\mathbf{E}X^2 = 1\cdot p + 0\cdot(1-p) = p, \quad \mathbf{E}X = p, \quad \mathbf{D}X = p - p^2 = p(1-p).$$

Пример 7.5. Если X имеет биномиальное распределение, то можно считать, что X есть число успехов в n испытаниях Бернулли (распределение то же самое). В этом случае X можно представить в виде суммы $X = X_1 + \ldots + X_n$, где все X_i независимы и распределены по закону Бернулли.

$$\mathbf{D}X = \mathbf{D}X_1 + \ldots + \mathbf{D}X_n = np(1-p).$$

Пример 7.6. Пусть случайная величина X имеет нормальное распределение с параметрами α, σ^2. Нахождение $\mathbf{D}X$ упростится, если мы сведем все к стандартному нормальному закону. Обозначим $Y = (X - \alpha)/\sigma$, тогда Y имеет стандартное нормальное распределение и $X = \sigma Y + \alpha$. В силу свойств дисперсии $\mathbf{D}X = \sigma^2 \mathbf{D}Y$, поэтому достаточно найти $\mathbf{D}Y$. Интегрируя по частям, получаем

$$\begin{aligned} \mathbf{E}Y^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \, e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \, d(-e^{-t^2/2}) = \\ &= \frac{1}{\sqrt{2\pi}} \left\{ -te^{-t^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-t^2/2} dt \right\} = \int_{-\infty}^{\infty} \varphi_{0,1}(t) dt = 1. \end{aligned}$$

Поскольку $\mathbf{E}Y = 0$, то $\mathbf{D}Y = 1$ и $\mathbf{D}X = \sigma^2$.

Домашнее задание 7

7.1. Найти математические ожидания и дисперсии случайных величин, введенных в задачах а) 5.1; б) 5.6; в) 5.8 а).

7.2. Найти математические ожидания и дисперсии декартовых координат точки в задаче 5.7.

7.3. Найти математические ожидания и дисперсии случайных величин с плотностями, введенными в задачах 5.3 а) и 5.4.

7.4. Найти математическое ожидание и дисперсию случайной величины Y в задаче 6.1.

7.5. Найти математическое ожидание и дисперсию случайной величины $X - Y$ в задаче 6.4.

7.6. Доказать, что $\mathbf{E}X = \sum\limits_{k=1}^{\infty} \mathbf{P}\,(X \geq k)$, если известно, что $\sum\limits_{k=1}^{\infty} \mathbf{P}\,(X = k) = 1$.

Задачи для решения в классе

7.7. Случайные величины X и Y независимы, X имеет стандартное нормальное распределение, Y имеет распределение Бернулли с параметром $1/3$. Найти математические ожидания и дисперсии случайных величин: а) $2X + 3Y$; б) $X - 9Y - 1$.

7.8. Вычислить математическое ожидание и дисперсию случайной величины, имеющей:
а) распределение Пуассона;
б) геометрическое распределение;
в) равномерное распределение на отрезке $[a; b]$;
г) показательное распределение с параметром α;
д) гамма-распределение.

7.9. Найти математические ожидания и дисперсии случайных величин Y_1 и Y_n, введенных в задаче 5.12.

7.10. Пусть случайная величина X принимает только целочисленные значения: $\sum_k \mathbf{P}(X = k) = 1$. Функция $\varphi(z) = \sum_k z^k \mathbf{P}(X = k)$ называется производящей. Выразить $\mathbf{E}X$ и $\mathbf{D}X$ через производные производящей функции.

§8. Моменты, ковариация, коэффициент корреляции

Определение. *Моментом* k-го порядка случайной величины X называется $\mathbf{E}X^k$, $k > 0$.

Определение. Коэффициентом корреляции называется

$$\rho(X,Y) = \frac{\mathbf{E}((X-\mathbf{E}X)(Y-\mathbf{E}Y))}{\sqrt{\mathbf{D}X\,\mathbf{D}Y}} = \frac{\mathbf{E}(XY)-\mathbf{E}X\mathbf{E}Y}{\sqrt{\mathbf{D}X\,\mathbf{D}Y}}.$$

Для вычисления коэффициента корреляции необходимо знать совместное распределение пары (X,Y). Если, к примеру, известна плотность $f_{X,Y}(u,v)$, то смешанный момент вычисляется по формуле

$$\mathbf{E}(XY) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} uv\, f_{X,Y}(u,v)\,du\,dv.$$

Одномерные плотности получаются из двумерной интегрированием:

$$f_X(u) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv, \quad f_Y(v) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,du.$$

Домашнее задание 8

8.1. Найти $\mathbf{E}X^{2009}$, если X имеет стандартное нормальное распределение.

8.2. Вычислить момент k-го порядка для случайной величины, имеющей:
а) равномерное распределение;
б) гамма-распределение.

8.3. Случайная величина X принимает натуральные значения с вероятностями $\mathbf{P}\,(X = k) = Ck^{-10}, \quad k = 1, 2, \ldots$. Как найти C? Какого порядка моменты существуют у этой случайной величины X?

8.4. Найти коэффициент корреляции $\rho\ (X, X + Y)$, где X и Y независимы, одинаково распределены и имеют конечную ненулевую дисперсию.

8.5. Найти коэффициент корреляции между координатами точки из задачи 5.7.

8.6. Точка произвольным образом бросается в круг единичного радиуса. Найти коэффициент корреляции между ее декартовыми координатами.

Задачи для решения в классе

8.7. Случайная величина X имеет показательное распределение с параметром α. Найти, для каких значений параметра β существует математическое ожидание случайной величины $Y = e^{\beta X}$.

8.8. Решая задачу, студент обнаружил, что случайная величина X имеет плотность распределения $f(t) = 3t^{-2}$ при $t \geq 1$. При дальнейшем вычислении моментов по формулам

$$\mathbf{E}X^{-1} = \int_1^\infty 3t^{-3}dt = 3/2, \quad \mathbf{E}X^{-2} = \int_1^\infty 3t^{-4}dt = 1$$

получилось, что $\mathbf{D}X^{-1} = 1 - (3/2)^2 < 0$. Найти ошибку в рассуждениях, поскольку дисперсия отрицательной быть не может.

8.9. Вычислить коэффициент корреляции $\rho\,(X, Y)$ в условиях задачи 5.10.

8.10. Найти коэффициент корреляции $\rho(X, X^2)$, где X имеет:
а) стандартное нормальное распределение;
б) показательное распределение.

§9. Предельные теоремы

Определение. Последовательность случайных величин $\{Y_n\}$ называется сходящейся с вероятностью единица к случайной величине Y, если

$$\mathbf{P}(\omega:\ Y_n(\omega) \to Y(\omega)) = \mathbf{P}(Y_n \to Y) = 1.$$

Обозначение: $Y_n \xrightarrow{1} Y$.

Теорема (усиленный закон больших чисел, УЗБЧ). *Пусть случайные величины $X_1, X_2, \ldots$ независимы и одинаково распределены, причем $\mathbf{E}|X_1| < \infty$. Обозначим $a = \mathbf{E}X_1$, $S_n = \sum_{i=1}^n X_i$. Тогда при $n \to \infty$*

$$\frac{S_n}{n} \xrightarrow{1} a.$$

Центральная предельная теорема (ЦПТ). *Пусть $X_1, X_2 \ldots$ — независимые одинаково распределенные случайные величины. Предположим, что $\mathbf{E}X_1^2 < \infty$. Обозначим $S_n = X_1 + \ldots + X_n$, $a = \mathbf{E}X_1$, $\sigma^2 = \mathbf{D}X_1$, и пусть $\sigma^2 > 0$. Тогда для любого y*

$$\mathbf{P}\left(\frac{S_n - na}{\sigma\sqrt{n}} < y\right) = F_{\frac{S_n - na}{\sigma\sqrt{n}}}(y) \to \Phi_{0,1}(y) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-t^2/2}\, dt$$

при $n \to \infty$.

В частности, эта теорема применима, когда S_n — число успехов в схеме Бернулли. Однако если вероятность успеха в одном испытании Бернулли очень мала, то лучшее приближение обеспечивает следующее утверждение.

Теорема (приближение Пуассона в схеме Бернулли). *Пусть S_n — число успехов в n испытаниях схемы Бернулли, p — вероятность успеха в одном испытании. Тогда для любого подмножества $B \subset \{0, 1, 2, \ldots, n\}$*

$$\left| \sum_{k \in B} \mathbf{P}(S_n = k) - \sum_{k \in B} \frac{(np)^k}{k\,!}\, e^{-np} \right| \leq \min(p,\, np^2).$$

Если приведенная оценка дает удовлетворительную погрешность приближения, то следует пользоваться приближением Пуассона в схеме Бернулли. В противном случае нужно использовать ЦПТ.

Пример 9.1. К чему сходится с вероятностью единица при $n \to \infty$ последовательность

$$Y_n = \cos \frac{X_1 + \ldots + X_n}{n},$$

если $X_1, \ldots, X_n$ — независимые случайные величины, распределенные равномерно на $[0; \boldsymbol{\pi}]$?

Решение. В силу закона больших чисел

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{1} \mathbf{E}X_1 = \frac{\boldsymbol{\pi}}{2}.$$

Функция $\cos t$ непрерывна, поэтому

$$Y_n = \cos \frac{X_1 + \ldots + X_n}{n} \xrightarrow{1} \cos(\boldsymbol{\pi}/2) = 0.$$

Пример 9.2. 1000 раз бросается игральная кость. Найти пределы, в которых с вероятностью 0,95 будет лежать сумма выпавших очков.

Решение. Обозначим через S_n сумму выпавших очков. S_n есть сумма независимых случайных величин, каждая из которых принимает значения от 1 до 6 с равными вероятностями. Нетрудно вычислить: $a = \mathbf{E}X_1 = 3,5$; $\mathbf{E}X_1^2 = 91/6$;

$\sigma^2 = \mathbf{D}X_1 = 35/12.$ В силу ЦПТ случайная величина $(S_n - 3500)\,/\sqrt{1000 \cdot 35/12}$ имеет почти стандартное нормальное распределение (число n велико!), поэтому

$$\mathbf{P}\quad -1,96 < \frac{S_n - 3500}{\sqrt{1000 \cdot 35/12}} < 1,96\Bigg) \simeq \frac{1}{\sqrt{2\pi}} \int\limits_{-1,96}^{1,96} e^{-t^2/2}\,dt = 0,95.$$

Последнее мы заранее находим из таблиц. Таким образом,

$$\mathbf{P}(|S_n - 3500)| < 1,96\sqrt{1000 \cdot 35/12}) \simeq 0,95,$$

$$1,96\sqrt{1000 \cdot 35/12} = 105,85\ldots.$$

Пример 9.3. Имеется производство спичек. Каждая спичка независимо от других с вероятностью 0,015 является бракованной и при употреблении не возгорается. В соответствии с требованиями стандарта спички должны расфасовываться в коробки по 100 штук в каждую. Ясно, что при этом в каждой коробке с большой вероятностью годных спичек окажется меньше 100. Чтобы избежать претензий со стороны потребителей, руководство решает класть в каждую коробку добавочно некоторое число спичек x так, чтобы с вероятностью не менее 0,95 годных спичек там оказалось не менее 100.

Какое наименьшее число спичек x нужно для этого положить в коробку?

Решение. Мы имеем здесь схему Бернулли с числом испытаний $n = 100 + x$ и вероятностью успеха $0,015$. Обозначим число бракованных спичек S_n. Тогда годных спичек будет в коробке не менее 100, если $S_n \leq x$. Из приведенной выше теоремы заключаем, что приближение Пуассона дает в нашем случае вполне удовлетворительную точность. Считая для простоты, что $np = (100 + x)0,015 \simeq 1,5$, получаем соотношение

$$\mathbf{P}(S_n \leq x) = \sum_{k=0}^{x} \mathbf{P}(S_n = k) \simeq e^{-1,5}\left(1 + 1,5 + \frac{1,5^2}{2} + \ldots + \frac{1,5^x}{x\,!}\right).$$

Требуется, чтобы эта вероятность была не менее 0,95. Нетрудно вычислить, что для этого достаточно взять $x = 4$ в правой части.

Домашнее задание 9

9.1. Доказать, что Y_1 сходится к нулю с вероятностью единица при $n \to \infty$ для последовательности случайных величин $Y_1 = Y_1(n) = \min(X_1, \ \ldots, \ X_n)$, введенных в задаче 5.12 (указание: см. задачу 9.8).

9.2. Вероятность выхода из строя за время T одного конденсатора равна $0,05 \cdot №$ (здесь № — номер студента по списку группы). Определить вероятность того, что за время T из 100 конденсаторов выйдут из строя: а) не менее 5№ конденсаторов; б) менее 5№+8 конденсаторов.

9.3. Студент получает на экзамене 5 с вероятностью 0,2, 4 с вероятностью 0,4, 3 с вероятностью 0,3 и 2 с вероятностью 0,1. За время обучения он сдает 100 экзаменов. Найти пределы, в которых с вероятностью 0,95 лежит средний балл.

9.4. Урожайность куста картофеля задается следующим распределением:

Урожай в кг	0	1	1,5	2	2,5
Вероятность	0,1	0,2	0,2	0,3	0,2

На участке высажено 900 кустов. В каких пределах с вероятностью 0,95 будет находиться урожай? Какое наименьшее число кустов нужно посадить, чтобы с вероятностью не менее 0,975 урожай был не менее тонны?

9.5. Игральная кость подбрасывается до тех пор, пока общая сумма очков не превысит 700. Оценить вероятность того, что для этого потребуется более 210 бросаний.

9.6. Пусть $X_1, X_2, \ldots$ — независимые одинаково распределенные случайные величины, $\mathbf{E}X_1 = 0, \quad \mathbf{D}X_1 < \infty$. Известно, что

$$\mathbf{P}\left(\frac{X_1+\ldots+X_n}{\sqrt{n}} \geq 1\right) \to \frac{1}{3}$$

при $n \to \infty$. Найти $\mathbf{D}X_1$.

Задачи для решения в классе

9.7. Случайные величины $X_1, X_2, \ldots$ независимы и одинаково распределены по закону Пуассона с параметром λ. К чему сходится с вероятностью единица последовательность

$$\frac{X_1^2+\ldots+X_n^2}{n} - \left(\frac{X_1+\ldots+X_n}{n}\right)^2 \quad ?$$

9.8. Доказать, что $Y_n \to a$ с вероятностью единица при $n \to \infty$ для последовательности случайных величин $Y_n = \max(X_1, \ldots, X_n)$, введенных в задаче 5.12 (указание: использовать тот факт, что для сходимости монотонной последовательности к константе a с вероятностью единица достаточно сходимости функций распределения во всех точках, отличных от точки a).

9.9. Известно, что вероятность рождения мальчика приблизительно равна 0,515. Какова вероятность того, что среди 10 тыс. новорожденных окажется мальчиков не больше, чем девочек?

9.10. Для лица, дожившего до двадцатилетнего возраста, вероятность смерти на 21-м году жизни равна 0,006. Застрахована группа 10000 лиц 20-летнего возраста, причем каждый застрахованный внес 1200 рублей страховых взносов за год. В случае смерти застрахованного родственникам выплачивается 100000 рублей. Какова вероятность того, что:

а) к концу года страховое учреждение окажется в убытке;

б) его доход превысит 6000000 рублей?

Какой минимальный страховой взнос следует учредить, чтобы в тех же условиях с вероятностью 0,95 доход был не менее 4000000 рублей?

9.11. Известно, что вероятность выпуска сверла повышенной хрупкости (брак) равна 0,02. Сверла укладываются в коробки по 100 шт. Чему равна вероятность того, что в коробке не окажется бракованных сверл? Какое наименьшее количество сверл нужно класть в коробку для того, чтобы с вероятностью, не меньшей 0,9, в ней было не менее 100 исправных?

9.12. Вероятность угадывания 6 номеров в спортлото (6 из 49) равна $7,2 \cdot 10^{-8}$. При подсчете оказались заполненными 5 млн. карточек. Какова вероятность, что никто не угадал все 6 номеров? Какое наименьшее количество карточек нужно заполнить, чтобы с вероятностью не менее 0,9 хотя бы один угадал 6 номеров?

Глава 3. Математическая статистика

§10. Выборка. Оценивание параметров

Выборка и вариационный ряд

Основным объектом исследования в математической статистике является **выборка** $\vec{X} = (X_1, X_2, ..., X_n)$, то есть набор значений случайной величины X, полученных в результате n независимых воспроизведений эксперимента. Иначе говоря, выборка представляет собой случайный вектор, координаты которого — **элементы выборки** $X_1, X_2, ..., X_n$ — независимые случайные величины, имеющие общее распределение с функцией распределения $F(t)$. Будем говорить в этом случае, что имеется **случайная выборка** $\vec{X}$ из распределения F, и обозначать сокращенно: $\vec{X} \Subset F$. Число n называется **объемом выборки**. Конкретный набор числовых значений случайных величин $X_1, X_2, ..., X_n$, полученный в результате эксперимента, будем называть **реализацией** выборки и обозначать $\vec{x} = (x_1, x_2, ..., x_n)$.

Если элементы выборки $X_1, \ldots, X_n$ упорядочить по возрастанию, то получится новый набор случайных величин, называемый **вариационным рядом**:

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

Случайная величина $X_{(k)}$, $k = 1, \ldots, n$ называется **k-м членом вариационного ряда**, или **k-й порядковой статистикой**. В частности, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Эмпирическая функция распределения, гистограмма

Эмпирической функцией распределения $F_n^*(t)$ называется частота элементов выборки, меньших заданного t. Эмпирическая функция распределения, соответствующая выборке

$\vec{X} = (X_1, X_2, ..., X_n)$, может быть построена по этой выборке с помощью любой из следующих формул:

$$F_n^*(t) = \frac{\{\text{количество } X_i : X_i < t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i < t),$$

где функция

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

— индикатор события $\{X_i < t\}$.

Заметим, что эмпирическая функция распределения, соответствующая случайной выборке $\vec{X}$, сама является случайной, поскольку определяется через элементы выборки $X_1,\ X_2, \ldots, X_n$, являющиеся случайными величинами. В то же время любая реализация $\vec{x} = (x_1, x_2, \ldots, x_n)$ выборки $\vec{X}$ порождает соответствующую реализацию эмпирической функции распределения (по той же формуле), которая является обычной (а не случайной) функцией распределения.

Эмпирическая функция распределения $F_n^*(t)$ является выборочным аналогом неизвестной теоретической функции распределения $F(t)$, ее называют также **оценкой** для $F(t)$. Выборочным аналогом для теоретической плотности распределения $f(t)$ является **гистограмма**, или **эмпирическая плотность распределения**, которая строится по выборке $\vec{X} = (X_1, X_2, ..., X_n)$ следующим образом.

Пусть $h > 0$ — произвольное число. Разобьем область значений изучаемой случайной величины (например, всю числовую ось) на промежутки $\Delta_k = [z_{k-1}, z_k)$ длины h и построим ступенчатую функцию $f_n^*(t)$, которая на каждом промежутке Δ_k принимает постоянное значение, вычисляемое по любой из формул:

$$f_n^*(t) = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{I}(X_i \in \Delta_k) = \frac{\boldsymbol{\nu}_k}{nh},\ t \in \Delta_k, \qquad (3)$$

где ν_k — число элементов выборки, попавших в промежуток Δ_k.

Иногда шаг гистограммы h выбирают следующим образом. Сначала расчитывают число интервалов K по *формуле Стеджеса*

$$K = [\log_2 n] + 1. \tag{4}$$

Здесь n — объем выборки, $[a]$ — целая часть числа a. Потом длина интервала расчитывается по формуле

$$h = \frac{X_{(n)} - X_{(1)}}{K}.$$

При построении гистограммы последний промежуток выбирается замкнутым: $\Delta_K = [z_{K-1};\ z_K]$. Величину

$$X_{(n)} - X_{(1)} = \max\{X_i\} - \min\{X_i\}$$

называют размахом выборки.

Пример 10.1. По данной реализации выборки $\vec{x} = (3;\ 8;\ 6;\ 4;\ 6;\ 1;\ 5;\ 4;\ 9;\ 4)$ построить реализацию вариационного ряда, графики реализаций эмпирической функции распределения и гистограммы. Число интервалов для построения гистограммы выбрать по формуле Стеджеса.

Решение. Реализацию вариационного ряда образуем из элементов данной реализации выборки, расположив их в порядке возрастания:

$$1;\ 3;\ 4;\ 4;\ 4;\ 5;\ 6;\ 6;\ 8;\ 9. \tag{5}$$

Объем выборки $n = 10$. График реализации эмпирической функции распределения строим с помощью полученной реализации вариационного ряда. График — ступенчатая функция со скачками в точках вариационного ряда, принимающая значение 0 в промежутке $(-\infty, 1]$ и имеющая скачки в точках $x_{(i)}$, равные частоте элемента $x_{(i)}$. Например, скачок в точке $x_{(1)} = 1$ равен $\frac{1}{10}$, скачок в точке $x_{(2)} = 3$ равен $\frac{1}{10}$, и т. д. График реализации эмпирической функции распределения изображен на рис. 3.

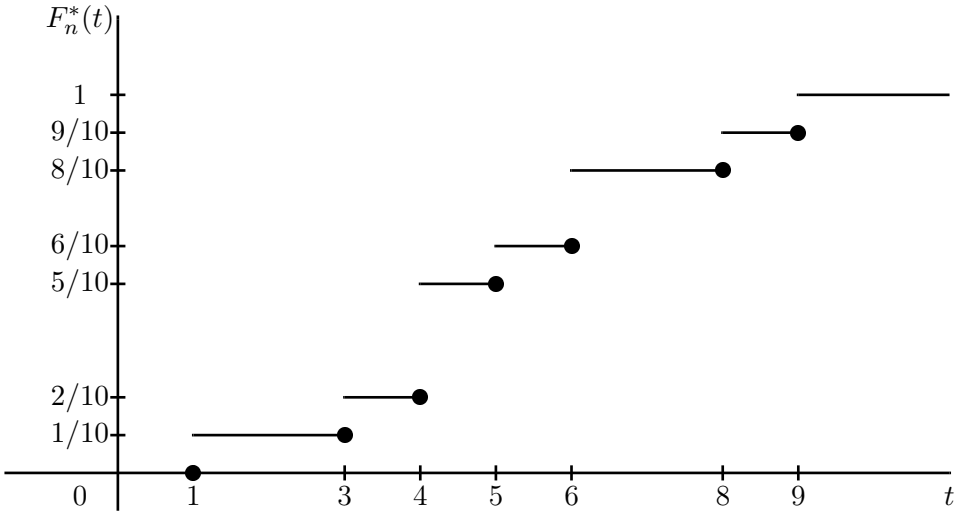

Рис. 3: Эмпирическая функция распределения $F_n^*(t)$

Расчитаем число промежутков по формуле Стеджеса: $K = [\log_2 10] + 1 = 3 + 1 = 4$. Размах выборки равен $9 - 1 = 8$, шаг гистограммы $h = 8/4 = 2$. Разобьем отрезок $[1;\ 9]$ на промежутки длины $h = 2$:

$$\Delta_1 = [1;\ 3);\ \Delta_2 = [3;\ 5);\ \Delta_3 = [5;\ 7);\ \Delta_4 = [7;\ 9].$$

Число элементов выборки, попавших в интервал Δ_1, равно $\nu_1 = 1$. Аналогично находим:

$$\nu_2 = 4;\ \ \nu_3 = 3;\ \ \nu_4 = 2.$$

Вычисляя значения функции $f_n^*(t) = \frac{\nu_k}{nh}, t \in \Delta_k$, на каждом из интервалов Δ_k, строим гистограмму (рис. 4).

Выборочные моменты

По выборке $\vec{X} = (X_1, X_2, ..., X_n)$ можно построить эмпирические (выборочные) аналоги числовых характеристик рас-

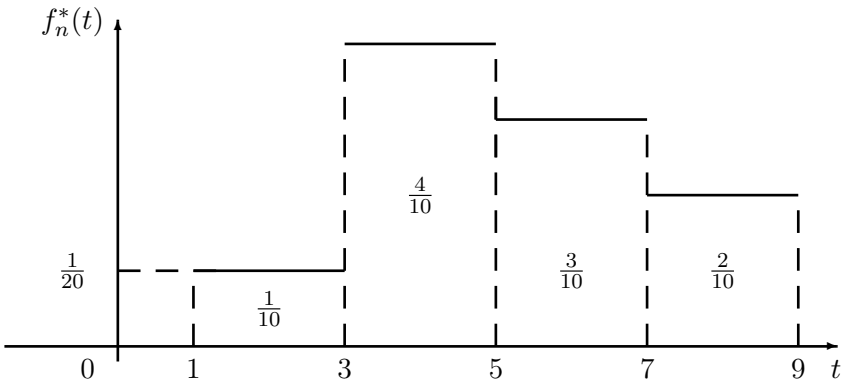

Рис. 4: Гистограмма $f_n^*(t)$

пределения. Наиболее употребительными являются выборочное математическое ожидание, или **выборочное среднее**, $\overline{X}$, и **выборочная дисперсия** S^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{6}$$

Подобно выборочным среднему и дисперсии определяются выборочные моменты порядка k

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

которые являются эмпирическими аналогами моментов $\alpha_k = \mathbf{E}X_i^k$. Отметим, что

$$\mathbf{E}\overline{X^k} = \alpha_k.$$

Приведенное соотношение означает, что математические ожидания эмпирических моментов совпадают с соответствующими теоретическими моментами. Это свойство называется **несмещенностью**. Эмпирические моменты являются **несмещенными оценками** для соответствующих теоретических.

В то же время центральные эмпирические моменты являются смещенными оценками для своих теоретических аналогов.

Отметим, что выборочная дисперсия вычисляется аналогично дисперсии.

$$S^2 = \overline{X^2} - (\overline{X})^2.$$

Для доказательства этой формулы достаточно раскрыть скобки в определении S^2. Вычислим математическое ожидание статистики S^2:

$$\mathbf{E}S^2 = \mathbf{E}\overline{X^2} - \mathbf{E}\overline{X}^2 = \mathbf{E}\overline{X^2} - (\mathbf{E}\overline{X})^2 - \mathbf{D}\overline{X} = \frac{n-1}{n}\mathbf{D}X_1.$$

Итак, эта оценка является асимптотически несмещенной.

Для того, чтобы получить несмещенную оценку дисперсии, делят S^2 на $\frac{n-1}{n}$.

Несмещенная выборочная дисперсия — это статистика

$$S_0{}^2 = \frac{n}{n-1}S^2.$$

Для нее выполнено свойство

$$\mathbf{E}S_0{}^2 = \mathbf{D}X_1.$$

Отметим, что корень из несмещенной выборочной дисперсии S_0 не является несмещенной оценкой для стандартного отклонения σ_X, так как $\mathbf{E}\sqrt{Y} \neq \sqrt{\mathbf{E}Y}$.

Статистики и оценки

Задача оценивания параметров возникает в ситуации, когда распределение F не является полностью неизвестным, а известен его математический вид $F = F(t, \theta)$, содержащий неизвестный параметр θ (или несколько, тогда θ — многомерный параметр). Задача состоит в том, чтобы по выборке $\vec{X}$ вычислить приближенное значение $\theta^*(\vec{X})$ для неизвестного параметра,

причем сделать это в том или ином смысле оптимальным образом. Это задача **точечного оценивания**. Другой подход состоит в построении по выборке $\mathbf{X}$ интервала $(\theta_-(\vec{X});\ \theta_+(\vec{X}))$, который накрывает неизвестное значение параметра θ с заданной (высокой) вероятностью. Этот подход называется **интервальным оцениванием**, а $(\theta_-(\vec{X});\ \theta_+(\vec{X}))$ называется **доверительным интервалом**.

Пусть $\vec{X} \Subset F(t,\theta)$, причем параметр θ может принимать значения из множества Θ, которое называется **параметрическим множеством**. Будем называть **статистикой** любую случайную величину вида $T(\vec{X})$, которая является функцией *только от элементов выборки*. **Оценкой параметра** θ называется статистика $\tilde{\theta} = \tilde{\theta}(\vec{X})$, которая принимает значения из параметрического множества Θ.

Оценка $\tilde{\theta}$ называется **несмещенной** оценкой параметра θ, если для любого $\theta \in \Theta$ выполнено

$$\mathbf{E}\tilde{\theta} = \theta. \tag{7}$$

Договоримся указывать в обозначении статистики объем выборки, если это необходимо подчеркнуть: $\tilde{\theta} = \tilde{\theta}_n$.

Оценка $\tilde{\theta}_n$ называется **(сильно) состоятельной оценкой параметра** θ, если для любого $\theta \in \Theta$ при $n \to \infty$ имеет место сходимость с вероятностью единица:

$$\tilde{\theta}_n \xrightarrow{1} \theta, \tag{8}$$

то есть $\mathbf{P}\{\tilde{\theta}_n \to \theta\} = 1$.

К следующему примеру мы будем часто возвращаться в дальнейшем.

Пример 10.2. Задача о расписании автобусов. Придя на остановку, пассажир пытается оценить длительность интервалов между автобусами выбранного им маршрута. Он анкетирует других пассажиров, ожидающих этот автобус, и у каждого из n пассажиров выясняет время, проведенное им на остановке, получая таким образом выборку $X_1, \ldots, X_n$. Предполагает-

ся, что $X_1, \ldots, X_n$ образуют выборку из равномерного распределения $U_{[0;\ \theta]}$, где $\theta > 0$ — неизвестный параметр — интервал времени между автобусами.

Первый пассажир предлагает для оценки параметра θ использовать выборочное среднее, т. е. получить оценку в виде $\widetilde{\theta}_1 = c_1\overline{X}$.

Второй пассажир предлагает использовать самое большое время ожидания, т. е. получить оценку в виде $\widetilde{\theta}_2 = c_2 X_{(n)}$.

Третий пассажир предлагает сложить самое большое и самое маленькое время ожидания: $\widetilde{\theta}_3 = X_{(n)} + X_{(1)}$.

Вычислить константы c_1, c_2, обеспечивающие несмещенность оценок $\widetilde{\theta}_1$, $\widetilde{\theta}_2$. Проверить несмещенность оценки $\widetilde{\theta}_3$.

Решение. Вычислим математическое ожидание оценки $\widetilde{\theta}_1$:

$$\mathbf{E}\widetilde{\theta}_1 = c_1\mathbf{E}\overline{X} = c_1\mathbf{E}X_1 = c_1\theta/2.$$

Условие несмещенности выполнено, если

$$\mathbf{E}\widetilde{\theta}_1 = c_1\theta/2 = \theta.$$

Отсюда с необходимостью $c_1 = 2$. Итак, мы получили первую несмещенную оценку $\widetilde{\theta}_1 = 2\overline{X}$. Ее сильная состоятельность следует из усиленного закона больших чисел: т. к. имеет место сходимость $\overline{X} \to \mathbf{E}X_1 = \theta/2$ с вероятностью 1, то $\widetilde{\theta}_1 = 2\overline{X} \to \theta$ с вероятностью 1 в силу непрерывности функции $y(t) = 2t$.

Исследование оценки $\widetilde{\theta}_2$ значительно более трудоемко. Найдем распределение статистики $X_{(n)}$. Ее функция распределения равна

$$F_{X_{(n)}}(y) = \mathbf{P}\{X_{(n)} < y\} = \mathbf{P}\{\max(X_1, \ldots, X_n) < y\} =$$

$$= \mathbf{P}\{\bigcap_{i=1}^{n}(X_i < y)\} = \prod_{i=1}^{n}\mathbf{P}(X_i < y) = F^n(y),$$

где

$$F(y) = \begin{cases} 0, & \text{если } y \le 0; \\ \frac{y}{\theta}, & \text{если } 0 < y \le \theta; \\ 1, & \text{если } y > \theta; \end{cases}$$

— функция распределения закона $U_{[0;\,\theta]}$. Дифференцируя $F_{X_{(n)}}(y)$, найдем плотность распределения случайной величины $X_{(n)}$:

$$f_{X_{(n)}}(y) = nF^{n-1}(y)F'(y) = nF^{n-1}(y)f(y).$$

Подставляя в последнее равенство функцию распределения закона $U_{[0;\,\theta]}$ и его плотность

$$f(y) = \begin{cases} \frac{1}{\theta}, & \text{если } y \in (0,\theta); \\ 0, & \text{если } y \notin [0,\theta]. \end{cases}$$

находим плотность распределения $X_{(n)}$:

$$f_{X_{(n)}}(y) = \begin{cases} \frac{ny^{n-1}}{\theta^n}, & \text{если } y \in (0,\theta); \\ 0, & \text{если } y \notin [0,\theta]. \end{cases}$$

Найдем математическое ожидание оценки:

$$\mathbf{E}\widetilde{\theta}_2 = c_2\mathbf{E}X_{(n)} = c_2\int\limits_0^\theta y\frac{ny^{n-1}}{\theta^n}dy = \frac{nc_2}{n+1}\frac{y^{n+1}}{\theta^n}\bigg|_0^\theta = \frac{nc_2}{n+1}\theta.$$

Отсюда следует, что оценка $\widetilde{\theta}_2 = c_2X_{(n)}$ является несмещенной для параметра θ при условии $c_2n\theta/(n+1) = \theta$, то есть при выполнении равенства $c_2 = (n+1)/n$. Мы получили вторую несмещенную оценку: $\widetilde{\theta}_2 = (n+1)X_{(n)}/n$.

Для доказательства несмещенности третьей оценки заметим, что минимум выборки $X_{(1)}$ распределен симметрично максимуму $X_{(n)}$ относительно середины отрезка $\theta/2$, т. е. для всех t выполнено

$$\mathbf{P}\{X_{(1)} < t\} = \mathbf{P}\{\theta - X_{(n)} < t\}.$$

Отсюда $\mathbf{E}X_{(1)} = \theta - \mathbf{E}X_{(n)} = \theta/(n+1)$, $\mathbf{E}\widetilde{\theta}_3 = \mathbf{E}X_{(1)} + \mathbf{E}X_{(n)} = \theta$ — оценка несмещенная.

Метод моментов (одномерный случай)

Наиболее распространенными методами нахождения оценок являются *метод моментов* и *метод максимального правдоподобия.*

Пусть $\theta \in \Theta$ — одномерный параметр, и $g : \mathbf{R} \longrightarrow \mathbf{R}$ — некоторая числовая функция. Тогда по данной выборке $\vec{X} = (X_1, X_2, ..., X_n)$ можно построить выборку

$$\vec{g}(X) = (g(X_1), g(X_2), ..., g(X_n)).$$

Обозначим

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

выборочное среднее этой выборки. С другой стороны, можно найти теоретическое среднее выборки $\vec{g}(X)$:

$$m_g(\theta) = \mathbf{E}g(X_i).$$

Оценкой метода моментов (ОММ) называется такое значение $\theta_g^* = \theta_g^*(\vec{X})$, при котором теоретическое среднее выборки $\vec{g}(X)$ совпадает с выборочным средним:

$$m_g(\theta_g^*) = \overline{g(X)},$$

то есть ОММ является решением уравнения относительно неизвестного θ_g^*.

Если при этом оказывается, что функция $m_g(\theta)$ непрерывна и строго монотонна, то для нее существует обратная m_g^{-1}, и ОММ имеет вид:

$$\theta_g^*(\vec{X}) = m_g^{-1}(\overline{g(X)}).$$

В качестве функции g чаще всего выбирают степенные функции: $g(t) = t^k$, где $k = 1, 2, \ldots$. В этом случае теоретическое среднее выборки $\vec{g}(X)$ совпадает с теоретическим моментом соответствующего порядка, например, если $g(t) = t$, то $m_g(\theta) = \mathbf{E}X_i = \alpha_1(\theta)$; если $g(t) = t^2$, то

$m_g(\theta) = \mathbf{E}X_i^2 = \alpha_2(\theta)$, и т. д. При этом уравнение для нахождения ОММ приобретает вид:

$$\alpha_k(\theta^*) = \overline{X^k}.$$

Оценка по методу моментов в этом случае называется *оценкой по k-тому моменту* и обозначается θ_k^*.

Отметим, что если функция $m_g(\theta) = \mathbf{E}g(X_1)$ непрерывна и строго монотонна, то оценка по методу моментов $\theta_g^*(\vec{X}) = m_g^{-1}(\overline{g(X)})$ сильно состоятельна.

Метод моментов (многомерный случай)

Пусть $\vec{X} \Subset \mathbf{F}_\theta$, где параметр $\theta \in \Theta$, подлежащий оцениванию, — многомерный. Рассмотрим для простоты двумерный случай, то есть $\theta = (\theta_1, \theta_2)$. Тогда для однозначного нахождения двух неизвестных θ_1, θ_2 одного уравнения недостаточно. Оценкой метода моментов в этом случае называется решение (θ_1^*, θ_2^*) системы уравнений вида:

$$\begin{cases} m_{g_1}(\theta_1, \theta_2) = \overline{g_1(X)}, \\ m_{g_2}(\theta_1, \theta_2) = \overline{g_2(X)}. \end{cases}$$

В качестве функций g_1, g_2 можно выбрать, как и раньше, степенные функции $g_i(t) = t^k$, где $k = 1, 2, \ldots$. Тогда уравнения системы получаются как результат приравнивания эмпирических моментов выборки $\vec{X}$ соответствующим теоретическим. Например, приравнивая первые два момента, получим систему:

$$\begin{cases} \alpha_1(\theta_1, \theta_2) = \overline{X}, \\ \alpha_2(\theta_1, \theta_2) = \overline{X^2}. \end{cases}$$

Как и раньше, вместо вторых моментов можно приравнивать дисперсии.

Пример 10.3. Пусть $\vec{X} \Subset \Pi_\lambda$, где $\lambda > 0$ — неизвестный параметр. Найти оценки параметра λ по а) первому и б) второму моментам.

Решение.

а) Так как для распределения Пуассона $\mathbf{E}X_i = \lambda$, то λ_1^* получается сразу: заменяя λ на λ_1^*, а $\mathbf{E}X_i$ на $\overline{X}$, получаем $\lambda_1^* = \overline{X}$.

б) В этом случае вычисляем второй момент распределения Пуассона:

$$\mathbf{E}X_i^2 = (\mathbf{E}X_i)^2 + \mathbf{D}X_i = \lambda + \lambda^2.$$

Приравнивая эту функцию второму выборочному моменту и заменяя λ на λ_2^*, получим уравнение:

$$\lambda_2^* + (\lambda_2^*)^2 = \overline{X^2},$$

из которого находим λ_2^* :

$$\lambda_2^* = -\frac{1}{2} \pm \sqrt{\frac{1}{4} + \overline{X^2}}.$$

Так как $\lambda_2^* > 0$, то из двух решений выбираем одно — неотрицательное, и ОММ имеет вид:

$$\lambda_2^*(X) = -\frac{1}{2} + \sqrt{\frac{1}{4} + \overline{X^2}}.$$

Замечание. Из двух найденных оценок λ_1^* представляется предпочтительней. Во-первых, она несмещенная, так как

$$\mathbf{E}_\lambda \lambda_1^* = \mathbf{E}_\lambda \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}_\lambda X_i = \frac{1}{n} n\lambda = \lambda;$$

во-вторых, она состоятельная в силу усиленного закона больших чисел (УЗБЧ).

В то же время, оценка λ_2^* менее удобна для исследования, хотя она является состоятельной (проверьте, используя закон больших чисел). Например, исследовать для нее свойство несмещенности — технически трудная задача.

Пример 10.4. Пусть $\vec{X} \Subset U_{[\theta_1;\, \theta_2]}$, где $\theta_1 < \theta_2$ — неизвестные параметры. Найти ОММ.

Решение. Вычислим моменты первых двух порядков равномерного распределения

$$\alpha_1(\theta_1, \theta_2) = \mathbf{E}X_i = \frac{\theta_1 + \theta_2}{2},$$

$$\mathbf{D}X_i = \frac{(\theta_2 - \theta_1)^2}{12}.$$

Составим систему уравнений, приравнивая теоретические и эмпирические математическое ожидание и дисперсию:

$$\begin{cases} \mathbf{E}X_1 = \overline{X}; \\ \mathbf{D}X_1 = S^2; \end{cases} \iff \begin{cases} \frac{\theta_1+\theta_2}{2} = \overline{X}; \\ \frac{(\theta_2-\theta_1)^2}{12} = S^2; \end{cases} \iff \begin{cases} \theta_1 + \theta_2 = 2\overline{X}; \\ \theta_2 - \theta_1 = \sqrt{12}S. \end{cases}$$

Решая последнюю систему относительно неизвестных θ_1, θ_2 (вычитая и складывая уравнения системы), получим оценки ММ:

$$\theta_1^* = \overline{X} - \sqrt{3}S; \quad \theta_2^* = \overline{X} + \sqrt{3}S.$$

Домашнее задание 10

10.1. Измерен рост (в см) студентов одной учебной группы. Результаты измерений дали выборку (171; 186; 164; 190; 158; 181; 176; 180; 174; 157; 176; 169; 164; 186).

а) Построить реализацию гистограммы.

б) Вычислить реализации выборочного среднего, выборочной дисперсии и выборочного стандартного отклонения S. На одном графике с гистограммой построить график плотности нормального закона с параметрами $\overline{X}$, S^2.

10.2. Пассажир маршрутного такси измерил 8 раз время ожидания такси и получил следующие результаты (в минутах): 8; 4; 5; 4; 2; 15; 1; 6. У него есть две гипотезы относительно графика движения такси: либо график движения соблюдается, и время ожидания имеет равномерное распределение на отрезке

$[0;\ \theta]$, либо график движения не соблюдается, и время ожидания имеет показательное распределение с параметром λ.

а) Вычислить реализации оценок параметров θ и λ, использовав оценки $\widetilde{\theta}_2 = (n+1)X_{(n)}/n$ и $\widetilde{\lambda}_2 = \frac{n-1}{n\overline{X}}$.

б) Построить на одном графике реализацию эмпирической функции распределения и теоретические функции распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок.

в) Построить на одном графике реализацию гистограммы и теоретические плотности распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок.

г) На основании проведенного исследования сделать вывод о том, какая из гипотез выглядит более соответствующей экспериментальным данным.

10.3. По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценки методом моментов:

а) параметра p по первому и по второму моменту при известном $m > 0$;

б) параметров p и m.

Исследовать состоятельность построенных оценок.

10.4. Используя метод моментов, построить бесконечную последовательность различных оценок параметра θ равномерного распределения на отрезке $[0;\ \theta]$. Будут ли полученные оценки состоятельными?

10.5. С помощью метода моментов построить оценку параметра $\theta > 0$, если распределение выборки имеет плотность:

а) $\theta t^{\theta-1}$ при $t \in [0;\ 1]$; б) $2t/\theta^2$ при $t \in [0;\ \theta]$.

Исследовать полученные оценки на состоятельность.

10.6. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ 1]; \\ 0, & t \notin [0;\ 1]. \end{cases}$$

Найти оценку параметра $\theta > 0$ методом моментов, исследовать ее на несмещенность и состоятельность.

10.7. Методом моментов найти оценку параметра $\alpha > 0$ по выборке из показательного распределения с плотностью $f_\alpha(t) = \alpha e^{-\alpha t}$, $t > 0$. Будет ли оценка несмещенной и состоятельной?

10.8. По выборке $(X_1, \ldots, X_n)$ методом моментов найти две различные оценки параметра $p \in (0, 1)$, если известно, что:

$P\{X_1 = 1\} = p/2$; $P\{X_1 = 2\} = p/2$; $P\{X_1 = 3\} = 1 - p$.

Будут ли полученные оценки несмещенными и состоятельными?

Задачи для решения в классе

10.9. По данной реализации выборки $\vec{x}$=(0; 0; 1; 1; 0; 0; 0; 0; 0; 1):

а) построить графики эмпирической функции распределения и гистограммы;

б) вычислить выборочные среднее и дисперсию.

10.10. Пусть $\vec{X} \in \Phi_{a,\sigma^2}$. Вычислить $\mathbf{E}\overline{X}$, $\mathbf{D}\overline{X}$. Какое распределение имеет случайная величина $\overline{X}$?

10.11. Дана выборка $\vec{X} \in \Pi_\lambda$, $\lambda > 0$ — неизвестный параметр. Проверить, что статистики

$$T_1 = \overline{X}, \quad T_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i = k), \quad T_3 = \frac{X_1 + X_n}{2}$$

являются несмещенными оценками соответственно для λ, $\frac{\lambda^k}{k!}e^{-\lambda}$ и λ. Являются ли эти оценки состоятельными?

10.12. По выборке $(X_1, \ldots, X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0; 1)$ построить оценки параметра p:

а) по первому моменту;

б) по второму моменту;

в) по произвольному k-му моменту.

Можно ли отдать предпочтение какой-либо из построенных оценок? Исследовать их состоятельность и несмещенность.

10.13. При каких значениях параметра $\theta > 0$ распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1; \\ 0, & t < 1 \end{cases}$$

существует оценка параметра по первому моменту? Можно ли построить состоятельную оценку методом моментов в случае, когда оценки по первому моменту не существует?

10.14. По выборке $(X_1, \ldots, X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2} e^{-\lambda|t|}, \quad t \in \mathbf{R}$, построить оценку параметра $\lambda > 0$ методом моментов.

10.15. Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод моментов, построить оценки:

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

10.16. Используя метод моментов, оценить параметр θ равномерного распределения на отрезке:

а) $[-\theta;\ \theta]$, $\theta > 0$; б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

§11. Оценки максимального правдоподобия.

Метод максимального правдоподобия

Пусть $\vec{X} \Subset F(t,\theta),\ \theta \in \Theta$. Предположим, что теоретическое распределение либо абсолютно непрерывно с плотностью $f(t,\theta) = f_{X_i}(t),$ либо дискретно, при этом для ряда распределения будем использовать то же обозначение: $f(t,\theta) = \mathbf{P}(X_i = t)$. **Функцией правдоподобия, соответствующей выборке** $\vec{X}$, называется функция

$$\Pi(\theta) = \Pi(\vec{X},\ \theta) = \prod_{i=1}^{n} f(X_i,\ \theta).$$

Оценкой максимального правдоподобия (ОМП) называется такое значение параметра $\theta = \hat{\theta}(\vec{X}),$ при котором функция правдоподобия принимает наибольшее значение, то есть

$$\Pi(\vec{X},\ \hat{\theta}) = \max_{\theta\in\Theta} \Pi(\vec{X},\ \theta).$$

Пример 11.1. Пусть $\vec{X} \Subset U_{[0,\theta]}$, где $\theta > 0$. Найти ОМП для параметра θ.

Решение. Найдем функцию правдоподобия, соответствующую выборке $\vec{X}$ из равномерного распределения $U_{[0,\theta]}$. Плотность распределения закона $U_{[0,\theta]}$ при $t = X_i$ равна:

$$f(X_i,\theta) = \begin{cases} \frac{1}{\theta}, & \text{если } X_i \in [0,\theta]; \\ 0, & \text{если } X_i \notin [0,\theta] \end{cases} \tag{9}$$

Тогда функция правдоподобия вычисляется следующим образом:

$$\Pi(\theta) = \Pi(\vec{X},\ \theta) = \prod_{i=1}^{n} f(X_i,\theta) =$$

$$= \begin{cases} \frac{1}{\theta^n}, & \text{если } X_i \in [0,\theta] \text{ для всех } i = 1,2,\ldots,n; \\ 0, & \text{иначе.} \end{cases}$$

Это соотношение можно переписать в следующем виде:

$$\Pi(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{если } \theta > X_{(n)}; \\ 0, & \text{иначе.} \end{cases}$$

Последнее задание функции $\Pi(\theta) = \Pi(\vec{X},\ \theta)$ позволяет изобразить ее график (см. рис. 5). Из графика видно, что своего наибольшего значения функция $\Pi(\theta)$ достигает при $\theta = X_{(n)}$. Следовательно, оценка максимального правдоподобия имеет вид: $\hat{\theta}(\vec{X}) = X_{(n)}$.

Если функция правдоподобия дифференцируема при всех $\theta \in \Theta$, то значение $\theta = \hat{\theta}$ должно быть решением уравнения

$$\Pi'(\theta) = 0,$$

которое называется уравнением правдоподобия, или эквивалентного уравнения

$$\frac{d}{d\theta} \ln \Pi(\theta) = 0 \iff \sum_{i=1}^{n} \frac{d}{d\theta} \ln f(X_i, \theta) = 0.$$

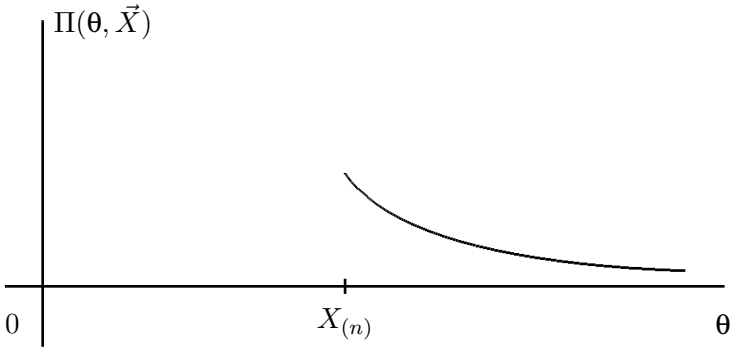

Рис. 5: Функция правдоподобия для $\vec{X} \Subset U_{[0,\theta]}$

Пример 11.2. Найти ОМП неизвестного параметра λ для выборки из распределения Пуассона.

Решение. Для распределения Пуассона Π_λ ряд распределения имеет вид:

$$f(t,\lambda) = \mathbf{P}(X_i = t) = e^{-\lambda}\frac{\lambda^t}{t!}.$$

Искомая оценка должна быть решением уравнения правдоподобия. Для решения этого уравнения вычислим последовательно:

$$f(X_i,\lambda) = e^{-\lambda}\frac{\lambda^{X_i}}{X_i!}; \quad \ln f(X_i,\lambda) = -\lambda + X_i\ln\lambda - \ln(X_i!);$$

$$\frac{d}{d\lambda}\ln f(X_i,\lambda) = -1 + \frac{1}{\lambda}X_i;$$

$$\sum_{i=1}^{n}\frac{d}{d\lambda}\ln f(X_i,\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n}X_i = -n + \frac{1}{\lambda}n\overline{X}.$$

Тогда уравнение правдоподобия и его решение имеют вид:

$$-n + \frac{1}{\lambda}n\overline{X} = 0 \iff \lambda = \overline{X}.$$

Заметим, что вторая производная логарифмической функции правдоподобия

$$\frac{d^2}{d\lambda^2}\ln\Pi(\lambda) = -\frac{1}{\lambda^2}n\overline{X} < 0$$

при всех λ, так как при нашем предположении $\vec{X} \Subset \Pi_\lambda$ все элементы выборки $X_1,...,X_n$, а значит, и выборочное среднее $\overline{X}$, с вероятностью единица неотрицательны. Значит, найденное решение $\lambda = \overline{X}$ уравнения правдоподобия является единственной точкой максимума функций $\Pi(\lambda)$ и $\ln\Pi(\lambda)$, а следовательно, статистика $\hat{\lambda} = \overline{X}$ является ОМП параметра λ.

Рассмотрим теперь случай многомерного параметра, предположив опять для простоты, что $\theta = (\theta_1,\theta_2)$ — двумерный параметр. Тогда для нахождения ОМП нужно найти точку

$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ наибольшего значения функции двух переменных $\Pi(\theta_1, \theta_2)$.

Пример 11.3. По выборке из равномерного распределения $U_{[\theta_1;\ \theta_2]}$ найти ОМП неизвестного параметра $\theta = (\theta_1, \theta_2)$.

Решение. Найдем функцию правдоподобия, соответствующую выборке $\vec{X}$ из равномерного распределения $U_{[\theta_1,\theta_2]}$. Так как плотность распределения закона $U_{[\theta_1,\theta_2]}$ при $t = X_i$ равна

$$f(X_i, \theta) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \text{если } X_i \in [\theta_1, \theta_2]; \\ 0, & \text{если } X_i \notin [\theta_1, \theta_2]; \end{cases}$$

то функция правдоподобия представляется в виде:

$$\Pi(\theta_1, \theta_2) = \prod_{i=1}^{n} f(X_i, \theta) = \begin{cases} \frac{1}{(\theta_2 - \theta_1)^n}, & \text{если все } X_i \in [\theta_1, \theta_2]; \\ 0, & \text{иначе.} \end{cases}$$

Или по-другому:

$$\Pi(\theta_1, \theta_2) = \begin{cases} \frac{1}{(\theta_2 - \theta_1)^n}, & \text{если } \theta_2 \geq X_{(n)},\ \theta_1 \leq X_{(1)}; \\ 0, & \text{иначе.} \end{cases}$$

Из последнего равенства видно, что функция правдоподобия отлична от нуля (более того, строго положительна) лишь при значениях (θ_1, θ_2), удовлетворяющих неравенствам:

$$\theta_1 \leq X_{(1)} \leq X_{(n)} \leq \theta_2.$$

Значит, своего наибольшего значения функция $\Pi(\theta_1, \theta_2)$ достигает лишь при таких (θ_1, θ_2). Однако при таких значениях (θ_1, θ_2) разность $(\theta_2 - \theta_1)$ принимает свое наименьшее значение $(X_{(n)} - X_{(1)})$ при $\theta_2 = X_{(n)},\ \theta_1 = X_{(1)}$. А значит, функция $\Pi(\theta_1, \theta_2)$ принимает свое наибольшее значение при тех же значениях (θ_1, θ_2), то есть искомая ОМП имеет вид: $\hat{\theta}_2 = X_{(n)}$, $\hat{\theta}_1 = X_{(1)}$.

Если функция правдоподобия дифференцируема, то для решения этой задачи вместо решения уравнения правдоподобия

нужно найти решение следующей системы уравнений:

$$\begin{cases} \frac{\partial \Pi(\theta_1,\theta_2)}{\partial \theta_1} = 0; \\ \\ \frac{\partial \Pi(\theta_1,\theta_2)}{\partial \theta_2} = 0. \end{cases}$$

Сравнение оценок: среднеквадратический подход

Пусть $\vec{X} \Subset F(t,\theta)$, $\theta \in \Theta$, и $\widetilde{\theta} = \widetilde{\theta}(\vec{X})$ — какая-нибудь оценка параметра θ. Так как оценка является случайной величиной, то даже свойство несмещенности не гарантирует близость ее конкретной реализации $\widetilde{\theta}(\vec{x})$ к оцениваемому параметру. Если оценка является состоятельной, то такая близость гарантируется с заданной вероятностью, но только при достаточно больших объемах выборки n. При фиксированном объеме выборки наиболее распространенной «мерой близости» оценки к оцениваемому параметру является **квадратическая характеристика**, или среднее значение квадрата отклонения $\mathbf{E}(\widetilde{\theta}-\theta)^2$.

Из двух оценок $\widetilde{\theta}_1$ считается *лучше*, чем $\widetilde{\theta}_2$, если при всех $\theta \in \Theta$ выполняется неравенство

$$\mathbf{E}(\widetilde{\theta}_1 - \theta)^2 \le \mathbf{E}(\widetilde{\theta}_2 - \theta)^2,$$

а хотя бы для одного θ неравенство оказывается строгим.

Заметим, что квадратическая характеристика оценки не меньше ее дисперсии, и равенство достигается для несмещенных оценок:

$$\mathbf{E}(\widetilde{\theta}-\theta)^2 = \Big(\mathbf{E}(\widetilde{\theta}-\theta)\Big)^2 + \mathbf{D}(\widetilde{\theta}-\theta) = \Big(\mathbf{E}\widetilde{\theta}-\theta\Big)^2 + \mathbf{D}\widetilde{\theta} \ge \mathbf{D}\widetilde{\theta}.$$

Если $\widetilde{\theta}$ — несмещенная оценка параметра θ, то есть $\mathbf{E}\widetilde{\theta} = \theta$, то для нее:

$$\mathbf{E}(\widetilde{\theta}-\theta)^2 = \Big(\mathbf{E}\widetilde{\theta}-\theta\Big)^2 + \mathbf{D}\widetilde{\theta} = \mathbf{D}\widetilde{\theta}.$$

Пример 11.4. Пусть $\vec{X} \Subset U_{[0;\,\theta]}$, $\theta > 0$. Сравнить с помощью среднеквадратического подхода оценки параметра θ: $\theta_1^* = 2\overline{X}$ и $\hat{\theta} = X_{(n)}$.

Решение. Проверим сначала свойство несмещенности обеих оценок. Вычисляем математические ожидания, используя результаты предыдущего параграфа (см. решение примера 10.2):

$$\mathbf{E}\theta_1^* = \mathbf{E}(2\overline{X}) = 2\mathbf{E}\overline{X} = 2\frac{\theta}{2} = \theta, \quad \mathbf{E}\hat{\theta} = \mathbf{E}X_{(n)} = \frac{n}{n+1}\theta.$$

Видим, что из двух оценок $\theta_1^* = 2\overline{X}$ является несмещенной, а $\hat{\theta} = X_{(n)}$ — смещенной. Чтобы выяснить, какая из оценок лучше, вычислим для каждой квадратичную характеристику. Для несмещенной оценки она совпадает с дисперсией:

$$\mathbf{E}(\theta_1^* - \theta)^2 = \mathbf{D}\theta_1^* = \mathbf{D}(2\overline{X}) = 4\mathbf{D}\frac{1}{n}\sum_{i=1}^{n} X_i = 4\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}_\theta X_i =$$

$$= 4\frac{1}{n^2} n\mathbf{D}_\theta X_1 = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}. \tag{10}$$

При вычислении квадратичной характеристики оценки $\hat{\theta}_n = X_{(n)}$ мы будем использовать плотность ее распределения, найденную при решении примера 10.2.

$$\mathbf{E}(X_{(n)}-\theta)^2 = \mathbf{E}X_{(n)}^2 - 2\theta\mathbf{E}X_{(n)} + \theta^2 = \int_0^\theta y^2 \frac{ny^{n-1}}{\theta^n} dy - 2\theta\frac{n\theta}{n+1} + \theta^2 =$$

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

Сравнивая квадратичные характеристики, вычисленные в (10) и (11), видим, что

$$\frac{\theta^2}{3n} \geq \frac{2\theta^2}{(n+1)(n+2)}$$

для всех $\theta > 0$ и для всех $n \geq 1$, а при $n > 1$ неравенство является строгим.

Следовательно, ОМП $\hat{\theta} = X_{(n)}$ лучше в среднеквадратичном, чем ОММ $\theta_1^* = 2\overline{X}$.

Отметим, что при среднеквадратическом подходе к сравнению оценок нельзя найти наилучшую в классе всех оценок (в частности, существуют несравнимые оценки). Доказательство этого факта основано на рассмотрении вырожденных оценок, равных константе независимо от значений выборки.

Для того, чтобы избежать необходимости сравнивать получаемые оценки с вырожденными оценками, нужно ограничить класс рассматриваемых оценок. Как правило, сравнивают только несмещенные оценки. Среди несмещенных оценок наилучшая оценка параметра для заданного параметрического семейства может существовать. Ее называют *эффективной* оценкой. Эффективная оценка имеет наименьшую дисперсию из всех несмещенных оценок.

Домашнее задание 11

11.1. С помощью метода максимального правдоподобия построить оценку параметра $\theta > 0$, если элементы выборки имеют плотность распределения:

а) $\theta t^{\theta-1}$ при $t \in [0;\ 1]$; б) $2t/\theta^2$ при $t \in [0;\ \theta]$.

Исследовать полученные оценки на состоятельность.

11.2. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ 1]; \\ 0, & t \notin [0;\ 1]. \end{cases}$$

Найти оценку параметра $\theta > 0$ методом максимального правдоподобия, исследовать ее на несмещенность и состоятельность.

11.3. По выборке $(X_1, \ldots, X_n)$ методом максимального правдоподобия найти оценку параметра $p \in (0, 1)$, если известно, что

$P\{X_1 = 1\} = p/2,\ P\{X_1 = 2\} = p/2,\ P\{X_1 = 3\} = 1 - p.$

Будет ли полученная оценка несмещенной и состоятельной?

11.4. Дана выборка $\vec{X} \Subset \mathrm{U}_{[0,\theta]};\ \theta > 0$ — неизвестный параметр. Сравнить, какая из оценок для параметра θ лучше в среднеквадратичном: $\theta_1^* = 2\overline{X}$, $\theta_2^* = \frac{n+1}{n} X_{(n)}$.

11.5. Пусть $\vec{X} \Subset F(t, \theta)$; где $\theta = \mathbf{E}_\theta X_1,\ \ \mathbf{D}X_1 < \infty$. Показать, что оценка $\theta_1^* = \overline{X}$ является наилучшей в среднеквадратичном среди всех несмещенных оценок вида:

$$\theta^* = C_1 X_1 + C_2 X_2 + \cdots + C_n X_n,\ \ C_1 + C_2 + \cdots + C_n = 1.$$

11.6. Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} e^{\theta - t}, & t \geq \theta; \\ 0, & t < \theta. \end{cases}$$

Найти оценку для θ:

а) методом моментов;

б) методом максимального правдоподобия.

Будут ли полученные оценки несмещенными и состоятельными?

11.7. Вычислить смещения оценок в задаче 11.6 и получить исправленные несмещенные оценки.

Задачи для решения в классе

11.8. По выборке $(X_1,\ \ldots,\ X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0;\ 1)$ построить оценку параметра p методом максимального правдоподобия. (Указание: показать, что вероятность попадания в точку t для элементов выборки равна $f(t,\ p) = p^t(1-p)^{1-t}$, где t может принимать только два значения — 0 и 1). Исследовать состоятельность и несмещенность полученной оценки.

11.9. По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценку максимального правдоподобия

параметра p при известном $m > 0$. Исследовать состоятельность и несмещенность оценки.

11.10. По выборке из показательного распределения E_α построить оценку максимального правдоподобия параметра $\alpha > 0$. Исследовать состоятельность оценки.

11.11. Построить оценку максимального правдоподобия по выборке из распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1; \\ 0, & t < 1. \end{cases}$$

Доказать состоятельность полученной оценки.

11.12. По выборке $(X_1, \ldots, X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2} e^{-\lambda|t|}, \quad t \in \mathbf{R}$, построить оценку параметра $\lambda > 0$ методом максимального правдоподобия.

11.13. Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод максимального правдоподобия, построить оценки:

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

11.14. Используя метод максимального правдоподобия, оценить параметр θ равномерного распределения на отрезке:

а) $[-\theta;\ \theta]$, $\theta > 0$;

б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

§12. Доверительные интервалы. Проверка статистических гипотез

Пусть имеется выборка объема n из распределения, известного с точностью до параметра: $\vec{X} \Subset F(t, \theta),\ \theta \in \Theta$. *Доверительным интервалом с уровнем доверия* γ (γ-доверительным интервалом) для неизвестного параметра θ называют случайный интервал $(\theta_-;\ \theta_+) \subset \Theta,$ построенный по выборке, который накрывает неизвестное значение параметра с вероятностью, равной $\gamma,$ или по крайней мере стремящейся к γ с ростом объема выборки, то есть:

$$\mathbf{P}\{\theta \in (\theta_-;\ \theta_+)\} \to \gamma$$

при $n \to \infty$.

В случае, когда вместо сходимости выполняется точное равенство, доверительный интервал называется *точным.*

θ_-, θ_+ — это оценки параметра $\theta,$ называемые *нижней и верхней доверительными границами.* Число $\gamma \in (0;\ 1)$ — уровень доверия, или доверительная вероятность, — выбирается заранее и отражает, как сказано в [7], «степень готовности мириться с возможностью ошибки». Чем менее мы готовы мириться с возможной ошибкой, тем большее (более близкое к единице) значение γ должны устанавливать.

Асимптотические доверительные интервалы

Если распределение не является нормальным, точный доверительный интервал, как правило, не удается построить. Поэтому строят асимптотический доверительный интервал, применяя центральную предельную теорему, которая утверждает, что для всех $t_1, t_2 \in \mathbf{R}\ (t_1 < t_2)$ выполнено:

$$\lim_{n\to\infty} \mathbf{P}\left(t_1 \le \frac{n\overline{X} - na}{\sigma\sqrt{n}} < t_2\right) = \Phi(t_2) - \Phi(t_1),$$

то есть центрированные и нормированные суммы случайных величин $n\overline{X} = X_1 + \ldots + X_n$ сходятся по распределению к случайной величине, имеющей стандартное нормальное распределение.

Здесь $a = \mathbf{E}X_1,\ \sigma^2 = \mathbf{D}X_1 > 0$ — математическое ожидание и дисперсия элементов выборки.

Если выбрать $t_2 = -t_1 = A$ и принять доверительный уровень равным γ, то:

$$\lim_{n\to\infty} \mathbf{P}\left(-A \le \frac{n\overline{X} - na}{\sigma\sqrt{n}} < A\right) = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = \gamma,$$

откуда получаем:

$$\Phi(A) = (\gamma + 1)/2.$$

По заданному γ можно найти A с помощью таблиц нормального распределения или программных приложений. Отметим следующее свойство сходимости по распределению: если Y_n сходится по распределению к Y, а Z_n сходится к 1 с вероятностью единица, то их произведение $Y_n Z_n$ сходится по распределению к Y. Выберем:

$$Y_n = \frac{n\overline{X} - na}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - a)}{\sigma}, \qquad Z_n = \frac{\sigma}{S}.$$

Вспомним, что $S = \sqrt{\overline{X^2} - (\overline{X})^2} \to \sigma$ с вероятностью 1, и, следовательно, $Z_n \to 1$ с вероятностью 1. Итак,

$$Y_n Z_n = \frac{\sqrt{n}(\overline{X} - a)}{\sigma} \cdot \frac{\sigma}{S} = \frac{\sqrt{n}(\overline{X} - a)}{S}$$

сходится по распределению к стандартной нормальной случайной величине, то есть:

$$\lim_{n\to\infty} \mathbf{P}\left(-A \le \frac{\sqrt{n}(\overline{X} - a)}{S} < A\right) = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = \gamma,$$

где константа A выбирается по формуле $\Phi(A) = (\gamma + 1)/2$.

Чтобы для неизвестного параметра θ найти доверительный интервал асимптотического уровня γ, нужно для исследуемого однопараметрического семейства распределений найти зависимость $\mathbf{E}X_1 = a = a(\theta)$ и решить относительно параметра θ двойное неравенство:

$$-A \le \frac{\sqrt{n}(\overline{X} - a(\theta))}{S} < A.$$

Для этого нужно, чтобы функция $a(\theta)$ была непрерывной и строго монотонной. Получившиеся границы дверительного интервала будем обозначать через θ_- и θ_+.

Пример 12.1. Пусть $\vec{X} \Subset E_\alpha$, $\alpha > 0$. Построить асимптотический доверительный интервал для параметра α.

Решение. Так как для показательного распределения с параметром α математическое ожидание равняется $\mathbf{E}X_1 = 1/\alpha$, то

$$-A \le \frac{\sqrt{n}(\overline{X} - 1/\alpha)}{S} < A.$$

Последовательно выразим:

$$-\frac{AS}{\sqrt{n}} \le \overline{X} - \frac{1}{\alpha} < \frac{AS}{\sqrt{n}};$$

$$\overline{X} - \frac{AS}{\sqrt{n}} < \frac{1}{\alpha} \le \overline{X} + \frac{AS}{\sqrt{n}};$$

$$\frac{1}{\overline{X} - \dfrac{AS}{\sqrt{n}}} > \alpha \ge \frac{1}{\overline{X} + \dfrac{AS}{\sqrt{n}}}.$$

Итак, мы получили доверительный интервал $(\alpha_-;\ \alpha_+)$, где

$$\alpha_- = \frac{1}{\overline{X} + \dfrac{AS}{\sqrt{n}}}, \qquad \alpha_+ = \frac{1}{\overline{X} - \dfrac{AS}{\sqrt{n}}}.$$

Статистические гипотезы

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ — выборка, $\vec{X} \Subset \mathbf{F}$, где $\mathbf{F}$ — полностью или частично неизвестное распределение отдельного наблюдения X_i.

Статистической гипотезой будем называть всякое утверждение о виде или свойствах неизвестного распределения $\mathbf{F}$.

Пусть $\mathbf{F}$ — полностью неизвестное распределение. Примерами гипотез являются
$\mathrm{H} : \mathbf{F} = \mathbf{F}_0$, где $\mathbf{F}_0$ — полностью определенное распределение;
$\mathrm{H} : \mathbf{F} \in \widehat{\mathbf{F}}_0$, где $\widehat{\mathbf{F}}_0$ — множество распределений (например, $\widehat{\mathbf{F}}_0 = \Phi_{a,\sigma^2}$ или $\widehat{\mathbf{F}}_0 = B_p$).

В этих примерах наблюдения имеют распределения из некоторого одно- или двухпараметрического семейства. Но могут быть непараметрические множества, например, $\{\mathbf{F} : \mathbf{E}X_i > 0\}$ — класс распределений с положительными математическими ожиданиями.

Пример 12.2. Пусть $\mathbf{F}$ — частично известное распределение. Например, $\mathbf{F} \in U_{[a;\, b]}$ (наблюдения имеют равномерное распределение). В этом случае примеры гипотез:
$\mathrm{H} : a = 0,\ b = 1$ (распределение равномерное на $[0;\ 1]$);
$\mathrm{H} : a = 0$ (распределение равномерное на $[0;\ b]$);
$\mathrm{H} : a < b - 1$ (распределение равномерное на отрезке длины более 1).

Гипотеза называется *простой*, если она однозначно определяет распределение $\mathbf{F}$, в противном случае гипотеза называется *сложной*. В приведенных выше примерах простыми являются гипотезы:
$\mathrm{H} : \mathbf{F} = \mathbf{F}_0$ и $\mathrm{H} : a = 0,\ b = 1$ (последняя в случае, когда известно, что распределение равномерное на $[a;\ b]$).

Остальные гипотезы являются сложными.

Мы будем рассматривать ситуацию, когда гипотез всего две. Одну из них называют *основной*, а другую — *альтернативной*, обозначая соответственно H_0 и H_1.

Статистические критерии

Статистическим критерием называют всякое правило, позволяющее на основании наблюдаемого выборочного вектора $\vec{X}$ принять одну из гипотез: основную или альтернативную.

При применении статистического критерия могут возникнуть ошибки двух родов. Ошибка нулевого рода состоит в том, что отвергается верная нулевая гипотеза. Ошибка первого рода — отвергается верная первая гипотеза. Вообще ошибка i-го рода состоит в том, что статистический критерий отвергает верную i-ю гипотезу.

принимаемая гипотеза	верна гипотеза H_0	верна гипотеза H_1
H_0	нет ошибки	ошибка 1-го рода
H_1	ошибка 0-го рода	нет ошибки

Критерий характеризуется вероятностями ошибок:

$$\alpha_0 = \mathbf{P}_{H_0}(H_0 \text{ отвергается}); \quad \alpha_1 = \mathbf{P}_{H_1}(H_1 \text{ отвергается}).$$

Здесь нижний индекс у символа вероятности указывает, при выполнении какой гипотезы подсчитывается вероятность. Из всевозможных критериев надо выбирать такие, у которых вероятности ошибок по возможности малы. Отметим, что, как правило, не существует критерия, для которого обе вероятности ошибок равны нулю, и чем меньше вероятность ошибки нулевого рода, тем больше вероятность ошибки первого рода.

Рассмотрим введенные понятия на следующем примере.

Пример 12.3. Студенты группы А считают, что они играют в шахматы вдвое лучше, чем студенты группы В. В свою оче-

редь, студенты группы B считают, что они играют в шахматы втрое лучше, чем студенты группы A. Для решения спора назначается шахматный матч между группами A и B. С каждой стороны участвуют 3 студента, выбираемые по жребию. Решено считать справедливым мнение группы, выигрывшей матч, то есть набравшей не менее 2 очков в 3 партиях. Предполагается, что ничьих нет. Найти, в чем состоят ошибки нулевого и первого рода. Вычислить вероятности этих ошибок.

Решение. Предполагаем, что нулевая гипотеза (соответствующая мнению студентов группы A) состоит в том, что вероятность выигрыша каждого студента группы A у студента группы B вдвое больше вероятности проигрыша, то есть вероятность выигрыша равна $2/3$. Согласно первой гипотезе (мнению студентов группы B), вероятность выигрыша каждого студента группы A втрое меньше вероятности проигрыша, то есть равняется $1/4$.

Итак, проводятся три испытания схемы Бернулли с вероятностью успеха p, гипотеза $H_0:\ p = 2/3$; гипотеза $H_1:\ p = 1/4$.

Критерий (исход матча) предписывает принять гипотезу H_0, если число успехов в схеме Бернулли равняется двум или трем, а в противном случае принять гипотезу H_1.

Ошибка нулевого рода состоит в том, что критерий предписывает считать вероятность выигрыша студента первой группы равной $1/4$, в то время как она равняется $2/3$. Ошибка первого рода описывает противоположную ситуацию: вероятность выигрыша студента первой группы равняется $1/4$, а критерий предписывает считать ее равной $2/3$.

Вычислим вероятности ошибок.

Вероятность ошибки нулевого рода α_0 — это вероятность отвергнуть верную нулевую гипотезу, то есть получить ноль или один успех в схеме Бернулли, которая предполагает 3 испытания с $p = 2/3$ в каждом. Вычислим эту вероятность на основании формулы Бернулли:

$$\alpha_0 = \mathbf{P}_{H_0}(H_0 \text{ отвергается}) =$$

$$= C_3^0(2/3)^0(1/3)^3 + C_3^1(2/3)^1(1/3)^2 = 1/27 + 2/9 \approx 0,25.$$

Вероятность ошибки первого рода α_1 — это вероятность получить два или три успеха в схеме Бернулли, которая предполагает 3 испытания с $p = 1/4$ в каждом.

$$\alpha_1 = \mathbf{P}_{H_1}(H_1 \text{ отвергается}) =$$

$$= C_3^2(1/4)^2(3/4)^1 + C_3^3(1/4)^3(3/4)^0 = 9/64 + 1/64 \approx 0,15.$$

Домашняя работа 12

12.1. Пусть элементы выборки $\vec{X}$ имеют плотность распределения

$$f(t) = \frac{1}{\pi(1 + (t - \theta)^2)}, \quad t \in \mathbf{R}.$$

Здесь θ — неизвестный параметр, $\theta \in \mathbf{R}$. Построить оптимальный точный доверительный интервал для параметра θ по одному наблюдению ($n = 1$).

12.2. $\vec{X} \Subset B_p$, $0 < p < 1$. Построить асимптотический доверительный интервал для параметра p на основе оценки $\theta^* = \overline{X}$.

12.3. Пусть $\vec{X} \Subset U_{[0;\ \theta]}$, где $\theta > 0$. С помощью статистик $\overline{X}$ и $\overline{X^2}$ построить асимптотические доверительные интервалы (соответственно (θ_1^-, θ_1^+) и (θ_2^-, θ_2^+)) уровня $1 - \varepsilon$ и показать, что случайный интервал (θ_2^-, θ_2^+) короче соответствующего (θ_1^-, θ_1^+).

12.4. Крупная партия товаров может содержать долю дефектных изделий. Поставщик полагает, что эта доля составляет 3%, а покупатель — 10%. Условия поставки: если при проверке 20 случайным образом отобранных товаров обнаружено не более одного дефектного, то партия принимается на условиях поставщика, в противном случае — на условиях покупателя. Требуется определить:

1) каковы статистические гипотезы, статистика критерия, область ее значений, критическая область;

2) какое распределение имеет статистика критерия, в чем состоят ошибки первого и второго рода и каковы их вероятности.

12.5. Имеется выборка объема 1 из нормального распределения $\Phi_{a,1}$. Проверяются простые гипотезы $H_0 : \ a = 0$, $H_1 : \ a = 1$. Используется следующий критерий (при заданной постоянной c):

$$H_0 \Leftrightarrow X_1 \leq c.$$

Вычислить, в зависимости от c, вероятности ошибок первого и второго рода.

12.6. Используя конструкции доверительного интервала, построить критерий точного уровня ε для проверки гипотезы $\mathrm{H} : \boldsymbol{\theta} = 1$, если:
а) $\vec{X} \Subset \Phi_{\boldsymbol{\theta},1}$;
б) $\vec{X} \Subset \Phi_{1,\boldsymbol{\theta}}$.

Задачи для решения в классе

12.7. $\vec{X} \Subset \Pi_{\boldsymbol{\lambda}}, \ \boldsymbol{\lambda} > 0$. Построить асимптотический доверительный интервал для параметра $\boldsymbol{\lambda}$ с помощью оценки $\boldsymbol{\lambda}^* = \overline{X}$.

12.8. Построить критерий, обладающий нулевыми вероятностями ошибок, для проверки гипотез $\mathrm{H}_0 : \vec{X} \Subset \Phi_{0,1}$ против $\mathrm{H}_1 : \vec{X} \Subset \Pi_{\boldsymbol{\lambda}}$.

12.9. Пусть $\vec{X} \Subset \Phi_{a,1}$. Для проверки гипотез $\mathrm{H}_0 : a = 0$ против $\mathrm{H}_1 : a = 1$ используется следующий критерий: H_0 принимается, если $X_{(n)} < 3$, и отвергается в противном случае. Найти вероятности ошибок.

12.10. Используя конструкции доверительного интервала, построить критерий асимптотического уровня ε для проверки гипотезы $\mathrm{H} : \boldsymbol{\theta} = 1$, если а) $\vec{X} \Subset \mathrm{E}_{\boldsymbol{\theta}}$; б) $\vec{X} \Subset \mathrm{B}_{\boldsymbol{\theta}/2}$; в) $\vec{X} \Subset \Pi_{\boldsymbol{\theta}}$.

Глава 4. Лабораторные работы

§13. Точечное и интервальное оценивание

Пакеты программ Microsoft Excel и Calc не являются специализированными пакетами статистического анализа, но широко распространены и снабжены набором функций, достаточным для решения большинства статистических задач.

Рассмотрим процедуры статистического анализа на примере искусственно сгенерированной выборки.

Пример 13.1. Сгенерировать реализацию выборки объема $n = 30$ по формуле $x_i = 1 - 100 \ln u_i$, где u_i — случайные числа — образуют реализацию выборки из равномерного распределения на $[0;\ 1]$. Построить реализацию вариационного ряда и гистограммы, выбрав число промежутков по формуле Стеджеса. Выдвинуть две двухпараметрических гипотезы о распределении выборочных значений. Оценить параметры распределений методом моментов (по первому и второму моментам) и методом максимального правдоподобия. На основании полученных реализаций оценок построить реализации оценок функций распределения. Сделать вывод о наиболее адекватной модели.

Решение. Получим реализацию выборки в столбике `A` электронной таблицы. Для этого в ячейку `A1` введем формулу

	A
1	=1-LN(СЛЧИС())*100

(здесь `СЛЧИС()` — *математическая* функция, реализующая независимые случайные числа, равномерно распределенные на отрезке от 0 до 1).

Скопируем содержимое ячейки в ячейки `A2-A30`. Скопируем *значения* столбика `A` в тот же столбик (для этого щелкнем правой кнопкой мыши по букве `A` и в выпадающем меню выберем *специальная вставка* ⇒ *значения*.

Копирование значений фиксирует реализацию выборки, сохраняя значения от последующего пересчета.

Вычислим количество промежутков по формуле Стеджеса: в ячейку B1 введем формулу

	A	B
1		=ЦЕЛОЕ(LOG(30;2))+1

В ячейках B2, B3, B4, B5 найдем последовательно наибольшее и наименьшее значения, размах реализации выборки и длину промежутка:

	A	B
1		=ЦЕЛОЕ(LOG(30;2))+1
2		=МАКС(A:A)
3		=МИН(A:A)
4		=B2-B3
5		=B4/B1

Последовательно прибавляя длину промежутка к минимальному значению, хранящемуся в ячейке A1, получаем в столбике C правые границы промежутков: 61,5; 117; 173; 229; 284 (округленно). Отметим, что здесь надо специально позаботиться о том, чтобы все элементы попали левее самой правой границы промежутка, для этого прибавим к самой правой границе 1, получив 285 вместо 284.

Подсчитаем количества элементов, попавших в каждый из промежутков. Для этого воспользуемся функцией ЧАСТОТА. Введем в ячейку D1 формулу

	D
1	=ЧАСТОТА(A1:A30;C1:C5)

Затем выделим ячейки D1:D5, нажмем клавишу F2 и введем формулу как формулу массива, нажав клавиши CTRL+SHIFT+ENTER.

В столбике F получим значения гистограммы, разделив значения столбика D на $n = 30$ и на длину промежутка,

хранящуюся в ячейке `B5`. Построим гистограмму по столбику `F` с помощью функции *диаграмма* (см. рис. 6).

	A	B	C	D	E	F	G	H
1	5,82	5	61,5	12		0,0072	82,89	**30,58**
2	10,6	284	117	5		0,003	113,5	**0,012**
3	12,6	5,82	173	5		0,003		
4	14,4	279	229	5		0,003		
5	16,4	55,7	285	3		0,0018		
6	29,7							
7	34,6							
8	35,1							
9	39,7							
10	48,8							
11	52							
12	55,7							
13	84,6							
14	89,5							
15	108							
16	112							
17	115							
18	117							
19	154							
20	163							
21	163							
22	168							

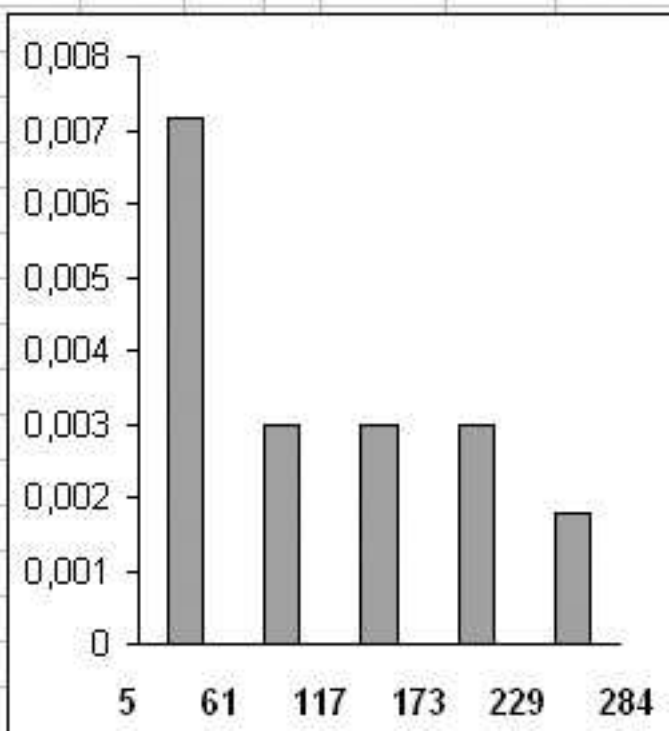

Рис. 6: Таблица Excel и гистограмма выборочных данных

По виду гистограммы нам предстоит решить, какие гипотезы о распределении выборки следует выдвинуть. Вспомним, как выглядят графики плотности распределения изученных нами двухпараметрических семейств распределений (равномерного, сдвинутого показательного, Парето, нормального). Заметим, что только сдвинутое показательное распределение и распределение Парето имеют плотности, похожие на полученную гистограмму (рис. 7). На рисунке слева изображен график плотности сдвинутого показательного распределения, справа — распределения Парето. Напомним, что формулы для плотностей распределений имеют следующий вид:

$$f_{\alpha,\theta}(t) = \begin{cases} \alpha e^{-\alpha(t-\theta)}, & t \geq \theta; \\ 0 & \text{иначе}; \end{cases} \qquad f_{\gamma,h}(t) = \begin{cases} \gamma h^{\gamma} t^{-(\gamma+1)}, & t \geq h; \\ 0 & \text{иначе}. \end{cases}$$

У сдвинутого показательного распределения параметр α положительный, а параметр θ — любое действительное число. У распределения Парето оба параметра γ и h положительны. Соответствующие функции распределения имеют вид:

$$F_{\alpha,\theta}(t) = \begin{cases} 1 - e^{-\alpha(t-\theta)}, & t \geq \theta; \\ 0 & \text{иначе}; \end{cases} \qquad F_{\gamma,h}(t) = \begin{cases} 1 - h^{\gamma} t^{-\gamma}, & t \geq h; \\ 0 & \text{иначе}. \end{cases}$$

Построим оценки параметров по первому и второму моментам. Для сдвинутого показательного распределения элементы выборки X_i равны $X_i = \theta + Y_i$, где Y_i образуют выборку из показательного распределения с параметром α, а θ — параметр сдвига. Как известно, $\mathbf{E}Y_i = 1/\alpha$, $\mathbf{D}Y_i = 1/\alpha^2$.

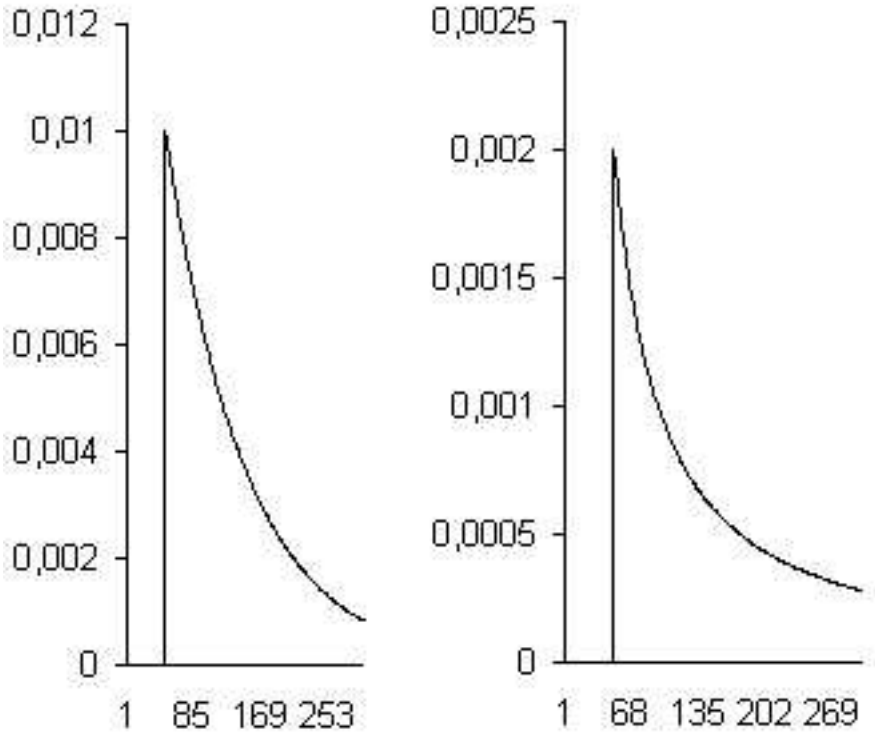

Рис. 7: Плотности сдвинутого показательного распределения и распределения Парето

Пользуясь свойствами математического ожидания и дисперсии, получаем систему уравнений:

$$\begin{cases} \mathbf{E}X_i = \theta + 1/\alpha; \\ \mathbf{D}X_i = 1/\alpha^2. \end{cases}$$

Выразим параметры:

$$\begin{cases} \alpha = (\mathbf{D}X_i)^{-1/2}; \\ \theta = \mathbf{E}X_i - (\mathbf{D}X_i)^{1/2}. \end{cases}$$

Заменим математическое ожидание и дисперсию на выборочное среднее $\overline{X}$ и выборочную дисперсию S^2, а параметры α и θ на их оценки α^* и θ^*. Получим оценки параметров:

$$\begin{cases} \alpha^* = S^{-1}, \\ \theta^* = \overline{X} - S. \end{cases}$$

Найдем реализации этих оценок. Выборочное стандартное отклонение S — это функция `СТАНДОТКЛОНП`, а выборочное среднее — функция `СРЗНАЧ`. Вычислим их значения в ячейках `G1` и `G2`, введя туда функции `=СТАНДОТКЛОНП(A:A)` и `=СРЗНАЧ(A:A)`. В ячейках `H1` и `H2` получим реализации оценок θ^* и α^*. Для того, чтобы понять, насколько хороши оценки методом моментов, построим графики реализаций параметрической оценки функции распределения $F(t, \alpha^*, \theta^*)$ и эмпирической функции распределения $F_n^*(t)$. Получим формулу для интервала дискретизации dt переменной t, исходя из того, чтобы dt было целой степенью числа 10, и множество выборочных значений делилось не менее чем на 100 интервалов. Обозначив через $R = X_{(n)} - X_{(1)}$ размах выборки, получаем:

$$R/100 \geq dt, \qquad dt = 10^k, \qquad dt \ \leq 10^{\lg R - 2}.$$

Выбирая в качестве dt наибольшее из таких чисел, приходим к формуле:

$$dt = 10^{[\lg R] - 2},$$

где $[a]$ — целая часть числа a.

Поскольку в нашем примере размах выборки равен 279, получаем $[\lg 279] = 2$, и $dt = 1$. Найдем значения оценки функции распределения по формуле

	I
1	=ЕСЛИ(СТРОКА()<H$1;0;1-EXP(-H$2*(СТРОКА()-H$1)))

и скопируем эту формулу в ячейки `I1:I285`.

Получим значения эмпирической функции распределения в тех же точках. Для этого создадим вспомогательный столбик `M`, содержащий границы промежутков дискретизации, скопировав функцию `=СТРОКА()` в ячейки `M1:M285`. Потом подсчитаем, сколько элементов выборки попало в каждый из промежутков. Для этого воспользуемся функцией `ЧАСТОТА`. Введем в ячейку `P1` формулу

	P
1	=ЧАСТОТА(A1:A30;M1:M285)

Затем выделим ячейки `P1:P285`, нажмем клавишу F2 и введем формулу как формулу массива, нажав клавиши `CTRL+SHIFT+ENTER`. Теперь получим значения эмпирической функции распределения в столбике `J`, введя в первую ячейку формулу:

	J
1	=СУММ(P$1:P1)/30

и скопировав ее в остальные ячейки. Здесь $30 = n$ — объем выборки.

Построим диаграмму по столбикам `I` и `J` (рис. 8).

Теперь получим оценки максимального правдоподобия для параметров α и θ сдвинутого показательного распределения. Заметим, что плотность распределения:

$$f_{\alpha,\theta}(t) = \begin{cases} \alpha e^{-\alpha(t-\theta)}, & \text{если } t \geq \theta; \\ 0 & \text{иначе;} \end{cases}$$

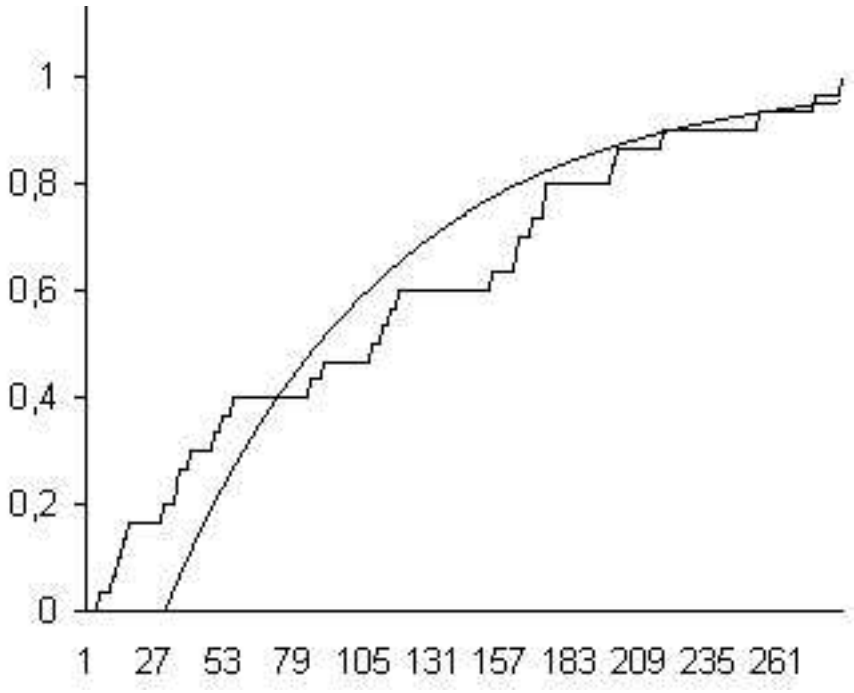

Рис. 8: Оценка функции сдвинутого показательного распределения методом моментов

непрерывна по параметру $\alpha > 0$ и разрывна по параметру θ. Сначала найдем оценку параметра θ непосредственно отысканием точки максимума функции правдопобобия. Функция правдоподобия равна:

$$\Pi(\vec{X}, \alpha, \theta) = \begin{cases} \prod_{i=1}^{n} (\alpha e^{-\alpha(X_i - \theta)}), & \text{если все } X_i \geq \theta; \\ 0 & \text{иначе;} \end{cases}$$

или

$$\Pi(\vec{X}, \alpha, \theta) = \begin{cases} \alpha^n e^{-\alpha(\sum_{i=1}^{n} X_i - n\theta)}, & \text{если } \theta \leq \min\{X_i\}; \\ 0 & \text{иначе.} \end{cases}$$

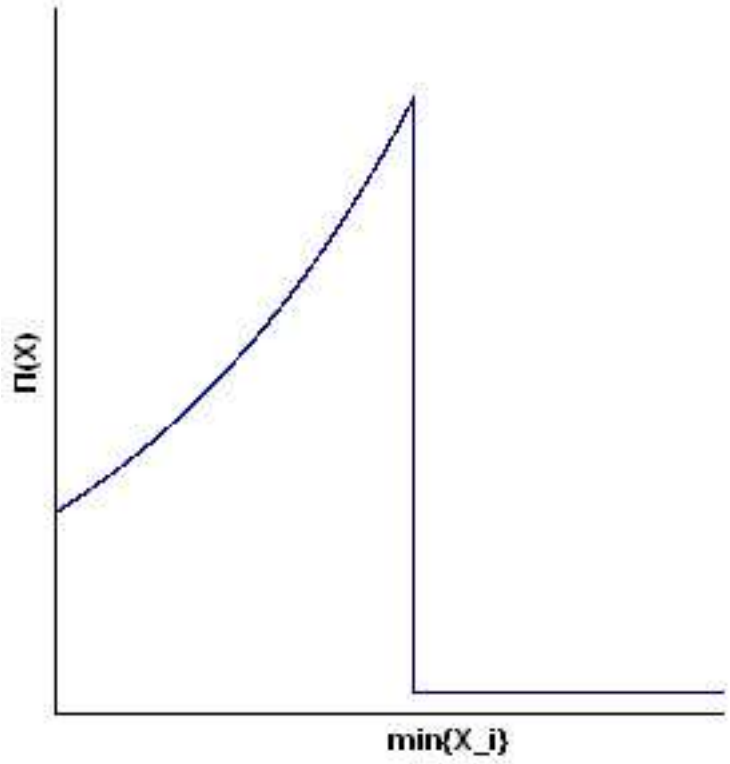

Рис. 9: Зависимость функции правдоподобия от параметра θ

Зависимость функции правдоподобия от параметра θ изображена на рис. 9. Ее максимум достигается в точке $\widehat{\theta} = \min\{X_i\}$, которая является оценкой максимального правдоподобия параметра θ.

Найдем оценку максимального правдоподобия параметра α. Для этого последовательно вычислим:

$$\ln f(t, \alpha, \theta) = \ln \alpha - \alpha(t - \theta) \text{ при } t \geq \theta;$$

$$\frac{\partial}{\partial \alpha} \ln f(t, \alpha, \theta) = \frac{1}{\alpha} - (t - \theta) \text{ при } t \geq \theta;$$

$$\frac{\partial}{\partial\alpha}\ln\Pi(\vec{X},\alpha,\theta)=\sum_{i=1}^{n}\frac{\partial}{\partial\alpha}\ln f(X_i,\alpha,\theta)=\sum_{i=1}^{n}\left(\frac{1}{\alpha}-(X_i-\theta)\right),$$

если все $X_i \geq \theta$. Приравнивая производную логарифма функции правдоподобия к нулю, получаем уравнение для определения оценки параметра α:

$$\sum_{i=1}^{n}\left(\frac{1}{\alpha}-(X_i-\theta)\right)=0,$$

решением которого является:

$$\alpha=\frac{n}{\sum_{i=1}^{n}X_i-n\theta}=\frac{1}{\overline{X}-\theta}.$$

Поскольку параметр θ неизвестен, заменим его на оценку максимального правдоподобия $\widehat{\theta}=\min\{X_i\}$ и получим:

$$\widehat{\alpha}=\frac{1}{\overline{X}-\min\{X_i\}}.$$

Условие $X_i \geq \widehat{\theta}$ оказывается выполненным автоматически.

Найдем реализации оценок максимального правдоподобия и построим графики реализаций параметрической оценки функции распределения $F(t,\widehat{\alpha},\widehat{\theta})$ (как для оценок методом моментов) и эмпирической функции распределения $F_n^*(t)$. График приведен на рисунке 10.

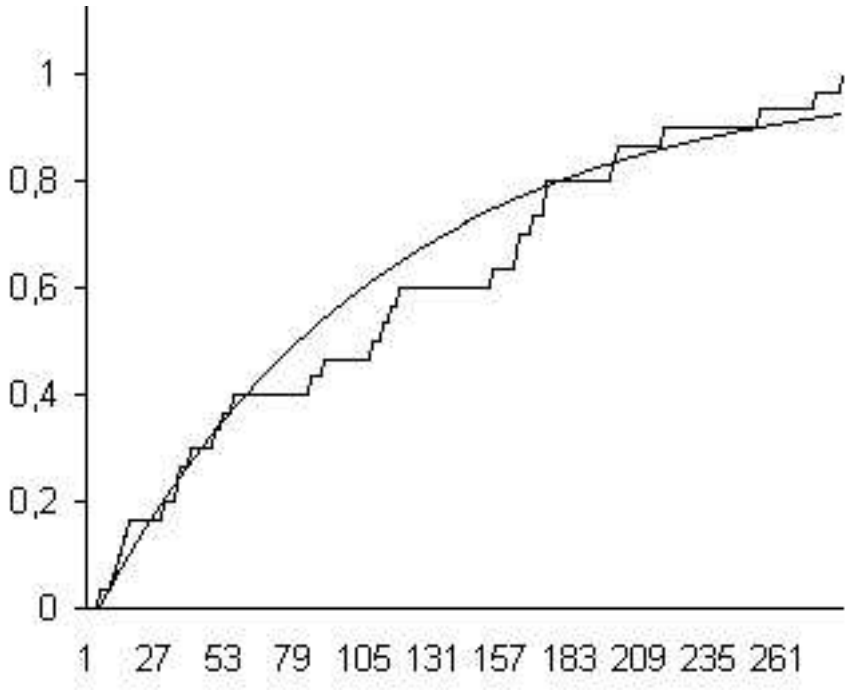

Рис. 10: Оценка функции сдвинутого показательного распределения методом максимального правдоподобия

Сравнивая результат с рис. 8, видим, что использование оценок максимального правдоподобия позволяет более точно приблизить эмпирическую функцию распределения.

Построим оценки параметров распределения Парето по первому и второму моментам. Вспомним, что плотность распределения задается формулой:

$$f_{\gamma,h}(t) = \begin{cases} \gamma h^{\gamma} t^{-(\gamma+1)}, & \text{если } t \geq h; \\ 0 & \text{иначе.} \end{cases}$$

Вычислим $\mathbf{E}X_i$ и $\mathbf{E}X_i^2$:

$$\mathbf{E}X_i = \int_{-\infty}^{\infty} t f_{\gamma,h}(t) dt = \int_{h}^{\infty} t\gamma h^{\gamma} t^{-(\gamma+1)} dt =$$

$$= \gamma h^\gamma \int_h^\infty t^{-\gamma} dt = \gamma h^\gamma \frac{t^{-\gamma+1}}{-\gamma+1}|_h^\infty = \frac{\gamma h^\gamma h^{-\gamma+1}}{\gamma-1} = \frac{\gamma h}{\gamma-1},$$

если $\gamma > 1$ (в противном случае математическое ожидание не существует). Аналогично:

$$\mathbf{E}X_i^2 = \int_{-\infty}^\infty t^2 f_{\gamma,h}(t) dt = \int_h^\infty t^2 \gamma h^\gamma t^{-(\gamma+1)} dt =$$

$$= \gamma h^\gamma \int_h^\infty t^{-\gamma+1} dt = \gamma h^\gamma \frac{t^{-\gamma+2}}{-\gamma+2}|_h^\infty = \frac{\gamma h^\gamma h^{-\gamma+2}}{\gamma-2} = \frac{\gamma h^2}{\gamma-2},$$

если $\gamma > 2$ (в противном случае второй момент не существует).

Получаем систему уравнений:

$$\begin{cases} \mathbf{E}X_i = \frac{\gamma h}{\gamma-1}, \\ \mathbf{E}X_i^2 = \frac{\gamma h^2}{\gamma-2}. \end{cases}$$

Выразим параметры:

$$h = \frac{\gamma-1}{\gamma}\mathbf{E}X_i;$$

$$\frac{\gamma(\gamma-1)^2}{\gamma^2(\gamma-2)}(\mathbf{E}X_i)^2 = \mathbf{E}X_i^2;$$

$$(\gamma-1)^2(\mathbf{E}X_i)^2 = \gamma(\gamma-2)\mathbf{E}X_i^2.$$

Получаем квадратное уравнение:

$$\mathbf{D}X_i\gamma^2 - 2\mathbf{D}X_i\gamma - (\mathbf{E}X_i)^2 = 0.$$

Решая его и выбирая положительный корень, получаем:

$$\gamma = 1 + \sqrt{1 + \frac{(\mathbf{E}X_i)^2}{\mathbf{D}X_i}}.$$

Заменим математическое ожидание и дисперсию на выборочное среднее $\overline{X}$ и выборочную дисперсию S^2, а параметры h и γ на их оценки h^* и γ^*. Получим оценки параметров:

$$\begin{cases} \gamma^* = 1 + \sqrt{1 + \frac{(\overline{X})^2}{S^2}}, \\ h^* = \frac{\gamma^* - 1}{\gamma^*}\overline{X}. \end{cases}$$

Отметим, что оценка параметра γ всегда не меньше числа 2, что соответствует требованию к параметру, обеспечивающему конечность второго момента. Найдем реализации оценок по выборке и построим графики реализаций параметрической оценки функции распределения $F(t, h^*, \gamma^*)$ по формуле:

$$=\text{ЕСЛИ}(\text{СТРОКА}() < R\$1; 0; 1 - (\text{СТРОКА}()/R\$1)\text{^}(-R\$2))$$

и эмпирической функции распределения $F_n^*(t)$. График приведен на рисунке 11.

Из рисунка видно, что приближение в этом случае оказывается очень неудачным.

Получим оценки параметров распределения Парето методом максимального правдоподобия.

Сначала найдем оценку параметра h непосредственно отысканием точки максимума функции правдопобобия. Для этого запишем функцию правдоподобия

$$\Pi(\vec{X}, h, \gamma) = \begin{cases} \prod_{i=1}^{n} (\gamma h^{\gamma} X_i^{-(\gamma+1)}), & \text{если все } X_i \geq \theta; \\ 0 & \text{иначе.} \end{cases}$$

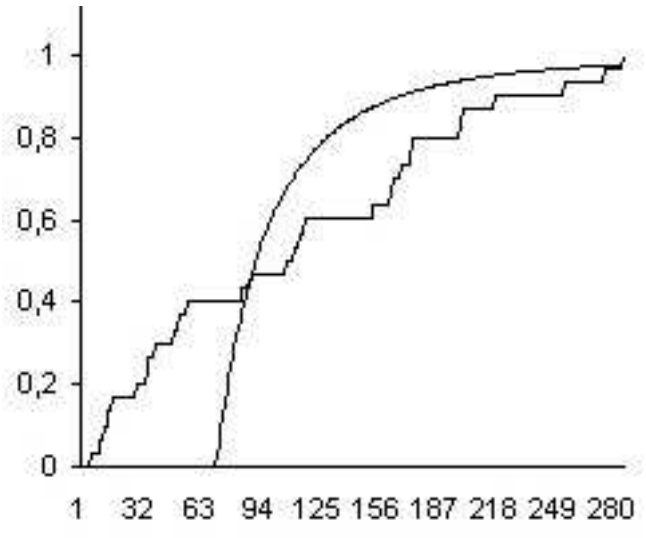

Рис. 11: Оценка функции распределения Парето
методом моментов

Зависимость функции правдоподобия от параметра h имеет тот же характер, что и в случае сдвинутого показательного распределения зависимость от параметра θ. Она изображена схематично на рисунке 9. Ее максимум достигается в точке $\widehat{h} = \min\{X_i\}$, которая является оценкой максимального правдоподобия параметра h.

Найдем оценку максимального правдоподобия параметра γ. Для этого последовательно вычислим:

$$\ln f(t, h, \gamma) = \ln \gamma + \gamma \ln h - (\gamma + 1) \ln t \text{ при } t \geq h;$$

$$\frac{\partial}{\partial \gamma} \ln f(t, h, \gamma) = \frac{1}{\gamma} + \ln h - \ln t \text{ при } t \geq h.$$

Приравнивая производную логарифма функции правдоподобия к нулю, получаем уравнение для определения оценки пара-

метра $\boldsymbol{\gamma}$:

$$\sum_{i=1}^{n}\left(\frac{1}{\boldsymbol{\gamma}}+\ln h-\ln X_i\right)=0,$$

решением которого является:

$$\boldsymbol{\gamma}=\frac{1}{\overline{\ln X}-\ln h}.$$

Поскольку параметр h неизвестен, заменим его на оценку максимального правдоподобия $\widehat{h}=\min\{X_i\}$ (так же мы поступали при нахождении оценок для сдвинутого показательного распределения) и получим:

$$\widehat{\boldsymbol{\gamma}}=\frac{1}{\overline{\ln X}-\ln(\min\{X_i\})}.$$

Условие $X_i\geq\widehat{h}$ оказывается выполненным автоматически.

Найдем реализации оценок максимального правдоподобия. Отметим, что для нахождения выборочного усреднения логарифма $\overline{\ln X}$ нужно предварительно в отдельном столбике вычислить логарифмы всех выборочных значений, и затем вычислить среднее из 30 значений логарифмов.

Построим графики реализаций параметрической оценки функции распределения $F(t,\widehat{h},\widehat{\boldsymbol{\gamma}})$ (как для оценок методом моментов) и эмпирической функции распределения $F_n^*(t)$.

График приведен на рис. 12.

Анализируя график, видим, что для распределения Парето оценки максимального правдоподобия также не дают хорошего приближения эмпирической функции распределения. На основании проведенного исследования можно заключить, что более адекватной моделью является модель сдвинутого показательного распределения, и лучший метод оценивания ее параметров — метод максимального правдоподобия. Оценки максимального правдоподобия здесь получаются смещенными, однако мы не будем обсуждать, как можно уменьшить смещение.

Распределения, связанные с нормальным

При построении доверительных интервалов для параметров нормального распределения мы будем использовать два специальных распределения, связанных с нормальным: распределение хи-квадрат и распределение Стьюдента. Название «распределение Стьюдента» связано с именем английского статистика К.Госсета, который подписывал свои работы псевдонимом «Стьюдент».

Случайная величина Z_n имеет *распределение хи-квадрат с n степенями свободы*, если

$$Z_n = {X_1}^2 + \ldots + {X_n}^2;$$

где $X_1, \ldots, X_n$ — независимые случайные величины со стандартным нормальным распределением.

Рис. 12: Оценка функции распределения Парето методом максимального правдоподобия

Отметим, что «число степеней свободы» — это просто традиционное название для параметра n распределения хи-квадрат.

Параметр n — положительное целое число. В частности, при $n = 1$ получаем квадрат одной случайной величины со стандартным нормальным распределением: $Z_1 = X^2$, где $X \in \Phi_{0,\,1}$.

Будем использовать следующее обозначение: $Z_n \in \chi^2_n$.

Отметим следующие свойства распределения хи-квадрат.

Пусть $Z_n \in \chi^2_n$. Тогда:

1) $\mathbf{E}Z_n = n$;

2) $Z_n/n \to 1$ с вероятностью единица при $n \to \infty$.

Случайная величина Y_n имеет *распределение Стьюдента с n степенями свободы*, если

$$Y_n = \frac{X}{\sqrt{Z_n/n}},$$

где случайные величины X и Z_n независимы, причем X имеет стандартное нормальное распределение, а Z_n имеет распределение хи-квадрат с n степенями свободы. Здесь, как и у распределения хи-квадрат, n — это просто положительный целый параметр.

Будем использовать следующее обозначение: $Y_n \in T_n$.

Отметим следующие свойства распределения Стьюдента.

Пусть $Y_n \in T_n$. Тогда:

1) для любого t выполнено $\mathbf{P}\{Y_n < -t\} = \mathbf{P}\{Y_n > t\}$, то есть распределение Стьюдента симметрично;

2) $Y_n \to X$ с вероятностью единица при $n \to \infty$, где X имеет стандартное нормальное распределение.

Точные доверительные интервалы

Наиболее распространенной ситуацией, когда возможно построение точных доверительных интервалов, является случай нормального распределения: $\vec{X} \in \Phi_{a,\sigma^2}$, когда хотя бы один из его параметров неизвестен. В этом случае известно совместное распределение наиболее употребительных оценок $\overline{X}$ и S^2 параметров a и σ^2, с помощью которого и строятся соответству-

ющие доверительные интервалы. Основные результаты содержатся в следующей теореме.

Теорема Фишера. Пусть $\vec{X} \Subset \Phi_{a,\sigma^2}$. Тогда верны следующие 4 факта.

1) $\dfrac{\sqrt{n}(\overline{X}-a)}{\sigma} \Subset \Phi_{0,1}$;

2) $\dfrac{\sum_{i=1}^{n}(X_i-a)^2}{\sigma^2} \Subset \chi^2_n$;

3) $\dfrac{nS^2}{\sigma^2} \Subset \chi^2_{n-1}$;

4) $\dfrac{\sqrt{n-1}\,(\overline{X}-a)}{S} \Subset T_{n-1}$.

Решение типовых примеров

Пример 13.2. Пусть $\vec{X} \Subset \Phi_{a,\sigma^2}$, $a \in R$. Построить доверительный интервал $(a_-;\ a_+)$ для параметра a, считая σ^2 известным. Вычислить реализацию доверительного интервала с уровнем доверия $\gamma = 0,95$, располагая данными: $n = 10,\ \overline{X} = 2,7,\ \sigma^2 = 4$.

Решение. Для построения доверительного интервала используем оценку $\overline{X}$, распределение которой известно. Для заданной доверительной вероятности γ найдем такое $A > 0$, что

$$\gamma = \mathbf{P}\left\{\left|\sqrt{n}\frac{\overline{X}-a}{\sigma}\right| < A\right\} = \mathbf{P}\left\{-\frac{\sigma A}{\sqrt{n}} < \overline{X}-a < \frac{\sigma A}{\sqrt{n}}\right\}. \quad (12)$$

Таким образом, нужно искать $\varepsilon_1 = -\frac{\sigma A}{\sqrt{n}}$, $\varepsilon_2 = \frac{\sigma A}{\sqrt{n}}$ такие, что выполняется равенство:

$$\mathbf{P}\left\{\varepsilon_1 < \overline{X}-a < \varepsilon_2\right\} = \gamma.$$

Для этого вернемся к (12). В силу теоремы Фишера, случайная величина, стоящая под знаком модуля, имеет стандартное

нормальное распределение, поэтому вероятность в правой части можно выразить через функцию распределения $\Phi(t)$ стандартного нормального закона, и тогда уравнение (12) приобретает вид:

$$2\Phi(A) - 1 = \gamma \iff \Phi(A) = \frac{1+\gamma}{2},$$

где $\Phi(t)$ — функция Лапласа, значения которой представлены в таблице приложения в конце книги.

Найдя значение A по таблице и подставив в (12), получим равенство:

$$\gamma = \mathbf{P}\left\{\left|\sqrt{n}\frac{\overline{X}-a}{\sigma}\right| < A\right\} = \mathbf{P}\left\{-A < \sqrt{n}\frac{a-\overline{X}}{\sigma} < A\right\} \iff$$

$$\iff \gamma = \mathbf{P}\left\{\overline{X} - \sigma\frac{A}{\sqrt{n}} < a < \overline{X} + \sigma\frac{A}{\sqrt{n}}\right\},$$

откуда искомый γ-доверительный интервал:

$$(a_-;\ a_+) = \left(\overline{X} - \sigma\frac{A}{\sqrt{n}},\ \overline{X} + \sigma\frac{A}{\sqrt{n}}\right).$$

Подставляя сюда конкретные данные из условия, вычисляем реализацию доверительного интервала:

$$(a_-;\ a_+) \approx \left(2{,}7 - 2\frac{1{,}96}{\sqrt{10}},\ 2{,}7 + 2\frac{1{,}96}{\sqrt{10}}\right) \approx (1{,}46; 3{,}94) \iff$$

$$\iff \mathbf{P}(1{,}46 < \theta < 3{,}94) = 0{,}95.$$

Замечание. Построенный доверительный интервал оказывается симметричным относительно выборочного среднего $\overline{X}$ и имеет длину $2A\frac{\sigma}{\sqrt{n}}$, пропорциональную значению A, которое было найдено из условия (12).

Пример 13.3. Пусть $\vec{X} \Subset \Phi_{a,\sigma^2}$, где $a \in \mathbf{R}$, $\sigma^2 > 0$ — два неизвестных параметра. Построить доверительный интервал для параметра σ^2. Вычислить реализацию доверительного

интервала с уровнем доверия $\gamma = 0,9$, располагая данными: $n = 10;\ S^2 = 4$.

Решение. Чтобы построить двусторонний доверительный интервал, используем следующее уравнение:

$$\gamma = \mathbf{P}\left\{x_1 < \frac{nS^2}{\sigma^2} < x_2\right\} \iff \mathbf{P}\left\{\frac{nS^2}{x_2} < \sigma^2 < \frac{nS^2}{x_1}\right\} = \gamma, \quad (13)$$

где $0 < x_1 < x_2$, удовлетворяющие (13), находим по известному распределению χ^2_{n-1} случайной величины $\frac{nS^2}{\sigma^2}$.

В общем случае эта задача не имеет единственного решения. Если обратиться к графику плотности распределения χ^2_{n-1}, представленному на рис. 13, то $x_1 < x_2$ следует выбирать таким образом, чтобы сумма вероятностей, представленных площадями заштрихованных областей под графиком плотности, равнялась $1 - \gamma$. Ясно, что это можно сделать многими способами. Чтобы сделать решение однозначным, выберем $x_1 < x_2$ так, чтобы каждая из заштрихованных площадей равнялась $(1-\gamma)/2$, тогда нетрудно видеть, что x_1, x_2 выражаются через квантили распределения χ^2_{n-1}:

$$x_1 = \chi^2_{(1-\gamma)/2,n-1}, \quad x_2 = \chi^2_{(1+\gamma)/2,n-1}.$$

Подставляя эти значения, находим искомый двусторонний доверительный интервал:

$$\mathbf{P}_\theta\left\{\frac{nS^2}{\chi^2_{(1+\gamma)/2,n-1}} < \sigma^2 < \frac{nS^2}{\chi^2_{(1-\gamma)/2,n-1}}\right\} = \gamma \iff$$

$$\iff (\sigma^2_-;\ \sigma^2_+) = \left(\frac{nS^2}{\chi^2_{\frac{1+\gamma}{2},n-1}},\ \frac{nS^2}{\chi^2_{\frac{1-\gamma}{2},n-1}}\right).$$

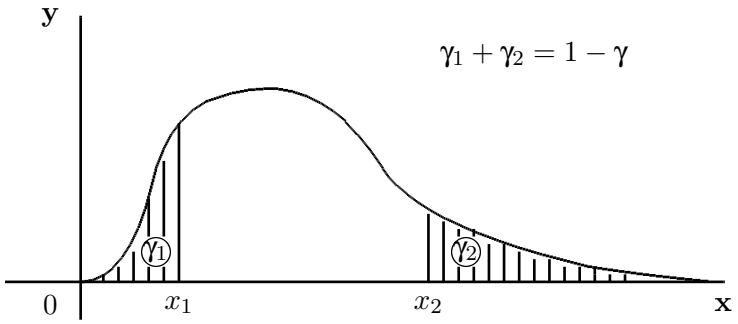

Рис. 13: Плотность распределения χ^2_{n-1}

Находя с помощью программных приложений значения $\chi^2_{(1-\gamma)/2,\ n-1}$, $\chi^2_{(1+\gamma)/2,\ n-1}$, используя распределение хи-квадрат с $n-1$ степенью свободы для конкретных значений $\gamma = 0,9$; $n = 10$, и данных в условии численных значений, находим реализацию доверительного интервала:

$$\chi^2_{0,95,\ 9} = 16,9;\ \chi^2_{0,05,\ 9} \approx 3,325;$$

$$(\sigma^2_-,\ \sigma^2_+) = (2,37;\ 12,03).$$

Задачи для решения в классе

13.1. В тексте задачи № обозначает номер студента по списку группы.

1) Для выборки $X_1, \ldots, X_n$ из равномерного распределения на $[0;\ \theta]$ получить оценки параметра θ методом моментов на основании первого, второго, №+2-го момента. Вычислить $E(X_1 + №)e^{X_1/№}$ и на этом основании получить оценку параметра θ через усреднение соответствующей функции по выборке.

2) Для той же выборки найти оценку максимального правдоподобия параметра θ, вычислить ее математическое ожидание и исправить ее, получив несмещенную оценку.

3) Генерировать реализацию выборки объема $n = 100+№$ из равномерного распределения на $[0;\ \theta]$, приняв $\theta = №$.

4) Вычислить реализации всех полученных оценок. Подсчитать абсолютные погрешности оценивания и ранжировать оценки по абсолютной погрешности.

13.2. Случайная величина имеет логарифмически нормальное распределение, если ее логарифм распределен по нормальному закону.

1) Для выборки $X_1,\ \ldots,\ X_n$ из логарифмически нормального распределения с параметрами $a,\ \sigma$ получить оценки параметров методом моментов и методом максимального правдоподобия.

2) Генерировать реализацию выборки объема $n = 100$ по формуле $U_1U_2U_3/U_4$, где $U_1, \ldots, U_4$ — случайные числа, равномерно распределенные на $[0;\ 1]$. Построить гистограмму, выбрав число промежутков группирования по формуле Стеджеса.

3) По реализации выборки вычислить реализации всех полученных оценок.

4) Найти теоретические значения параметров. Подсчитать абсолютные погрешности оценивания и ранжировать оценки по абсолютной погрешности.

13.3. Пусть $\vec{X} \Subset \Phi_{\theta_1,\theta_2},\ \ \theta_1 \in \mathbf{R},\ \theta_2 > 0$. Построить центральный доверительный интервал для параметра θ_1. Вычислить реализацию доверительного интервала с уровнем $\gamma = 0,95$, располагая данными: $n = 10;\ \overline{X} = 2,7;\ S^2 = 4$.

13.4. Пусть $\vec{X} \Subset \Phi_{a,\theta},\ \ \theta > 0$, a — известно. Построить точный доверительный интервал для параметра θ на основе статистики $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$. Вычислить реализацию построенного интервала с уровнем $\gamma = 0,9$, располагая данными: $n = 10;\ S_1^2 = 4$.

§14. Критерии согласия

Удобно представлять статистический критерий как функцию $\delta(\vec{X})$ от выборочного вектора, принимающую два значения: H_0 и H_1. Наиболее общий подход для построения статистических критериев состоит в следующем.

Пусть $T = T(\vec{X})$ — некоторая статистика, характеризующая отклонение эмпирических данных, представленных выборкой, от теоретических, соответствующих проверяемой гипотезе H_0. Если распределение статистики $T(\vec{X})$ известно (точно или хотя бы приближенно), то для любого $\alpha > 0$ можно найти такое множество T_α значений T, для которого будет выполнено неравенство

$$\mathbf{P}(T \in \mathrm{T}_\alpha / \mathrm{H}_0) \leq \alpha.$$

Пусть $\alpha > 0$ настолько мало, что событие, имеющее вероятность, не превосходящую α, может считаться практически невозможным. Тогда статистический критерий можно задать следующим образом:

$$\delta(\vec{X}) = \begin{cases} \mathrm{H}_1, & \text{если } T(\vec{X}) \in \mathrm{T}_\alpha; \\ \mathrm{H}_0, & \text{если } T(\vec{X}) \notin \mathrm{T}_\alpha. \end{cases}$$

Это правило основано на здравом смысле; оно предписывает отвергнуть гипотезу H_0 (то есть принять H_1), если происходит событие $\{T(\vec{X}) \in \mathrm{T}_\alpha\}$, которое не должно произойти, будь гипотеза H_0 справедлива. Число $\alpha > 0$, которое фигурирует в формулах, называется *уровнем критерия*, или *уровнем значимости*, статистика $T(\vec{X})$ называется *статистикой критерия*, а множество T_α — *критическим множеством*.

Достигаемый уровень значимости

От статистики $T = T(\vec{X})$ требуют следующих свойств:

1) при выполнении гипотезы H_0 статистика T имеет известное распределение или, по крайней мере, сходится по распреде-

лению к некоторой случайной величине J с известным распределением;

2) при выполнении гипотезы H_1 статистика T сходится почти наверное к бесконечности с ростом объема выборки.

Для того, чтобы получить критерий уровня α, задают критическое множество в виде

$$\mathrm{T}_\alpha = \{T \geq C\},$$

где C — константа, определяемая условием

$$\mathbf{P}\{J \geq C\} = \alpha,$$

то есть $F_J(C) = 1 - \alpha$.

Ясно, что при таком выборе константы C вероятность ошибки нулевого рода α_0 либо равна уровню критерия α (в случае, когда статистика T при верной нулевой гипотезе распределена в точности как J), либо, по крайней мере, сходится к α с ростом объема выборки.

Сходимость статистики T почти наверное к бесконечности при выполненной первой гипотезе гарантирует *состоятельность* критерия, то есть сходимость вероятности ошибки первого рода α_1 к нулю с ростом объема выборки.

Для каждой конкретной выборки $\vec{X}$ можно найти предельное значение уровня $\alpha^* = \alpha^*(\vec{X})$, при котором гипотеза H_0 еще может быть принята. Такое значение называется *реально достигаемым уровнем значимости*, или просто *достигаемым уровнем значимости*. Достигаемый уровень значимости α^* имеет смысл вероятности получить худшее согласие с проверяемой гипотезой, чем реально полученное, если гипотеза H_0 верна. Поэтому чем меньше α^*, тем более это говорит против гипотезы H_0.

Достигаемый уровень значимости вычисляется с помощью распределения статистики J:

$$\alpha^* = \mathbf{P}\{J \geq T(\vec{X})\} = 1 - F_J(T(\vec{X})).$$

В терминах достигаемого уровня значимости критическая область имеет вид

$$\mathrm{T}_\alpha = \{\alpha^* \leq \alpha\},$$

то есть нулевая гипотеза отвергается на уровне α в случае, когда $\alpha^* \leq \alpha$.

Каждый критерий согласия использует свою статистику, предназначенную для различения нулевой гипотезы и альтернативы и обладающую нужными свойствами: сходимостью к фиксированному распределению при выполнении нулевой гипотезы и сходимостью почти наверное к бесконечности при ее невыполнении.

В качестве важных примеров критериев согласия рассмотрим критерии Колмогорова и хи-квадрат Пирсона.

Критерии согласия Колмогорова и χ^2 Пирсона

Рассмотрим выборку $\vec{X} \Subset \mathrm{F}$ объема n с неизвестной функцией распределения F и простую гипотезу $\mathrm{H}_0 : \mathrm{F} = \mathrm{F}_0$. Альтернативной для H_0 является сложная гипотеза $\mathrm{H}_1 : \mathrm{F} \neq \mathrm{F}_0$.

Критерий Колмогорова применяется в случае, когда функция распределения $\mathrm{F}_0(t)$ непрерывна. Рассматривается следующее расстояние между эмпирической и теоретической функциями распределения:

$$D_n = D(F_n^*, F_0) = \sup_{-\infty<t<\infty} |F_n^*(t) - F_0(t)| = \max_{-\infty<t<\infty} |F_n^*(t) - F_0(t)|.$$

В качестве статистики критерия Колмогорова выбирается это расстояние, умноженное на $\sqrt{n}$, где n — объем выборки:

$$T_n = \sqrt{n} D_n = \sqrt{n} \max_{-\infty<t<\infty} |F_n^*(t) - F_0(t)|.$$

А. Н. Колмогоров доказал следующие свойства статистики T_n:

1) если гипотеза H_0 верна, то T_n с ростом n сходится к случайной величине J с функцией распределения, называемой функцией распределения Колмогорова:

$$F_J(t) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2t^2};$$

2) если гипотеза H_0 неверна, то T_n сходится почти наверное к $+\infty$ при $n \to \infty$. Таким образом, достигаемый уровень значимости критерия Колмогорова равен:

$$\alpha^* = 1 - F_J(T_n) = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2T_n^2} = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2}. \tag{14}$$

Отметим, что для расчетов по этой формуле нужно брать не всю бесконечную сумму, а только несколько слагаемых, при этом ошибка вычислений не превосходит последнего отброшенного слагаемого. Критерий Колмогорова отвергает гипотезу H_0 на уровне α, если $\alpha^* \leq \alpha$.

Для практического вычисления статистики D_n можно использовать следующую формулу:

$$D_n = \max_{1\leq i\leq n} \max\left(\left|F(X_{(i)}) - \frac{i}{n}\right|;\ \left|F(X_{(i)}) - \frac{i-1}{n}\right|\right).$$

Здесь $X_{(i)}$ — это элементы *вариационного ряда*, то есть для этих вычислений выборку следует предварительно *упорядочить по возрастанию*.

Если гипотетическая функция распределения $\mathrm{F}_0(x)$ не является непрерывной, то критерий Колмогорова неприменим. В этом случае можно воспользоваться χ^2—*критерием Пирсона*. Статистика критерия Пирсона строится после предварительного «группирования» выборочных данных. Для этого все множество S возможных значений случайных величин X_i разбивается на конечное число непересекающихся частей:

$$S = S_1 \cup S_2 \cup \cdots \cup S_r, \quad S_i \cap S_j = \emptyset, i \neq j.$$

Обозначим ν_j — число элементов выборки $\vec{X}$, попавших в множество S_j, а p_j — вероятность попадания случайной величины X_i в множество S_j, вычисленная с помощью гипотетической функции распределения $\mathrm{F} = \mathrm{F}_0$. Тогда в качестве статистики критерия χ^2 рассматривают следующую предложенную Пирсоном меру отклонения эмпирического распределения от предполагаемого теоретического:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j}.$$

Справедлива следующая теорема, позволяющая находить распределение статистики χ^2 при больших значениях n, а стало быть, и строить статистический критерий.

Если гипотеза H_0 однозначно фиксирует вероятности $p_1,\ p_2,\ \ldots,\ p_r$, где $p_j = \mathbf{P}(X_i \in S_j)$, то при выполнении этой гипотезы статистика $\chi^2(\vec{X})$ слабо сходится к распределению χ^2_{r-1}:

$$\chi^2 \Longrightarrow \chi^2_{r-1}, \quad n \to \infty.$$

При невыполнении нулевой гипотезы статистика $\chi^2(\vec{X})$ сходится почти наверное к $+\infty$.

Для построения критерия, основанного на статистике χ^2, используем распределение χ^2_{r-1}, и по найденному значению $\chi^2(\vec{X})$ отыскиваем достигаемый уровень значимости:

$$\alpha^* = 1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X}))$$

по таблице 5 распределения хи-квадрат или с помощью математических пакетов. В пакете Microsoft Excel достигаемый уровень значимости вычисляется формулой

=ХИ2РАСП(ячейка; r-1)

— в качестве ячейки надо подставить адрес ячейки, в которой вычислена статистика хи-квадрат, а $r-1$ — число степеней свободы.

Тогда критерий Пирсона имеет следующий вид:

$$\mathrm{H}_0 \Leftrightarrow \ \alpha^* > \alpha.$$

Заметим, что для практического применения рекомендуется разбиение производить таким образом, чтобы выполнялось условие $np_j \geq 10$. При нарушении этого условия нужно объединить соседние множества S_j. Вероятности p_j надо выбирать по возможности равными.

Критерий хи-квадрат часто используют для проверки сложных гипотез о принадлежности распределения к некоторому параметрическому семейству (например, к нормальному). При этом вместо известных вероятностей p_j подставляют их оценки p_j^*, полученные путем оценивания неизвестных парметров распределения. Важно понимать, что в этом случае предельное распределение статистики $\chi^2(\vec{X})$ уже не будет распределением χ^2_{r-1}, а будет близко к распределению χ^2_{r-1-s}, где s — число оцениваемых параметров ($s = 2$ для нормального распределения). Более точно, предельная функция распределения заключена между функциями распределения χ^2_{r-1-s} и χ^2_{r-1}.

Достигаемый уровень значимости α^* удовлетворяет неравенству:

$$1 - F_{\chi^2_{r-1-s}}(\chi^2(\vec{X})) \leq \alpha^* \leq 1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X})),$$

где s — число оцениваемых параметров.

Для того, чтобы получить в точности распределение хи-квадрат с $r-1-s$ степенями свободы, следует оценивать неизвестные параметры методом максимального правдоподобия по *группированной* выборке, но это приводит, как правило, к сложным вычислительным процедурам.

Решение типовых примеров

Пример 14.1. Вариационный ряд выборки имеет вид (1; 2; 3; 4; 5; 6; 7; 8; 9; 10). Проверить гипотезу о равномерности распределения элементов выборки на отрезке от 0 до 10

с помощью критерия Колмогорова: найти реализацию достигаемого уровня значимости и сделать вывод о принятии гипотезы на уровнях 0,1 и 0,01.

Решение. Построим на одном графике эмпирическую $F_n^*(t)$ и теоретическую $F_0(t)$ функции распределения.

Эмпирическая функция распределения — это ступенчатая функция, высота ступеньки равна $1/10$ в точках $1;\ \ldots;\ 10$.

Теоретическая функция распределения равномерного закона на отрезке от 0 до 10 равна:

$$\mathrm{F}_0(t) = \begin{cases} 0, & \text{если } t \le 0; \\ t/10, & \text{если } 0 < t \le 10; \\ 1, & \text{если } t > 10. \end{cases}$$

Так как функция распределения $\mathrm{F}_0(t)$ непрерывна, то можно применять критерий Колмогорова. Найдем по графику значение D_n — наибольшую по модулю разность между эмпирической и теоретической функциями распределения. Эта разность достигается в точках разрыва эмпирической функции распределения и равна $1/10$. Вычислим реализацию достигаемого уровня значимости, вспоминая, что $n = 10$:

$$\alpha^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2} \approx$$

$$\approx 2e^{-0,2} - 2e^{-4\cdot 0,2} + 2e^{-9\cdot 0,2} - 2e^{-16\cdot 0,2} + 2e^{-25\cdot 0,2} - 2e^{-36\cdot 0,2} + 2e^{-49\cdot 0,2} \approx$$

$$\approx 0,99997.$$

Достигаемый уровень значимости оказался близким к 1; это означает, что нет оснований отвергать гипотезу о равномерности выборочных значений. Эту гипотезу следовало бы отвергнуть только в случае, когда достигаемый уровень значимости оказался бы близким к нулю.

В частности, в нашем случае выполнено неравенство $\alpha^* > 0,1$. Следовательно, гипотеза о равномерности принимается на уровне 0,1. Тем более она будет приниматься на уровне 0,01.

Пример 14.2. Решить пример 14.1 для реализации выборки (10; 0; 0; 10; 10; 10; 0; 0; 0; 10).

Решение. Упорядочив реализацию выборки по неубыванию, получим реализацию вариационного ряда: (0; 0; 0; 0; 0; 10; 10; 10; 10; 10). Как и в предыдущем примере, построим на одном графике эмпирическую $F_n^*(t)$ и теоретическую $F_0(t)$ функции распределения. В отличие от предыдущего примера, эмпирическая функция распределения здесь имеет всего две ступеньки в точках 0 и 10, высотой по $5/10 = 0,5$. Теоретическая функция распределения остается той же самой. Значение D_n достигается в точках разрыва эмпирической функции распределения и равняется $0,5$. Вычислим реализацию достигаемого уровня значимости:

$$\alpha^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2} \approx 2e^{-2\cdot 10\cdot 0,5^2} = 2e^{-5} \approx 0,0135.$$

Здесь мы взяли только одно слагаемое суммы, так как остальные слагаемые гораздо меньше.

В этом примере достигаемый уровень значимости оказался близким к 0, что говорит против гипотезы H_0. В частности, $\alpha^* < 0,1$, то есть гипотеза однородности отвергается на уровне 0,1. Однако она принимается на более низком уровне 0,01, так как $\alpha^* > 0,01$.

Пример 14.3. Проверить гипотезу о равномерности на отрезке от 0 до 10 для выборок из двух предыдущих примеров с помощью критерия хи-квадрат Пирсона: найти реализации достигаемых уровней значимости и сделать выводы о принятии гипотезы на уровнях 0,1 и 0,01. Число промежутков группирования выбрать по формуле Стеджеса.

Решение.

Согласно формуле Стеджеса, вычисляем целую часть логарифма по основанию 2 от объема выборки и прибавляем 1:

$$r = [\log_2 n] + 1 = [\log_2 10] + 1 = 3 + 1 = 4;$$

т. к. $2^3 = 8 < 10 < 2^4 = 16$.

Итак, множество допустимых выборочных значений — отрезок $[0;\ 10]$ — следует разбить на 4 промежутка равной длины:

$$S_1 = [0;\ 2,5);\quad S_2 = [2,5;\ 5);\quad S_3 = [5;\ 7,5);\quad S_4 = [7,5;\ 10].$$

Согласно нулевой гипотезе, распределение равномерное на отрезке от 0 до 10. Следовательно, равны вероятности попадания элемента выборки в отрезки равной длины:

$$p_1 = p_2 = p_3 = p_4 = 1/4 = 0,25.$$

Значения статистики хи-квадрат Пирсона различны для примеров 14.1 и 14.2:

1) В примере 14.1 количества элементов, попавших в каждый из промежутков, равны соответственно

$$\nu_1 = 2;\quad \nu_2 = 2;\quad \nu_3 = 3;\quad \nu_4 = 3.$$

Вычислим статистику хи-квадрат:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} =$$

$$= \frac{(2 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(2 - 10 \cdot 0,25)^2}{10 \cdot 0,25} +$$

$$+ \frac{(3 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(3 - 10 \cdot 0,25)^2}{10 \cdot 0,25} = 0,4.$$

Найдем достигнутый уровень значимости, используя функцию ХИ2РАСП и подставляя значение 0,4 и число степеней свободы, равное $r - 1 = 4 - 1 = 3$:

$$\text{ХИ2РАСП}(0{,}4;3) \approx 0,94.$$

Итак, здесь достигнут уровень значимости 0,94, что не дает оснований отвергать гипотезу о равномерности ни на уровне $0,1 < 0,94$, ни тем более на уровне 0,01.

2) В примере 14.2 количества элементов, попавших в каждый из промежутков, принимают значения:

$$\nu_1 = 5; \quad \nu_2 = 0; \quad \nu_3 = 0; \quad \nu_4 = 5.$$

Как и в пункте (1), вычислим статистику хи-квадрат и найдем достигнутый уровень значимости:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} = \frac{(5 - 10 \cdot 0,25)^2}{10 \cdot 0,25} +$$

$$+ \frac{(0 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(5 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(0 - 10 \cdot 0,25)^2}{10 \cdot 0,25} = 10;$$

$$\text{ХИ2РАСП}(10;3) \approx 0,0186.$$

В этом примере достигнут низкий уровень значимости 0,0186, что дает основания отвергать гипотезу о равномерности на уровне $0,1 > 0,0186$, но не на уровне 0,01.

Отметим, что для рассмотренных примеров критерии Колмогорова и хи-квадрат Пирсона дают похожие результаты — достигнутые уровни значимости для обоих критериев оказались довольно близкими. В случае, когда основная гипотеза предполагает дискретное распределение, критерий Колмогорова неприменим, и мы будем пользоваться только критерием хи-квадрат Пирсона.

Пример 14.4. При 4040 бросаниях монеты Бюффон получил $\nu_1 = 2048$ выпадений герба и $\nu_2 = n - \nu_1 = 1992$ выпадений решетки. Согласуется ли это с гипотезой о том, что монета правильная, при уровне значимости $\alpha = 0,1$? С каким предельным уровнем значимости может быть принята эта гипотеза?

Решение. Можно считать, что мы имеем дело со статистической моделью $\vec{X} \in \mathrm{B}_p$, где неизвестен параметр p — вероятность выпадения герба. Проверяемая гипотеза $\mathrm{H}_0 : p = 0,5$. Поскольку выборочные данные уже сгруппированы ($\nu_1 = 2048$ —

число значений $X_i = 1$, ν_2 — число значений $X_i = 0$), то можем вычислить наблюдаемое значение статистики χ^2:

$$p_1 = \mathbf{P}_{H_0}(X_i = 1) = 0,5; \quad p_2 = \mathbf{P}_{H_0}(X_i = 0) = 0,5;$$

$$\frac{(\nu_1 - np_1)^2}{np_1} = \frac{(2048 - 2020)^2}{2020} = 0,285;$$

$$\frac{(\nu_2 - np_2)^2}{np_2} = \frac{(1992 - 2020)^2}{2020} = 0,388; \quad \chi^2 = 0,285{+}0,388 = 0,673.$$

Число множеств разбиения $r = 2$, поэтому достигнутый уровень значимости

$$\text{ХИ2РАСП}(0{,}673;1) \approx 0,412.$$

Достигнутый уровень значимости довольно высок. В частности, $0,412 > 0,1$, то есть гипотеза о симметричности монеты принимается на уровне 0,1.

Задачи для решения в классе

14.1. При $n = 4000$ независимых испытаний события $A1$, $A2$, $A3$, составляющие полную группу, осуществились соответственно 1905, 1015 и 1080 раз. Проверить, согласуются ли эти данные при уровне значимости 0,05 с гипотезой H_0: $p_1 = 1/2$, $p_2 = p_3 = 1/4$, где $p_j = \mathbf{P}(A_j)$. Найти достигнутый уровень значимости.

14.2. В экспериментах с селекцией гороха Мендель наблюдал частоты различных видов семян, полученных при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей по теории наследственности приведены в следующей таблице:

Семена	Частота	Вероятность
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16
Σ	n=556	1

Следует проверить гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями (на уровне значимости 0,1) и найти достигнутый уровень значимости.

14.3. В таблице приведены числа m_i участков равной площади 0,25 км^2 южной части Лондона, на каждый из которых приходилось по i попаданий самолетов-снарядов во время второй мировой войны. Проверить согласие опытных данных с законом распределения Пуассона, приняв за уровень значимости $\alpha = 0,05$:

i	0	1	2	3	4	5 и более	Итого
m_i	229	211	93	35	7	1	$\Sigma m_i = 576$

Таблица нормального распределения

Значения функции $\Phi(t) = \frac{1}{\sqrt{2\pi}} \cdot \int\limits_{-\infty}^{t} e^{-\frac{u^2}{2}} du$ и функции $\overline{\Phi}(t) = \Phi(-t) = 1 - \Phi(t)$.

t	$\Phi(-t)$	$\Phi(t)$
4,75	0,000001	0,999999
4,26	0,00001	0,99999
3,72	0,0001	0,9999
3,09	0,001	0,999
2,58	0,005	0,995
2,33	0,01	0,99
2,05	0,02	0,98
1,96	0,025	0,975
1,88	0,03	0,97
1,75	0,04	0,96
1,64	0,05	0,95
1,28	0,1	0,9
1,04	0,15	0,85
0,84	0,2	0,8
0,67	0,25	0,75
0,52	0,3	0,7
0,39	0,35	0,65
0,25	0,4	0,6
0,13	0,45	0,55
0,00	0,5	0,5

Для $|t| > 4,75$ можно использовать аппроксимацию $\overline{\Phi}(t) \sim \frac{e^{-t^2/2}}{t\sqrt{2\pi}}$.

Список литературы

1. *Боровков А.А.* Теория вероятностей. — М.: Эдиториал УРСС, 1999. — 470 с.

2. *Боровков А.А.* Математическая статистика. — Новосибирск: Наука, 1997. — 772 с.

3. *Бородин А.Н.* Элементарный курс теории вероятностей и математической статистики. — СПб., 1999. — 223 с.

4. *Бородихин В.М.* Теория вероятностей и математическая статистика: Практикум. — Новосибирск, 2000. — Ч. 1. — 159 с.

5. *Бородихин В.М.* Теория вероятностей и математическая статистика: Практикум. — Новосибирск, 2001. — Ч. 2. — 105 с.

6. *Бородихин В.М., Ковалевский А.П.* Высшая математика. — Т. 4.2: Теория вероятностей и математическая статистика. — Новосибирск: НГТУ, 2005. — 256 с.

7. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. — М.: Высшая школа, 1984. — 248 с.

8. *Ивченко Г.И., Медведев Ю.И., Чистяков А.В.* Сборник задач по математической статистике. — М.: Высшая школа, 1989. — 255 с.

9. *Коршунов Д.А., Фосс С.Г., Эйсымонт И.М.* Сборник задач и упражнений по теории вероятностей. — СПб., 2004. — 192 с.

10. *Коршунов Д.А., Чернова Н.И.* Сборник задач и упражнений по математической статистике. — Новосибирск, 2001. — 120 с.

11. *Лотов В.И.* Теория вероятностей и математическая статистика. — Новосибирск: НГУ, 2006. — 128 с.

12. *Свешников А.А. и др.* Сборник задач по теории вероятностей, математической статистике и теории случайных функций. — М., 1970. — 656 с.

13. *Чистяков В.П.* Курс теории вероятностей. — М.: Наука, 1987. — 240 с.

14. *Чернова Н.И.* Теория вероятностей. — Новосибирск: НГУ, 2007. — 160 с.

15. *Чернова Н.И.* Математическая статистика. — Новосибирск: НГУ, 2007. — 148 с.

Источники Интернет

1. *Лотов В.И.* Лекции по теории вероятностей и математической статистике.
http://www.nsu.ru/mmf/tvims/lotov/tv&ms_ff.pdf

2. *Коршунов Д.А., Фосс С.Г.* Сборник задач и упражнений по теории вероятностей.
http://www.math.nsc.ru/LBRT/v1/dima/ExerciseProbability2.pdf

3. *Коршунов Д.А., Чернова Н.И.* Сборник задач и упражнений по математической статистике.
http://www.math.nsc.ru/LBRT/v1/dima/ExerciseStatistics2.pdf

4. *Чернова Н.И.* Лекции по теории вероятностей.
http://www.nsu.ru/mmf/tvims/chernova/tv/index.html

5. *Чернова Н.И.* Лекции по математической статистике.
http://www.nsu.ru/mmf/tvims/chernova/ms/lec/ms.html

Учебное издание

Быстров Александр Александрович
Ковалевский Артем Павлович
Лотов Владимир Иванович

ПРАКТИКУМ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ

Учебное пособие

Редактор Е. В. Дубовцева

Подписано в печать 3.02.2009 г.
Формат 60 × 84 1/16. Офсетная печать.
Уч.-изд. л. 7,5 Усл.-печ. л. 7,0 Тираж 220 экз.

Заказ №36

Редакционно-издательский центр НГУ.
630090, Новосибирск-90, ул. Пирогова, 2.