Елена Анатольевна Лукьянова, выпускница факультета вычислительной математики и кибернетики МГУ им. Ломоносова (1997 г., прикладная математика), старший преподаватель кафедры медицинской информатики, ведет курсы медицинской статистики и медицинской информатики со студентами медицинского факультета РУДН.

Учебные программы по **медицинской информатике и статистике** доступны на сайте медицинского факультета *http://med.pfu.edu.ru*

Е.А. Лукьянова

МЕДИЦИНСКАЯ СТАТИСТИКА

Издательство Российского университета дружбы народов
Москва 2002

Е.А. Лукьянова

МЕДИЦИНСКАЯ СТАТИСТИКА

Учебное пособие

Москва
Издательство Российского университета дружбы народов
2002

ББК 5
Л 84

Утверждено
РИС Ученого совета
Российского университета
дружбы народов

Рецензент —
кандидат технических наук, заведующий лабораторией биостатистики
Гематологического научного центра РАМН *С.М. Куликов*

Л 84 **Лукьянова Е.А.**
Медицинская статистика: Учеб. пособие. — М.: Изд-во РУДН, 2002. — 255 с.: ил.

ISBN 5–209–01344–8

В настоящее время потребность в знании статистики и методов обработки данных клинического и экспериментального исследований очевидна как для студентов, так и преподавателей медицинских учебных заведений. Диагностика и прогнозирование должны быть основаны на доказанном. Получение доказательства при исследовании случайных процессов, протекающих в популяциях и сообществах, может быть получено лишь при правильном использовании математического аппарата. В данном пособии изложены основные виды статистического анализа, предусмотренные учебным планом по медицинской статистике.

Пособие адаптировано к нетехнической аудитории и предназначено для студентов, стажеров, ординаторов и преподавателей медицинских вузов.

ISBN 5–209–01344–8 ББК 5

ПРЕДИСЛОВИЕ

Назначение данного учебного пособия – дать в руки медику-исследователю инструмент для объективного познания внутреннего мира человека и явлений внешней среды, в которой человек обитает. И сам человек, и среда его обитания пронизаны множеством случайных процессов и непредсказуемых воздействий. Гибкий арсенал методов математической статистики, используемых в медицине, позволяет выявить закономерности в потоках случайных событий, сделать выводы и прогнозы, основанные на доказанном, дать оценки вероятностей их выполнения или невыполнения.

Распространение персональных компьютеров, их совершенствование, а также наличие дружественного программного обеспечения для статистического анализа данных не только не снижает потребности в знании основ математической статистики, но и предъявляет к пользователю новые требования выбора и грамотного использования необходимых статистических процедур, корректной интерпретации и правильных заключений по результатам статистического анализа.

Настоящее пособие рассчитано на учащихся высших медицинских заведений и преподавателей. Многие вопросы, освещенные в пособии, будут интересны стажерам, ординаторам и аспирантам, участвующим в клинических и экспериментальных исследованиях. При составлении пособия автор использовал разнообразный материал из разделов математической статистики и адаптировал его специально для медицинской аудитории.

В книге содержатся основные определения и краткое описание статистических методов, наиболее широко применяемых в медико-биологических исследованиях: планирование исследования, сбор данных, первичная обработка материала исследования, постановка гипотез и их проверка. Рассмотренные примеры в виде типовых задач, таблиц, графиков, диаграмм хорошо иллюстрируют излагаемый материал.

Пособие построено по плану лекций и может быть использовано для теоретической подготовки к лабораторным занятиям, контрольным, зачетам и экзаменам.

В работе над книгой и подготовке ее к печати большую и неоценимую помощь оказал заведующий кафедрой медицинской информатики РУДН доцент Владимир Данилович Проценко, за что автор выражает ему глубокую благодарность и признательность, а также благодарит заведующего лабораторией биостатистики Гематологического научного центра РАМН Сергея Михайловича Куликова, который внимательно прочел рукопись книги, сделал замечания и дал ценные советы, которые способствовали улучшению данного пособия.

Автор будет признателен всем преподавателям и научным работникам, специализирующимся в области применения теории вероятностей и математической статистики в биологии, химии, медицине, за критические замечания и возможные предложения, направленные на улучшение и усовершенствование пособия.

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

СЛУЧАЙНЫЕ СОБЫТИЯ

События

Теория вероятностей – математическая наука, изучающая закономерности в случайных явлениях. Испытание или опыт, событие, вероятность события, случайная величина и случайный процесс – все это основные понятия теории вероятностей.

Испытанием называется осуществление на практике какого-нибудь комплекса условий.

Под *комплексом условий* понимают обстоятельства, от которых что-то зависит. Например, при проведении химического опыта результат может зависеть от температуры, давления.

Событием называется всякое явление, о котором имеет смысл говорить, что оно происходит или что оно не происходит (в настоящем, прошедшем или в будущем) при наличии комплекса условий. События обозначаются заглавными буквами латинского алфавита — A, B, C.

Если при определенных условиях событие обязательно произойдет, то такое событие называется *достоверным*. Если же событие не может произойти при данных условиях, то оно называется *невозможным*. Достоверное событие обозначается Ω (омега), невозможное — $\varnothing$ (знак пустого множества).

Если событие при реализации определенного комплекса условий может произойти, а может и не произойти, то оно называется *случайным*. Случайные события представляют различные возможные исходы испытания.

Если в условиях испытания появление одного события исключает появление других событий, то такие события называются *несовместными*.

Пример.

Пусть испытание заключается в вытаскивании карты из колоды. Событие A – появление карты пиковой масти, событие B – появление Дамы. Эти события являются совместными, так как может появиться Дама пик.

Пример.

Пусть испытание заключается в бросании игральной кости. Событием является появление на верхней грани определенного числа очков от 1 до 6. События $A_1=\{1\}$, $A_2=\{2\}$, $A_3=\{3\}$, $A_4=\{4\}$, $A_5=\{5\}$, $A_6=\{6\}$ являются несовместными, так как появление одного из них исключает появление других.

Событие называется *сложным (составным)*, если оно может быть разбито на несколько событий. Если событие невозможно разбить, то оно является простым и называется *элементарным* событием.

Пример.

Пусть испытание заключается в бросании игральной кости. Событие B – выпадение четного числа очков – является составным и может быть разбито на три элементарных события: A_1 – выпадение 2, A_2 – выпадение 4, A_3 – выпадение 6.

Для каждого испытания существует множество всех возможных элементарных событий. Такое множество называется *пространством элементарных событий* (данного испытания). Элементарные события взаимно исключают друг друга, и в результате данного опыта обязательно произойдет одно из них. Пространство элементарных событий образует полную группу попарно несовместных событий, так как наступление хотя бы одного из событий полной группы есть достоверное событие. Пространство элементарных событий обозначается так же, как и достоверное событие Ω.

Два несовместных события, образующих полную группу, называются *противоположными*. Событие, противоположное событию, A обозначается $\bar{A}$ (не A).

Пример.

Пусть испытание заключается в подбрасывании монеты. Событие A – выпадение «герба», $\bar{A}$ – выпадение «решки».

Соотношения между событиями

Следование ($A\subset B$). Если при каждом испытании, в результате которого происходит событие A, происходит также и событие B, то говорят, что событие A влечет за собой событие B (из A следует B, или A есть частный случай события B).

Пример.

Бросание игральной кости. $A=\{2\}$, $B=\{$«четное число»$\}$. При наступлении события A наступает и событие B.

Объединение ($A\cup B$) – событие, состоящее из тех элементарных событий, которые входят либо в событие A, либо в событие B, либо одновременно – в оба события.

Свойства:

1) $A\cup A=A$ – объединение события A с самим собой – есть событие A.

2) $A\cup\Omega=\Omega$ – объединение события A с достоверным событием – есть достоверное событие.

3) $A\cup\varnothing=A$ – объединение события A с невозможным событием – есть событие A.

4) $A\cup\bar{A}=\Omega$ – объединение события A с противоположным событием – есть достоверное событие.

5) $A\cup B=B\cup A$ – коммутативный закон.

Пересечение ($A\cap B$) – событие, состоящее из тех элементарных событий, которые входят в оба события в A и в B одновременно.

Свойства:

1) $A\cap A=A$ – пересечение события A с самим собой – есть событие A.

2) $A \cap \Omega = A$ – пересечение события A с достоверным событием – есть событие A.

3) $A \cap \bar{A} = \emptyset$ – пересечение события A с противоположным событием – есть невозможное событие.

4) $A \cap \emptyset = \emptyset$ – пересечение события A с невозможным событием – есть невозможное событие.

5) $A \cap B = B \cap A$ – коммутативный закон.

Разность $(A - B)$ – событие, состоящее из тех элементарных событий, которые входят в A и не входят в B.

Свойства

1) $\Omega - A = \bar{A}$ – разность достоверного события и события A – есть событие, противоположное событию A.

2) $A - A = \emptyset$

Симметрическая разность $A \Delta B = (A-B) \cup (B-A)$ – событие, которое состоит из тех элементарных событий, которые входят в A или B, но не входят в их пересечение.

Пример.

В испытании с игральной костью событие A означает, число выпавших очков меньше пяти ($A=\{1,2,3,4\}$), а событие B означает, что число выпавших очков больше трех ($B=\{4,5,6\}$). Тогда объединение этих событий ($A \cup B$)= $\{1,2,3,4,5,6\}$, пересечение ($A \cap B$)=$\{4\}$, разность событий ($A-B$)=$\{1,2,3\}$, симметрическая разность $A \Delta B$=$\{1,2,3,5,6\}$.

Для лучшего понимания соотношения между событиями обычно используют условные графические изображения, представляя достоверное событие как прямоугольник, а другие события как круги. Тогда введенные выше соотношения могут быть представлены в виде диаграмм Вьенна.

Пользуясь введенными соотношениями между событиями, уточним данные ранее определения несовместных, противоположных событий и полной группы событий.

События A и B называются несовместными, если $A \cap B = \emptyset$.

События A и $\bar{A}$ называются противоположными, если $A \cap \bar{A} = \emptyset$ и $A \cup \bar{A} = \Omega$.

События A_1, A_2,…, A_n образуют полную группу событий, если

$A_1 + A_2 + \ldots + A_n = \Omega$ и $A_i \cap A_j = \emptyset$; i=1,2,…,n; j=1,2,…,n ($i \neq j$).

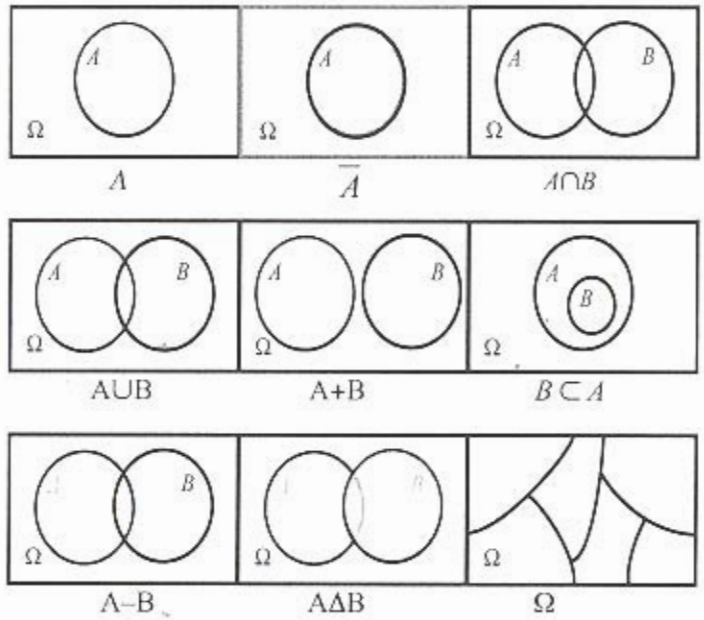


Рис.1. Диаграммы Вьенна

Задачи

1. Монета подбрасывается три раза подряд. Под исходом опыта будем понимать последовательность (X_1, X_2, X_3), где каждый из X_i обозначает выпадение "герба" (Г) или "цифры" (Ц). а) Построить пространство Ω элементарных событий. б) Описать событие A, состоящее в том, что выпало не менее двух гербов.

2. Событие B является частным случаем события A. Чему равны их сумма и произведение?

3. Пусть на плоскость наудачу бросается точка, и пусть события A и B состоят в том, что эта точка попадает соответственно в круг A и в круг B. Какой смысл имеют события $\overline{A}, \overline{B}, A \cup B, \overline{A \cup B}, AB, \overline{AB}$?

4. Пусть A, B, C – случайные события. Выяснить смысл равенств: а) $ABC=A$, б) $A \cup B \cup C = A$.

5. Пусть A, B, C – три произвольных события. Найти выражения для событий, состоящих в том, что из A, B, C:
   а) произошло только событие A;
   б) произошли A и B, но C не произошло;
   в) все три события произошли;
   г) произошло хотя бы одно из этих событий;
   д) произошли хотя бы два события;
   е) произошло одно и только одно событие;
   ж) произошли два и только два события;
   з) ни одно событие не произошло;
   и) произошли не более двух событий.

6. Относительно событий, перечисленных в каждом примере, указать, образуют ли они в данном опыте полную группу событий (да, нет):
   а) опыт – бросание монеты, события:
   A_1={герб}; A_2={цифра}.
   б) опыт – бросание двух монет; события:
   B_1={два герба}; B_2={две цифры}.

   в) опыт – бросание двух игральных кубиков; события:
   C_1={на обоих кубиках шестерки}; C_2={ни на одном кубике нет шестерки}; C_3={на одном из кубиков шестерка, на другом - нет}.

7. Относительно каждой группы событий ответить на вопрос, являются ли они в данном опыте несовместными (да, нет).
   а) опыт – бросание монеты; события:
   A_1={герб}; A_2={цифра};
   б) опыт – бросание двух монет; события:
   B_1={герб на первой монете}; B_2={герб на второй монете};
   в) опыт – вынимание двух карт из колоды; события:
   C_1={обе карты черной масти}; C_2={среди вынутых карт есть дама треф}; C_3={среди вынутых карт есть туз пик}.

8. Опыт состоит в бросании двух монет. Рассматриваются следующие события:
   A={герб на первой монете};
   B={цифра на первой монете};
   C={герб на второй монете};
   D={цифра на второй монете};
   E={хотя бы один герб};
   F={хотя бы одна цифра};
   G={один герб и одна цифра};
   H={ни одного герба};
   K={два герба}.
   Определить каким событиям этого списка равносильны следующие события: 1) $A+C$; 2) AC; 3) EF; 4) $G+E$; 5) GE; 6) BD; 7) $E+K$.

ВЕРОЯТНОСТЬ СЛУЧАЙНОГО СОБЫТИЯ

Аксиомы Колмогорова

В 1933 г. А. Н. Колмогоров в книге «Основные понятия теории вероятностей» дал аксиоматическое обоснование теории вероятностей. «Это означает, что, после того как даны названия изучаемым объектам и их основным отношениям, а также аксиомы, которым эти отношения подчиняются, все дальнейшее изложение должно основываться исключительно лишь на этих аксиомах…», – писал Колмогоров. По мнению ученого, наиболее целесообразным представлялось аксиоматизирование понятий случайного события и его вероятности.

Пусть Ω – множество элементов ω, которые мы будем называть элементарными событиями, а $\mathfrak{I}$ – множество подмножеств из Ω. Элементы множества $\mathfrak{I}$ будем называть случайными событиями (или просто событиями), а Ω – пространством элементарных событий.

Аксиомы.

I. $\mathfrak{I}$ является алгеброй множеств.

II. Каждому множеству A из $\mathfrak{I}$ поставлено в соответствие неотрицательное число $P(A)$. Это число называется вероятностью события A.

III. $P(\Omega)=1$.

IV. Если A и B не пересекаются, то $P(A+B)=P(A)+P(B)$.

V. Для убывающей последовательности $A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n$ событий из $\mathfrak{I}$ такой, что $\bigcap_n A_n = \varnothing$ имеет место равенство

$$\lim_{n\to\infty} P(A_n) = 0.$$

Поясним некоторые аксиомы. Система $\mathfrak{I}$ подмножеств множества Ω называется алгеброй, если $\Omega\in\mathfrak{I}$, объединение, пересечение и разность двух множеств системы опять принадлежит этой системе. Другими словами, если A и B – события, то $A\cup B$, AB, $A-B$ тоже события. Вероятность достоверного события равна единице.

Следствие 1.

Вероятность события, противоположного событию A равна единице минус вероятность события A: $P(\overline{A}) = 1 - P(A)$.

Доказательство:

По определению противоположных событий $A\cap\overline{A}=\varnothing$ и $A\cup\overline{A}=\Omega$. По аксиоме III $P(\Omega)=1$. По аксиоме IV $P(A+\overline{A}) = P(A)+P(\overline{A})$. Из вышеперечисленного имеем $P(\Omega) = P(A+\overline{A}) = P(A)+P(\overline{A}) = 1$, следовательно, $P(\overline{A}) = 1 - P(A)$.

Следствие 2.

Вероятность невозможности события равна нулю ($P(\varnothing)= 0$).

Доказательство:

Так как $\overline{\Omega}=\varnothing$, то из следствия 1 вытекает, что

$$P(\varnothing)=1 - P(\overline{\Omega})=1 - 1 = 0.$$

Следствие 3.

Если события $A_1,A_2,\ldots,A_n$ попарно несовместны, то $P(A_1+A_2+A_3+\ldots+A_n) = P(A_1) + P(A_2)+\ldots+P(A_n)$. Это следует из аксиомы IV.

Понятие вероятности случайного события

Рассматривая события, мы видим, что каждое из них обладает какой-то степенью возможности: одни – большей, другие – меньшей, причем для некоторых из этих событий мы сразу же можем решить, какое из них более, а какое менее возможно. Чтобы количественно сравнить между собой события по степени их возможности, очевидно, нужно с каждым событием связать определенное число, которое тем больше, чем более возможно событие. Такое число мы назовем вероятностью события.

О вероятности того или иного события можно говорить только в рамках определенного испытания.

Выше были введены понятия полной группы, несовместности и равновозможности событий.

Существуют группы событий, обладающие всеми тремя свойствами: они образуют полную группу, несовместны и равновозможны. События, образующие такую группу, называют *случаями*.

Случай называется *благоприятным* некоторому событию, если появление этого случая влечет за собой появление данного события.

Если опыт сводится к схеме случаев, то вероятность события A в данном опыте можно оценить по относительной доле благоприятных случаев.

Пусть результатом некоторого испытания может быть конечное число n несовместных и равновозможных исходов (случаев), и что в m исходах (случаях), называемых благоприятными, осуществляется событие A. Тогда вероятность события A равна отношению числа m благоприятных исходов (случаев) к общему числу возможных исходов (случаев).

$$P(A) = \frac{m}{n}.$$

Если нарушается хотя бы одно из трех указанных условий (конечность, несовместность и равновозможность исхода), формула не приемлема.

Теорема .

Значения вероятности случайного события принадлежат отрезку [0;1], т.е. $0 \le P(A) \le 1$.

Доказательство:

Количество благоприятых исходов лежит в пределах от 0 до n. Для невозможного события $m = 0$, для достоверного $m = n$.

$$0 \le m \le n$$

$$\frac{0}{n} \le \frac{m}{n} \le \frac{n}{n}.$$

Задачи

9. Брошены два игральных кубика. Найти вероятность того, что сумма очков на выпавших гранях – четная, причем на грани хотя бы одной из костей появится шестерка.

10. Задумано двузначное число. Найти вероятность того, что задуманным числом окажется: а) случайно названное двузначное число; б) случайно названное двузначное число, цифры которого различны.

11. Указать ошибку «решения» задачи: брошены два игральных кубика; найти вероятность события A = {сумма выпавших очков равна трем }.

    Решение:

    Возможны два исхода испытания: сумма выпавших очков равна трем, сумма выпавших очков не равна трем. Событию A благоприятствует один исход; общее число исходов равно двум. Следовательно, искомая вероятность $P(A)=1/2$.

12. Брошены два игральных кубика. Найти вероятности следующих событий: а) сумма выпавших очков равна семи; б) сумма выпавших очков равна восьми, а разность – четырем; в) сумма выпавших очков равна восьми, если известно, что их разность равна четырем; г) сумма выпавших очков равна пяти, а произведение – четырем.

13. Монета брошена два раза. Найти вероятность того, что хотя бы один раз появится «герб».

14. В коробке шесть одинаковых, занумерованных кубиков. Наудачу по одному извлекают все кубики. Найти вероятность того, что номера извлеченных кубиков появятся в возрастающем порядке.

15. Найти вероятность того, что при бросании трех игральных кубиков шестерка выпадает на одном (безразлично каком) кубике, если на гранях двух других кубиков выпадут числа очков, не совпадающие между собой (и не равные шести).

16. В коробке имеется десять шаров: три белых и семь черных. Из коробки наугад вынимается шар. Какова вероятность того, что этот шар: а) белый; б) черный?

17. Из слова НАУГАД выбирается одна буква. Какова вероятность, что это буква *А*? Какова вероятность того, что это гласная?

18. Брошены три монеты. Найти вероятность того, что выпадут ровно два герба?

19. Бросают игральный кубик. Какова вероятность выпадения номера «4» на верхней грани упавшего на стол кубика? Какова вероятность выпадения номера большего четырех?

20. Брошены две игральные кости. Какова вероятность выпадения на двух костях в сумме не менее девяти очков? Какова вероятность выпадения единицы хотя бы на одной из костей?

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

22. Из пяти карточек с буквами А, Б, В, Г, Д наугад одна за другой выбираются три карточки и располагаются в ряд в порядке появления. Какова вероятность, что получится слово ДВА?

23. В коробке три белых и семь черных шаров. Какова вероятность того, что вынутые наугад два шара окажутся черными?

24. Набирая номер телефона, абонент забыл последние три цифры и, помня лишь, что эти цифры различны, набрал наудачу. Найти вероятность того, что набраны нужные цифры.

25. При наборе телефонного номера абонент забыл две последние цифры и набрал их наудачу, помня только, что эти цифры нечетные и разные. Найти вероятность того, что номер набран правильно.

26. В лотерее R билетов, из которых T – выигрышных. Участник лотереи покупает K билетов. Определить вероятность того, что он выиграет хотя бы на один билет.

27. В партии из 50 изделий 5 бракованных. Из партии выбирается наугад шесть изделий. Определить вероятность того, что среди этих шести изделий два окажутся бракованными.

28. Найти вероятность того, что дни рождения двенадцати человек придутся на разные месяцы года.

29. В партии из N деталей имеется S стандартных. Наудачу отобраны L деталей. Найти вероятность того, что среди отобранных деталей ровно K – стандартных.

30. В группе 12 студентов, среди которых 8 отличников. По списку наудачу отобраны 9 студентов. Найти вероятность того, что среди отобранных студентов пять отличников.

31. В конверте среди 100 фотокарточек находится одна разыскиваемая. Из конверта наудачу извлечены 10 карточек. Найти вероятность того, что среди них окажется нужная.

32. В коробке пять одинаковых изделий, причем три из них окрашены. Наудачу извлечены два изделия. Найти вероятность того, что среди двух извлеченных изделий

окажутся: а) одно окрашенное изделие; б) два окрашенных изделия; в) хотя бы одно окрашенное изделие.

33. В ящике имеется 15 деталей, среди которых 10 – окрашенных. Сборщик наудачу извлекает три детали. Найти вероятность того, что извлеченные детали окажутся окрашенными.

34. В ящике 100 деталей, из них 10 – окрашенных. Наудачу извлечены 4 детали. Найти вероятность того, что среди извлеченных деталей: а) нет окрашенных; б) нет неокрашенных.

35. Относительно каждой из групп событий ответить на вопрос, равновозможны ли они в данном опыте (да, нет):

а) опыт – бросание монеты; события:
A_1={герб}; A_2={цифра};

б) опыт – бросание двух монет; события:
B_1={два герба}; B_2={две цифры}; B_3={один герб и одна цифра};

в) опыт – вынимание наугад одной карты из колоды; события:
$C1$={черви}; $C2$={бубны}; $C3$={трефы}; $C4$={пики};

г) опыт – бросание игрального кубика; события:
$D1$={не менее 3-х очков}; $D2$={не более 3-х очков}.

36. В ящике a белых и b черных шаров. Из ящика вынимают наугад 1 шар. Найти вероятность того, что этот шар – белый.

37. В ящике a белых и b черных шаров. Из ящика вынимают 1 шар и откладывают в сторону. Этот шар оказался белым. После этого из ящика берут еще 1 шар. Найти вероятность того, что этот шар тоже будет белым.

38. В ящике a белых и b черных шаров. Из ящика вынимают один шар и, не глядя ,откладывают в сторону. После этого из ящика взяли еще 1 шар. Он оказался белым. Найти вероятность того, что первый шар, отложенный в сторону, тоже белый.

39. Из ящика, содержащего a белых и b черных шаров, вынимают подряд все находящиеся в нем шары, кроме одного. Найти вероятность того, что последний, оставшийся в ящике шар, будет белым.

40. Из ящика, содержащего a белых и b черных шаров, вынимают подряд все находящиеся в нем шары. Найти вероятность того, что вторым по порядку будет вынут белый шар.

41. В ящике a белых и b черных шаров ($a \geq 2$). Из ящика вынимают сразу два шара. Найти вероятность того, что оба шара будут белыми.

42. В ящике a белых и b черных шаров ($a \geq 2$, $b \geq 3$). Из ящика вынимают сразу пять шаров. Найти вероятность того, что два из них будут белыми, а три черными.

43. В ящике a белых и b черных шаров ($a \geq 2$, $b \geq 2$). Из ящика одновременно вынимают два шара. Какое событие более вероятно: A={шары одного цвета}; B={шары разных цветов}.

44. Из ящика, содержащего n перенумерованных шаров, наугад вынимают один за другим все находящиеся в нем шары. Найти вероятность того, что номера вынутых шаров будут идти по порядку: 1,2,…, n.

45. Тот же ящик, что и в задаче 44, но каждый шар после вынимания вкладывается обратно и перемешивается с другими, а его номер записывается. Найти вероятность

того,что будет записана естественная последовательность номеров: 1,2,…,n.

46. Некто купил карточку Спортлото и отметил в ней 6 из имеющихся 49 номеров, после чего в тираже разыгрываются 6 «выигрышных» номеров из 49. Найти вероятности следующих событий:

A_1 ={верно угаданы 3 выигрышных номера из 6};
A_2 ={верно угаданы 4 выигрышных номера из 6};
A_3 ={верно угаданы 5 выигрышных номера из 6};
A_4 ={верно угаданы все 6 номеров}.

47. На девяти карточках написаны цифры: 0, 1, 2, 3, 4, 5, 6, 7, 8. Две из них вынимаются наугад и укладываются на стол в порядке появления, затем читается полученное число, например 07 (семь), 14 (четырнадцать) и т.п. Найти вероятность того, что число будет четным.

48. На пяти карточках написаны цифры: 1,2,3,4,5. Две из них, одна за другой, вынимаются. Найти вероятность того, что число на второй карточке будет больше, чем на первой.

49. N человек случайным образом рассаживаются за круглым столом (N>2). Найти вероятность того, что два фиксированных лица A и B окажутся рядом.

50. В первом ящике находятся шары с номерами от 1 до 5, а во втором – с номерами от 6 до 10. Из каждого ящика вынули по одному шару. Какова вероятность того, что сумма номеров вынутых шаров: а) не меньше 7; б) равна 11; в) не больше 11.

51. Бросают два кубика. Пусть A – событие, состоящее в том, что сумма очков нечетная; B – событие, заключающееся в том, что хотя бы на одном из кубиков выпала единица. Описать события AB, $A\cup B$, $A\overline{B}$, $\overline{A}B$, $\overline{A}\cup\overline{B}$ и найти их вероятности.

52. В кармане имеется несколько монет достоинством в 2 коп. и 10 коп. (на ощупь неразличимых). Известно, что двухкопеечных монет втрое больше, чем десятикопеечных. Наугад вынимается одна монета. Какова вероятность того, что она будет десятикопеечная.

53. Общество состоит из 5 мужчин и 10 женщин. Найти вероятность того, что при случайной группировке их на пять групп по три человека в каждой группе будет мужчина.

54. Бросают n кубиков. Найти вероятность получить сумму очков равную n, n+1.

55. Два игрока по очереди бросают кубик, каждый по одному разу. Выигравшим считается тот, кто получит большее число очков. Найти вероятность выигрыша первого игрока.

56. Два игрока бросают монету по 2 раза каждый. Выигравшим считается тот, кто получит больше гербов. Найти вероятность того, что выиграет первый игрок.

ОСНОВНЫЕ ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Мы познакомились с классической формулой для вероятности события. Однако на практике обычно требуется определить вероятности событий, непосредственное экспериментальное воспроизведение которых затруднено. Поэтому для определения вероятностей событий применяются методы, позволяющие по известным вероятностям одних событий определить вероятности других событий, с ними связанные. Такие методы позволяют свести необходимый эксперимент к минимуму. Применяя их, мы пользуемся основными теоремами теории вероятностей. Это теорема сложения и теорема умножения вероятностей.

Теорема сложения вероятностей

Напомним, что суммой двух событий A и B называется событие C, состоящее в появлении события A или события B, или обоих вместе. Другими словами, суммой двух событий A и B называется событие C, состоящее в появлении хотя бы одного из событий A и B.

Итак, суммой нескольких событий называется событие, состоящее в появлении хотя бы одного из этих событий.

Теорема:

Вероятность суммы двух несовместных событий равна сумме вероятностей этих событий.

Доказательство:

Пусть результатом опыта (эксперимента) могут быть n элементарных событий (случаев). Предположим, что из этих случаев m – благоприятны событию A, а k – событию B. Тогда:

$$P(A) = \frac{m}{n}; \quad P(B) = \frac{k}{n}$$

По условию события A и B несовместны, значит $m+k$ случаев благоприятны событию $(A+B)$. Имеем:

$$P(A + B) = \frac{m + k}{n} = \frac{m}{n} + \frac{k}{n} = P(A) + P(B).$$

Теорема доказана.

Обобщим теорему сложения на случай трех событий. Обозначая событие $A+B$ буквой D и присоединяя к сумме еще одно событие C, которому благоприятны l случаев. Тогда

$$P(C) = \frac{l}{n}, \qquad P(D) = \frac{m + k}{n},$$

докажем, что $P(A+B+C)=P(A)+P(B)+P(C)$.

Действительно

$$P((A+B)+C)=P(D+C) = \frac{m + k}{n} + \frac{l}{n} = \frac{m}{n} + \frac{k}{n} + \frac{l}{n} =$$
$$= P(A)+P(B)+P(C).$$

Методом индукции можно обобщить теорему сложения на произвольное число событий.

Таким образом, теорема сложения вероятностей применима к любому числу несовместных событий. Ее удобно записать в виде:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Из теоремы сложения вытекает несколько следствий.

Следствие 1.

Если события A_1, A_2,…,A_n образуют полную группу несовместных событий, то сумма их вероятностей равна единице:

$$\sum_{i=1}^{n} P(A_i) = 1.$$

Доказательство:

Так как события A_1, A_2, … A_n образуют полную группу, то появление хотя бы одного из них – достоверное событие:

$P(A_1+A_2+\ldots+A_n)=1$

Так как события A_1, A_2, … A_n – несовместные события, то к ним применима теорема сложения вероятностей:

$$P(A_1+A_2+\ldots+A_n) = P(A_1)+P(A_2)+\ldots+P(A_n) = \sum_{i=1}^{n} P(A_i)$$

$$\sum_{i=1}^{n} P(A_i) = 1,$$

что и требовалось доказать.

Следствие 2.

Сумма вероятностей противоположных событий равна единице:

$$P(A) + P(\overline{A}) = 1.$$

Поскольку противоположными событиями называются два несовместных события, образующих полную группу, то это следствие есть частный случай следствия 1.

На практике часто оказывается легче вычислить вероятность противоположного события $\overline{A}$, чем вероятность прямого события А. В этих случаях вычисляют $P(\overline{A})$ и находят

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

Теорема сложения вероятностей справедлива только для несовместных событий. В случае, когда события A и B совместны, вероятность суммы этих событий выражается формулой:

$$P(A + B) = P(A) + P(B) - P(AB).$$

Таким образом, имеем обобщенную теорему для двух событий: вероятность суммы двух событий равна сумме их вероятностей минус вероятность совместного их появления в данном испытании.

Аналогично вероятность суммы трех совместных событий вычисляется по формуле:

$$P(A+B+C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

Методом индукции доказывается общая формула для вероятности суммы любого числа совместных событий:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i,j} P(A_i A_j) + \sum_{i,j,k} P(A_i A_j A_k) - \ldots$$

$$\ldots + (-1)^{n-1} P(A_1 A_2 \ldots A_n),$$

где суммы распространяются на различные значения индексов i; i,j; i,j,k и т.д.

Условная вероятность

Введем понятие независимых и зависимых событий.

Событие A называется *независимым* от события B, если вероятность события A не зависит от того, произошло событие B или нет.

Событие A называется *зависимым* от события B, если вероятность события A меняется в зависимости от того, произошло событие B или нет.

Вероятность события A, вычисленная при условии, что имело место другое событие B, называется условной вероятностью события А и обозначается $P(A|B)$ или $P_B(A)$.

Условная вероятность события A при условии наступления события B равна

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

Эта формула называется формулой условной вероятности.

Условие независимости события A от события B можно записать в виде:

$$P(A|B) = P(A),$$

а условие зависимости:

$$P(A|B) \neq P(A).$$

Теорема умножения вероятностей

Теорема

Вероятность произведения двух событий равна произведению вероятности одного из них на условную

вероятность другого, вычисленную при условии, что первое имело место:

$$P(AB) = P(A)P(B|A).$$

Доказательство:

Пусть возможные исходы опыта сводятся к n элементарным событиям.

Предположим, что событию A благоприятны m элементарных событий, а событию B благоприятны k элементарных событий. Так как мы не предполагали события A и B несовместными, то вообще существуют элементарные события, благоприятные и событию A, и событию B одновременно. Пусть число таких элементарных событий l. Тогда

$$P(AB) = \frac{l}{n}, \qquad P(A) = \frac{m}{n}.$$

Вычислим $P(B|A)$, т.е. условную вероятность события B в предположении, что A имело место. Если известно, что событие A произошло, то из ранее возможных n элементарных событий остаются возможными только m, которые благоприятствовали событию A. Из них l элементарных событий благоприятны событию B. Следовательно,

$$P(B \mid A) = \frac{l}{m}.$$

Подставим полученные выражения в $P(AB)=P(A)P(B|A)$

$$\frac{l}{n} = \frac{m}{n} \times \frac{l}{m}.$$

Получили тождество. Теорема доказана.

Рассмотрим следствия, вытекающие из теоремы умножения.

Следствие 1.

Если событие A не зависит от события B, то и событие B не зависит от события A.

Доказательство:

Дано, что событие A не зависит от события B, т.е.

$$P(A) = P(A \mid B)$$

Требуется доказать, что и событие B не зависит от A, т.е.

$$P(B) = P(B \mid A).$$

При доказательстве будем предполагать, что $P(A)\neq 0$.

Напишем теорему вероятностей в двух формах:

$$P(AB)=P(B)P(A|B),$$
$$P(AB)=P(A)P(B|A).$$

Откуда

$$P(A)P(B|A)=P(B)P(A|B)$$

Или так как событие A не зависит от B, т.е. $P(A) = P(A \mid B)$, то $P(A)P(B|A)=P(B)P(A)$.

Разделим обе части равенства на $P(A)$. Получим:

$$P(B|A)=P(B).$$

Что и требовалось доказать.

Итак, два события называются независимыми, если появление одного из них не изменяет вероятности появления другого.

Следствие 2.

Вероятность произведения двух независимых событий равна произведению вероятностей этих событий.

При вычислении условной вероятности сужается пространство элементарных событий. Поясним это. Пусть задана некоторая полная группа событий $\omega_1, \omega_2,\omega_3,\ldots,\omega_n$, несовместных и равновероятных между собой, и на основании ее установлены вероятности событий в множестве допустимых событий A, B и т.д. При вычислении условной вероятности решается вопрос, как должны измениться вероятности, если предположить, что станет достоверным наступление одного из допустимых событий, например A. Предположение, что стало достоверным событие A, нарушает заданную полную группу несовместных между событий $\omega_1, \omega_2,\omega_3,\ldots,\omega_n$ в том смысле, что благодаря этому предположению все события заданной группы уже не могут больше считаться равновероятными. А именно, предположение, что стало достоверным событие A, означает, что из всех событий $\omega_1, \omega_2,\omega_3,\ldots,\omega_n$ остаются возможными только те, на которые подразделяется A, остальные же становятся невозможными. При этом те

из событий ω_1, ω_2,ω_3,…,ω_n , которые остаются возможными, остаются равновероятными между собой.

Формула полной вероятности

Пусть дана группа несовместных событий B_1, B_2… B_n и некоторое событие A, подразделяющееся на частные случаи AB_1, AB_2… AB_n. И пусть даны вероятности $P(B_1)$, $P(B_2)$,…,$P(B_n)$ и условные вероятности $P(A|B_1)$, $P(A|B_2)$,…,$P(A|B_n)$. Требуется определить вероятность $P(A)$.

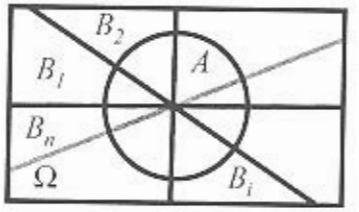


Рис. 2. Пояснение к формуле полной вероятности

Так как

$$A = AB_1 + AB_2 + \ldots + AB_n,$$

То

$$P(A) = P(AB_1 + AB_2 + \ldots + AB_n).$$

События B_1, B_2,…,B_n несовместные, следовательно, события AB_1, AB_2,…,AB_n тоже несовместные. Воспользуемся теоремой сложения для несовместных событий.

$$P(A) = P(AB_1 + AB_2 + \ldots + AB_n) = P(AB_1) + P(AB_2) + \ldots + P(AB_n)$$

По теореме умножения для каждого слагаемого имеем

$$P(AB_i) = P(B_i)P(A|B_i).$$

Следовательно

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n).$$

Или

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

Эта формула называется формулой полной вероятности.

В рассмотренной схеме событие A осуществляется только вместе с каким-либо из событий B_1, B_2, …, B_n. Последние, таким образом, выступают как единственно возможные и взаимно ис-

ключающие условия, определяющие появление события A, или как гипотезы, в предположении которых (и только их) может произойти событие A.

Формула Байеса

Пусть дана группа несовместных событий B_1, B_2 ,… B_n и некоторое событие A, подразделяющееся на частные случаи AB_1, AB_2,… AB_n. И пусть даны вероятности $P(B_1)$, $P(B_2)$,…,$P(B_n)$ и условные вероятности $P(A|B_1)$, $P(A|B_2)$,…,$P(A|B_n)$. Требуется определить условные вероятности $P(B_1|A)$, $P(B_2|A)$,…,$P(B_n|A)$.

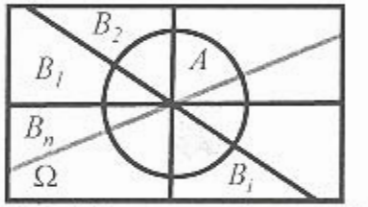


Рис. 3. Пояснение к формуле Байеса

По теореме умножения

$$P(AB) = P(B)P(A|B) \quad \text{или} \quad P(AB) = P(A)P(B|A).$$

Следовательно,

$$P(B)P(A|B) = P(A)P(B|A).$$

Воспользуемся последним равенством и выразим $P(B|A)$ в общем случае

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{P(A)}.$$

$P(A)$ находим по формуле полной вероятности

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

Итак,

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

Эта формула называется формулой Байеса.

Вероятность $P(B_i)$ осуществления события B_i (i = 1,…,n), вычисления безотносительно к событию A, называется *априорной вероятностью* (a priori). Условная вероятность $P(B_i|A)$ выполнения события B_i (i = 1,…,n), вычисленная в предположении, что событие A осуществилось, называется *апостериорной вероятностью* (a posteriori).

События B_i называют гипотезами, а теорему Байеса теоремой гипотез.

Формулы полной вероятности и Байеса связаны между собой и дают прямое и обратное решения одной и той же проблемы. Первая прогнозирует возможность появления события A по известным до опыта вероятностям осуществления гипотез. Последняя оценивает вероятность осуществления каждой гипотезы, если событие A произошло.

Задачи

57. Доказать, что если события A и B не зависимы, то события A и $\overline{B}$, $\overline{A}$ и B, $\overline{A}$ и $\overline{B}$ также независимы.

58. Пусть события A и B_1 независимы и независимы также события A и B_2, при этом $B_1B_2 = \varnothing$. Доказать, что события A и $B_1 + B_2$ независимы.

59. Бросили монету и игральный кубик. Определить, зависимы или независимы события: A = {выпал «герб»}; B = {выпало четное число очков}.

60. Брошены последовательно 3 монеты. Определить зависимы или независимы события: A = {выпадение герба на первой монете}; B = {выпадение хотя бы одной решетки}.

61. Опыт состоит в бросании двух монет. Рассматриваются следующие события:

    A={герб на первой монете};
    E={хотя бы один герб};
    F={хотя бы одна цифра};
    C={герб на второй монете}.

    Определить зависимы или независимы пары событий: 1) A и E; 2) A и F; 3) D и E; 4) D и F. Определить условные и безусловные вероятности событий в каждой паре.

62. Доказать, что если A и B – независимые события с положительными вероятностями, то они совместны.

63. Бросили игральный кубик. Какова вероятность того, что выпало простое число очков, если известно, что число выпавших очков нечетно?

64. В ящике лежат 12 красных, 8 зеленых и 10 синих шаров. Наугад вынимаются 2 шара. Какова вероятность, что

вынутые шары разного цвета, если известно, что не вынут синий шар?

65. В одном ящике 5 белых и 10 красных шаров, в другом ящике 10 белых и 5 красных шаров. Найти вероятность того, что хотя бы из одного ящика будет вынут 1 белый шар, если из каждого ящика вынуто по одному шару.

66. Вероятность того, что в течение одной смены возникнет неполадка станка, равна 0,05. Какова вероятность того, что не произойдет ни одной неполадки за три смены?

67. Гардеробщица выдала одновременно номерки четырем лицам, сдавшим в гардероб свои шляпы. После этого она перепутала все шляпы и повесила их наугад. Найти вероятность следующих событий:

A={каждому из четырех лиц гардеробщица выдаст его собственную шляпу};

B={ровно три лица получат свои шляпы};

C={ровно два лица получат свои шляпы};

D={ровно одно лицо получит свою шляпу};

E={ни одно из четырех лиц не получит своей шляпы}.

68. На полке в библиотеке в случайном порядке расставлены 15 учебников, причем пять из них – в переплете. Библиотекарь берет три учебника. Найти вероятность того, что хотя бы один из взятых учебников окажется в переплете.

69. В ящике 10 шаров, из которых четыре – окрашены. Взяли три шара. Найти вероятность того, что хотя бы один из взятых шаров окрашен.

70. Доказать, что если событие A влечет за собой событие B, то $P(B) \geq P(A)$.

71. Вероятности каждого из двух независимых событий A_1 и A_2 соответственно равны p_1 и p_2. Найти вероятность появления только одного из этих событий.

72. Вероятность попадания в цель первым стрелком p_1, а вторым – p_2. Стрелки выстрелили одновременно. Какова

вероятность того, что один из них попадет в цель, а другой не попадет?

73. Для сигнализации об аварии установленны два независимо работающих сигнализатора. Вероятность того, что при аварии сигнализатор сработает, равна 0,95 для первого сигнализатора и 0,9 для второго. Найти вероятность того, что при аварии сработает только один сигнализатор.

74. Два стрелка стреляют по мишени. Вероятность попадания в мишень при одном выстреле для первого стрелка равна 0,7, а для второго – 0,8. Найти вероятность того, что при одном выстреле в мишень попадает только один из стрелков.

75. В читальном зале имеется шесть учебников по теории вероятностей, из которых три – в переплете. Библиотекарь наудачу взял два учебника. Найти вероятность того, что оба учебника окажутся в переплете.

76. Среди 100 лотерейных билетов есть 5 выигрышных. Найти вероятность того, что два наудачу выбранные билета окажутся выигрышными.

77. Вероятность попадания в цель при одном выстреле из 2-х орудий равна 0,38. Найти вероятность поражения цели при одном выстреле первым из орудий, если известно, что для второго орудия эта вероятность равна 0,8.

78. Отдел контроля проверяет изделия (лекарства) на стандартность. Вероятность того, что изделие стандартно, равна 0,9. Найти вероятность того, что из двух проверенных изделий только одно стандартное.

79. Из партии изделий товаровед отбирает изделия высшего сорта. Вероятность того, что наудачу взятое изделие окажется высшего сорта, равна 0,8. Найти вероятность того, что из трех проверенных изделий только два изделия высшего сорта.

80. Студент знает 20 из 25 вопросов программы. Найти вероятность того, что студент знает предложенные ему экзаменатором три вопроса.

81. Найти вероятность $P(A)$ по данным вероятностям: $P(AB)=0{,}72$; $P(A\overline{B})=0{,}18$.

82. Найти вероятность $P(A\overline{B})$ по данным вероятностям: $P(A)=a$; $P(B)=b$; $P(A+B)=c$.

83. Найти вероятность $P(\overline{A}\,\overline{B})$ по данным вероятностям: $P(A)=a$; $P(B)=b$; $P(A+B)=c$.

84. Имеются 3 одинаковые с виду коробки. В первой a белых шаров и b черных; во второй c белых и d черных; в третьей только белые шары. Некто подходит наугад к одной из коробок и вынимает из нее 1 шар. Найти вероятность того, что этот шар белый.

85. Имеются 2 одинаковых ящика с шарами. В первом ящике 2 белых и 1 черный шар, во втором – 1 белый и 4 черных шара. Наудачу выбирают один ящик и вынимают из него 1 шар. Какова вероятность, что вынутый шар окажется белым?

86. В коробку, содержащую два шара, опущен белый шар, после этого из нее наудачу извлечен один шар. Найти вероятность того,что извлеченный шар окажется белым, если равновозможны все возможные предположения о первоначальном составе шаров.

87. В коробку, содержащую n шаров, опущен белый шар, после чего наудачу извлечен один шар. Найти вероятность того, что извлеченный шар окажется белым, если равновозможны все возможные предположения о первоначальном составе шаров (по цвету).

88. В пирамиде пять винтовок, три из которых снабжены оптическим прицелом. Вероятность того, что стрелок поразит мишень при выстреле из винтовки с оптическим прицелом, равна 0,95; для винтовки без оптического прицела эта вероятность равна 0,7. Найти вероятность

того, что мишень будет поражена, если стрелок произведет один выстрел из наудачу взятой винтовки.

89. Имеются две коробки: в первой 3 белых шара и 2 черных, во второй 4 белых и 4 черных шара. Из первой коробки во вторую перекладывают, не глядя, два шара. После этого из второй коробки берут 1 шар. Найти вероятность того, что этот шар будет белым.

90. Студент знает не все экзаменационные билеты. В каком случае вероятность вытащить неизвестный билет будет для него наименьшей, когда он тащит билет первым или последним?

91. Вероятности того, что во время работы цифровой электронной машины произойдет сбой в арифметическом устройстве, в оперативной памяти, в остальных устройствах, относятся как 3:2:5. Вероятности обнаружения сбоя в арифметическом устройстве, в оперативной памяти и в остальных устройствах соответственно равны 0.8, 0.9, 0.9. Найти вероятность того, что возникший в машине сбой будет обнаружен.

92. Группа студентов состоит из a отличников, b хорошо успевающих и c занимающихся слабо. Отличники на предстоящем экзамене могут получить только отличные оценки. Хорошо успевающие студенты могут получить с равной вероятностью хорошие и отличные оценки. Слабо занимающиеся могут получить с равной вероятностью хорошие, удовлетворительные и неудовлетворительные оценки. Для сдачи экзамена вызывается наугад один студент. Найти вероятность события A={студент получит хорошую или отличную оценку}.

93. Имеются 3 ящика: в первом a белых шаров и b черных; во втором c белых шаров и d черных; в третьем k белых шаров (черных нет). Некто выбирает наугад один ящик и

вынимает из нее шар. Этот шар оказался белым. Найти вероятность того, что: а) этот шар вынут из первого ящика; б) этот шар вынут из второго ящика; в) этот шар вынут из третьего ящика.

94. У рыбака имеется три излюбленных места для ловли рыбы, которые он посещает с равной вероятностью каждое. Если он закидывает удочку в первом месте, рыба клюет с вероятностью p_1; во втором месте – с вероятностью p_2; в третьем – с вероятностью p_3. Известно, что рыбак, выйдя на ловлю рыбы, три раза закинул удочку, и рыба клюнула только один раз. Найти вероятность того, что он удил рыбу в первом месте.

95. 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).

96. Два стрелка, независимо один от другого, стреляют по одной мишени, делая каждый по одному выстрелу. Вероятность попадания в мишень для первого стрелка 0,8, для второго – 0,4. После стрельбы в мишени обнаружена одна пробоина. Найти вероятность того, что в мишень попал первый стрелок.

97. Число грузовых автомашин, проезжающих по шоссе, на котором стоит бензоколонка, относится к числу легковых машин, проезжающих по тому шоссе, как 3:2. Вероятность того, что будет заправляться грузовая машина, равна 0,1; для легковой машины эта вероятность равна 0,2. К бензоколонке подъехала для заправки машина. Найти вероятность того, что это грузовая машина.

98. Событие A может появиться при условии появления лишь одного из несовместимых событий $B_1, B_2, \ldots, B_n$, образующих полную группу событий. После появления события A были переоценены вероятности гипотез, т.е.

были найдены условные вероятности $P(B_i|A)$ (i = 1,2,…,n). Доказать, что $\sum_{i=1}^{n} P(B_i \mid A) = 1$.

99. Событие A может появиться при условии появления одного из несовместимых событий (гипотез) B_1, B_2, B_3, образующих полную группу событий. После появления события A были переоценены вероятности гипотез, т.е. были найдены условные вероятности этих гипотез,причем оказалось, что $P_A(B_1)=0,6$ и $P_A(B_2)=0,3$. Чему равна условная вероятность $P_A(B_3)$ гипотезы B_3?

100. Имеются три партии деталей по 20 деталей в каждой. Число стандартных деталей в первой, второй и третьей партиях соответственно равно 20, 15, 10. Из наудачу выбранной партии наудачу извлечена деталь, оказавшаяся стандартной. Деталь возвращают в партию и вторично из той же партии наудачу извлекают деталь, которая также оказывается стандартной. Найти вероятность того, что детали были извлечены из третьей партии.

101. В специализированную больницу поступают в среднем 50% больных с заболеванием K, 30% – с заболеванием L, 20% – с заболеванием M. Вероятность полного излечения болезни K равна 0,7; для болезней L и M эти вероятности соответственно равны 0,8 и 0,9. Больной, поступивший в больницу, был выписан здоровым. Найти вероятность того, что этот больной страдает заболеванием K.

102. Два из трех, независимо работающих, элементов вычислительного устройства отказали. Найти вероятность того, что отказали первый и второй элементы, если вероятности отказа первого, второго и третьего элементов соответственно равны 0,6; 0,4 и 0,3.

103. В трех ящиках имеются белые и черные шары. Известно, что во втором и третьем ящиках число шаров одинаково, причем в два раза больше, чем в первом. Про ящики также известно, что во втором ящике черных и белых шаров поровну, в первом ящике белых шаров в 4 раза больше, чем черных, а в третьем ящике черных шаров столько же, сколько и в первом. Из наугад выбранного ящика

случайным образом вынимают шар. Какова вероятность того, что он белый.

104. В ящике лежат 20 теннисных мячей, в том числе 12 новых и 8 игранных. Из ящика извлекаются наугад два мяча для игры и после игры возвращаются обратно. После этого из ящика вынимают два мяча для следующей игры. Найти вероятность того, что эти оба мяча окажутся новыми.

105. Прибор может работать в двух режимах: A и B. Режим A наблюдается в 80% всех случаев работы прибора; режим B – в 20%. Вероятность выхода прибора из строя за время t в режиме A равна 0,1, в режиме B – 0,7. Найти полную вероятность выхода прибора из строя за время t.

106. Имеются два ящика: в первом a белых шаров и b черных; во втором c белых и d черных. Из первого ящика во второй перекладывают, не глядя, один шар. После этого из второго ящика берут один шар. Найти вероятность того, что этот шар будет белым.

107. Приборы одного наименования изготовляются двумя заводами; первый завод поставляет 2/3 всех изделий, поступающих на производство; второй – 1/3. Надежность (вероятность безотказной работы) прибора, изготовленного первым заводом, равна $p1$; второго $p2$. Определить полную надежность прибора, поступившего на производство.

108. Имеются две партии лекарственных препаратов; первая партия состоит из N препаратов, среди которых n просроченных; вторая партия – из M препаратов, среди которых – m дефектных. Из первой партии берется K препаратов, а из второй L ($K<N$, $L<M$); эти $K+L$ препараты смешиваются и образуется новая партия. Из новой партии берется наугад один препарат. Найти вероятность того, что он будет просроченным.

109. У рыбака имеется три излюбленных места для ловли рыбы, которые он посещает с равной вероятностью каждое. Если он закидывает удочку на первом месте, рыба клюет с вероятностью p_1; во втором месте – с вероятностью p_2; на третьем – с вероятностью p_3. Известно, что рыбак, выйдя

на ловлю рыбы, три раза закинул удочку и рыба клюнула только один раз. Найти вероятность того, что он удил рыбу во втором месте.

110. В группе 10 студентов, пришедших на экзамен, 3 – подготовленных отлично, 4 – хорошо, 2 – удовлетворительно и 1 – плохо. В экзаменационных билетах имеются 20 вопросов. Отлично подготовленный студент может ответить на все 20 вопросов, хорошо подготовленный – на 16, удовлетворительно подготовленный – на 10, плохо подготовленный – на 5. Вызванный наугад студент ответил на три произвольно заданных вопроса. Найти вероятность того, что студент подготовлен отлично (плохо).

111. Пассажир может обратиться за билетом в одну из трех касс. Вероятности обращения в каждую кассу зависят от их места расположения и равны соответственно p_1, p_2, p_3. Вероятность того, что к моменту прихода пассажира имеющиеся в кассе билеты будут распроданы, равна для первой кассы p_1, для второй p_2, для третьей p_3. Пассажир направился за билетом в одну из касс и приобрел билет. Найти вероятность того, что это была первая касса.

ПОВТОРНЫЕ ИСПЫТАНИЯ

Схема независимых испытаний Бернулли

До сих пор мы в основном разбирали задачи нахождения вероятности события в единичном испытании, т.е. когда эксперимент производится один раз. Теперь мы рассмотрим случай, когда одно и то же испытание повторяется несколько раз, т.е. проводится серия испытаний.

В этой задаче нас не интересует результат отдельного испытания, а интересует результат серии опытов.

В каждом отдельном испытании событие A может либо появиться, либо не появиться.

Так вот, в рассматриваемых задачах нас будет интересовать (т.е. мы должны определить) вероятность появления события A ровно k раз (или любое наперед заданное количество раз), в серии из n опытов.

Мы будем рассматривать случай, когда испытания независимы, т.е. каждое последующее испытание не зависит от предыдущего.

Примеры независимых испытаний.

Пример.

Несколько последовательных бросаний монеты.

Пример.

Несколько последовательных выниманий карты из колоды, при условии, что карта возвращается каждый раз и колода перемешивается, т.е. выборка с возвращением (иначе эти испытания – зависимые).

Независимые испытания могут производиться в одинаковых или различных условиях. В первом случае вероятность появления события A во всех опытах одна и та же. Во втором случае вероятность события A меняется от испытания к испытанию.

Мы будем рассматривать первый случай, т.е. вероятность появления события A не изменяется от испытания к испытанию. Решение таких задач сводится к схеме независимых испытаний Бернулли.

Схема

1. Испытания независимые (друг от друга).
2. Каждое отдельное испытание имеет только два возможных исхода (например, "Успех"= A, "Неудача"=$\bar{A}$).
3. Вероятности исходов остаются неизменными для всех испытаний.

В отдельном испытании вероятность события A определяют равной p ($P(A) = p$), при этом $P(\bar{A}) = 1 - P(A) = 1 - p = q$ (обозначают q) (т.е. вероятность успеха равна p, неудачи – $(1-p) = q$).

Формула Бернулли

Итак, пусть проводится n независимых испытаний. Обозначим A_i появление события A в i-ом испытании, а вероятность этого события обозначим p_i ($P(A_i) = p_i$). Непоявление события в i-ом испытании обозначим соответственно $\bar{A}_i$, а вероятность непоявления $P(\bar{A}_i) = 1 - p_i = q_i$. Поскольку мы будем рассматривать схему, в которой вероятности исходов не меняются от испытания к испытанию, то обозначим все $p_i = p$, а $q_i = q$ для всех значений i.

Рассмотрим несколько примеров:

1) два испытания, B: «только в одном успех»

$$B = A_1\bar{A}_2 + \bar{A}_1A_2 \rightarrow P(B) = P(A_1\bar{A}_2 + \bar{A}_1A_2) =$$

= {так как $A_1\bar{A}_2$ и $\bar{A}_1A_2$ несовместны, то по теореме сложения}=

$$= P(A_1\bar{A}_2) + P(\bar{A}_1A_2) =$$

={так как A_1 и A_2 независимы, то независимы A_1 и $\bar{A}_2$, $\bar{A}_1$ и A_2, тогда по теореме умножения имеем}=

$$= P(A_1)P(\bar{A}_2) + P(\bar{A}_1)P(A_2) = p(1-p) + (1-p)p = pq + pq = 2pq;$$

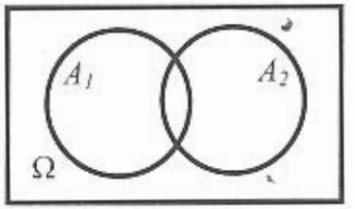


Рис. 4. Два испытания, B:" только в одном успех" (закрашенная область соответствует событию B)

2) три испытания, B: «только в двух успех»

$$B=A_1\bar{A}_2A_3 + \bar{A}_1A_2A_3 + A_1A_2\bar{A}_3;$$

$$P(B)=p(1-p)p+(1-p)pp+pp(1-p)=3p^2(1-p)=3p^2q;$$

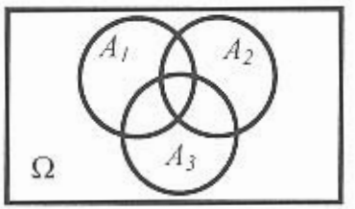


Рис. 5. Три испытания, B:" только в двух успех" (закрашенная область соответствует событию B)

3) четыре испытания, B: «только в двух успех»

$$B=A_1A_2\bar{A}_3\bar{A}_4+A_1\bar{A}_2A_3\bar{A}_4+A_1\bar{A}_2\bar{A}_3A_4+$$
$$+\bar{A}_1A_2A_3\bar{A}_4+\bar{A}_1A_2\bar{A}_3A_4+\bar{A}_1\bar{A}_2A_3A_4;$$

$$P(B)=pp(1-p)(1-p)+p(1-p)p(1-p)+p(1-p)(1-p)p+$$
$$+(1-p)pp(1-p)+(1-p)p(1-p)p+(1-p)(1-p)pp;$$

$$P(B)=6p^2(1-p)^2=6p^2q^2;$$

4) пять испытаний, B: «только в двух успех»

$$B=A_1A_2\bar{A}_3\bar{A}_4\bar{A}_5+A_1\bar{A}_2A_3\bar{A}_4\bar{A}_5+A_1\bar{A}_2\bar{A}_3A_4\bar{A}_5+$$
$$+A_1\bar{A}_2\bar{A}_3\bar{A}_4A_5+\bar{A}_1A_2A_3\bar{A}_4\bar{A}_5+\bar{A}_1A_2\bar{A}_3A_4\bar{A}_5+$$
$$+\bar{A}_1A_2\bar{A}_3\bar{A}_4A_5+\bar{A}_1\bar{A}_2A_3A_4\bar{A}_5+\bar{A}_1\bar{A}_2A_3\bar{A}_4A_5+$$
$$+\bar{A}_1\bar{A}_2\bar{A}_3A_4A_5;$$

$$P(B)=10p^2(1-p)^3=10p^2q^3;$$

5) n испытаний. B: «только в k успех»

$$B=A_1A_2A_3A_4\ldots A_{k-1}A_k\bar{A}_{k+1}\bar{A}_{k+2}\ldots\bar{A}_{n-2}\bar{A}_{n-1}\bar{A}_n+\ldots$$
$$+\bar{A}_1\bar{A}_2\bar{A}_3\ldots\bar{A}_n\ldots A_{n-1}A_n;$$

$$P(B)=p^k(1-p)^{n-k}+\ldots+p^k(1-p)^{n-k}.$$

Итак

$$P(B)=Kp^k(1-p)^{n-k}=Kp^kq^{n-k},$$

где K--это количество слагаемых. Как это количество найти? Количество слагаемых равно числу способов, какими можно из n опытов выбрать k, в которых произошло событие A. Из комбинаторики мы знаем, что число таких комбинаций равно

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

Теорема

Если производится n независимых испытаний, в каждом из которых событие A появится с вероятностью p, то вероятность того, что событие A появится ровно k раз в n испытаниях, выражается формулой

$$P_n(k)=C_n^kp^k(1-p)^{n-k}$$

или

$$P_n(k)=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}.$$

Эта формула называется формулой Бернулли.

Формула описывает, как распределяются вероятности между возможным числом появления события A в n испытаниях.

Пример.

Допустим, в некоторой местности частота заболевания определенной болезнью A среди коров равна 25% (p=1/4). На ферме случайным образом отбирают корову и смотрят, есть ли у нее заболевание или нет, после этого животное возвращается в стадо. Затем процедура повторяется 7 раз. Найти вероятность того, что из 7-ми проверенных

животных ровно k окажется с заболеванием A (k = 0,1,2,3,4,5,6,7).

Решение:

Здесь $P(A)=p=1/4$, $n=7$, k *– изменяется от 0 до 7. Используя формулу Бернулли, выпишем следующую таблицу.*

k	Формула	Численная формула	$P_n(k)$
0	$C_7^0 p^0 (1-p)^{7-0}$	$\frac{7!}{0!(7-0)!}\left(\frac{1}{4}\right)^0\left(1-\frac{1}{4}\right)^{7-0}$	0,134
1	$C_7^1 p^1 (1-p)^{7-1}$	$\frac{7!}{1!(7-1)!}\left(\frac{1}{4}\right)^1\left(1-\frac{1}{4}\right)^{7-1}$	0,312
2	$C_7^2 p^2 (1-p)^{7-2}$	$\frac{7!}{2!(7-2)!}\left(\frac{1}{4}\right)^2\left(1-\frac{1}{4}\right)^{7-2}$	0,312
3	$C_7^3 p^3 (1-p)^{7-3}$	$\frac{7!}{3!(7-3)!}\left(\frac{1}{4}\right)^3\left(1-\frac{1}{4}\right)^{7-3}$	0,173
4	$C_7^4 p^4 (1-p)^{7-4}$	$\frac{7!}{4!(7-4)!}\left(\frac{1}{4}\right)^4\left(1-\frac{1}{4}\right)^{7-4}$	0,058
5	$C_7^5 p^5 (1-p)^{7-5}$	$\frac{7!}{5!(7-5)!}\left(\frac{1}{4}\right)^5\left(1-\frac{1}{4}\right)^{7-5}$	0,012
6	$C_7^6 p^6 (1-p)^{7-6}$	$\frac{7!}{6!(7-6)!}\left(\frac{1}{4}\right)^6\left(1-\frac{1}{4}\right)^{7-6}$	0,0009
7	$C_7^7 p^7 (1-p)^{7-7}$	$\frac{7!}{7!(7-7)!}\left(\frac{1}{4}\right)^7\left(1-\frac{1}{4}\right)^{7-7}$	0,0001

Значение вероятности $P_n(k)$ при данном n сначала увеличивается при изменении k от 0 до некоторого значения m, а затем уменьшается при изменении k от m до n. Поэтому m называют наивероятнейшим числом наступления успеха в n опытах. Это число m равно целой части числа $(n+1)p$. Если $(n+1)p$ – целое, то наивероятнейшим является также и число (m–1) с той же вероятностью.

Пример.

В нашем случае мы видим, что значение вероятности увеличивается от $k=0$ до $k=2$ и уменьшается от $k=3$ до $k=7$. Выясним какое значение является наивероятнейшим. $m=(n+1)p=(7+1)0,25=2$, поскольку m оказалось целым, то здесь два наивероятнейших значения $k=2$ и $k=1$.

Приведем еще несколько формул.

Вероятность того, что в n испытаниях событие наступит:

менее K раз $P_n(<K)=P_n(0)+P_n(1)+\ldots+P_n(K-1)$

более K раз $P_n(>K)=P_n(K+1)+P_n(K+2)+\ldots+P_n(n)$

не менее K раз $P_n(\geq K)=P_n(K)+P_n(K+1)+\ldots+P_n(n)$

не более K раз $P_n(\leq K)=P_n(0)+P_n(1)+\ldots+P_n(K)$

Пример.

Переформулируем нашу задачу. Пусть в ней требуется найти вероятность того, что из отобраных 7-ми коров не менее 3 страдают заболеванием (т.е. из семи страдают заболеванием либо 3, либо 4, либо 5, либо 6, либо 7 коров). По формуле имеем:

$P_n(\geq K)=P_n(k)+P_n(k+1)+\ldots+P_n(n)$;

$P_n(\geq 3)=P_7(3)+P_7(4)+P_7(5)+P_7(6)+P_7(7)$;

$P_n(\geq 3)=0,173+0,057+0,0115+0,0012+0,00006=0,24276$.

Использование формулы Бернулли при больших значениях n затруднительно. В этом случае в теории вероятностей применяют асимптотические (предельные) формулы, которыми можно заменить формулу Бернулли, не вызывая этим заметных погрешностей в вычислениях.

Функция $\varphi(x)$ называется асимптотическим приближением функции $f(x)$, если

$$\lim_{x \to \infty} \frac{f(x)}{\varphi(x)} = 1 .$$

Формула Пуассона

Часто интерес представляет случай большого числа n и малой вероятности p успеха в одном отдельном испытании. В этом случае удобно воспользоваться приближением Пуассона.

Теорема

Если вероятность p наступления события A в каждом испытании постоянна, близка к нулю, а число независимых испытаний n достаточно велико, то вероятность $P_n(k)$ того, что в n независимых испытаниях событие A наступит k раз, приближенно равна:

$$P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

где $\lambda = np$.

Эта формула называется формулой Пуассона. Обычно приближенную формулу Пуассона применяют, когда $p<0,1$, а $npq<10$.

Функция $\frac{\lambda^k e^{-\lambda}}{k!}$ затабулирована, т.е. имеет таблицу.

Значения функции $\frac{\lambda^k e^{-\lambda}}{k!}$

k	λ			
	1	2	5	8
1	0,36	0,27	0,0337	0,0027
2	0,18	0,27	0,084	0,0107
3	0,06	0,18	0,14	0,028
5	0,0031	0,036	0,175	0,0916

Формула Пуассона используется в задачах, относящихся к редким событиям.

Пример.

Пусть известно, что при изготовлении некоторого препарата брак (количество упаковок, не соответствующих стандарту) составляет 0,2%. Оценить приближенно вероятность того, что среди 1000 наугад выбранных упаковок окажутся три упаковки, не соответствующие стандарту.

Решение:

Выбор каждой очередной упаковки можно рассматривать как независимое испытание. Из условий задачи следует, что $n=1000$ (т.е. велико), а $p=0,002$ (т.е. мало), следовательно, A можно считать редким событием. $\lambda=np=1000*0,002=2<10$.

Воспользуемся приближенной формулой Пуассона или таблицей.

$$P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!};$$

$$P_{1000}(3) = \frac{2^3 e^{-2}}{3!} \approx 0,18 .$$

По таблице: находим ячейку пересечения столбца $\lambda=2$ и строки $k=3$.

Локальная теорема Муавра-Лапласа

Локальная теорема Муавра-Лапласа дает ассимптотическое приближение при большом количестве испытаний и достаточно больших вероятностях события A.

Теорема

Если вероятность p появления события A в каждом испытании постоянна и отлична от нуля и единицы ($0<p<1$), то вероятность $P_n(k)$ того, что событие A появится в n испытаниях ровно k раз, приближенно равна (тем точнее, чем больше n) значению функции

$$P(x) = \frac{1}{\sqrt{npq}}\varphi(x),$$

где

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x = \frac{k - np}{\sqrt{npq}}.$$

Значение функции

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

находят по таблицам.

Функция $\varphi(x)$ четная, поэтому для отрицательных и положительных значений аргумента пользуются одними таблицами, т.е.

$$\varphi(-x) = \varphi(x).$$

Пример.

400 студентов проходят диспансеризацию. Найти вероятность того, что у 80 из них будут найдены отклонения в артериальном давлении, если вероятность появления этого события в каждом случае равна 0,2.

Решение:

$n = 400$; $k = 80$; $p = 0,2$; $q = 0,8$.

По асимптотической формуле находим:

$$P_{400}(80) \approx \frac{1}{\sqrt{400 \times 0.2 \times 0,8}} \times \phi(x) = \frac{1}{8}\phi(x).$$

Вычислим значение x

$$x = \frac{k - np}{\sqrt{npq}} = \frac{80 - 400 \times 0,2}{8} = 0.$$

По таблице находим значение функции $\varphi(0)=0,3989$. Искомая вероятность равна

$$P_{400}(80) \approx \frac{1}{8} \times 0,3989 = 0,04986.$$

Найденная по формуле Бернулли вероятность выявления 80 случаев повышенного артериального давления в 400 наблюдениях составит 0,0498.

Интегральная теорема Лапласа

Если нас интересует вопрос, с какой вероятностью будет обнаружено отклонение АД у группы студентов от 70 до 100 человек, то следует применять интегральную теорему Лапласа.

Теорема

Если вероятность p наступления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность $P_n(k_1,k_2)$ того, что событие A появится в n испытаниях от k_1 до, k_2 раз, приближенно равна определенному интегралу

$$P_n(k_1,k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-\frac{z^2}{2}} dz = \Phi(x'') - \Phi(x'),$$

где $x' = \frac{k_1 - np}{\sqrt{npq}}$, $\quad x'' = \frac{k_2 - np}{\sqrt{npq}}$;

$\Phi(x)$ – функция Лапласа:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz.$$

Для решения задач с применением интегральной теоремы Лапласа пользуются таблицами. В таблицах определяют некоторую функцию $\Phi(x)$. Следует помнить, что данная функция нечетна, а это значит, что $\Phi(-x) = -\Phi(x)$.

Пример.

Рассмотрим наш пример: $k_1=70$; $k_2=100$; $n=400$; $p=0,2$; $q=1-p=0,8$. Следовательно,

$$x' = \frac{70 - 400 \times 0{,}2}{\sqrt{400 \times 0{,}2 \times 0{,}8}} = -1{,}25\,;$$

$$x'' = \frac{100 - 400 \times 0{,}2}{\sqrt{400 \times 0{,}2 \times 0{,}8}} = 2{,}5$$

$$P_n(k_1,k_2) = \Phi(2{,}5) - \Phi(-1{,}25) = \{\text{так как } \Phi(-x) = -\Phi(x)\} =$$
$$= \Phi(2{,}5) + \Phi(1{,}25) = 0{,}49 + 0{,}39 = 0{,}88.$$

Задачи

112. В некоторой местности в сентябре в среднем бывает 12 дождливых дней. Какова вероятность, что из случайно взятых в этом месяце восми дней три дня окажутся дождливыми?

113. Два равносильных шахматиста играют в шахматы. Что вероятнее: выиграть две партии из четырех или три партии из шести (ничьи во внимание не принимаются).

114. Что вероятнее, выиграть у равносильного противника (ничейный исход партии исключен) три партии из четырех или пять из восьми?

115. Монету бросают пять раз. Найти вероятность того, что «герб» выпадет: а) менее двух раз, б) не менее двух раз.

116. Вероятность рождения мальчика равна 0,515, девочки – 0,485. В некоторой семье шестеро детей. Найти вероятность, что среди них не больше 2-х девочек.

117. Изделия некоторого производства содержат 5% брака. Найти вероятность того, что среди пяти, взятых наугад, изделий: а) нет ни одного испорченного; б) будут два испорченных.

118. Всхожесть семян данного сорта растений оценивается с вероятностью, равной 0,8. Какова вероятность того, что из пяти посеянных семян взойдут не менее четырех?

119. В семье пятеро детей. Найти вероятность того, что среди этих детей: а) два мальчика; б) не более двух мальчиков; Вероятность рождения мальчика принять, равной 0,51.

120. Вероятность получения удачного результата при производстве сложного химического опыта равна 2/3. Найти наивероятнейшее число удачных опытов, если общее их количество равно 7.

121. Вероятность попадания в цель при каждом выстреле из орудия равна 0,8. Сколько нужно произвести выстрелов, чтобы наивероятнейшее число попаданий было равно 20?

122. Вероятность попадания в цель при каждом выстреле равна 0,001. Найти вероятность попадания в цель двух пуль и более, если число выстрелов равно 5000.

123. Вероятность того, что любой абонент позвонит на станцию в течение часа, равна 0,01. Телефонная станция обслуживает 800 абонентов. Какова вероятность, что в течение часа позвонят 5 абонентов.

124. Имеется общество из 500 человек. Найти вероятность того, что у двух человек день рождения придется на Новый год. Считать, что вероятность рождения в фиксированный день равна 1/365.

125. Учебник издан тиражом 100000 экземпляров. Вероятность того, что учебник сброшюрован неправильно, равна 0,0001. Найти вероятность того, что тираж содержит ровно пять бракованных книг.

126. Завод отправил на базу 500 изделий. Вероятность повреждения изделия в пути равна 0,002. Найти вероятность того, что в пути будет повреждено изделий: а)ровно три; б) менее трех; в) более трех; г) хотя бы одно.

127. Магазин получил 1000 бутылок минеральной воды. Вероятность того, что при перевозке бутылка окажется разбитой, равна 0,003. Найти вероятность того, что магазин получит разбитых бутылок: а) ровно две; б) менее двух; в)более двух; г)хотя бы одну (принять $e^{-3}=0{,}04979$).

128. Вероятность появления успеха в каждом испытании равна 0,25. Какова вероятность, что при 300 испытаниях успех наступит: а) ровно 75 раз, б) ровно 85 раз?

129. В первые классы должно быть принято 200 детей. Определить вероятность того, что среди них окажется 100 девочек, если вероятность рождения мальчика 0,515.

130. Найти вероятность того, что событие A наступит ровно 70 раз в 243 испытаниях, если вероятность появления этого события в каждом испытании равна 0,25.

131. Вероятность появления события A в каждом из 100 независимых испытаний постоянна и равна $p = 0{,}8$. Найти вероятность того, что событие появится: а) не менее 75 раз и не более 90 раз; б) не менее 75 раз; в) не более 74 раз.

132. Всхожесть семян данного растения равна 0,9. Найти вероятность того, что из 900 посаженных семян число проросших будет заключено между 790 и 830.

133. Вероятность появления успеха в каждом из 400 независимых испытаний равна 0,8. Найти такое положительное число ε, что с вероятностью 0,9876 абсолютная величина отклонения частоты появления успеха от его вероятности 0,8 не превысит ε.

134. Игральную кость бросают 80 раз. Найти приближенно границы, в которых число μ выпадений шестерки будет заключено с вероятностью 0,9973.

135. Найти вероятность того, что событие A наступит 1400 раз в 2400 испытаниях, если вероятность появления этого события в каждом испытании равна 0,6.

136. Вероятность рождения мальчика равна 0,51. Найти вероятность того, что среди 100 новорожденных окажется 50 мальчиков.

137. Вероятность появления положительного результата в каждом из η опытов равна 0,9. Сколько нужно произвести опытов, чтобы с вероятностью 0,98 можно было ожидать, что не менее 150 опытов дадут положительный результат?

138. Два равносильных противника играют в шахматы. Что вероятнее: а) выиграть одну партию из двух или две партии из четырех? б) выиграть не менее двух партий из четырех или не менее трех партий из пяти? Ничьи во внимание не принимаются.

139. а) Найти вероятность того, что событие A появится не менее трех раз в четырех независимых испытаниях, если вероятность появления события A в одном испытании равна 0,4; б) событие B появится в случае, если событие A наступит не менее четырех раз. Найти вероятность наступления события B, если будет произведено пять независимых испытаний, в каждом из которых вероятность появления события A равна 0,8.

140. В семье пятеро детей. Найти вероятность того, что среди этих детей: а) более двух мальчиков; б) не менее двух и не более трех мальчиков. Вероятность рождения мальчика принять равной 0,51.

141. Рабочий обслуживает 1000 машин. Вероятность поломки одной машины в течение одной минуты равна 0,004. Найти вероятность того, что в течение одной минуты сломаются пять машин.

142. Производство дает 1% брака. Какова вероятность того, что из взятых на исследование 1100 изделий забраковано не больше 17?

143. Вероятность появления события в каждом из 2100 независимых испытаний равна 0,7. Найти вероятность того, что событие появится: а) не менее 1470 и не более 1500; б) не менее 1470 раз; в) не более 1469 раз.

144. Монета брошена $2N$ раз (N велико!). Найти вероятность того, что «герб» выпадет ровно N раз.

145. Монета брошена $2N$ раз (N велико!). Найти вероятность того, что «герб» выпадет на $2m$ раз больше, чем цифра.

146. Вероятность появления события в каждом из 21 независимых испытаний равна 0,7. Найти вероятность того, что событие появится в большинстве испытаний.

147. Монета брошена $2N$ раз (N велико!). Найдите вероятность того, что число выпадений "герба" будет заключено между числами $N - \frac{\sqrt{2N}}{2}$ и $N + \frac{\sqrt{2N}}{2}$.

148. Вероятность появления события в каждом из независимых испытаний равна 0,8. Сколько нужно произвести испытаний, чтобы с вероятностью 0,9 можно было ожидать, что событие появится не менее 75 раз?

149. Средняя плотность болезнетворных микробов в одном кубическом метре воздуха равна 100. На пробу взяли 2 дм^3 воздуха. Найти вероятность того, что в нем будет обнаружен хотя бы один микроб.

150. Какова вероятность того, что в столбике из 100 наугад отобранных монет число монет, расположенных "гербом" вверх, будет от 45 до 55?

151. Игральная кость бросается пять раз. Найти вероятность того, что два раза появится число очков, кратное 3.

152. Изделия некоторого производства содержат 5% брака. Найти вероятность того, что среди пяти взятых наугад изделий: а) не окажется ни одного испорченного; б) будут два испорченных изделия.

153. Сколько нужно взять случайных цифр, чтобы вероятность появления среди них цифры 5 была не менее 0,9.

154. В некотором обществе имеется 1% дальтоников. Каков должен быть объем случайной выборки (с возвращением), чтобы вероятность встретить в ней хотя бы одного дальтоника была не менее 0,95?

155. Монета бросается 20 раз. Найти наивероятнейшее число появлений герба.

156. Игральная кость бросается 16 раз. Найти наивероятнейшее число появлений числа очков, кратного трем.

157. На факультете 731 студент. Вероятность рождения студента в данный день равна 1/365. Найти наиболее вероятное число студентов, родившихся 1 января, и вероятность того, что найдутся три студента с одним и тем же днем рождения.

158. В камере хранения ручного багажа 80% всей клади составляют чемоданы, которые вперемешку с другими вещами хранятся на стеллажах. Через окно выдачи были получены

все вещи одного стеллажа в количестве 50 мест. Найти вероятность того, что среди выданных вещей было 38 чемоданов.

159. Если в среднем левши составляют 1%, каковы шансы на то, что среди 200 человек окажется ровно четверо левшей.

160. В некоторой местности в среднем на каждые 100 выращиваемых арбузов приходится один весом не менее 10 кг. Найти вероятность того, что в партии арбузов из этой местности, содержащей 4000 штук, будет: а) ровно три арбуза весом не менее 1 кг каждый; б) не менее двух таких арбузов.

161. В страховом обществе застрахованы 10000 лиц одного возраста и одной социальной группы. Вероятность смерти в течение года для каждого лица равна 0,006. Каждый застрахованный вносит 1 января 12 руб. страховых, и в случае смерти его родственники получают от общества 1000 руб. Найти вероятность того, что: а) общество потерпит убыток; б) общество получит прибыль, не меньше 40 000, 60 000, 80 000 руб.

СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Понятие случайной величины

Мы переходим к изучению еще одного важного понятия теории вероятностей, к понятию "случайная величина".

Чтобы лучше понять это, приведем несколько примеров.

Примеры случайных величин:

1.Число мальчиков, родившихся в течение суток в определенном городе.

2.Значение температуры при неоднократном измерении.

3.Число выпавших "гербов" при пятикратном бросании монеты.

4.При однократном бросании кости количество выпавших очков также является случайной величиной.

В примерах мы можем заметить, что одни случайные величины принимают отдельные изолированные значения, которые мы можем занумеровать или перечислить (1,3,4).

Другие величины, например температура, принимают несчетные значения (значения, которые нельзя пересчитать), т.е. эти значения заполняют некоторый интервал.

Случайной называется величина, которая в результате опыта может принять то или иное возможное значение, неизвестное заранее, но обязательно одно, т.е. величина, значение которой меняется от опыта к опыту случайным образом, называется случайной.

Случайные величины принято обозначать большими латинскими буквами X, Y, а принимаемые ими значения – малыми латинскими буквами x, y.

Различают случайные величины следующих типов: дискретные и непрерывные.

Дискретной случайной величиной называют такую случайную величину, множество возможных значений которой либо конечное, либо бесконечное, но счетное.

Непрерывной случайной величиной называют такую случайную величину, которая может принять любое значение из некоторого конечного или бесконечного интервала.

Операции над случайными величинами

Введем теперь операции над случайными величинами. Пусть имеются две случайные величины X и Y, возможными значениями которых являются $x_1, x_2,\ldots, x_n$ и $y_1, y_2,\ldots, y_n$ соответственно.

Суммой $X + Y$ случайных величин X и Y называется случайная величина Z, возможное значение которой есть x_1+y_1; $x_1+y_2;\ldots;x_i+y_j;\ldots;x_n+y_n$.

Произведением $X \times Y$ случайных величин X и Y называется случайная величина Z, возможное значение которой есть: x_1y_1; x_1y_2; $x_1y_3;\ldots;$ $x_iy_j;\ldots;$ x_ny_m.

Произведением $C \times X$ случайных величин X на постоянную C называется такая случайная величина Z, возможные значения которой есть: Cx_1, $Cx_2,\ldots,$ $Cx_i,\ldots,$ Cx_n.

Пример.

При бросании двух кубиков значения случайной величины Z– «сумма выпавших очков» – находим по формуле $Z = X + Y$.

$X=\{1,2,3,4,5,6\}$, $Y=\{1,2,3,4,5,6\}$,

$Z = \{2,3,4,5,6,7,8,9,10,11,12\}$.

Значения случайной величины Z находятся в неодинаковых условиях. Действительно, значение 2 может получится из пары (1,1), а значение 8 из пяти пар: (2,6), (3,5), (4,4), (5,3), (6,2). Это значит, что 2 и 8 имеют неодинаковую вероятность появления.

$$P(Z = 2) = \frac{1}{36}, \quad \text{а } P(Z = 8) = \frac{5}{36}.$$

Появление тех или иных значений случайной величины можно рассматривать как события, а событиям, как нам известно, соответствуют вероятности. Поэтому возможные значения случайных величин отличаются между собой с вероятностной точки зрения.

Таким образом, перечисление всех возможных значений случайной величины не дает достаточно полного представления о ней. Необходимо знать, как часто могут появляться те или иные значения в результате испытаний, проводящихся в одинаковых условиях, т.е. следует знать вероятности их появления.

Связь между случайной величиной и соответствующей вероятностью называют законом распределения.

Законом распределения случайной величины называется любое правило, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями. Закон распределения может быть представлен в табличном, аналитическом или графическом виде.

Дискретные случайные величины

Для начала рассмотрим дискретную случайную величину.

Пусть случайная величина X имеет n возможных значений $x_1,\ldots,x_n$, каждое из этих значений возможно, но не достоверно, и случайная величина X может принять каждое из них с некоторой вероятностью. В результате опыта величина X примет одно из этих значений, т.е. произойдет одно из полной группы несовместимых событий:

$$X = x_1$$
$$X = x_n$$
$$\ldots$$
$$X = x_n.$$

Обозначим вероятности этих событий: p_1, $p_2,\ldots,p_n$ соответственно, т.е.

$$P(X = x_1) = p_1$$
$$P(X = x_2) = p_2$$
$$\ldots$$
$$P(X = x_n) = p_n;$$

Так как события несовместимые и образуют полную группу, то

$$\sum_{i=1}^{n} p_i = 1,$$

т.е. сумма вероятностей всех возможных значений случайной величины равна 1. Эта суммарная вероятность распределена каким-то образом между отдельными значениями.

Случайная величина будет полностью описана с вероятностной точки зрения, если мы зададим это распределение, т.е. в точности укажем, какой вероятностью обладает каждое из событий, этим мы установим так называемый закон распределения случайных величин.

Ряд распределения дискретной случайной величины

Установим форму, в которой может быть задан закон распределения дискретной случайной величины X.

Простейшей формой задания этого закона является таблица, в которой перечислены возможные значения случайных величин (упорядоченные по возрастающей) и соответствующие им вероятности, где $x_1<x_2<\ldots<x_n$.

X	x_1	x_2	…	x_n
P(X)	p_1	p_2	…	p_n

Такая таблица называется рядом распределения случайных величин X или вариационным рядом.

Чтобы придать ряду распределения более наглядный вид, часто используют его графическое изображение: по оси абсцисс (OX) откладываются возможные значения случайных величин, а по оси ординат (OY) – вероятности этих значений. Для наглядности полученные точки соединяют отрезками прямых.

Такая фигура называется *многоугольником* распределения или *полигоном*.

Многоугольник распределения так же, как и ряд распределения, полностью характеризует случайную величину; он является одной из формул закона распределения.

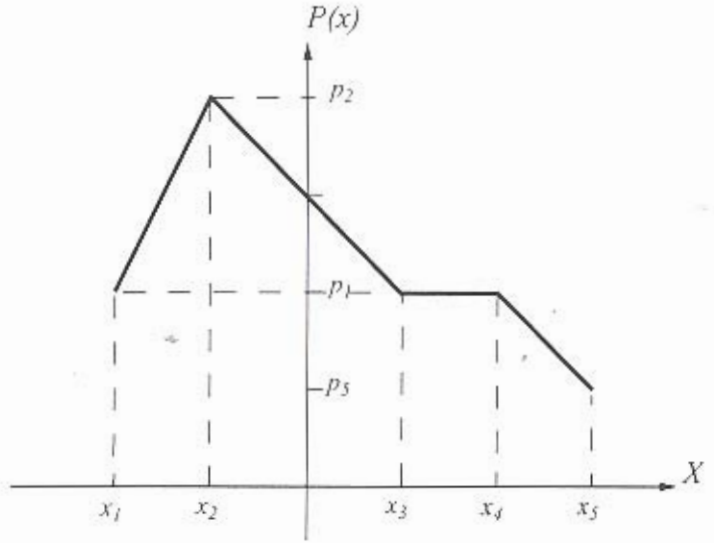


Рис. 6. Полигон распределения

Пример.

Рассмотрим случайную величину Z (сумму выпавших очков при бросании 2-х кубиков), принимающую следующие значения:

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}
Z	2	3	4	5	6	7	8	9	10	11	12
P	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Это вариационный ряд. Просуммируем полученные вероятности.

$$\sum_{i=1}^{11} p_i = \frac{1}{36}+\frac{2}{36}+\ldots+\frac{1}{36}=1.$$

Построим полигон.

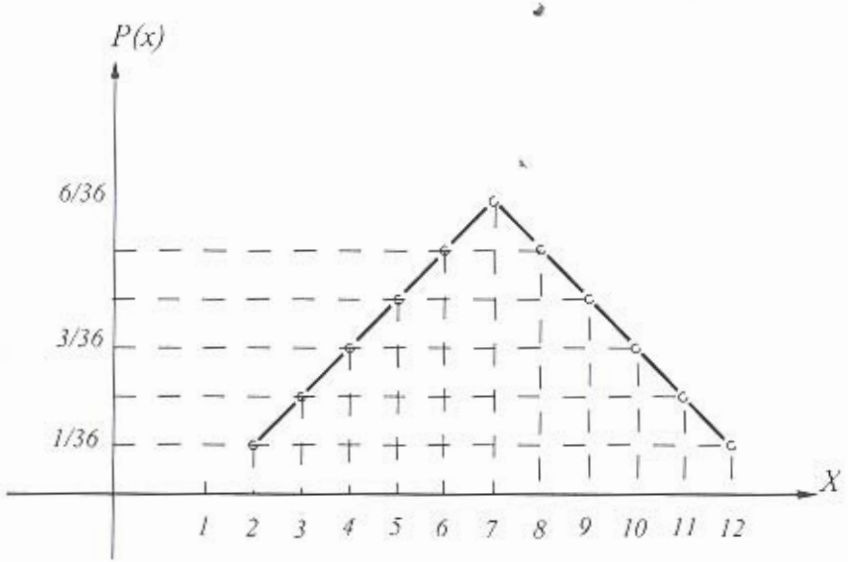


Рис. 7. Полигон распределения к примеру

Как уже говорилось, закон распределения полностью характеризует случайную величину. Выше были рассмотренны табличный и графический способы представления.

Функция распределения случайной величины

Однако непрерывную случайную величину сложно представить с помощью ряда, для этого нужно более универсальное средство. Рассмотрим *аналитический* способ задания.

Функцией распределения случайной величины X называется функция $F(x)$, задающая вероятность того, что случайная величина X принимает значения меньше x, т.е.

$$F(x) = P(X < x).$$

Функция распределения является одной из форм закона распределения. Она – самая универсальная характеристика случайной величины и существует для всех случайных величин как дискретных, так и непрерывных.

Сформулируем некоторые общие свойства функции распределения.

Свойства:

1. Функция $F(x)$ есть неубывающая функция своего аргумента, т.е. $F(x_2) \geq F(x_1)$ при $x_2 > x_1$.
2. На минус бесконечности функция $F(x) = 0$, $F(-\infty) = 0$.
3. На плюс бесконечности функция $F(x) = 1$, $F(+\infty) = 1$.

Последние два свойства вытекают из определения $F(x)$ как вероятности $P(X<x)$.

Для дискретных случайных величин: $x_i < x \ \forall$ i, тогда

$$P(X<x) = P(X=x_1)+...+P(X=x_i) = \sum_{X_i<X} P(X = x_i),$$

(неравенство под знаком суммы означает, что суммирование касается всех тех значений x_i, величина которых меньше x.)

Построим функцию распределения дискретной случайной величины, представленной рядом:

X	x_1	x_2	x_3	....	x_{n-1}	x_n
$P(x)$	p_1	p_2	p_3	....	p_{n-1}	p_n

$$F(x) = \begin{cases} 0 & x \leq x_1 \\ p_1 & x_1 < x \leq x_2 \\ p_1 + p_2 & x_2 < x \leq x_3 \\ p_1 + p_2 + p_3 & x_3 < x \leq x_4 \\ \ldots\ldots\ldots & \ldots\ldots\ldots \\ p_1 + p_2 + p_3 + \ldots + p_{n-1} & x_{n-1} < x \leq x_n \\ \ldots\ldots\ldots & \ldots\ldots\ldots \\ p_1 + p_2 + p_3 + \ldots + p_{n-1} + p_n = 1 & x > x_n \end{cases}$$

Изобразим графическую интерпрстацию.

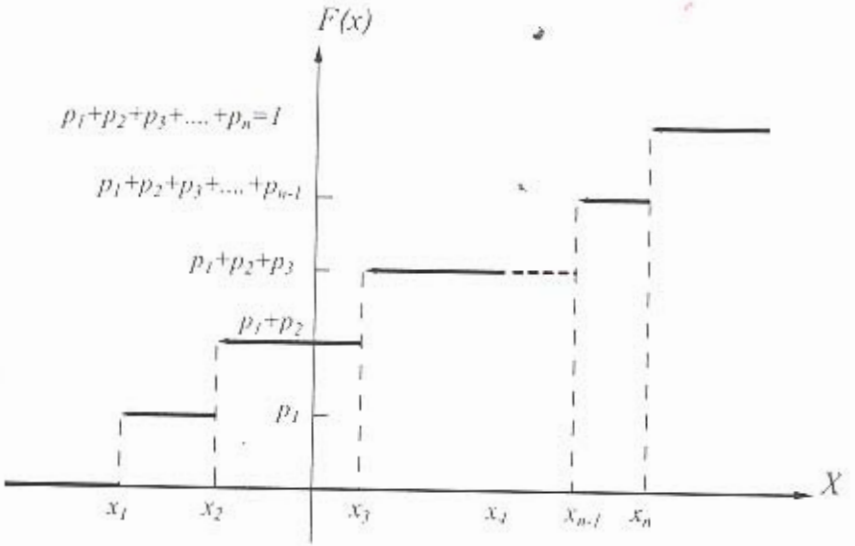


Рис. 8. График функции распределения

Пример.

Эксперимент: подбрасывают монету два раза, случайная величина X – это "число выпадения Герба при двух бросках монеты". Надо построить вариационный ряд, полигон и функцию распределения:

X	x_1	x_2	x_3
P	p_1	p_2	p_3

$$p_1 = C_2^0 (\frac{1}{2})^0 (\frac{1}{2})^2 = \frac{1}{4},$$

$$p_2 = C_2^1 (\frac{1}{2}) (\frac{1}{2}) = 2\frac{1}{4},$$

$$p_3 = C_2^2 (\frac{1}{2})^2 (\frac{1}{2})^0 = \frac{1}{4}.$$

X	0	1	2
P	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Проверяем: $p_1 + p_2 + p_3 = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = \frac{4}{4} = 1$

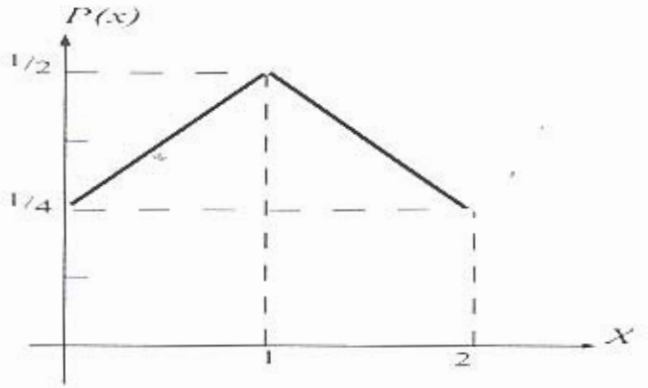


Рис. 9. Полигон

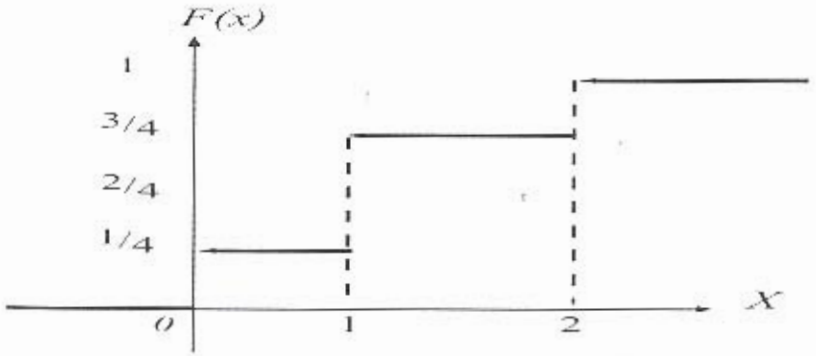


Рис. 10. Функция

Плотность распределения непрерывной случайной величины

Для непрерывной случайной величины нельзя построить ряд распределения. Одним из возможных способов задания непрерывной случайной величины является функция распределения. Если функция распределения непрерывной случайной величины имеет непрерывную производную, то для задания такой величины наряду с функцией распределения можно использовать плотность вероятности.

Плотностью распределения вероятностей непрерывной случайной величины X называют функцию $f(x)$ – первую производную от функции распределения:

$$f(x) = F'(x).$$

Плотность называют *дифференциальной функцией* распределения вероятности, а саму $F(x)$ – *интегральной*.

Свойства плотности распределения вероятностей:

1. Плотность распределения – неотрицательная функция, т.е.

$$f(x) \geq 0 \quad \forall x \in R.$$

2. Интеграл от плотности распределения (несобственный) в пределах от -∞ до +∞ = 1

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

3. Плотность и функция распределения связаны между собой следующим соотношением

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

На практике часто требуется определить вероятность того, что случайная величина примет значение, принадлежащее заданному интервалу (a, b), зная либо плотность $f(x)$, либо функцию $F(x)$.

Если знаем плотность $f(x)$, то вероятность равна

$$P(a \leq x < b) = \int_{a}^{b} f(x)dx.$$

Если знаем функцию распределения $F(x)$, то вероятность равна $P(a \leq x < b) = F(b) - F(a)$.

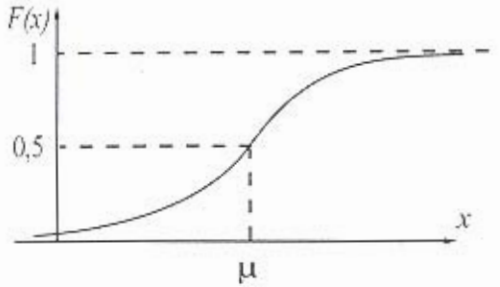


Рис. 11. Функция распределения

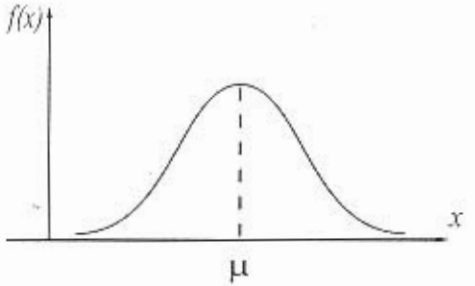


Рис. 12. Функция плотности распределения

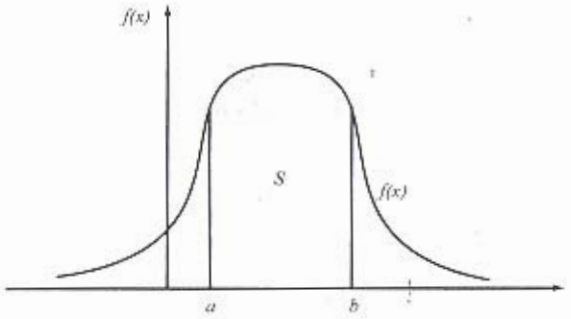


Рис. 13. Значение площади области S соответствует значению вероятности попаданияслучайной величины в интервал (a;b)

Отсюда следует лемма:

Лемма

Вероятность того, что непрерывная случайная величина X примет одно определенное значение равна нулю.

$$P(a \le X \le a) = \int_a^a f(x)dx = F(a) - F(a) = 0.$$

Геометрически полученный результат можно истолковать так: вероятность того, что непрерывная случайная величина примет значение, принадлежащее интервалу (a,b), равное площади криволинейной трапеции, ограниченной осью Ox, кривой $f(x)$ и $x=a$, $x=b$.

Пример.

Функция распределения случайной величины X задана выражением

$$F(x) = \begin{cases} 0 & при \quad x < -\frac{\pi}{4} \\ \frac{1}{2}\sin(x - \frac{\pi}{4}) + \frac{1}{2} & при \quad -\frac{\pi}{4} \le x \le \frac{3}{4}\pi \\ 1 & при \quad x > \frac{3}{4}\pi \end{cases}$$

а) найти вероятность попадания значения случайной величины X в результате опыта в интервал $(a;b)$

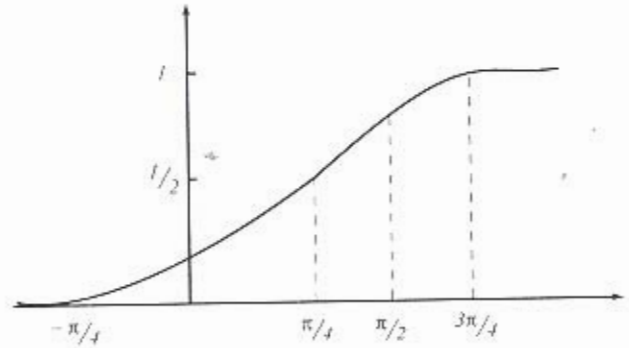
$$P(\frac{\pi}{4} < x < \frac{3}{4}\pi) = \int_{\pi/4}^{3/4\pi} \frac{1}{2}\cos(x - \frac{\pi}{4})dx = \int_0^{\pi/2} \frac{1}{2}\cos y dy = \frac{1}{2}\sin y \Big|_{\frac{\pi}{4}}^{\frac{\pi}{2}} =$$

$$= \frac{1}{2}(\sin(\frac{\pi}{2}) - \sin(0)) = \frac{1}{2}(1-0) = \frac{1}{2}.$$

Мы воспользовались заменой $x - \pi/4$.

$$P(\frac{\pi}{4} < x < \frac{3}{4}\pi) = \frac{1}{2}.$$

б) *Построить график функции в интервале $(\pi/4; 3\pi/4)$*



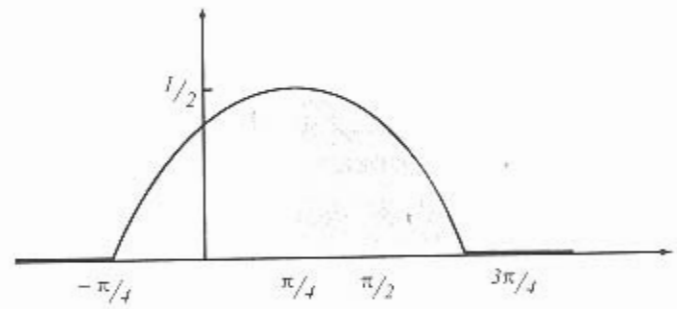


Рис. 14. Графики функции распределения и функции плотности распределения к примеру

Задачи

162. Дискретная случайная величина X задана законом распределения:

X	1	3	6	8
P	0,2	0,1	0,4	0,3

Построить полигон распределения.

163. Дискретная случайная величина X задана законом распределения:

а)

X	2	4	5	6
P	0,3	0,1	0,2	0,4

б)

X	10	15	20
P	0,1	0,7	0,2

Построить полигон распределения, функцию распределения и начертить ее график.

164. Устройство состоит из трех независимо работающих элементов. Вероятность отказа каждого элемента в одном опыте равна 0,1. Составьте закон распределения числа отказавших элементов в одном опыте.

165. В партии из 10 деталей имеются 8 стандартных. Наудачу отобраны 2 детали. Составьте закон распределения числа стандартных деталей среди отобранных.

166. Построить ряд и функцию распределения числа попадания мячом в корзину при 2-х бросках, если вероятность попадания равна 0,4.

167. Из партии в 25 изделий, среди которых имеются 6 бракованных, выбраны случайным образом 3 изделия. Построить ряд распределения случайного числа X бракованных изделий, содержавшихся в выборке.

168. Дан ряд распределения

X	-2	-1	0	1	2
P	0,1	0,2	0,2	0,4	0,1

Требуется: а) построить полигон распределения; б) построить функцию распределения и начертить ее график; в) найти вероятность того, что величина X примет значение, непревосходящее по абсолютной величине 1.

169. В партии 10% нестандартных деталей. Наудачу отобраны 4 детали. Написать биномиальный закон распределения дискретной случайной величины X – числа нестандартных деталей среди 4-х отобранных – и построить полигон полученного распределения.

170. Написать биномиальный закон распределения дискретной случайной величины X – числа появлений герба при двух бросаниях монеты.

171. Случайная величина X принимает значения –1; 0; 1 с вероятностями, соответственно равными ¼; ½; ¼. Написать выражение и построить график функции распределения случайной величины X.

172. Случайная величина задана функцией распределения:

$$f(x) = \begin{cases} 0, & \text{если } x < 2 \\ (x-2)^2, & \text{если } 2 \le x \le 3 \\ 0, & \text{если } x > 3 \end{cases}.$$

Найти: а) плотность вероятности $f(x)$; б) вероятность попадания случайной величины в интервал (1; 2,5).

173. Функция распределения случайной величины задана формулой: $F(x)=A+B*\operatorname{arctg}(x)$ $(-\infty < x < \infty)$. Найти: а) постоянные A и B; б) плотность вероятности $f(x)$; в) вероятность того, что случайная величина попадет в отрезок $[-1;1]$

174. Плотность вероятности непрерывной случайной величины равна:

$$f(x)=\begin{cases} 0, & \text{если} \quad x<1 \\ x-\frac{1}{2}, & \text{если } 1\le x\le 2 \\ 0, & \text{если} \quad x>2 \end{cases}.$$

Построить функцию распределения $F(x)$ и начертить ее график.

175. Непрерывная случайная величина X задана плотностью распределения $f(x) = (3/2)\sin 3x$ в интервале $(0,\frac{\pi}{3})$; вне этого интервала $f(x)= 0$. Найти вероятность того, что X примет значение, принадлежащее интервалу $(\frac{\pi}{6},\frac{\pi}{4})$.

176. Задана плотность распределения непрерывной случайной величины X:

$$f(x)=\begin{cases} 0, & \text{если} \quad x<0 \\ \cos x, & \text{если} \quad 0\le x\le \frac{\pi}{2} \\ 0, & \text{если} \quad x>\frac{\pi}{2} \end{cases}.$$

Найти функцию распределения.

177. Плотность распределения непрерывной случайной величины X в интервале $(0, \frac{\pi}{2})$ равна $f(x)=C\sin 2x$; вне этого интервала $f(x)= 0$. Найти постоянный параметр C.

178. Плотность вероятности непрерывной случайной величины равна:

$$f(x)=\begin{cases} 0, & \text{если} \quad x<0 \\ \frac{1}{2}\sin x, & \text{если} \quad 0\le x\le \pi \\ 0, & \text{если} \quad x>\pi \end{cases},$$

требуется: а) построить функцию распределения $F(x)$; б) найти вероятность того, что в результате испытания случайная величина примет значение, заключенное в интервале $(0,\frac{\pi}{4})$.

179. Задана плотность распределения непрерывной случайной величины X:

$$f(x)=\begin{cases} 0, & \text{если} \quad x<\frac{\pi}{6} \\ 3\sin 3x, & \text{если} \quad \frac{\pi}{6}\le x\le \frac{\pi}{3} \\ 0, & \text{если} \quad x>\frac{\pi}{3} \end{cases}.$$

Найти функцию распределения.

180. Плотность распределения непрерывной случайной величины X задана на всей оси Ox равенством $f(x)=2C/(1+x^2)$. Найти постоянный параметр C.

181. Случайная величина X подчинена закону распределения с плотностью $f(x)$, причем

$$f(x) = \begin{cases} 0, & \text{если} \quad x < 0 \\ a(3x - x^2), & \text{если} \quad 0 \le x \le 3 \\ 0, & \text{если} \quad x > 3 \end{cases},$$

Требуется: а) Найти коэффициент а; б) построить график распределения плотности y = $f(x)$; в) найти вероятность попадания X в промежуток (1,2).

182. Случайная величина X задана функцией распределения (интегральной функцией):

$$F(x) = \begin{cases} 0, & \text{если} \quad x < 1 \\ \dfrac{x-1}{2}, & \text{если} \quad 1 \le x \le 3 \\ 1, & \text{если} \quad x > 3 \end{cases}.$$

Вычислить вероятность попадания случайной величины X в интервалы (1,5; 2,5) и (2,5; 3,5).

183. Случайная величина X задана функцией распределения (интегральной функцией):

$$F(x) = \begin{cases} 0, & \text{если} \quad x < 2 \\ (x-2)^2, & \text{если} \quad 2 \le x \le 3 \\ 1, & \text{если} \quad x > 3 \end{cases}.$$

Найти дифференциальную функцию распределения случайной величины. Вычислить вероятность попадания случайной величины X в интервалы (1;2,5) и (2,5;3,5).

ОСНОВНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Случайные величины измеряются и анализируются в терминах их статистических и вероятностных свойств, главным выразителем которых является функция распределения. Хотя число потенциально возможных моделей распределения велико, относительно небольшое их число находится на особом положении – либо потому, что они обладают желательными математическими свойствами, либо потому, что хорошо описывают какую-то часть действительности, либо в силу обеих причин.

Биномиальное распределение B(n,p)

Дискретная случайная величина X, которая может принимать только целые неотрицательные значения с вероятностью:

$$B(n,x) = C_n^x p^x (1-p)^{n-x},$$

где p>0, $0 \le x \le n$ называется распределенной по биномиальному закону, а p – параметром биномиального распределения.

Биномиальная случайная величина $B(n,p)$ есть число успехов в n независимых испытаниях Бернулли с вероятностью успеха в каждом испытании, равной p.

Ряд распределения случайной величины, подчиненной биномиальному закону, можно представить в следующем виде.

X=x	0	..	x	..	n
B(n,p,x)	$C_n^0 p^0 (1-p)^{n-0}$		$C_n^x p^x (1-p)^{n-x}$		$C_n^n p^n (1-p)^{n-n}$

Функция распределения в этом случае определяется формулой

$$F(x) = \begin{cases} 0 & при \quad x \le 0 \\ \sum\limits_{i<x} C_n^i p^i (1-p)^{n-i} & при \quad 0 < x \le n \\ 1 & при \quad x > n \end{cases}$$

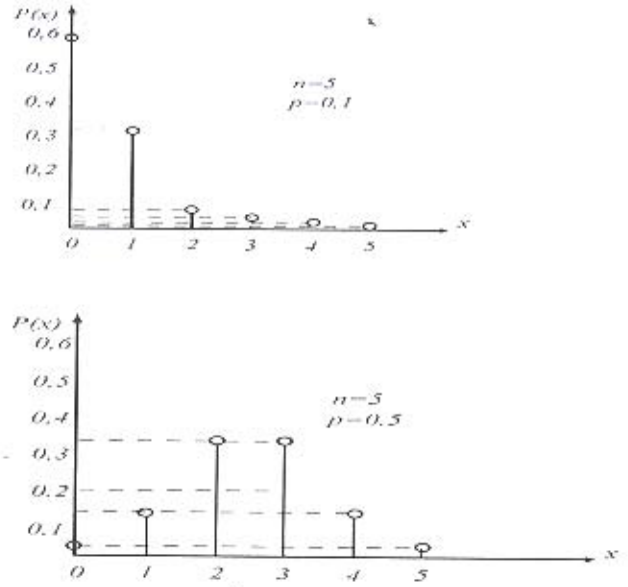


Рис. 15. Биномиальное распределение

Распределение Пуассона

Дискретная случайная величина X, которая может принимать только целые неотрицательные значения с вероятностями

$$P(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

называется распределенной по закону Пуассона с параметром λ, где $0 \le x <= +\infty$ и $\lambda > 0$.

Ряд распределения случайной величины, подчиненной закону Пуассона, можно представить в следующем виде:

X=x	0	1	2	…	x	…
P(x, λ)	$e^{-\lambda}$	$\frac{\lambda e^{-\lambda}}{1!}$	$\frac{\lambda^2 e^{-\lambda}}{2!}$	…	$\frac{\lambda^x e^{-\lambda}}{x!}$	…

Функция распределения в этом случае определяется формулой

$$F(x) = \sum_{i=0}^{(+\infty)} \frac{\lambda^i e^{-\lambda}}{i!}.$$

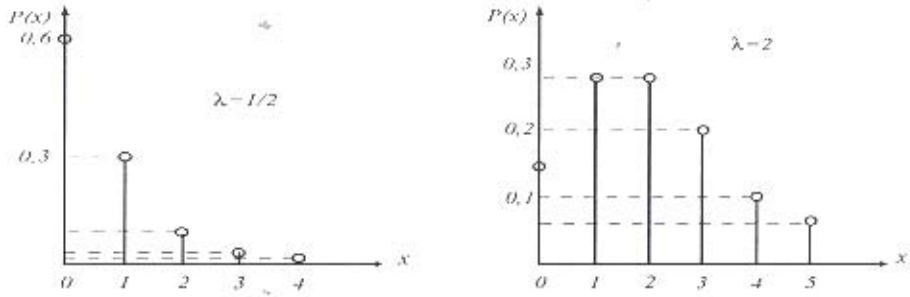


Рис. 16. Распределение Пуассона

Равномерное распределение

В некоторых практических задачах встречаются непрерывные случайные величины, о которых заранее известно, что их возможные значения лежат в пределах некоторого определенного интервала; кроме того, известно, что в пределах этого интервала все значения случайной величины одинаково вероятны (точнее, обладают одной и той же плотностью вероятности). О таких случайных величинах говорят, что они распределены равномерно.

Итак, непрерывная случайная величина X имеет равномерное распределение на отрезке $[a, b]$, если на этом отрезке плотность распределения вероятности случайной величины постоянна, т.е., если функция плотности распределения $f(x)$ имеет следующий вид:

$$f(x;a,b) = \begin{cases} 0 & при \quad x < a \\ C & при \quad a \le x \le b \\ 0 & при \quad x > b \end{cases}.$$

Найдем значение постоянной C.

Воспользуемся нормирующим свойством плотности:

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

$$\int_{-\infty}^{\infty} f(x) = \int_{a}^{b} Cdx = Cx\Big|_{a}^{b} = C(b-a) = 1.$$

$$C = \frac{1}{b-a}.$$

Тогда функцию f(x) можно представить в виде:

$$f(x;a,b) = \begin{cases} 0 & при \quad x < a \\ \dfrac{1}{b-a} & при \quad a \le x \le b \\ 0 & при \quad x > b \end{cases}.$$

График плотности вероятности $f(x)$ равномерного распределения:

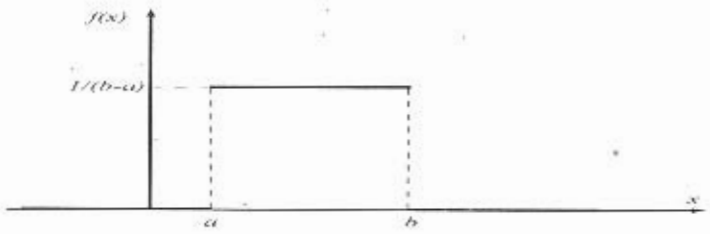


Рис. 17. Равномерное распределение (плотность)

Построим теперь функцию распределения: $F(x)$

$$F(x;a,b) = \begin{cases} 0 & при \quad x < a \\ \dfrac{x-a}{b-a} & при \quad a \le x \le b \\ 1 & при \quad x > b \end{cases}$$

График функции $F(x)$:

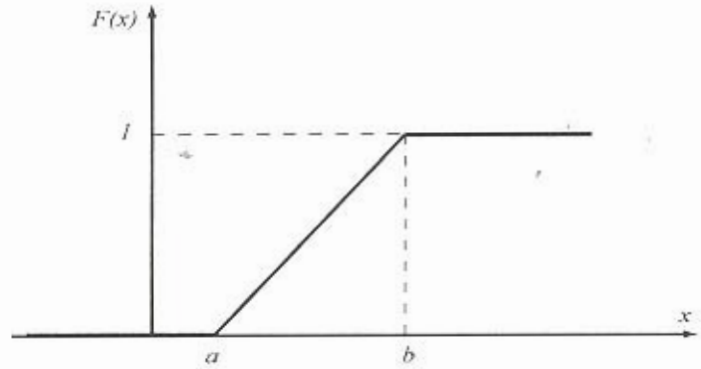


Рис. 18. Равномерное распределение (функция)

Нормальное распределение

Нормальный закон распределения (часто называемый законом Гаусса) играет важную роль в теории вероятностей и занимает среди других законов распределения особое положение. Это – наиболее часто встречающийся на практике закон распределения. Главная особенность, выделяющая нормальный закон среди других законов, состоит в том, что он является предельным законом, к которому приближаются другие законы распределения при весьма часто встречающихся типичных условиях.

Непрерывная случайная величина X имеет нормальное распределение, если плотность распределения вероятности $f(x)$ имеет вид:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ и σ – некоторые постоянные, называемые парамет рами нормального распределения. Обычно обозначают $N(x; \mu, \sigma)$; $N(\mu, \sigma)$.

Кривая плотности распределения нормального закона имеет симметричный холмообразный вид. Максимальная ордината кривой, равная $\frac{1}{\sigma\sqrt{2\pi}}$, соответствует точке μ; по мере удаления от точки μ плотность распределения падает, и при $x \to \pm\infty$ кривая асимптотически приближается к оси абсцисс.

График плотности имеет следующий вид:

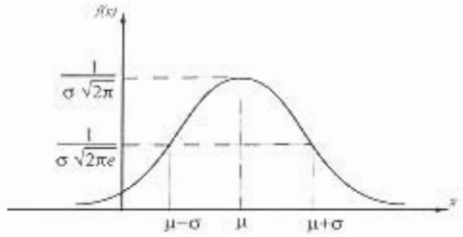


Рис. 19. Нормальное распределение (плотность)

Выясним смысл параметров μ и σ нормального распределения. Центром симметрии распределения является μ. Это ясно из того, что при изменении знака разности (x−μ) на обратный выражение

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

не меняется. Если изменять μ, кривая распределения будет смещаться вдоль оси абсцисс, не изменяя своей формы. Параметр μ характеризует положение распределения на оси абсцисс.

Рассмотрим как изменяется график функции в зависимости от изменения параметров:

а) если изменяется параметр μ, а параметр σ остается постоянным ($\mu_1 < \mu_2 < \mu_3$)

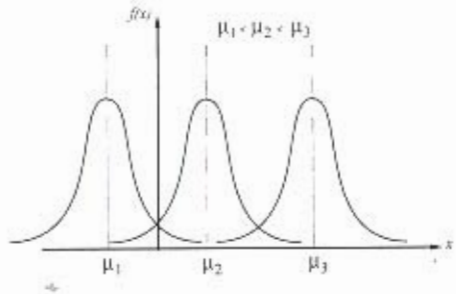


Рис. 20. Нормальное распределение (изменение μ)

Если параметр μ увеличивается (уменьшается), то график сдвигается влево (вправо).

Параметр σ характеризует не положение, а саму форму кривой распределения. Эт.е. характеристика рассеивания. Наибольшая ордината кривой распределения обратно пропорциональна σ; при увеличении σ максимальная ордината уменьшается. Так как площадь кривой распределения всегда должна оставаться равной единице, то при увеличении σ кривая распределения становится более плоской, растягиваясь вдоль оси абсцисс; напротив, при уменьшении σ кривая распределения вытягивается вверх, одновременно сжимаясь с боков, и становится более игловидной. При изменении параметра σ изменяется форма нормальной кривой:

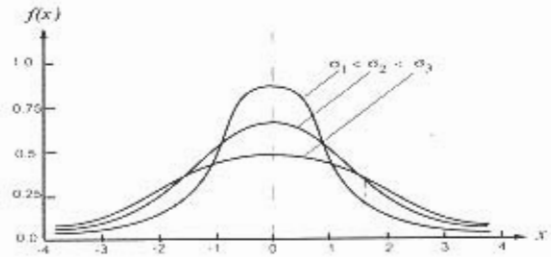


Рис. 21. Нормальное распределение (изменение σ)

На рисунке показаны 3 нормальные кривые при μ=0;

Изменение параметра σ равносильно изменению масштаба кривой распределения – увеличению масштаба по одной оси и такому же уменьшению по другой.

Функция плотности нормального распределения $f(x)$ с параметрами $\mu = 0$, $\sigma = 1$ называется плотностью стандартной нормальной случайной величины, а его график – стандартной кривой Гаусса.

$$N(x;0;1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

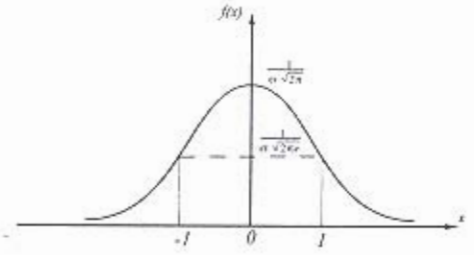


Рис. 22. Функция плотности стандартного нормального распределения

Функция распределения имеет вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Распределение Хи-квадрат (χ^2)

Пусть независимые случайные величины υ_1, υ_2, …, υ_k являются стандартными нормально распределенными величинами (т.е. $\upsilon_i = N(x;0,1)$ для i = 1, 2, …, k). Тогда случайная величина (сумма квадратов)

$$\chi_k^2 = \upsilon_1^2 + \upsilon_2^2 + \ldots + \upsilon_k^2$$

имеет распределение Хи-квадрат с k степенями свободы.

График функции плотности для $k = 4$

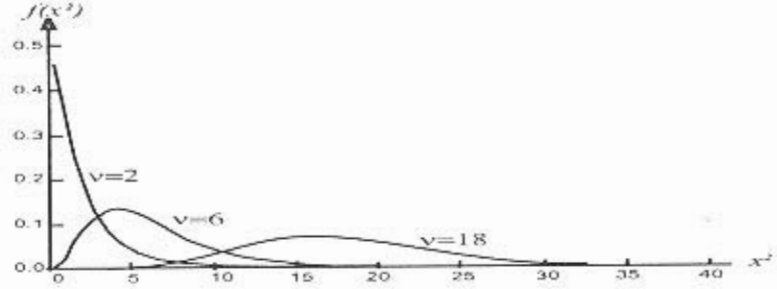


Рис. 23. Функция плотности распределения χ^2

Распределение Фишера (F)

Пусть χ_k^2, χ_l^2 независимые случайные величины, имеющие Хи-квадрат распределения с k и l степенями свободы соответственно. Тогда распределение случайной величины

$$F_{k,l} = \frac{\frac{x_k^2}{k}}{\frac{x_l^2}{l}} = \frac{l}{k} \cdot \frac{x_k}{x_l}$$

называется F-распределением с k и l степенями свободы.

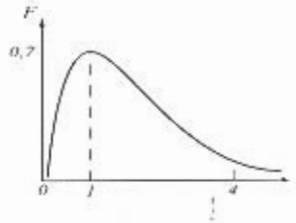


Рис. 24. Функция плотности распределения Фишера

Распределение Стьюдента (t)

Пусть U – является стандартной нормально распределенной случайной величиной, т.е. $U \sim N(x;0,1)$, а χ_k^2 имеет хи-квадрат распределение с k степенями свободы, U и χ_k^2 – независимые величины. Тогда распределение случайной величины

$$t_k = \frac{U}{\sqrt{\frac{\chi_k^2}{k}}},$$

называется t-распределением Стьюдента с k-степенями свободы

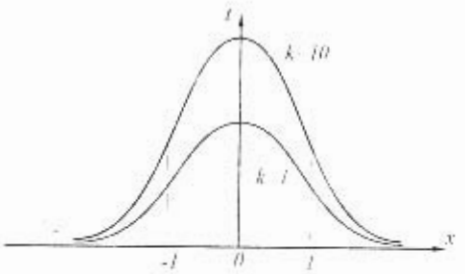


Рис. 25. Функция плотности распределения Стьюдента

ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

Для решения многих практических задач совсем не обязательно знать все возможные значения случайной величины и соответствующие им вероятности, а достаточно указать отдельные параметры, которые позволяют в компактной форме отразить существенные особенности случайной величины.

Эти характеристики случайной величины, являющиеся не функциями, а числами, называют числовыми характеристиками случайной величины. Их назначение – в сжатой форме выразить наиболее важные черты распределения.

Рассмотрим некоторые наиболее важные числовые характеристики и изучим их свойства.

Математическое ожидание

Возможные значения случайной величины могут быть сосредоточены вокруг некоторого центра. Для характеристики такой особенности распределения случайной величины служит математическое ожидание.

Сначала рассмотрим дискретную случайную величину.

Пусть дискретная случайная величина X задана рядом распределения:

X	x_1	x_2	…	x_i	…	x_n
$P(x)$	p_1	p_2	…	p_i	…	p_n

Математическим ожиданием $M[x]$ дискретной случайной величины X называется сумма произведений всех возможных значений случайной величины на соответствующие вероятности значений, т.е.:

$$M[X] = x_1 p_1 + \ldots + x_n p_n = \sum_{i=1}^{n} x_i p_i .$$

Пример.

Дан ряд распределения

X	2	5	8	19
P	0,2	0,3	0,4	0,1

Найдем математическое ожидание случайной величины X.

$$M[X] = 2\cdot0{,}2 + 5\cdot0{,}3 + 8\cdot0{,}4 + 19\cdot0{,}1 = 7.$$

Математическим ожиданием $M[X]$ непрерывной случайной величины X, возможные значения которой распределены по всей оси Ox, называется несобственный интеграл:

$$M[X] = \int_{-\infty}^{\infty} xf(x)dx,$$

где $f(x)$ – плотность вероятности.

Если возможные значения случайных величин распределены в отрезке [a, b], то

$$M[X] = \int_{a}^{b} xf(x)dx .$$

Пример.

Непрерывная случайная величина X задана плотностью распределения:

$$f(x) = \begin{cases} 0 & x < 1 \\ x-1 & 1 \le x < 2 \\ -x+3 & 2 \le x < 3 \\ 0 & x \ge 3 \end{cases}$$

Найти математическое ожидание M[X].

$$M[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{1} x0dx + \int_{1}^{2} x(x-1)dx + \int_{2}^{3} x(3-x)dx + \int_{3}^{+\infty} x0dx =$$

$$= 0 + \left(\frac{x^3}{3} - \frac{x^2}{2}\right)\Big|_1^2 + \left(\frac{3x^2}{2} - \frac{x^3}{3}\right)\Big|_2^3 = 2.$$

Свойства математического ожидания

1. Математическое ожидание постоянной величины равно самой постоянной
$$M[const] = const.$$
2. Постоянный множитель случайной величины может быть вынесен за знак математического ожидания
$$M[const\,X] = const\ M[X].$$
3. Математическое ожидание алгебраической суммы двух случайных величин X и Y равно алгебраической сумме их математических ожиданий
$$M[X \pm Y] = M[X] + M[Y].$$
4. Математическое ожидание произведения двух независимых случайных величин X и Y равно произведению их математических ожиданий
$$M[XY] = M[X]M[Y].$$
5. Математическое ожидание отклонения $X - M[X]$ случайной величины X от ее математического ожидания $M[X]$ равно нулю
$$M[X - M[X]] = 0.$$

Действительно: $M[X - M[X]] = M[X] - M[M[X]] = M[X] - M[X] = 0$.

Дисперсия

На практике встречаются случайные величины, имеющие одинаковые математические ожидания, однако принимающие резко отличающиеся значения. У одних из этих величин отклонения значений от математического ожидания небольшие, а у других, наоборот, значительные, т.е. для одних рассеивание значений случайной величины вокруг математического ожидания мало, а для других – велико. Таким образом, математическое ожидание не полностью характеризует поведение случайной величины. Рассмотрим пример:

X	2	3	4	5
$P(X)$	0,1	0,2	0,3	0,4

$$M[X] = 2\cdot0{,}1+3\cdot0{,}2+4\cdot0{,}3+5\cdot0{,}4 = 4.$$

Y	-1	3	8	11
P(Y)	0,2	0,5	0,2	0,1

$$MY=-1\cdot0{,}2+3\cdot0{,}5+8\cdot0{,}2+11\cdot0{,}1=4.$$

Отложим значения этих величин на числовых осях с одинаковым масштабом

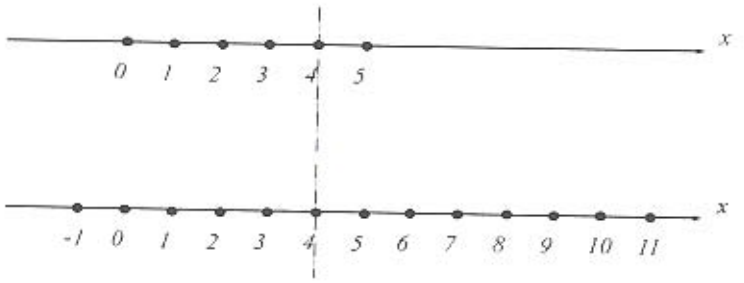


Рис. 26. Случайные величины имеют одинаковое математическое ожидания, но их дисперсии различны.

Рассматриваемые величины имеют одинаковые математические ожидания = 4. Однако рассеивание значений случайной величины X вокруг математического ожидания значительно меньше, чем у величины Y.

Дисперсия является такой характеристикой случайной величины, которая оценивает меру рассеивания значения случайной величины вокруг ее математического ожидания.

Математическое ожидание квадрата отклонения случайной величины X от ее математического ожидания M[X] называют дисперсией случайной величины X и обозначают D[X], т.е.

$$D[X] = M[(X-M[X])^2].$$

Для дискретной случайной величины:

$$D[X] = \sum_{i=1}^{n} (x_i - M[x])^2 P_i .$$

Для непрерывной:

$$D[X] = \int_{-\infty}^{\infty} (x - M[x])^2 \cdot f(x)dx .$$

Среднее квадратическое отклонение

Случайная величина и ее математическое ожидание имеют одну и ту же размерность, но дисперсия имеет размерность квадрата случайной величины. Для наглядной характеристики рассеяния удобнее пользоваться величиной, размерность которой совпадает с размерностью случайной величины. Для этого из дисперсии извлекают квадратный корень. Полученная величина называется средним квадратическим отклонением (иначе «стандартом») случайной величины X. Среднее квадратическое отклонение будем обозначать σ[X]:

$$\sigma[X] = \sqrt{D[X]} .$$

Пример.

Посчитаем дисперсию и среднее квадратическое отклонение СКО для ряда:

X	2	5	8	19
$P(X)$	0,2	0,3	0,4	0,1

Выше было найдено математическое ожидание M[X] = 7.

Найдем дисперсию:

$$D[X] = (2-7)^2\cdot0{,}2+(5-7)^2\cdot0{,}3+(8-7)^2\cdot0{,}4+(19-7)^2\cdot0{,}1 =$$

$$= 5 + 1{,}2 + 0{,}4 + 14{,}4 = 21$$

$$\sigma[X] = \sqrt{21} \approx 4{,}6 .$$

Для непрерывной случайной величины заданной плотностью вероятностей

$$f(x) = \begin{cases} 0 & x < -1 \\ 3x^2 & -1 \le x \le 0 \\ 0 & x > 0 \end{cases}$$

найти дисперсию и СКО

$$M[X] = \int_{-\infty}^{\infty} x f(x)dx = \int_{-1}^{0} x \cdot 3x^2 dx = \int_{-1}^{0} 3x^3 dx = \frac{3}{4}x^4 \Big|_{-1}^{0} = -\frac{3}{4}.$$

$$D[X] = \int_{-\infty}^{\infty} (x - M[X])^2 f(x)dx = \int_{-1}^{0} (x + \frac{3}{4})^2 \cdot 3 \cdot x^2 dx =$$

$$= 3\int_{-1}^{0} (x^4 + \frac{3}{2}x^3 + \frac{9}{16}x^2)dx = 3(\frac{1}{5}x^5 + \frac{3}{8}x^4 + \frac{3}{16}x^3)\Big|_{-1}^{0} = 0{,}037.$$

$$\sigma[X] = \sqrt{0{,}037} = 0{,}19.$$

Свойства дисперсии

1. Дисперсия постоянной величины равна нулю

$$D[const] = 0.$$

2. Постоянный множитель случайной величины можно выносить за знак дисперсии, предварительно возведя его в квадрат

$$D[const\,X] = const^2 D[X].$$

3. Дисперсия алгебраической суммы двух независимых случайных величин X и Y равна сумме дисперсий этих величин

$$D[X \pm Y] = D[X] + D[Y].$$

4. Дисперсия случайной величины X равна разности между математическим ожиданием квадрата случайной величины и квадратом ее математического ожидания.

$$D[X] = M[X^2] - (M[X])^2.$$

Мода

Модой – называется наиболее вероятное значение x_i случайной величины, т.е. это такое значение для которого:

дискретное распределение: $p(x_i)$ – имеет наибольшее значение;

непрерывное распределение: плотность вероятностей $f(x)$ принимает наибольшее значение.

Если случайная величина имеет единственную моду, то распределение называют унимодальным, если две, то бимодальным, если много, то полимодальным.

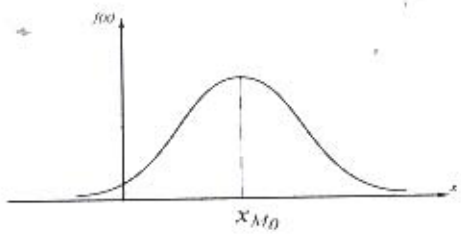


Рис. 27. Унимодальное распределение

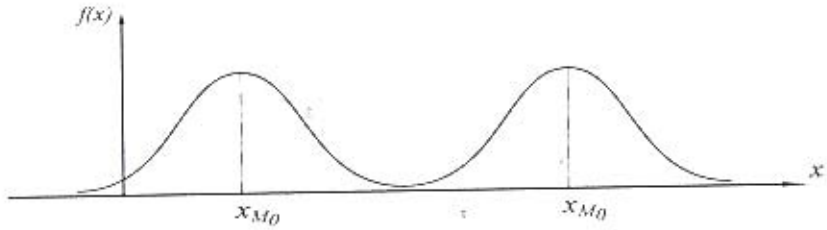


Рис. 28. Бимодальное распределение

Медиана

Этой характеристикой пользуются в основном для непрерывных случайных величин.

Медиана – это такое значение случайной величины X, для которого функция распределения равна $\frac{1}{2}$

$$\int_{-\infty}^{x} f(x)dx = \frac{1}{2}, \text{ где } x=M_e \text{ (медиана).}$$

$$\int_{M_e}^{+\infty} f(x)dx = \int_{-\infty}^{M_e} f(x)dx = \frac{1}{2}.$$

Это означает, что вероятность случайной величины X принять значение меньшее медианы, в точности равна вероятности принять значение большее медианы, т.е. $P(x<M_e) = P(x>M_e)$.

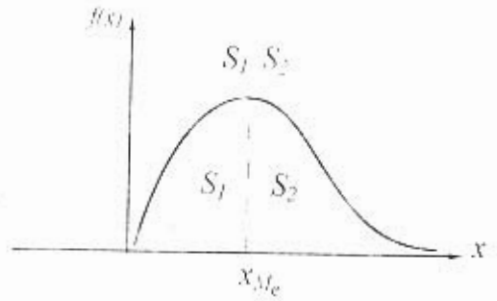


Рис. 29. Медиана делит площадь под кривой распределения на две равные части

Квантили

При описании непрерывных распределений часто используют квантили.

Квантилем, отвечающим заданному уровню вероятности p, называется такое значение x_p случайной величины X, при котором функция распределения принимает значение равное p, т.е.

$$F(x_p)=p \text{ или } \int_{-\infty}^{x_p} f(x)dx = p$$

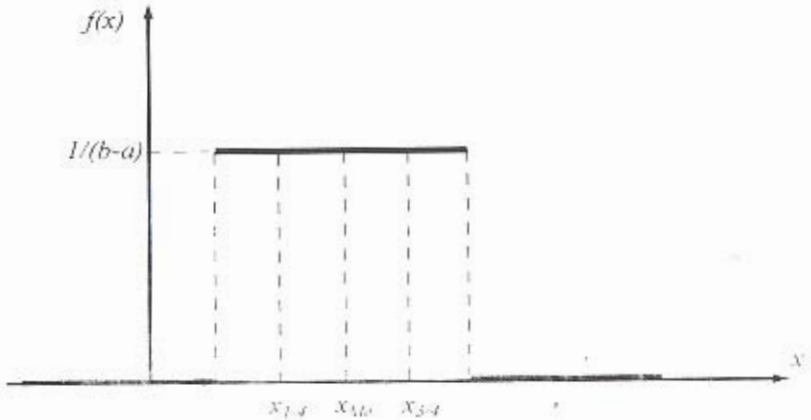


Рис. 30. Квантили распределения

Частные случаи квантилей:

медиана $p = \frac{1}{2}$;

нижний квартиль $p = \frac{1}{4}$;

верхний квартиль $p = \frac{3}{4}$;

процентиль $p = 0,01$; $p = 0,09$ и т.д.

Начальные и центральные моменты

Начальным моментом k-го порядка ν_k случайной величины X называется математическое ожидание X^k, т.е. $\nu_k = M(X^k)$.

Для дискретной случайной величины

$$\nu_k = \sum_{i=1}^{n} x_i^k p_i.$$

Для непрерывной случайной величины

$$\nu_k = \int_{-\infty}^{\infty} x^k f(x)dx.$$

Частные случаи начальных моментов:

начальный момент нулевого порядка равен единице

$$k = 0,\ \nu_0 = 1;$$

начальный момент первого порядка есть математическое ожидание X

$$k = 1, \quad \nu_1 = M[X];$$

центральным моментом k-го порядка μ_k случайной величины X называется математическое ожидание величины $(X - MX)^k$, т.е.

$$\mu_k = M[(X - MX)^k].$$

Для дискретной случайной величины

$$\mu_k = \sum_{i=1}^{n} (x_i - \mu x)^k p_i$$

Для непрерывной случайной величины

$$\mu_k = \int_{-\infty}^{\infty} (x_i - \mu x)^k f(x)dx$$

Частные случаи центральных моментов:

центральный момент нулевого порядка равен единице

$$k = 0,\ \mu_0 = 1;$$

центральный момент первого порядка равен нулю

$$k = 1,\ \mu_1 = 0;$$

центральный момент первого порядка есть дисперсия случайной величины X

$$k = 2, \quad \mu_2 = D[X].$$

Коэффициенты асимметрии и эксцесса

Часто применяются такие числовые характеристики, как асимметрия и эксцесс.

Асимметрия (скошенность).

Коэффициент асимметрии $\gamma_1 = \dfrac{\mu_3}{\sigma^3}$.

Для симметричных распределений $\gamma_1 = 0$, $\gamma_1 > 0$, если мода предшествует медиане; $\gamma_1 < 0$, если мода следует за медианой.

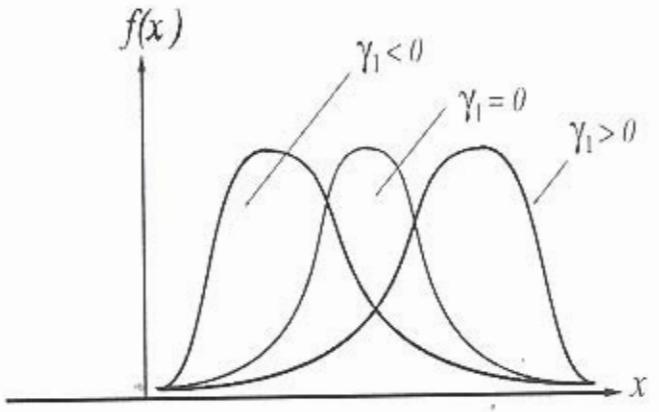


Рис. 31. Изменение коэффициента асимметрии

Эксцесс (крутость) γ_2

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

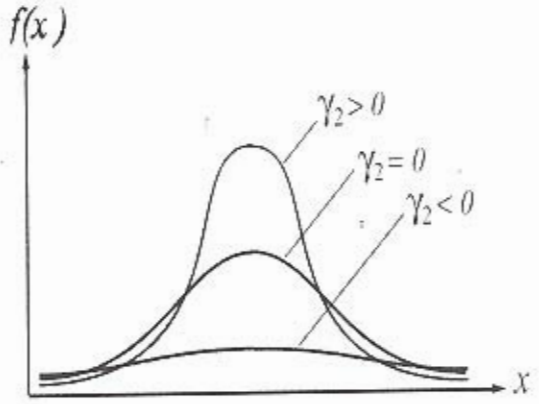


Рис. 32. Изменение коэффициента эксцесса

Распределения и их характеристики.

Числовые характеристики	Математическое ожидание	Дисперсия.	СКО	Мода. Медиана.
Распределение.	M[X]	D[X]	σ_x	
1.Биноминальное.	np	np(1-p)	$\sqrt{n \cdot p(1-p)}$	
2. Пуассона	λ	λ	$\sqrt{\lambda}$	
3. Равномернос	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	$\frac{\lvert b-a \rvert}{\sqrt{12}}$	Медиана $\frac{a+b}{2}$
4) Нормальное Гауссово	μ	σ^2	σ	Мода=μ Медиана=μ
5) Хи-квадрат	k	$2k$	$\sqrt{2k}$	Мода=k-2, $k \geq 2$
6) t-распределение	0	$\frac{k}{k-2}$ k>2	$\sqrt{\frac{k}{k-2}}$	Мода 0

Задачи

184. Найти математическое ожидание дискретной случайной величины X, заданной законом распределения:

а)

X	-4	6	10
P	0,2	0,3	0,5

б)

X	0,21	0,54	0,61
P	0,1	0,5	0,4

185. Найти математическое ожидание случайной величины Z, если известны математические ожидания X и Y. а) $Z = X+2Y$, $M(X) = 5$, $M(Y) = 3$; б) $Z = 3X+4Y$, $M(X) = 2$, $M(Y) = 6$.

186. Дискретная случайная величина X принимает три возможных значения: $x_1 = 4$ с вероятностью $p_1 = 0,5$, $x_2 = 6$ с вероятностью $p_2 = 0,3$ и x_3 с вероятностью p_3. Найти x_3 и p_3, зная, что $M(X) = 8$.

187. В партии из 10 деталей содержится 3 нестандартных. Наудачу отобраны 2 детали. Найти математическое ожидание дискретной случайной величины X – числа нестандартных деталей среди 2-х отобранных.

188. Найти математическое ожидание дискретной случайной величины X, распределенной по закону Пуассона:

X	0	1	2	...	k
P	$e^{-\lambda}$	$\frac{\lambda \cdot e^{-\lambda}}{1!}$	$\frac{\lambda^2 \cdot e^{-\lambda}}{2!}$	...	$\frac{\lambda^k \cdot e^{-\lambda}}{k!}$

189. Случайные величины X, Y независимы. Найти дисперсию случайной величины $Z=3X+2Y$, если известно, $D(X)=5$, $D(Y)=6$.

190. Найти дисперсию и среднее квадратическое отклонение дискретной случайной величины X, заданной законом распределения:

X	-5	2	3	4
P	0,4	0,3	0,1	0,2

191. Найти дисперсию и среднее квадратическое отклонение дискретной случайной величины X, заданной законом распределения:

а)

X	4,3	5,1	10,6
P	0,2	0,3	0,5

б)

X	131	140	160	180
P	0,05	0,10	0,25	0,60

192. Найти среднее квадратическое отклонение случайной величины, заданной законом распределения

ξ	3	5	7	9
P	0,4	0,3	0,2	0,1

193. Найти дисперсию дискретной случайной величины X – числа появлений события A в 5-ти независимых испытаниях, если вероятность появления событий A в каждом испытании равна 0,2.

194. Найти дисперсию дискретной случайной величины X – числа появлений события A в 2-х независимых

испытаниях, если вероятности появления события в этих испытаниях одинаковы и известно, что $M(X) = 1,2$.

195. Найти дисперсию дискретной случайной величины X, распределенной по закону Пуассона :

X	0	1	2	...	k
P	$e^{-\lambda}$	$\frac{\lambda \cdot e^{-\lambda}}{1!}$	$\frac{\lambda^2 \cdot e^{-\lambda}}{2!}$	...	$\frac{\lambda^k \cdot e^{-\lambda}}{k!}$

196. Случайная величина X задана плотностью распределения $f(x)=2x$ в интервале $(0,2)$ и $f(x) = 0$ вне этого интервала. Найти математическое ожидание величины X.

197. Случайная величина X в интервале $(-c, c)$ задана плотностью распределения $f(x) = \frac{1}{\pi\sqrt{c^2 - x^2}}$; вне этого интервала $f(x)=0$. Найти математическое ожидание величины X.

198. Случайная величина X задана плотностью распределения $f(x)=c(x^2+2x)$ в интервале $(0,1)$; вне этого интервала $f(x)=0$. Найти: а) параметр c; б) математическое ожидание величины X.

199. Найти математическое ожидание случайной величины X, заданной функцией распределения

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{4} & 0 \le x \le 4 \\ 1 & x > 4 \end{cases}$$

200. Случайная величина X в интервале $(-c, c)$ задана плотностью распределения $f(x)=\frac{1}{\pi\sqrt{c^2-x^2}}$, вне этого интервала $f(x)=0$. Найти дисперсию X.

201. Найти дисперсию случайной величины X, заданной функцией распределения.

$$F(x)=\begin{cases}0 & x<-2\\ \frac{x}{4}+\frac{1}{2} & -2\le x\le 2\\ 1 & x>2\end{cases}$$

202. Математическое ожидание и дисперсия случайной величины ξ равны соответственно 2 и 10. Найти математическое ожидание и дисперсию величины $2\xi+5$.

203. Случайные величины X, Y не зависимы. Найти дисперсию случайной величины $Z=3X+3Y$, если известно, что $D(X)=4$, $D(Y)=5$.

204. Найти математическое ожидание и дисперсию числа очков, выпадающих при бросании игральной кости.

205. Найти дисперсию дискретной случайной величины X – числа появлений события A в двух независимых испытаниях, если вероятности появления события в этих испытаниях одинаковы и известно, что $M(X)=0,9$.

206. Случайная величина X задана плотностью распределения $f(x)= 0,5x$ в интервале $(0,2)$ и $f(x) = 0$ вне этого интервала. Найти математическое ожидание величины X.

207. Случайная величина X в интервале $(-3; 3)$ задана плотностью распределения $f(x)=\frac{1}{\pi\sqrt{9-x^2}}$; вне этого интервала $f(x) = 0$. а)Найти дисперсию $D[X]$; б) что вероятнее: в результате испытания окажется $X<1$ или $X>1$?

208. Случайная величина X в интервале $(0, \pi)$ задана плотностью распределения $f(x)=(1/2)\sin x$; вне этого интервала $f(x)=0$. Найти дисперсию X.

209. Найти среднее квадратическое отклонение случайной величины, заданной законом распределения.

ξ	3	5	7	9
P	0,4	0,3	0,2	0,1

210. Дан перечень возможных значений дискретной случайной величины X: $x_1 = -1$; $x_2 = 0$; $x_3 = 1$, а также известны математические ожидания этой величины и ее квадрата: $M(X) = 0,1$; $M(X^2)=0,9$. Найти вероятности p_1, p_2, p_3 соответствующие возможным значениям x_1, x_2, x_3.

211. Дан перечень возможных значений дискретной случайной величины X: $x_1=1$; $x_2=2$; $x_3=3$, а также известны математические ожидания этой величины и ее квадрата: $M(X)=2,3$; $M(X^2)=5,9$. Найти вероятности p_1, p_2, p_3, соответствующие возможным значениям x_1, x_2, x_3.

212. Дискретная случайная величина X имеет только два возможных значения x_1 и x_2, причем равновероятных. Доказать, что дисперсия величины X равна квадрату полуразности возможных значений:

$$D(X)=\left[\frac{x_2-x_1}{2}\right]^2.$$

213. Дискретная случайная величина X имеет только два возможных значения: x_1 и x_2, причем $x_2>x_1$. Вероятность того, что X примет значение x_1, равна 0,6. Найти закон распределения величины X, если математическое ожидание и дисперсия известны: $M(X) = 1,4$; $D(X) = 0,24$.

214. Дискретная случайная величина X имеет только два возможных значения: x_1 и x_2, причем $x_2>x_1$. Вероятность того, что X примет значение x_1, равна 0,2. Найти закон распределения величины X, если математическое ожидание $M(X) = 2,6$ и среднее квадратическое отклонение $\sigma(X) = 0,8$.

215. Дискретная случайная величина X имеет только три возможных значения: $x_1=1$, x_2 и x_3, причем $x_1<x_2<x_3$. Вероятность того, что X примет значения x_1 и x_2 соответственно равны 0,3 и 0,2. Найти закон распределения величины X, если ее математическое ожидание и дисперсия известны: $M(X)=2,2$; $D(X)=0,76$.

216. Плотность равномерного распределения сохраняет в интервале (a,b) постоянное значение, равное C; вне этого интервала $f(x) = 0$. Найти значение постоянного параметра C.

217. Закон равномерного распределения задан плотностью вероятности $f(x) = \frac{1}{b-a}$ в интервале (a,b); вне этого интервала $f(x) = 0$. Найти функцию распределения $F(x)$.

218. Найти математическое ожидание случайной величины X, равномерно распределенной в интервале (a,b).

219. Найти дисперсию и среднее квадратическое отклонение случайной величины X, равномерно распределенной в интервале (a,b).

220. Равномерно распределенная случайная величина X в интервале $(a-l,\ a+l)$ задана плотностью распределения

$f(x) = \frac{1}{2l}$; вне этого интервала $f(x)=0$. Найти математическое ожидание и дисперсию X.

221. Случайные величины X и Y независимы и распределены равномерно: X – в интервале (a,b), Y – в интервале (c,d). Найти математическое ожидание произведения XY.

222. Математическое ожидание нормально распределенной случайной величины X равно 3 и среднее квадратическое отклонение равно 2. Найти плотность вероятности X.

223. Написать плотность вероятности нормально распределенной случайной величины X, зная, что $M(X)=3$, $D(X)=16$.

224. Нормально распределенная случайная величина X задана плотностью:

$$f(x) = \frac{1}{5\sqrt{2\pi}} e^{\frac{-(x-1)^2}{50}}.$$

Найдите математическое ожидание, дисперсию и среднее квадратическое отклонение.

225. Найти математическое ожидание случайной величины X, равномерно распределенной в интервале (2, 8).

226. Найти дисперсию и среднее квадратическое отклонение случайной величины X, равномерно распределенной в интервале (2, 8).

227. Случайные величины X и Y независимы и распределены равномерно: X – в интервале (a,b), Y – в интервале (c,d). Найти дисперсию произведения XY.

228. Нормально распределенная случайная величина X задана плотностью: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}$. Найти моду и медиану.

СИСТЕМА СЛУЧАЙНЫХ ВЕЛИЧИН

В практических применениях теории вероятностей часто приходится сталкиваться с задачами, в которых результат опыта описывается не одной, а двумя или более случайными величинами. Эти величины образуют систему случайных величин. Например, параметры-показатели, результаты анализов пациента. Случайными величинами в таком исследовании будут: рост, вес, возраст, температура, давление, содержание химических элементов в крови. Условимся систему нескольких случайных величин обозначать $(X_1, X_2, \ldots, X_n)$.

Многомерная случайная величина. Функция распределения многомерной случайной величины

Систему случайных величин также называют многомерной случайной величиной.

Как и в одномерном случае, для того, чтобы полностью описать многомерную случайную величину существует функция распределения, n – мерной случайной величины, которая определяется формулой:

$$F(x_1, x_2, \ldots, x_n) = \mathrm{P}(X_1 < x_1; X_2 < x_2, \ldots, X_n < x_n).$$

Многомерная функция распределения обладает следующими свойствами:

1. $F(x_1, x_2, \ldots, x_n)$ не убывает по каждому аргументу.
2. Непрерывна слева по каждому аргументу.
3. Стремится к 0, если хотя бы один аргумент стремится к $-\infty$.

$$\lim_{x_i \to -\infty} F(x_1, x_2, \ldots, x_n) = 0, \ (1 \le i \le n).$$

4. Стремится к 1, если все аргументы одновременно стремятся к $+\infty$

$$\lim F(x_1, x_2, \ldots, x_n) = 1, \qquad x_1 \to +\infty; \ldots, x_n \to +\infty.$$

5. Если часть аргументов функции стремится к $+\infty$, то получается функция остальных аргументов

$$\lim F(x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_n) = F(x_1, x_2, \ldots, x_k),$$

если $x_{k+1} \to +\infty; \ldots, x_n \to +\infty$.

Распределение непрерывной многомерной случайной величины можно охарактеризовать плотностью вероятностей:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F}{\partial x_1 \cdot \partial x_n}.$$

Это – n-я смешанная частная производная от функции распределения.

При этом функция распределения выражается через плотность вероятности формулой

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) dt_1, \ldots, dt_n.$$

Свойство функции плотности

Неотрицательная для любых $x_1, x_2, \ldots, x_n$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f(t_1, \ldots, t_n) dt_1, \ldots, dt_n = 1.$$

Двумерные случайные величины

Функцией распределения системы двух случайных величин (X,Y) называется вероятность совместного выполнения двух неравенств $X < x$, $Y < y$

$$F(x, y) = \mathrm{P}[(X < x)(Y < y)].$$

Для понимания удобно воспользоваться геометрической интерпретацией системы. Систему двух случайных величин можно изобразить случайной точкой на плоскости с координатами (X, Y). Тогда функция распределения $F(x, y)$ есть не что иное, как вероятность попадания случайной точки (x, y) в бесконечный квадрант с вершиной в точке (x, y), лежащий левее и ниже ее.

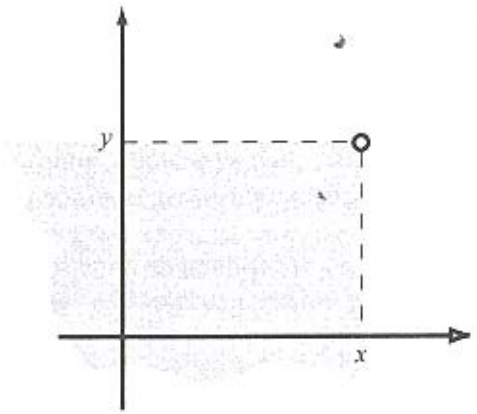


Рис. 33. бесконечный квадрант с вершиной в точке (x,y)

Сформулируем основные свойства для функции двумерной случайной величины:

1. Функция распределения $F(x)$ есть неубывающая функция обоих своих аргументов, т.е.

$$F(x_2, y) \geq F(x_1, y) \quad \text{при } x_2 > x_1,$$
$$F(x, y_2) \geq F(x, y_1) \quad \text{при } y_2 > y_1.$$

2. Если хотя бы один аргумент стремиться к $-\infty$, то $F(x,y)$ стремится к 0

$$F(x,-\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0.$$

3. При одном из аргументов, равном $+\infty$, функция распределения системы превращается в функцию распределения случайной величины, соответствующей другому аргументу

$$F(x, +\infty) = F_2(x),$$
$$F(+\infty, y) = F_2(y),$$

где $F_1(x)$ и $F_2(y)$ – соответственно функции распределения случайной величины X и Y.

4. Если оба аргумента равны $+\infty$, функция распределения системы равна единице

$$F(+\infty, +\infty) = 1.$$

5. Из определения функции и ее свойств можно заключить следующее:

$$0 \leq F(x, y) \leq 1.$$

Как и одномерную, двумерную случайную величину можно задать различными способами: табличным, графическим, аналитическим. Как и одномерные, многомерные случайные величины делятся на дискретные, непрерывные и смешанные.

Для начала рассмотрим дискретную двумерную случайную величину.

Законом распределения дискретной двумерной случайной величины называют перечень возможных значений этой величины, т.е. пар чисел (x,y) и их вероятностей $p(x_i,y_i)$ $(i = 1,\ldots,n;\ j = 1, 2,\ldots,m)$.

Обычно закон распределения дискретной двумерной случайной величины задают в виде таблицы с двойным входом:

Y	X						$F_2(y)$
	x_1	x_2	…	x_i	…	x_n	
y_1	$p(x_1,y_1)$	$p(x_2,y_1)$	…	$p(x_i,y_1)$	…	$p(x_n,y_1)$	$p_{\bullet 1}$
…	…	…	…	…	…	…	…
y_j	$p(x_1,y_j)$	$p(x_2,y_j)$	…	$p(x_i,y_j)$	…	$p(x_n,y_j)$	$p_{\bullet j}$
…	…	…	…	…	…	…	…
y_m	$p(x_1,y_m)$	$p(x_2,y_m)$	…	$p(x_i,y_m)$	…	$p(x_n,y_m)$	$p_{\bullet m}$
$F_1(x)$	$p_{1\bullet}$	$p_{2\bullet}$	…	$p_{i\bullet}$	…	$p_{n\bullet}$	1

где $F_1(x)$, $F_2(y)$ – одномерные функции распределения случайных величин X и Y соответственно, а $p_{i\bullet} = \sum_{j=1}^{m} p(x_i y_i)$ и $p_{\bullet j} = \sum_{i=1}^{n} p(x_i y_j)$.

Так как события $(X=x_i)$, $(Y=y_j)$ i = 1,2, ,n), j = 1, 2, , m) образуют полную группу, то сумма вероятностей $p(x_i, y_j) = 1$, т.е.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i y_j) = 1.$$

Зная закон распределения двумерной дискретной случайной величины, можно найти законы распределения каждой из составляющих. Для этого надо сложить вероятности по строкам или по столбцам соответственно.

Пример.

Найти законы распределения составляющих двумерной случайной величины, заданной законом распределения.

Y	X		
	x_1	x_2	x_3
y_1	0,1	0,3	0,2
y_2	0,06	0,18	0,16

Решение: Найдем закон распределения случайной величины X:

$$p(x_1) = 0,1 + 0,06 = 0,16$$
$$p(x_2) = 0,3 + 0,18 = 0,48$$
$$p(x_3) = 0,2 + 0,16 = 0,36$$

Проверим: $p(x_1)+p(x_2)+p(x_3) = 0,16+0,48+0,36 = 1.$

Теперь найдем закон распределения случайной величины Y:

$$p(y_1) = 0,1+0,3+0,2=0,6,$$
$$p(y_2) = 0,06+0,18+0,16=0,4;$$

проверим $p(y_1)+p(y_2) = 0,6+0,4 = 1$

Итак, мы нашли законы распределения составляющих двумерной случайной величины

X	x_1	x_2	x_3
$p(x)$	0,16	0,48	0,32

Y	y_1	y_2
$p(y)$	0,6	0,4

Непрерывную двумерную случайную величину можно задать, используя плотность распределения.

Плотностью совместного распределения вероятностей $f(x,y)$ двумерной непрерывной случайной величины (X,Y) называют вторую смешанную частную производную от функции распределения.

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \cdot \partial y} = F''_{xy}.$$

Соответственно, если известна плотность распределения, то можно найти функцию

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv.$$

Пример.

Найти плотность совместного распределения f(x,y) системы случайных величин (X,Y) по известной функции распределения

$$F(x,y) = \sin x \sin y \ (0 \le x \le \frac{\pi}{2} \text{ и } 0 \le y \le \frac{\pi}{2}).$$

Найдем частную производную по x от функции распределения:

$$\frac{\partial F}{\partial x} = F'_x = \cos x \cdot \sin y.$$

Найдем от полученного результата частную производную по y. В итоге получаем искомую плотность:

$$\frac{\partial^2 F}{\partial x \partial y} = \cos x \cdot \cos y; \ (0 \le x \le \frac{\pi}{2}; \ 0 \le y \le \frac{\pi}{2}).$$

Зная плотность совместного распределения $f(x,y)$, можно найти функцию распределения $F(x, y)$.

Пример.

Найти функцию распределения двумерной случайной величины по данной плотности совместного распределения.

$$f(x, y) = \frac{1}{\pi^2(1+x^2)(1+y^2)}.$$

Решение:

$$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}\frac{1}{\pi^2(1+x^2)(1+y^2)}dx\cdot dy=$$

$$=\frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}\int_{-\infty}^{x}\frac{1}{(1+x^2)}\,\mathrm{d}x\mathrm{d}y=$$

$$=\frac{1}{\pi^2}\int_{-\infty}^{y}\frac{1}{1+y^2}(\operatorname{arctg}x+\frac{\pi}{2})\mathrm{d}y=\frac{1}{\pi^2}(\operatorname{arctg}x+\frac{\pi}{2})\int_{-\infty}^{y}\frac{1}{(1+y^2)}dy=$$

$$=\frac{1}{\pi^2}(\operatorname{arctg}x+\frac{\pi}{2})(\operatorname{arctg}y+\frac{\pi}{2})=(\frac{1}{\pi}\operatorname{arctg}x+\frac{1}{2})(\frac{1}{\pi}\operatorname{arctg}y+\frac{1}{2}).$$

Свойства плотности.

1. Двумерная плотность вероятности неотрицательная
$$f(x, y)\geq 0.$$
2. Двойной несобственный интеграл с бесконечными пределами от двумерной плотности равен 1
$$\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\mathrm{d}y = 1.$$

Отыскание плотностей вероятности составляющих непрерывной двумерной случайной величины

Пусть известна плотность совместного распределения вероятностей системы 2-х случайных величин. Найдем плотности распределения каждой из составляющих.

Плотность распределения составляющей X:

$$f_1(x)=\int_{-\infty}^{+\infty} f(x, y)dy\,.$$

Плотность распределения составляющей Y:

$$f_2(y)=\int_{-\infty}^{+\infty} f(x, y)dx\,.$$

Итак, плотность распределения одной из составляющих равна несобственному интегралу с бесконечными пределами от плотности совместного распределения системы, причем переменная интегрирования соответствует другой составляющей.

Числовые характеристики случайных величин, входящих в двумерную величину

Числовая характеристика		Для дискретных	Для непрерывных
Математическое ожидание:	MX	$\sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij}$	$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\cdot f(x, y)\mathrm{d}x\mathrm{d}y$
	MY	$\sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij}$	$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\cdot f(x, y)\mathrm{d}x\mathrm{d}y$
Дисперсия	DX	$\sum_{i=1}^{n}\sum_{j=1}^{m} (x_i-MX)^2 p_{ij}$	$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-MX)^2 f(x,y)\mathrm{d}x\mathrm{d}y$
	DY	$\sum_{i=1}^{n}\sum_{j=1}^{m} (y_i-MY)^2\cdot p_{ij}$	$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y-MY)^2 f(x,y)\mathrm{d}x\mathrm{d}y$
Среднее квадратическое отклонение		$\sigma_x=\sqrt{DX}$	
		$\sigma_y=\sqrt{DY}$	

Точка (MX, MY) называется центром рассеивания двумерной случайной величины (X, Y).

Условные законы распределения

Для того, чтобы исчерпывающим образом охарактеризовать систему, недостаточно знать распределение каждой из величин, входящих в систему. Нужно еще знать зависимость между величинами, входящими в систему. Эта зависимость может быть охарактеризована с помощью условных законов распределения.

Условным законом распределения величины X, входящей в систему (X,Y), называется ее закон распределения, вычисленный при условии, что другая случайная величина Y приняла определенное значение y.

Зная закон распределения двумерной случайной величины, можно вычислить условные законы распределения составляющих.

Для дискретных случайных величин:

$$p(x_i|y_j) = \frac{p(x_i y_j)}{p(y_j)}, \quad \text{где} \quad p(y_j) = \sum_{i=1}^{n} p(x_i, y_j),$$

$$p(y_j|x_i) = \frac{p(x_i y_j)}{p(x_i)}, \quad \text{где} \quad p(x_i) = \sum_{j=1}^{m} p(x_i, y_j).$$

Сумма условных вероятностей распределения равна 1.

Пример.

Дискретная двумерная случайная величина задана таблицей

Y	X		
	x_1	x_2	x_3
y_1	0,1	0,3	0,2
y_2	0,06	0,18	0,16

Найти условный закон распределения составляющей X, при условии, что составляющая $Y = y_1$.

Решение:

$$P(x_i|y_1) = \frac{p(x_i y_1)}{p(y_1)}.$$

Найдем $p(y_1)$:

$$p(y_1) = \sum_{i=1}^{3} p(x_i y_1) = 0{,}1+0{,}3+0{,}2=0{,}6,$$

$$p(x_1|y_1) = \frac{0{,}1}{0{,}6} = \frac{1}{6},$$

$$p(x_2|y_1) = \frac{p(x_2 y_1)}{p(y_1)} = \frac{0{,}3}{0{,}6} = \frac{3}{6},$$

$$p(x_3|y_1) = \frac{p(x_3 y_1)}{p(y_1)} = \frac{0{,}2}{0{,}6} = \frac{2}{6}.$$

Проверка:

$$\frac{1}{6}+\frac{1}{3}+\frac{2}{6}=\frac{6}{6}=1.$$

Условный закон распределения.

	$(x_1\|y_1)$	$(x_2\|y_1)$	$(x_3\|y_1)$
$P(X\|y_1)$	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

Для непрерывных случайных величин условный закон распределения можно задавать как функцией распределения, так и плотностью. Условная функция распределения обозначается $F(x|y)$, условная плотность $f(x|y)$.

Если известна плотность совместного распределения $f(x, y)$, то условные плотности составляющих могут быть найдены по формулам:

$$f(x|y) = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dx} = \frac{f(x,y)}{f_2(y)},$$

$$f(y|x) = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy} = \frac{f(x,y)}{f_1(x)},$$

где $f_1(x)$ и $f_2(y)$-функции плотностей составляющих.

Если мы запишем формулы в виде:

$$f(x_1 y) = f_2(y)\cdot f(x|y)$$
$$f(x_1 y) = f_1(x)\cdot f(y|x),$$

то можно заключить, что умножая закон распределения одной из составляющих на условный закон распределения другой составляющей, найдем закон распределения системы случайной величины.

Свойства условной плотности.

1. $f(x|y) \geq 0$ и $f(y|x) \geq 0$.

2. $\int_{-\infty}^{+\infty} f(x \mid y)dx = 1$ и $\int_{-\infty}^{+\infty} f(y \mid x)dy = 1$.

Условное математическое ожидание

Условным математическим ожиданием дискретной случайной величины Y при $X=x$ (x–определенное возможное значение X) называют произведение возможных значений Y на их условные вероятности:

$$M(Y|X = x) = \sum_{j=1}^{m} y_j p(y_j \mid x).$$

Для непрерывных величин

$$M(Y|X = x) = \int_{-\infty}^{+\infty} yf(y \mid x)dy,$$

где $f(y|x)$ – условная плотность случайной величины Y при $X=x$.

Условное математическое ожидание $M(Y|x)$ есть функция от x, которую называют функцией регрессии Y на X.

Аналогично определяются условное математическое ожидание случайной величины X и функция регрессии $M(X|y)$ X на Y.

Пример.

Дискретная двумерная случайная величина задана таблицей:

Y	X			
	$x_1 = 1$	$x_2 = 3$	$x_3 = 4$	$x_4 = 8$
$y_1 = 3$	0,15	0,06	0,20	0,09
$y_2 = 6$	0,3	0,1	0,08	0,02

Найти условное математическое ожидание при $X = x_1 = 1$.

Построить линию регрессии.

Решение:

1) найдем $p(x_1) = \sum_{j=1}^{2} p(x_1 y_j) = 0{,}15+0{,}3 = 0{,}45$.

2) найдем условное распределение вероятности:

$$p(y_1|x_1) = \frac{p(x_1y_1)}{p(x_1)} = \frac{0{,}15}{0{,}45} = \frac{1}{3},$$

$$p(y_2|x_1) = \frac{p(x_1y_2)}{p(x_1)} = \frac{0{,}3}{0{,}45} = \frac{2}{3}.$$

3) условное математическое ожидание:

$$M(Y|x_1) = \sum_{j=1}^{2} y_j p(y_j \mid x_1) = 3 \cdot \frac{1}{3} + 6 \cdot \frac{2}{3} = 5.$$

Аналогично находятся условные математические ожидания для всех значений случайной величины X.

Для построения линии регрессии составим таблицу значений случайной величины X и соответствующие им условные математические ожидания (УМО):

X	1	3	4	8
УМО	5	4,8	3,8	3,5

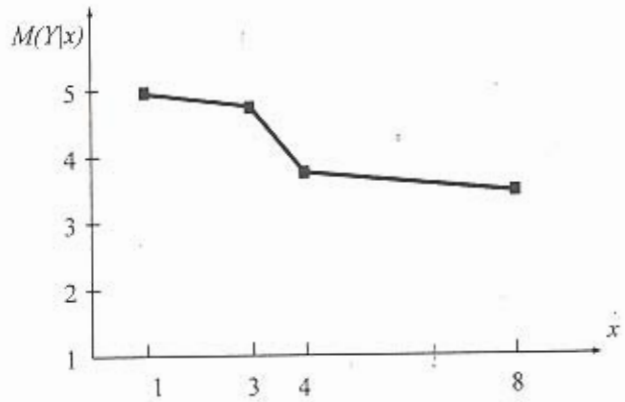


Рис. 34. Линия регрессии Y на X

Зависимые и независимые случайные величины

Необходимые и достаточные условия:

Теорема 1.

Для того, чтобы случайные величины X и Y были независимыми, необходимо и достаточно, чтобы функция распределения системы (X,Y) была равна произведению функций распределения составляющих:

$$F(x,y)=F_1(x)F_2(w).$$

Теорема 2.

Для того, чтобы непрерывные случайные величины X и Y были независимыми, необходимо и достаточно, чтобы плотность совместного распределения системы (X,Y) была равна произведению плотностей распределения составляющих:

$$f(x,y)=f_1(x)f_2(y).$$

Если же случайные величины зависимы, то нам надо каким-то образом охарактеризовать эту зависимость. Такими характеристиками являются ковариация и корреляция.

Коэффициенты ковариации и корреляции

Ковариацией случайной величины X и Y называют число σ_{xy}= cov(X,Y), равное математическому ожиданию произведения отклонений этих величин от своих математических ожиданий.

$$\sigma_{xy} = \mathrm{cov}(X,Y) = \mathrm{M}[(X-\mathrm{M}[X])(Y-\mathrm{M}[Y])].$$

Для дискретных случайных величин X:$(x_1,\ldots,x_n)$, Y:$(y_1,\ldots,y_m)$

$$\mathrm{cov}(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - M[X])(y_j - M[Y])p_{ij}.$$

Для непрерывных случайных величин:

$$\mathrm{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - M[X])\cdot(y - M[Y])\cdot f(x,y)dxdy,$$

где $f(x,y)$ – плотность распределения двумерной случайной величины.

Ковариация может быть также найдена по формуле:

$$\mathrm{cov}(X,Y)=\mathrm{M}(X,Y) - \mathrm{M}[X]\cdot\mathrm{M}[Y].$$

Если X и Y независимы, то cov$(x,y) = 0$. Обратное, однако, не верно.

Свойства ковариации.

1) cov(X,Y) = cov(Y,X);
2) cov(X,X) = D$[X]$;
3) cov$(X+c, Y+c)$=cov(X,Y);
4) cov$(c_1X+c_2Y,Z) = c_1$cov$(X,Z)+c_2$cov(Y,Z), где c_1 и c_2 – константы.

Коэффициент ковариации cov(X,Y) линейно зависит от выбранного масштаба измерения исходных параметров. Нам, однако, нужна характеристика, которая не связана с масштабом измерения исходных параметров. Для получения такой характеристики переходим от исходных случайных величин к нормированным:

$$Z_x = \frac{X - M[X]}{\sqrt{D[X]}}.$$

В качестве безразмерной характеристики зависимости случайных величин X,Y используют коэффициент корреляции ρ_{xy}, равный ковариации нормированных случайных величин:

$$Z_1 = \frac{X - M[X]}{\sqrt{D[X]}};\quad Z_2 = \frac{Y - M[Y]}{\sqrt{D[Y]}};$$

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{D[X]}\cdot\sqrt{D[Y]}} = \frac{\mathrm{cov}(X,Y)}{\sigma[X]\sigma[Y]}.$$

Другими словами коэффициент корреляции – это отношение коэффициента ковариации к произведению среднеквадратических отклонений случайных величин.

Свойство коэффициента корреляции:

$$-1 \le \rho_{xy} \le 1.$$

Для независимой случайной величины $\rho_{xy} = 0$ (обратное не верно).

Пример.

Для предыдущего примера посчитаем коэффициенты ковариации и корреляции.

Y	X				$F(y)$
	$x_1 = 1$	$x_2 = 3$	$x_3 = 3$	$x_4 = 8$	
$y_1 = 3$	0,25	0,2	0,13	0,1	0,68
$y_2 = 6$	0,15	0,1	0,05	0,02	0,32
$F(x)$	0,4	0,3	0,18	0,12	1

$M[X] = 1\cdot0,4+3\cdot0,3+4\cdot0,18+8\cdot0,12=2,98$; $D[X] = 4,8$; $\sigma_x = 2,18$;
$M[Y] = 3\cdot0,68=6\cdot0,32=3,96$; $D[Y] = 1,95$; $\sigma_y = 1,4$;

$$\mathrm{cov}(X,Y) = \sum_{i=1}^{4}\sum_{j=1}^{2}(x_i - M[X])(y_j - M[Y])\cdot p_{ij} = (x_1 - M[X])\cdot(y_1 - M[Y])\cdot p_{11} + (x_1 -$$

$$-M[X])\cdot(y_2 - M[Y])\,p_{12} + (x_2 - M[X])\cdot(y_1 - M[Y])\,p_{21} +$$
$$+(x_2 - M[X])(y_2 - M[Y])\,p_{22} + \ldots$$

$\mathrm{cov}(X,Y) = -0,43$.

$$\rho_{xy} = \frac{\mathrm{cov}(x,y)}{\sigma_y \sigma_x} = \frac{-0,43}{1,4\cdot 2,18} = -0,14\,.$$

Если $\rho_{xy} < 0$, то говорят, что корреляция отрицательная или обратная.

Если $\rho_{xy} > 0$, то корреляция положительная или прямая.

Ковариационная и корреляционная матрицы

Ковариационной матрицей случайных величин $x_1, x_2, \ldots, x_n$ называется матрица Σ, элементами которой являются ковариации $\sigma_{ij} = \mathrm{cov}(x_i x_j)$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}.$$

Из свойств cov следует, что ковариационная матрица является симметричной $\sigma_{ij} = \sigma_{ji}$, а ее диагональные элементы равны дисперсиям.

$$\Sigma = \begin{pmatrix} D[X_1] & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & D[X_2] & \ldots & \sigma_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{n1} & \sigma_{n2} & \ldots & D[X_n] \end{pmatrix}.$$

Если случайные величины не зависимы, то

$$\Sigma = \begin{pmatrix} D[X_1] & 0 & \ldots & 0 \\ 0 & D[X_2] & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & D[X_n] \end{pmatrix}.$$

Корреляционной матрицей случайной величины $x_1, \ldots, x_n$ называется матрица R, элементами которой является коэффициент корреляции ρ_{ij}

$$R = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{pmatrix}.$$

Корреляционная матрица является симметричной $\rho_{ij} = \rho_{ji}$.

Диагональные элементы корреляционной матрицы равны единице.

Задачи

Система случайных величин. Регрессия. Ковариация и корреляция.

229. Задано распределение вероятностей дискретной двумерной случайной величины:

Y	X		
	3	10	12
4	0,17	0,13	0,25
5	0,10	0,30	0,05

Найти законы распределения составляющих X и Y.

230. Задана функция распределения двумерной случайной величины:

$$F(x,y)=\begin{cases}1-2^{-x}-3^{-y}+3^{-x-y} & x\ge 0, y\ge 0\\ 0 & x<0, y<0\end{cases}$$

Найти двумерную плотность вероятности системы.

231. Задана функция распределения двумерной случайной величины:

$$F(x,y)=\begin{cases}(1-e^{-4x})(1-e^{-2y}) & x>0, y>0\\ 0 & x<0, y<0\end{cases}$$

Найти двумерную плотность вероятности системы.

232. Задана дискретная двумерная случайная величина (X,Y):

Y	X	
	3	6
10	0,25	0,10
14	0,15	0,05
18	0,32	0,13

Найти: а) условный закон распределения составляющей X при условии, что составляющая Y приняла значение $y = 10$; б) условный закон распределения составляющей Y при условии, что составляющая X приняла значение $x = 6$.

233. Задана дискретная двумерная случайная величина (X,Y):

Y	X		
	2	5	8
0,4	0,15	0,3	0,35
0,8	0,05	0,12	0,03

Найти: а) безусловные законы распределения составляющих; б) условный закон распределения составляющей X при условии, что составляющая Y приняла значение $y = 0,4$; в) условный закон распределения составляющей Y при условии, что составляющая X приняла значение $x = 5$; г) найти условные математические ожидания $M(X|y = 0,4)$ и $M(Y|x = 5)$; д) построить регрессию случайной величины Y на X; е) найти коэффициенты ковариации и корреляции.

234. Задано распределение вероятностей дискретной двумерной случайной величины:

Y	X			
	26	30	41	50
2,3	0,05	0,12	0,08	0,04
2,7	0,09	0,3	0,11	0,21

Найти: а) безусловные законы распределения составляющих; б) условный закон распределения составляющей X при условии, что составляющая Y приняла значение y=2,3; в) условный закон распределения составляющей Y при условии, что составляющая X приняла значение x = 41; г) найти условные математические ожидания $M(X|y = 2,3)$ и $M(Y|x= 41)$; д) построить регрессию случайной величины Y на X; е) найти коэффициенты ковариации и корреляции.

МЕДИЦИНСКАЯ СТАТИСТИКА

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

Цель науки – описание, объяснение и предсказание явлений действительности на основе установленных законов. В основе научных знаний лежит наблюдение. Для обнаружения закономерности, которой подчиняется явление, необходимо многократно наблюдать это явление в одинаковых условиях.

Многие явления взаимно связаны и влияют одно на другое. Проследить все связи и определить влияния каждого фактора на явление не всегда представляется возможным. Поэтому ограничиваются изучением влияния лишь основных факторов, определяющих течение явления. Тогда под одинаковыми условиями наблюдений понимается соблюдение практически одинаковых значений основных факторов.

Статистическое исследование состоит из следующих стадий: наблюдение; сводка и группировка результатов наблюдения; анализ полученных обобщающих показателей.

Все стадии связаны между собой, и на каждой из них используются специальные методы, объясняемые содержанием выполняемой работы.

Статистика – это наука, изучающая методы обработки результатов наблюдений массовых случайных явлений, обладающих закономерностью, с целью выявления этой закономерности.

Исходя из характера и основных черт предмета статистики как науки, можно сформулировать следующие ее задачи: изучение структуры, взаимосвязей и динамики массовых явлений и процессов.

Статистика, как наука, исследует не отдельные факты, а массовые явления и процессы, выступающие как множество от-

дельных факторов, обладающих как индивидуальными, так и общими признаками.

Выводы о закономерностях, которым подчиняются явления, изучаемые статистикой, всегда основываются на ограниченном, выборочном числе наблюдений. При большем или меньшем числе наблюдений эти выводы могут оказаться иными. Для вынесения более определенного заключения о закономерностях явлений статистика опирается на теорию вероятностей.

Статистика имеет дело с результатами наблюдений случайных явлений, а теория вероятностей логически изучает закономерности и имеет дело с моделями случайных явлений. Обработав результаты наблюдений, исследователь выдвигает ряд гипотез (предположений) о том, что рассматриваемое явление можно описать той или иной вероятностной теоретической моделью. Далее, используя математико-статистические методы, можно дать ответ на вопрос, какую из гипотез или моделей следует принять. Именно эта модель и считается закономерностью изучаемого явления. Правомерен такой выбор или нет покажет практика использования выбранной модели.

Статистика, опираясь на вероятностные модели, влияет на развитие теории вероятностей. Окружающий мир многообразен, и задачи, возникающие при изучении случайных явлений, при обработке результатов наблюдения над ними требуют разработки новых вероятностных моделей. Статистика и теория вероятностей – это две неразрывно связанные науки, они влияют друг на друга, развивают друг друга.

Генеральная совокупность и выборка

Объект статистического исследования в статистике называют статистической совокупностью.

Статистическая совокупность – это множество единиц, обладающих однородностью. Каждый отдельно взятый элемент данного множества называется единицей совокупности. Единицы статистической совокупности характеризуются общими свойствами, именуемыми в статистике признаками. Под однородностью совокупности понимается сходство единиц (объектов, явлений, процессов) по каким-либо существенным признакам, но различающихся по каким-либо другим признакам. По форме внешнего

выражения признаки делятся на качественные и количественные. Единицы совокупности наряду с общими для всех единиц признаками, обусловливающими качественную определенность совокупности, также обладают индивидуальными особенностями и различиями, отличающими их друг от друга. Именно наличие вариации предопределяет необходимость статистики.

Зачастую реально существующую совокупность объектов можно мысленно дополнить любым количеством таких же однородных объектов. Возьмем, к примеру, лекарства, выпущенные в первом квартале. Эту совокупность можно дополнить лекарствами, выпущенными во втором, третьем и так далее кварталах. Такие совокупности называются генеральными.

Итак, совокупность всех мыслимо возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины, или совокупность результатов всех мыслимых наблюдений, проводимых в неизменных условиях над одной из случайных величин, связанных с данным видом объектов, называется генеральной совокупностью.

Генеральная совокупность может быть конечной или бесконечной в зависимости от того, конечна или бесконечна совокупность составляющих ее элементов.

Не следует смешивать понятие генеральной совокупности с реально существующими совокупностями. Если на склад поступила продукция некоторого фармацевтического предприятия – это является реально существующей совокупностью, которую нельзя назвать генеральной, поскольку выпуск этого лекарства можно мысленно продолжить сколь угодно долго.

Статистическое наблюдение – это источник первичной статистической информации. Оно сводится к сбору данных о массовых явлениях путем регистрации их признаков. Статистическое наблюдение должно проводиться по заранее составленному плану: должны быть определены цели, объект, единица наблюдения, программа (перечнь вопросов, на которые надо получить ответы, и набор гипотез, которые надо проверить). От правильно организованного и хорошо продуманного наблюдения зависят полнота получаемых данных и точность выводов в результате обработки собранных данных. Следует особое внимание обратить на составление статистического формуляра бланка, в котором регистрируются сведения о единицах наблюдения, и на составления ин-

струкций – письменных разъяснений по вопросам заполнения статистических формуляров и организации наблюдения.

Зачастую невозможно провести сплошное обследование (это либо дорого, либо приводит к уничтожению исследуемого объекта). Поэтому приходится из всей совокупности объектов для обследования отбирать только часть, т.е. проводить выборочное обследование.

Например, на фармацевтическом предприятии надо проверить партию лекарства на качество. Каждое лекарство приходится вскрывать, т.е. портить товар. Следовательно, сплошное обследование невозможно. Поэтому берут небольшую часть лекарственной продукции и проверяют на качество. По полученным результатам можно судить о качестве всей продукции, не приводя в негодность всю партию лекарств.

Часть отобранных объектов генеральной совокупности называется выборочной совокупностью или выборкой.

Число N объектов генеральной совокупности называют объемом генеральной совокупности, а число n объектов выборочной совокупности – объемом выборки. N значительно больше, чем n.

Однако не всякая выборка может быть действительным представлением о генеральной совокупности.

Для того, чтобы по выборке можно было сделать правильные выводы о всей генеральной совокупности, она должна быть репрезентативной. Это значит, что все пропорции генеральной совокупности должны быть представлены в выборке. Репрезентативность выборки обеспечивается случайностью отбора. Это означает, что любой объект выборки отобран случайно, при этом все объекты имеют одинаковую вероятность попасть в выборку.

Существуют несколько способов отбора, обеспечивающих репрезентативность. Обычно поступают следующим образом, все объекты генеральной совокупности нумеруют (по возможности), после чего карточки с номерами перемешиваются и из полученной пачки выбирают одну наудачу. Объект, номер которого совпал с номером карточки, считается попавшим в выборку. Такую операцию повторяют до тех пор пока не образуется необходимая выборка. При этом существуют два различных варианта выборки: случайная повторная и случайная бесповторная.

При случайной повторной выборке каждая вынутая карточка возвращается в пачку. При случайной бесповторной выборке карточки в пачку не возвращаются.

При большом объеме генеральной совокупности применение карточек для организации случайной выборки затруднено. В таких случаях используют таблицы или датчик случайных чисел.

Если объем генеральной совокупности велик, то различие между выборками с возвратом и без возврата незначительно и практически не сказывается на окончательных результатах.

Эмпирическая функция распределения и гистограмма

Полученные результаты представляют собой множество беспорядочных данных. Для изучения их подвергают обработке.

Следующим этапом статистического исследования является сводка, суть которой в обработке первичных материалов наблюдения в целях получения итоговых или упорядоченных определенным образом числовых характеристик изучаемой совокупности. Основным моментом сводки является группировка, т.е. объединение статистических данных в однородные по определенным признакам группы. Группировки помогают изучать структуру совокупности, взаимосвязь между явлениями.

Изучение структуры совокупности достигается построением рядов распределения, характеризующих распределение единиц совокупности по одному признаку. Распределение единиц совокупности по количественному признаку называют вариационным рядом. Ряд может быть построен как по дискретному, так и по непрерывному признаку.

Дискретным называется признак, который может принимать определенные значения из конечного набора таких значений, выражаемых, как правило, целыми числами, например, число детей в семье.

Непрерывный признак может принимать любые промежуточные значения. Как правило, при построении вариационных рядов по непрерывному признаку последний указывается в виде интервалов «от и до», и ряд называется интервальным.

Кроме обычных частот в вариационном ряду можно рассчитывать нарастающим итогом накопленные (кумулятивные)

частоты, по которым строим суждение о том, какое число единиц в совокупности обладает значением признака «не более» или «не менее» определенного.

Для наглядности вариационные ряды изображают графически с помощью полигона (преимущественно дискретные ряды) и гистограммы (интервальные ряды).

Операция, заключающаяся в том, что результаты наблюдений над случайной величиной, т.е. наблюдаемые значения, располагают в порядке неубывания, называется ранжированием опытных данных. После ранжирования опытные данные легко объединить в группы, т.е. сгруппировать так, что в каждой отдельной группе значения случайной величины будут одинаковые. Значение случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется вариантом, а изменение этого значения – варьированием. Варианты обозначают буквами конца латинского алфавита *x,y,z*.

Для каждой группы сгруппированного ряда данных можно подсчитать их численность, т.е. определить число, которое показывает, сколько раз встречается соответствующий вариант в ряде наблюдений.

Численность отдельной группы сгруппированного ряда наблюдаемых данных называется частотой или весом соответствующего варианта и обозначается m_i.

Практический интерес представляет относительная частота варианта.

Отношение частоты данного варианта к общей сумме частот всех вариантов называется долей этого варианта и обозначается p_i, где i – индекс варианта.

$$\overline{p}_i = \frac{m_i}{\sum_{j=1}^{\nu} m_j},$$

где ν – число вариантов. Так как объем выборки $n = \sum_{j=1}^{\nu} m_j$, то

$$\overline{p}_i = \frac{m_i}{n}.$$

Заметим, что доля p_i является статистической вероятностью появления варианта x_i.

Подсчитав частоты и доли для каждого варианта, представим наблюдения в виде таблицы, где в первой строке расположены индексы вариантов i, во второй – варианты x_i, в третьем – частоты m_i, в четвертой доли p_i.

Индекс	i	1	2	3	…	i	…	n
Вариант	x_i	x_1	x_2	x_3	…	x_i	…	x_n
Частота	m_i	m_1	m_2	m_3	…	m_i	…	m_n
Доля	$\overline{p}_i$	$\overline{p}_1$	$\overline{p}_2$	$\overline{p}_3$	…	$\overline{p}_i$	…	$\overline{p}_n$

Полученная таблица называется дискретным вариационным рядом. Причем варианты расположены в порядке возрастания.

Дискретным вариационным рядом распределения называется ранжированная совокупность вариантов x_i с соответствующими им частотами m_i и долями p_i.

Данный ряд считается выборочным аналогом ряда распределения и $\sum_{i=1}^{\nu} \overline{p}_i = 1$.

Если изучаемая случайная величина является непрерывной, то ранжирование и группировка затруднены. Нецелесообразно это и для дискретных случайных величин, число возможных значений которой велико. В подобных случаях следует построить интервальный ряд. Для построения такого ряда весь интервал варьирования наблюдаемых значений случайной величины разбивают на частичные интервалы и подсчитывают частоту попадания значений величины в каждый частичный интервал.

Интервальным вариационным рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами (или долями) попаданий в каждый из них значений величины.

Для построения интервального ряда необходимо определить величину частичных интервалов, на которые разбивается весь интервал варьирования наблюдаемых значений случайной величины. Длину частичного интервала Δx выбирают следующим образом: находят размах варьирования

$$R = x_{\max} - x_{\min},$$

затем делят размах на количество интервалов ν, т.е. получается

$$\Delta x = \frac{x_{\max} - x_{\min}}{\nu}.$$

Количество интервалов может быть выбрано по усмотрению исследователя. При графическом представлении распределения наблюдений нашей целью является выбор интервалов группировки таким образом, чтобы основные, характерные черты распределения оказались выделенными, а случайные колебания были бы сглажены. Если длина интервала группировки мала, то влияние случайных колебаний начинает преобладать, так как каждый интервал содержит при этом лишь небольшое число наблюдений, если же длина интервала велика, то скрадываются основные характерные черты распределения.

Иногда интервальный вариационный ряд для простоты исследования условно заменяют дискретным. В этом случае серединное значение i-го интервала принимают за вариант x_i, а соответствующую интервальную частоту – за частоту варианта.

Мы разобрали выборочный аналог теоретического вариационного ряда. Теперь разберем выборочные аналоги для интегральной и дифференциальной функций распределения, а также рассмотрим полигон и гистограмму.

Пусть имеется выборочная совокупность объема n значений некоторой случайной величины X. И каждому варианту в этой совокупности поставлена в соответствие его доля $\overline{p}$. Пусть далее x – некоторое действительное число, а m_x – количество выборочных значений случайной величины X, меньших x. Тогда число $\frac{m_x}{n}$ является долей наблюдаемых в выборке значений величины X, меньших x, т.е. долей появления события A = (X<x). При изменении x в общем случае будет изменяться и величина $\frac{m_x}{n}$. Это означает, относительная частота $\frac{m_x}{n}$ является функцией аргумента x. А так как эта функция находится по выборочным данным, которые были получены в результате опытов, то ее называют выборочной, или эмпирической.

Выборочной функцией распределения называется функция $F(x)$, задающая для каждого значения x относительную частоту события (X<x).

Итак, по определению $\tilde{F}(x) = \frac{m_x}{n}$, где m_x – число выборочных значений, меньших x, n – объем выборки.

Функцию генеральной совокупности $F(x)$ называют теоретической функцией распределения, а функцию выборки $\tilde{F}(x)$ – эмпирической. Их отличие в том, что $F(x)$ определяет вероятность события (X<x), а выборочная $\tilde{F}(x)$ – относительную частоту этого события.

$\tilde{F}(x)$ – обладает теми же свойствами, что и $F(x)$:

1. $0 \le \tilde{F}(x) \le 1$;
2. $\tilde{F}(x)$ – неубывающая;
3. $\tilde{F}(-\infty) = 0$;
4. $\tilde{F}(+\infty) = 1$.

$\tilde{F}(x)$ можно задать и аналитически

$$\tilde{F}(x) = \begin{cases} 0 & x \le x_{\min} \\ \sum\limits_{j=1}^{i-1} \tilde{p}_j & x_{i-1} < x \le x_i \quad i = 1, \cdots, \nu \\ 1 & x > x_{\max} \end{cases}$$

Частоты $\sum\limits_{j=1}^{i-1} \tilde{p}_j$ называются накопленными (кумулятивными) частотами.

Выборочным аналогом плотности $f(x)$ считают функцию

$$\tilde{f}(x) = \frac{\tilde{F}(x + \Delta x) - \tilde{F}(x)}{\Delta x},$$

где $\tilde{F}(x + \Delta x) - \tilde{F}(x)$ частость попадания наблюдаемых значений случайной величины X в интервал $[x; x + \Delta x)$.

Выборочную функцию плотности можно задать соотношением

$$\tilde{f}(x)=\begin{cases}0 & x\le x_1\\ \dfrac{\tilde{p}_i}{n} & x_i<x\le x_{i+1},\quad i=1,\cdots,\nu,\\ 0 & x>x_{\nu+1}\end{cases}$$

где $\tilde{p}_i$ – доля попадания случайной величины в интервал $[x;x+\Delta x)$, где Δx – длина частичного интервала, $x_{\nu+1}$ – конец последнего ν-го интервала.

Также наблюдаемые данные можно представить в виде графических изображений: полигона, гистограммы, графика функции.

Полигон обычно используется для дискретного ряда- это точки с координатами $(x_i; m_i)$ или $(x_i; \tilde{p}_i)$, затем точки соединяются. Это выборочный аналог полигону теоретического распределения.

Гистограмма – обычно используется для интервальных вариационных рядов. Для построения гистограммы в прямоугольной системе координат на оси OX откладывают отрезки, изображающие частичные интервалы варьирования, и на этих отрезках, как на основаниях строят прямоугольники с высотами, равными частотам или долям соответствующих интервалов. В результате получаем ступенчатую фигуру, которую называем гистограммой.

Заметим, что если мы используем доли, то сумма площадей всех прямоугольников, построенных на частичных интервалах, равна единице.

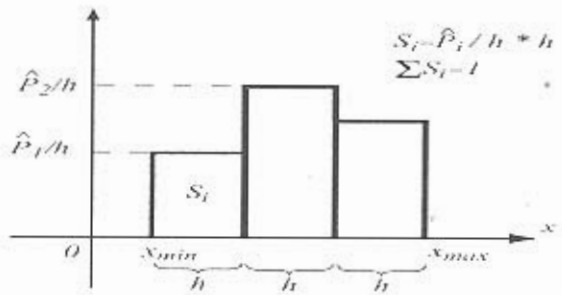


Рис. 35. Гистограмма. Сумма площадей прямоугольников равна 1

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Точечное оценивание

Как уже говорилось, наиболее полной и исчерпывающей характеристикой для случайной величины является закон распределения: функция распределения, ряд распределения, плотность. Некоторые законы распределения имеют параметры, которые однозначно их определяют. Естественно, возникает задача оценки этих параметров.

В общем случае задача оценки параметров распределения сводится к нахождению статистик. *Статистики* – это функции от выборочных значений. Статистики могут быть использованы для приближенного определения значений параметров, по которым судят о виде распределения.

Существуют два метода оценки параметров распределения случайных величин: точечный и интервальный.

Точечными называют оценку, которая определяется одним числом. *Интервальная* определяется двумя числами (концами интервала).

Рассмотрим точечные оценки: обозначим через θ оцениваемый параметр теоретической функции распределеня, а через $\tilde{\theta}$ оценку этого параметра. Для точечных оценок сформулирован ряд требований. Они должны быть состоятельными, несмещенными и эффективными.

Состоятельность – это сходимость по вероятности оценки к оцениваемому параметру при неограниченном возрастании объема наблюдения

$$\lim_{n\to+\infty} P\left(\left|\tilde{\theta}-\theta\right|<\varepsilon\right)=1\quad\forall\varepsilon>0.$$

Другими словами, чем больше объем выборки, тем ближе мы к истине.

Несмещенность – это отсутствие систематической погрешности. Математическое ожидание несмещенной оценки должно быть равно оцениваемому параметру:

$$M[\tilde{\theta}] = \theta .$$

Вообще говоря, не всякая состоятельная оценка будет несмещенной. Требование несмещенности особенно важно при малом объеме выборки.

Эффективной называется оценка, которая имеет минимальную дисперсию в классе несмещенных оценок. Эффективная оценка всегда состоятельна.

Приведем примеры точечных оценок для математического ожидания и дисперсии.

Пусть $x_1, x_2, x_3, \ldots, x_n$ – это n независимых наблюдений величины X, тогда для оценки математического ожидания случайной величины используют статистику $\overline{X}$:

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

Эта оценка является состоятельной, несмещенной и эффективной.

Приведем три статистики для оценки дисперсии случайной величины:

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

эта оценка является состоятельной, но смещенной;

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

эта оценка является состоятельной, несмещенной, но она не является эффективной;

$$s_0^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n},$$

эта оценка является состоятельной, несмещенной и эффективной, однако эта оценка практически не используется так как в нее входит математическое ожидание μ, которое, как правило, заранее не известно.

Для получения точечных оценок используют метод максимального правдоподобия и метод наименьших квадратов.

Также по выборочным данным можно оценить моду, медиану и другие характеристики случайной величины.

Отличие медианы и моды от средней арифметической заключается в том, что эти величины не зависят от крайних вариант и степени рассеяния ряда.

Медиана – это серединная варианта, делящая вариационный ряд пополам, на две равные части. Таким образом, медиана находится на центральном месте, от которого отстоит одинаковое число и больших, и меньших, чем медиана вариант. В ряду с четным числом наблюдений в центре находятся две варианты, тогда за медиану принимается их полусумма.

Мода – чаще всего встречающаяся или наиболее часто повторяющаяся величина.

В симметричном ряду (т.е. теоретически правильном, имеющем одинаковое число вариант, отличающихся от средней в большую и меньшую сторону, чаще применяются в санитарной статистике) средняя арифметическая, мода и медиана совпадают, поэтому нет необходимости вычислять все три, достаточно вычислить среднюю арифметическую. Прибегать к медиане и моде приходится при наличии асимметричных рядов (чаще встречаются в экспериментальных и клинических исследованиях).

На рисунке изображена резко асимметричная кривая распределения по длительности болезни умерших от рака прямой кишки. У подавляющего числа больных летальные исходы наступили в ранние сроки, но в отдельных случаях продолжительность болезни составила 96 – 104 и более месяцев. Эти нетипичные случаи «отягощают» среднюю арифметическую, которая равняется 25,6 мес., в то время как мода составила 10,4, а медиана 20,7 мес. Очевидно, что наиболее характерной для данного явления средней величиной служит мода.

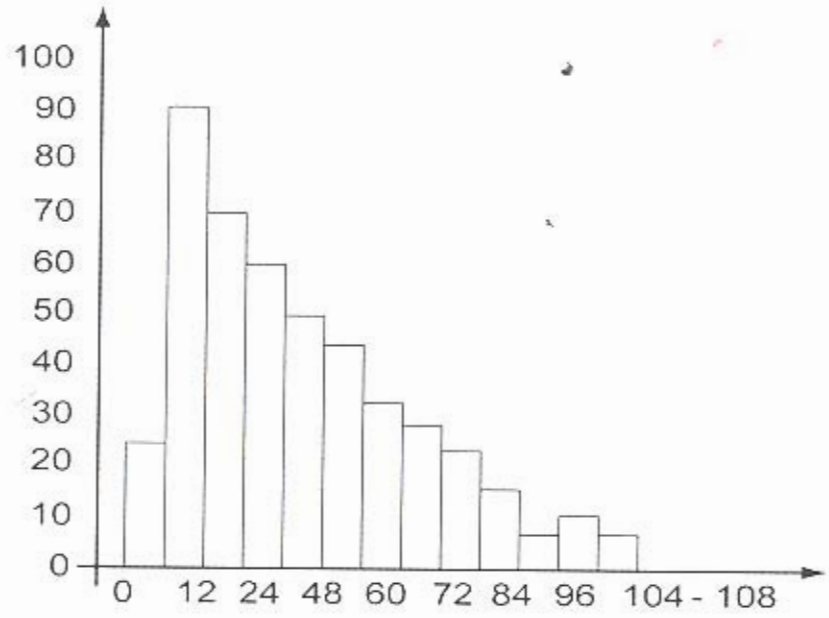


Рис. 36. Асимметричное распределение

Таким образом, различия в применяемых средних могут быть отражены в следующих определениях: средняя арифметическая является результативной суммой всех влияний, в ее формировании принимают участие все варианты, без исключения, в том числе и крайние варианты, имеющие нетипичный характер. Медиана и мода в отличие от средней арифметической не зависят от величины всех индивидуальных значений, т.е. всех членов вариационного ряда, а обусловливаются относительным расположением или распределением вариант. Поэтому, медиану и моду даже называют описательными или позиционными средними, так как они характеризуют главнейшие свойства данного распределения. Средняя арифметическая характеризует всю массу наблюдений, без исключений; медиана и мода – основную массу, без учета воздействия крайних вариант, зависящих иногда от случайных причин.

В примере нас интересует не столько средний срок длительности течения болезни, сколько тот срок, до которого практически остается в живых наибольшее число больных, т.е. модальный срок.

Бимодальный ряд распределения внушает подозрение в его неоднородности, в том, что две вершины ряда образовались в результате смешения качественно различных совокупностей. Так, например, две моды могут быть получены при изучении признаков физического развития школьников без учета их пола (одна мода характеризует мальчиков, другая – девочек).

Интервальное оценивание параметров распределений

Интервальная оценка определяется двумя числами (концами интервала).

Теория точечных оценок не дает возможности сделать заключение об их точности. В этом отношении оценки неизвестных параметров существенно дополняются результатами интервального оценивания с помощью доверительных интервалов.

Всякая статистическая оценка параметров, определенная по данным выборки с помощью выбранной статистической характеристики, может быть только приближенной. Поэтому она может иметь определенный смысл лишь в том случае, когда указываются границы возможной погрешности оценки или, иначе говоря, указывается интервал, который с известной вероятностью (надежностью) покрывает оцениваемое постоянное значение параметра.

Интервальные оценки в основном используются для выборок небольшого объема.

Процедура построения интервальной оценки:

Обозначим через θ оцениваемый параметр, $\overline{\theta}$ - точечная оценка для θ.

1. По сделанной выборке находится точечная оценка $\overline{\theta}$ неизвестной характеристики θ.
2. Затем задаются вероятностью γ (обычно 0,95; 0,99 и т.д.), которая отражает надежность нашей оценки.

3.По определенным правилам находят такое положительное число ε, чтобы выполнялось соотношение

$$P(\overline{\theta} - \varepsilon < \theta < \overline{\theta} + \varepsilon) = \gamma$$

или

$$P(|\overline{\theta} - \theta| < \varepsilon) = \gamma .$$

Число ε называется точностью оценки, γ – доверительной вероятностью, а интервал $(\overline{\theta} - \varepsilon; \overline{\theta} + \varepsilon)$ - интервальной оценкой.

Соотношение $P(\overline{\theta} - \varepsilon < \theta < \overline{\theta} + \varepsilon) = \gamma$ следует читать так: «Вероятность того, что доверительный интервал $(\overline{\theta} - \varepsilon; \overline{\theta} + \varepsilon)$ накроет характеристику θ, равна γ».

Поскольку довольно часто встречаются нормально распределенные случайные величины, построим интервальные оценки для параметров нормального распределения: математического ожидания и среднего квадратического отклонения.

Обозначим через X случайную величину, имеющую нормальный закон распределения с параметрами μ и σ ($X=N(\mu,\sigma)$). Будем предполагать, что наблюдения этой величины не зависимы и проводятся в одинаковых условиях.

Интервальная оценка математического ожидания нормального распределения при известной дисперсии

По наблюдениям найдем точечную оценку математического ожидания

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

Зададимся вероятностью γ.

Найдем такое число ε, чтобы выполнялось соотношение

$$P(\overline{X} - \varepsilon < \mu < \overline{X} + \varepsilon) = \gamma .$$

Из-за сложности выкладки опускаются. Приведем готовый результат

$$\varepsilon = \frac{u_\gamma \cdot \sigma}{\sqrt{n}},$$

здесь u_γ находится из соотношения

$$\Phi(u_\gamma) = \frac{\gamma}{2},$$

где $\Phi(u_\gamma)$ – функция Лапласа,

$$\Phi(x) = \int_0^x e^{-\frac{z^2}{2}} dz .$$

Таблица значений функции Лапласа.

γ	0,9	0,91	0,92	0,93	0,94
u_γ	1,65	1,7	1,76	1,81	1,88

γ	0,95	0,96	0,97	0,98	0,99
u_γ	1,96	2,06	2,18	2,34	2,58

Интервальная оценка математического ожидания нормального распределения при неизвестной дисперсии

По наблюдениям найдем точечные оценки математического ожидания и дисперсии

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ и } s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}.$$

Зададимся вероятностью γ.

Найдем такое число ε, чтобы выполнялось соотношение

$$P(\overline{X} - \varepsilon < \mu < \overline{X} + \varepsilon) = \gamma .$$

Снова приведем готовый результат

$$\varepsilon = \frac{t_\gamma \cdot s}{\sqrt{n}},$$

здесь t_γ находится из соотношения $t_{n-1}(t_\gamma) = \gamma$, где $t_{n-1}(t_\gamma)$ – распределение Стьюдента с $n-1$ степенями свободы.

Таблица значений функции распределения Стьюдента.

n	γ	
	0,95	0,99
5	2,78	4,60
10	2,26	3,25
15	2,15	2,98

n	γ	
	0,95	0,99
30	2,045	2,756
100	1,984	2,627
∞	1,96	2,57

Интервальная оценка квадратического отклонения и дисперсии нормального распределения

По наблюдениям найдем точечную оценку математического ожидания и дисперсии

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

за оценку среднего квадратического отклонения примем

$$s = +\sqrt{s^2}.$$

Зададимся вероятностью γ.

Найдем такое число ε, чтобы выполнялось соотношение

$$P(s - \varepsilon < \sigma < s + \varepsilon) = \gamma.$$

Среднее квадратическое отклонение всегда положительно, поэтому ε разумнее находить из условия

$$P[\max(0; s - \varepsilon) < \sigma < s + \varepsilon] = \gamma.$$

Снова приведем готовый результат

$$\varepsilon = s \cdot q_\gamma,$$

здесь q_γ находится из соотношения $\chi^2_{n-1}(q_\gamma) = \gamma$, где $\chi^2_{n-1}(q_\gamma)$ – хи-квадрат распределение с n-1 степенями свободы.

Таблица значений функции распределения Хи-квадрат.

n	γ	
	0,95	0,99
5	1,37	2,67
10	0,65	1,08
15	0,46	0,73

n	γ	
	0,95	0,99
30	0,28	0,43
100	0,143	0,198
250	0,089	0,1200

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Понятие статистической гипотезы

Статистическая гипотеза – это предположение о виде распределения или о величинах неизвестных параметров генеральной совокупности, которая может быть проверена на основании выборочных показателей.

Примеры статистических гипотез:

Генеральная совокупность распределена по закону Гаусса (нормальному закону).

Дисперсии двух нормальных совокупностей равны между собой.

Гипотеза «на Марсе есть жизнь» не является статистической, поскольку в ней не идет речь ни о виде, ни о параметрах распределения.

Вместе с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место альтернативная ей гипотеза. Целесообразно их различать. *Нулевой* (H_0) называют выдвинутую гипотезу. *Альтернативной* (H_1) – гипотезу, противоречащую нулевой.

Различают гипотезы, которые содержат только одно и более одного предположений. Гипотезу, содержащую только одно предположение называют *простой*, а гипотезу, которая состоит из конечного или бесконечного числа простых гипотез – *сложной*.

Примеры.

1. *H_0: математическое ожидание нормального распределения равно 3 ($\mu = 3$, σ – известно) – это простая гипотеза;*

2. *H_0: математическое ожидание нормального распределения меньше 3 ($\mu < 3$, σ – известно) – сложная (она состоит из бесконечного множества простых вида H_i: $\mu = b_i$, где b_i – любое число, меньшее 3).*

Ошибки I и II рода. Критерий значимости. Уровень значимости. Критическая область

Решение об отклонении или принятии статистической гипотезы принимается по выборочным данным. Поэтому приходится считаться и с возможностью ошибочного решения. Различают ошибки I и II рода.

Ошибка I рода состоит в том, что будет отвергнута правильная гипотеза (т.е. будет отвергнута нулевая гипотеза, в то время, когда она верна)

Ошибка II рода состоит в том, что будет принята неправильная гипотеза (т.е. будет принята нулевая гипотеза, в то время, когда она не верна)

При отбрасывании нулевой гипотезы есть вероятность того, что она все-таки верна (т.е. мы совершаем ошибку I-го рода), эту вероятность обозначают α. Вероятность α называется уровнем значимости.

Уровень значимости α – это вероятность совершить ошибку I рода. Вероятность ошибки II рода обозначают β, а величину $1 - \beta$ называют *мощностью критерия*.

Чем больше мощность, тем вероятность ошибки II рода меньше.

	H_0 принимается	H_1 принимается
H_0 верна	$P(H_0\|H_0) = 1 - \alpha$	$P(H_1\|H_0) = \alpha$ – уровень значимости
H_1 верна	$P(H_0\|H_1) = \beta$	$P(H_1\|H_1) = 1-\beta$ – мощность критерия

Допустимый процент возможных ошибок первого рода – вопрос взаимной договоренности, кроме всего прочего здесь должны приниматься во внимание возможные последствия принятия ошибочного решения. Ложные решения, например при экспертизе, могут иметь более серьезные последствия, чем ошибочно декларированная чистота химического реактива. Поэтому в

первом случае должны быть предусмотрены более высокая достоверность и, следовательно, более низкое число возможных ошибок I рода, чем во втором случае.

Обычно придерживаются следующих правил.

Проверяемая гипотеза отбрасывается, если ошибка I рода может появиться в менее чем $100\alpha = 1\%$ всех случаев (т.е. $\alpha \le 0{,}01$). Тогда рассматриваемое различие считается значимым.

Проверяемая гипотеза принимается, когда ошибка I рода возможна в более чем $100\alpha = 5\%$ всех случаев ($\alpha \ge 0{,}05$). Тогда рассматриваемое различие считается незначимым.

Рассматриваемую гипотезу надо обсуждать дальше, если число возможных ошибок I рода лежит в интервале между 5% и 1% ($0{,}01 \le \alpha \le 0{,}05$). Обнаруженная разность интерпретируется как спорная. Часто дополнительные измерения могут прояснить ситуацию. Если по каким-либо причинам дополнительных измерений окажется недостаточно, то полученные данные следует интерпретировать в расчете на самый неблагоприятный случай.

Выбор α - дело договорное, иногда достаточно выбрать $100\alpha = 10\%$, в отдельных случаях, практически, должна быть исключена возможность ошибочного решения (например, при оценке токсического действия фармацевтического препарата). Тогда проверяемая гипотеза отбрасывается, как только число возможных ошибок I рода достигает такого пренебрежительно малого уровня, как, например, $100\alpha = 0{,}1\%$.

Ошибки I и II рода зависят друг от друга. Чем меньше будет α, тем больше будет β (и наоборот). Поэтому, нет никакого смысла для проверки значимости выбирать слишком малое значение α, так как из-за этого очень вырастает неизвестное β. Выбор α относится к фазе планирования эксперимента!

После того, как задались уровнем значимости, находят правило, в соответствии с которым принимается или отклоняется данная гипотеза. Такое правило называется *статистическим критерием.*

Статистический критерий – правило, в соответствии с которым принимается или отклоняется нулевая гипотеза.

Построение критерия заключается в выборе подходящей функции $T = T(x_1,\dots,x_n)$ от результатов наблюдений $x_1,\dots,x_n$, которая служит мерой расхождения между опытными и гипотетическими значениями.

Эта функция, являющаяся случайной величиной, называется *статистикой критерия.*

Статистика критерия – специально выработанная случайная величина, функция распределения которой известна.

При этом предполагается, что распределение вероятностей $T = T(x_1,\dots,x_n)$ может быть вычислено при допущении, что проверяемая гипотеза верна и что это распределение не зависит от характеристик гипотетического распределения.

После выбора определенного критерия множество всех возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, а другая – при которых она принимается, т.е. на критическую область и область принятия гипотезы.

Критическая область – совокупность значений критерия, при которых нулевую гипотезу отвергают.

Область принятия гипотезы – совокупность значений критерия, при которых нулевую гипотезу принимают.

Основной принцип проверки гипотез можно сформулировать так: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

Поскольку критерий $T = T(x_1,\dots,x_n)$ – одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы также являются интервалами, и, следовательно, существуют точки, которые их разделяют. Такие точки называются критическими.

Критические значения критерия – это точки, отделяющие критическую область от области принятия гипотезы.

Критическое значение $T_{кр}$ находится по распределению статистики T такое, что если гипотеза верна, то вероятность события ($T \in$ критической области) равна α, α заранее заданный уровень значимости, т.е. это значение $T_{кр}$ статистики T для которого $P(T \in$ критической области$) = \alpha$.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критическую области. Они определяются из следующих выражений:

правосторонняя – $P(T > T_{кр}) = \alpha$;

левосторонняя – $P(T<T_{кр}) = \alpha$;

двусторонняя – $P(T<T_{кр1})+P(T>T_{кр2}) = \alpha$ $T_{кр1}<T_{кр2}$.

Если распределение критерия симметрично относительно нуля, то $P(T<-T_{кр}) = P(T>T_{кр})$, отсюда получаем $P(T>T_{кр})= \alpha/2$.

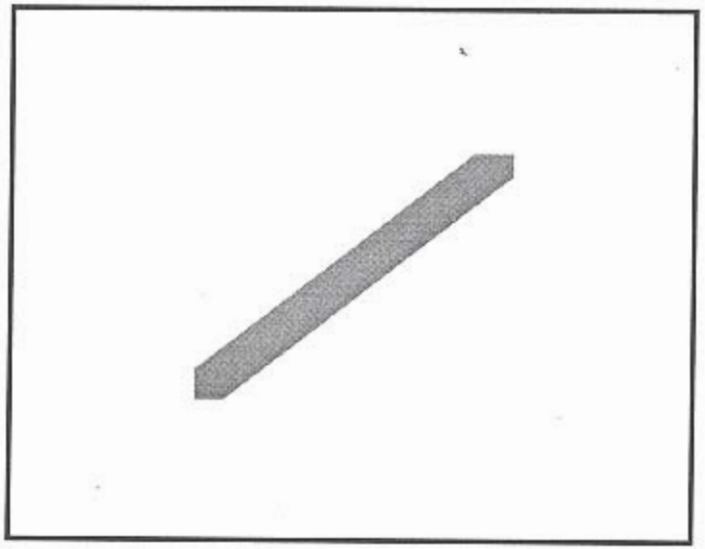


Рис. 37. Критические области: левосторонняя, правосторонняя, двусторонняя

Критические точки находят по таблицам, соответствующим распределению критерия.

Критерии значимости делят на параметрические и непараметрические. Первые строятся на основе параметров выборочной совокупности и представляют функции этих параметров, вторые – функции от вариант данной совокупности с их частотами. Параметрические критерии применимы лишь в тех случаях, когда генеральная совокупность, из которой взята выборка, распределяется нормально. Непараметрические критерии применимы к распределениям самых различных форм.

Последние имеют определенные преимущества по сравнению с параметрическими, благодаря меньшим требованиям к их применению, большему диапазону возможностей и, часто, большей простоте реализации. Конечно, нужно считаться и с часто более низкой точностью этих критериев по сравнению с параметрическими.

Результаты статистических методов проверки часто бывают неудобны для аналитиков. Во многих случаях они делают незначимые ($\alpha>0,05$) или спорные различия, хотя на основе субъективного опыта уже установлено «истинное» различие. В подобных случаях часто помогают дополнительные измерения. Чем больше получено результатов, тем меньшие различия будут достоверно фиксироваться. Ни в коем случае нельзя соблазняться заменой точных данных сомнительными на основании субъективной оценки.

Общая схема проверки гипотез

1. Исходя из содержания задачи формулируют нулевую и альтернативную гипотезы.
2. Задают величину уровня значимости критерия α, т.е. вероятностью отвергнуть основную гипотезу, когда она верна.
3. Выбирают некоторую функцию – статистику от результатов наблюдений – и при обеих гипотезах (основной и конкурирующей) находят законы ее распределения. Это самый сложный этап с теоретической точки зрения.
4. С помощью закона распределения на основе выбранного уровня значимости область возможных значений статистики разбивают на две или три части (на две части – при односторонней альтернативе, на три – при двусторонней).
5. Делают выборку и по ее результатам вычисляют статистику. Выясняют, в какую из областей попадает ее значение. Если величина находится в области, где правдоподобна основная гипотеза, то считают что эксперимент не противоречит основной гипотезе.

ПРОВЕРКА ГИПОТЕЗЫ О ВИДЕ РАСПРЕДЕЛЕНИЯ

Критерии согласия

При получении той или иной выборки встречаются случаи, когда закон распределения заранее не известен, но есть основания предположить, что он имеет определенный вид (назовем его вид A). В таких случаях исследователь формулирует нулевую гипотезу следующим образом: H_0 – генеральная совокупность распределена по закону A.

Критерий согласия – это критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Простая гипотеза прямо указывает некий закон вероятностей, по которому возникли выборочные значения. Сложная гипотеза указывает семейство распределений.

С помощью критерия согласия мы проверяем, согласуются эмпирические данные с нашим гипотетическим предположением относительно теоретической функции распределения или нет.

Рассмотрим критерии согласия Пирсона («хи-квадрат») и критерий согласия Колмогорова.

Критерий согласия Хи-квадрат Пирсона

Пусть дан ряд из n измерений. Важно установить, можно ли описать эти n значений с помощью принятой теоретической модели. Наиболее часто используют модель нормального (Гауссова) распределения или распределения Пуассона.

Для проверки выдвигают нулевую гипотезу – H_0: «между эмпирическим распределением и теоретической моделью нет никакого различия».

Из n значений (n>50) оценивают среднее $\overline{x}$ и стандартное отклонение s, а затем разбивают n значений на $m \approx \sqrt{n}$ классов.

Для каждого полученного класса определяют абсолютную частоту h попавших в него значений и сопоставляют ее с теоретической частотой h_t.

Из эмпирических и теоретических частот составляют выражение

$$\chi^2 = \sum_{i=1}^{m} \frac{(h_i - h_{ti})^2}{h_{ti}}.$$

Найденное выражение будет следовать хи-квадрат распределению с $m - k$ степенями свободы. Где m – количество классов, а k представляет число параметров, необходимых для описания выборки. Для нормального распределения k = 3 (среднее, стандартное отклонение и объем выборки), для распределения Пуассона k = 2 (среднее и объем выборки). Исходя из уровня значимости α и числа степеней свободы находим критическую точку.

Если при проверке получается, что $\chi^2 > \chi^2_{кр}$, то проверяемая гипотеза отбрасывается; между эмпирическим и теоретическим распределением существует значимое различие. Различие не значимо, если $\chi^2 < \chi^2_{кр}$.

Условием использования критерия "хи-квадрат" является достаточно большое число измерений (n>50).

Критерий согласия Колмогорова – Смирнова

Критерий нормальности Колмогорова – Смирнова обладает достаточной чувствительностью даже при малом числе значений. Его можно применять также для проверки соответствия любому распределению (например, равномерному). Однако следует иметь в виду, что функция распределения, установленная гипотезой, должна быть непрерывной.

Для проверки нормируют значения x_i по формуле

$$u_i = \frac{\left(x_i - \overline{x}\right)}{s}$$

и отыскивают значения гауссова интеграла, соответствующие u_i.

Затем находят разности $a_i = r_i - r(u_i)$ и сравнивают максимальную из них с критическим значением $d_{крит}$ из таблицы.

Процентные точки (α=0,05) для проверки на нормальность по Колмогорову и Смирнову.

n	3	4	5	6	7	8
d	0.376	0.375	0.343	0.323	0.304	0.288
n	9	10	11	12	13	14
d	0.274	0.261	0.251	0.242	0.234	0.226
n	15	16	17	18	19	20
d	0.219	0.213	0.207	0.202	0.197	0.192

Пример.

В результате 8-ми титрований получились значения объема V(мл):

20,23	20,12	20,21	20,17	20,13	20,07	20,24	20,19

Надо проверить, следуют ли они Гауссовому распределению.

Решение.

Воспользуемся критерием согласия Колмогорова. Вычислим точечные оценки для математического ожидания и среднеквадратического отклонения $\overline{x} = 20{,}17$ *и* $s = 0{,}06$. *Ранжируем данные и записываем их в порядке возрастания. Составим таблицу.*

V (мл)	Частота			u_i	$Y(u_i)$	$\lvert d_i \rvert$
	Абсолютная.	Относительная.	Кумулятивная Y_i			
1	2	3	4	5	6	7
20,07	1	0,125	0,125	-1,667	0,048	0,077
20,12	1	0,125	0,250	-0,833	0,203	0,047
20,13	1	0,125	0,375	-0,677	0,252	0,123
20,17	1	0,125	0,500	0	0,5	0,000
20,19	1	0,125	0,625	0,333	0,629	0,004
20,21	1	0,125	0,750	0,667	0,748	0,002
20,23	1	0,125	0,875	1,000	0,841	0,016
20,24	1	0,125	1,000	1,167	0,867	0,133

Максимальное значение статистики Колмогорова

$$\max(|d|)=0{,}133.$$

Сравним полученное максимальное значение с табличным $d(\alpha = 0{,}05,\ n = 8) = 0{,}288$.

Получили $d_{max} < d$, *следовательно, нет оснований отбрасывать гипотезу о нормальном распределении.*

ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ НОРМАЛЬНО РАСПРЕДЕЛЕННЫХ СОВОКУПНОСТЕЙ

Проверка гипотезы о равенстве среднего исследуемой нормальной совокупности определенному числовому значению при известной дисперсии

Пусть проверяется гипотеза о том, что независимые результаты наблюдений $x_1,\ldots,x_n$ подчиняются нормальному распределению со средним значением $\mu=\mu_0$ при известной дисперсии σ^2.

Итак, мы имеем:

нулевая гипотеза H_0: $\mu = \mu_0$;

альтернативная H_1: $\mu \neq \mu_0$;

уровень значимости α.

Статистика критерия $U = \sqrt{n}\dfrac{\overline{x} - \mu_0}{\sigma}$ распределена нормально с параметрами $N(0,1)$, где

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

– арифметическое среднее результатов наблюдений.

Существует связь между α и критическим значением $U_{кр}$

$$\frac{\alpha}{2} = \int_{U_{кр}}^{\infty} \phi(u)\,du$$

Из этого уравнения при заданном α мы можем найти $U_{кр}$. Для этого можно использовать таблицу значений стандартного нормального распределения.

Далее мы должны сделать выводы относительно полученного значения U.

Если $|U| > U_{кр}$, то мы отвергаем нулевую гипотезу.

Если $|U| \le U_{кр}$, то у нас нет оснований отвергать нулевую гипотезу (это, однако, не означает, что нулевая гипотеза подтверждается).

Проверка гипотезы о равенстве среднего исследуемой нормальной совокупности определенному числовому значению при неизвестной дисперсии

Если дисперсия σ^2 не известна, то вместо данного критерия для проверки гипотезы $\mu = \mu_0$ можно воспользоваться критерием Стьюдента, основанном на статистике

$$t = \sqrt{n}\frac{\left|\overline{x} - \mu_0\right|}{s},$$

которая включает несмещенную оценку дисперсии

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i - \overline{X}\right)^2$$

и подчинена распределению Стьюдента с $n-1$ степенями свободы. Для проверки гипотезы о неизвестном значении σ_2 используется хи-квадрат критерий.

Гипотеза о равенстве средних значений двух нормально распределенных совокупностей при неизвестных дисперсиях

Рассмотрим две независимые выборки X и Y, объемы которых равны n и m соответственно, причем известно, что они извлечены из нормальных генеральных совокупностей с равными дисперсиями $\sigma_x^2 = \sigma_y^2 = \sigma^2$, причем как величина дисперсии σ^2, так и средние μ_x и μ_y неизвестны. Требуется проверить следующую гипотезу:

H_0: $\mu_x = \mu_y$ $(\mu_x - \mu_y)=0$;

H_1: $\mu_x \neq \mu_y$.

Гипотеза H_0 – сложная, но может быть сведена к простой, если рассматривать разность средних ($\mu_x - \mu_y$). В этом случае естественно рассматривать и разность оценок ($\bar{x} - y$), распределение разности нормальное, поскольку нормальны сами оценки $\bar{x}$ и y.

Получаем статистику Стьюдента

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left[(n-1)\cdot s_x^2 + (m-1)\cdot s_y^2\right]/(n+m-2)}}\sqrt{\frac{nm}{n+m}},$$

где

$$s_X^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i - \bar{X}\right)^2 \quad \text{и} \quad s_y^2 = \frac{1}{m-1}\sum_{k=1}^{n}\left(Y_i - \bar{Y}\right)^2.$$

Эта статистика не зависит от неизвестных μ_x, μ_y и σ^2 и распределена по закону Стьюдента с $(n+m-2)$ степенями свободы.

Из изложенного можно получить следующий критерий с уровнем значимости $\alpha/2$ для проверки гипотезы H_0: $\mu_x = \mu_y$.

Гипотеза H_0 отвергается, если: $|t| > t_\alpha(n+m-2)$, где $t_\alpha(n+m-2)$ – критическая граница распределения Стьюдента, соответствующая уровню значимости $\alpha/2$.

Приведенные критерии являются точными и наилучшими, поэтому их можно использовать как при малых, так и больших объемах выборок (при нормальном распределении генеральных совокупностей).

Пример.

Две рабочие группы методом микроанализа определяли содержание азота в одном органическом соединении (цинхонине). Были получены следующие значения:

1-ая группа (%)	9,29	9,38	9,35	9,43
2-ая группа (%)	9,53	9,48	9,61	9,68

Посчитаем средние значения и среднее квадратическое отклонение.

1-ая группа (%)	$\bar{x}_1 = 9{,}363$	$s_1 = 0{,}058$
2-ая группа (%)	$\bar{x}_2 = 9{,}575$	$s_2 = 0{,}088$

Мы видим, что средние слегка различаются. Надо проверить, можно ли объяснить это различие только

случайной ошибкой или здесь есть систематическая ошибка. Сравним средние значения.

H_0: оба средних принадлежат одной и той же генеральной совокупности со средним μ.

По t-критерию стьюдента имеем $t=4{,}03$.

При $\alpha=0{,}01$ и $(8-2)$ =6-и степенях свободы $t_{кр}=3{,}71$.

$t > t_{кр}$, следовательно, гипотеза H_0 отвергается, т.е. выборки принадлежат различным совокупностям. Это означает, что в одной из групп есть систематическая ошибка.

Попробуем выяснить, в какой именно. Для этого проверим гипотезу о равенстве среднего определенному числовому значению. Из предыдущих исследований известно, что для исследуемого соединения теоретическое содержание азота равно $\mu_0 = 9{,}517\%$.

Сравним отдельно две выборки:

$$t_1 = \frac{|9{,}363 - 9{,}517|}{0{,}058}\sqrt{4} = 5{,}31,$$

$$t_2 = \frac{|9{,}575 - 9{,}517|}{0{,}088}\sqrt{4} = 1{,}32,$$

t_3 $(\alpha=0{,}05)=3{,}18$, где t_3 – распределение Стьюдента с 3-я степенями свободы.

Поскольку $t_1 > t(\alpha=0{,}05)$ с 3-я степенями свободы, можно предположить, что в результаты именно первой группы вкралась систематическая ошибка. Во второй группе отклонение от теоретического значения можно считать случайным.

Гипотеза о равенстве дисперсий двух нормально распределенных совокупностей при неизвестных средних

Предполагается, что генеральные совокупности имеют нормальные распределения. Выборку из первой генеральной совокупности будем обозначать через $X=(x_1,\ldots,x_n)$, а из второй $Y=(y_1,\ldots,y_m)$.

Итак, имеем две выборки $x_1,\ldots,x_n$ и $y_1,\ldots,y_m$ из нормальных совокупностей с неизвестными средними и дисперсиями σ_x^2 и σ_y^2 соответственно.

Следует проверить гипотезу H_0: $\sigma_x^2 = \sigma_y^2$

против альтернативной H_1: $\sigma_x^2 \neq \sigma_y^2$ $(\frac{\sigma_X^2}{\sigma_y^2} = 1)$.

Так как средние неизвестны, то наилучшими оценками дисперсий σ_x^2 и σ_y^2 являются:

$$s_X^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \text{и} \quad s_y^2 = \frac{1}{m-1}\sum_{k=1}^{n}\left(Y_i - \overline{Y}\right)^2 .$$

Рассмотрим статистику:

$$F = \frac{s_X^2}{s_y^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2},$$

которая при справедливости основной гипотезы не зависит от неизвестных параметров нормального распределения.

Гипотеза H_0 отвергается, если: $F>F_\alpha(n-1;m-1)$, где $F_\alpha(n-1;m-1)$ критическая граница распределения Фишера, соответствующая уровню значимости α.

Гипотеза о равенстве дисперсий двух нормально распределенных совокупностей при известных средних

Эта гипотеза проверяется аналогично предыдущей, но в данном случае:

$$s_X^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_i - \mu_X\right)^2 \quad \text{и} \quad s_y^2 = \frac{1}{m}\sum_{k=1}^{n}\left(Y_i - \mu_y\right)^2 ,$$

где μ_x и μ_y – известные средние.

Если верна гипотеза H_0: $\sigma_x^2 = \sigma_y^2 = \sigma^2$, то статистика

$$F = \frac{s_X^2}{s_y^2}$$

распределена по закону Фишера с n и m степенями свободы.

РЕГРЕССИОННЫЙ АНАЛИЗ

Пусть у нас есть серии значений двух параметров. Подразумевается, что у одного и того же объекта измерены два параметра. Нам надо выяснить есть ли значимая связь между этими параметрами.

Существуют следующие формы связи – функциональная и статистическая.

При функциональной зависимости каждому конкретному значению одной величины будет соответствовать определенное значение другой величины. Например, при фиксированной скорости пройденный путь линейно зависит от времени движения, или из (физики) объем газа зависит от давления, скорость движения частиц жидкости зависит от площади сечения трубы.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой.

Особенность статистической связи заключается в том, что каждому значению одного признака может соответствовать не одно единственное значение, а некоторое количество значений другого признака, варьирующих в определенных пределах.

Рассмотрим рост и вес человека. Каждому значению роста соответствуют различные значения веса. И эти значения имеют распределения (отличные друг от друга) для каждого значения роста, например, нормальное распределение с разными математическими ожиданиями.

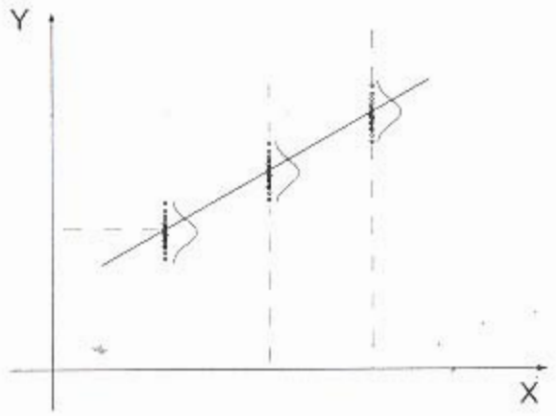


Рис. 38. Распределения случайной величины Y для определенных значений X

Линейная регрессия

Пусть при обследовании n объектов (человек, животное, природное явление) измерили два параметра и получили набор из n пар (n>2) значений (x_i,y_i).

Нанесем их на координатную плоскость.

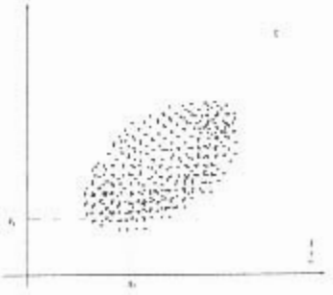


Рис. 39. Облако точек

Получилось облако точек. Теперь перед исследователем стоит вопрос: «Как описать это облако?». Общая задача, которую требуется решить, состоит в подгонке линии (желательно прямой) к этому набору точек.

Как нам известно, уравнение прямой линии имеет вид:

$$Y=b_0+b_1X.$$

Но (в зависимости от параметров b_0 и b_1) эти точки можно описать различными прямыми линиями. Как из всего множества нам выбрать наилучшую? Таким образом, задача сводится к подбору наилучших коэффициентов b_0 и b_1.

Метод наименьших квадратов (МНК)

Принято считать, что наилучшие оценки коэффициентов даст метод наименьших квадратов (МНК).

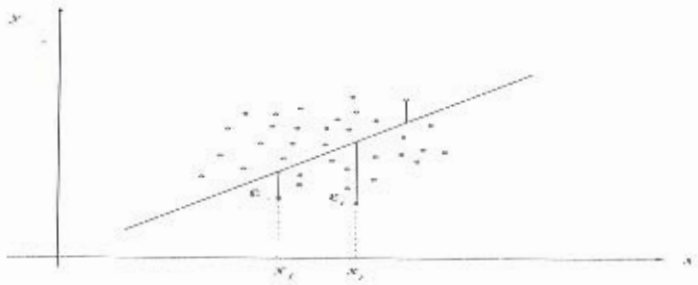


Рис. 40. Пояснение к оценке коэффициентов методом наименьших квадратов

Обозначим: Y_i – значение, вычисленное по уравнению $Y_i=b_0+b_1 x_i$, где y_i – измеренное значение, $\varepsilon_i= y_i- Y_i$ –разность, $y_i = Y_i + \varepsilon_i = b_0 + b_1 x_i + \varepsilon_i$ или $\varepsilon_i = y_i - b_0 - b_1 x_i$.

В методе наименьших квадратов требуется, чтобы ε_i, разность между измеренными y_i и вычисленными по уравнению значениям Y_i, была минимальной (вернее, сумма квадратов ε_i^2).

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \to \min . \quad (*)$$

Будем подбирать значения оценок b_0 и b_1 так, чтобы они давали минимальное значение S. Мы можем определить b_0 и b_1 путем дифференцирования уравнения (*) сначала по b_0, затем по b_1 (минимум функции достигается только в точках нуля производной):

$$\frac{\partial S}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i),$$

(**)

$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 \cdot x_i).$$

Приравнивая полученные результаты к нулю для оценок b_0 и b_1, получим

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i) = 0,$$

(***)

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 \cdot x_i) = 0.$$

Из (***) имеем:

$$b_0 n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i ,$$

(****)

$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i .$$

Эти уравнения называются *нормальными*.

Получаем:

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n}}.$$

Воспользуемся тем, что

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n},$$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n},$$

а также

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ и } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Решение уравнения относительно b_1 дает:

$$b_1 = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

b_1 называют коэффициентом регрессии или угловым коэффициентом;

b_0 называют свободным членом уравнения регрессии и вычисляют по формуле:

$$b_0 = \overline{y} - b_1 \overline{x}.$$

Полученная прямая является оценкой для теоретической линии регрессии. Имеем:

$$Y = b_0 + b_1 x = \overline{y} - b_1 \overline{x} + b_1 x = \overline{y} + b_1 (x - \overline{x}).$$

Регрессия может быть прямой ($b_1>0$) и обратной ($b_1<0$).

Прямая регрессия означает, что при росте одного параметра, значения другого параметра тоже увеличиваются. А *обратная*, что при росте одного параметра, значения другого параметра уменьшаются.

Не всегда можно утверждать, что предполагаемая линейная зависимость действительно имеет место.

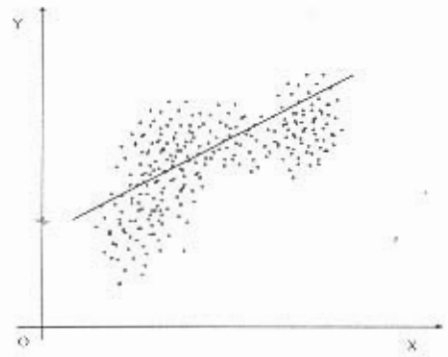


Рис. 41. Модель линии регрессии подобрана не верно. Данное облако должно быть описано нелинейной моделью

Проверка гипотезы о значимости коэффициента регрессии b_1

Мы вывели модель, описывающую наши измерения, теперь нам надо определить, верна ли она. Для решения этого вопроса нужно проверить гипотезы:

Модель регрессионного анализа выглядит следующим образом:

$$y_i = b_0 + b_1 x_i + \varepsilon_i \text{ где } i = 1,2,\ldots,n,$$

где b_0 – параметр, характеризующий смещение по Y;

b_1 – коэффициент регрессии – параметр, характеризующий смещение графика функции по X;

ε_i (эпсилон) – некоррелированные ошибки случайной переменной.

В регрессионном анализе проверяют гипотезы о значимости свободного члена b_0 и о значимости коэффициента регрессии b_1.

Проверка гипотезы о значимости коэффициента регрессии b_1.

1.Определим гипотезы H_0 и H_1:

H_0: $b_1 = 0$ (между переменными нет линейной зависимости);

H_1: $b_1 \neq 0$.

2.Зададим уровень значимости α.

3.Статистика критерия

$$F = \frac{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2},$$

где $s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-2}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ и $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Статистика F имеет распределение Фишера с 1 и $(n–2)$ степенями свободы.

4.Критические точки и критическая область область $|F|>F_{\alpha,1,n-2}$.

5.Если $|F|>F_{\alpha,1,n-2}$, то H_0 отвергается, т.е можно сделать вывод, что линейная зависимость – значима.

Если $|F|>F_{\alpha,1,n-2}$, то у нас нет оснований отвергать H_0 , т.е можно сделать вывод, что линейная зависимость – незначима или что наши данные нельзя описать моделью линейной регрессии.

Если в результате проверки оказывается, что линейная зависимость невозможна (незначима), то пытаются преобразовать результаты в удобную форму. Во многих случаях целесообразно логарифмическое преобразование. Для простоты в обращении всегда стремятся получить линейную зависимость с помощью удобного преобразования. Однако важно помнить, что после подобных преобразований необходимо критически перепроверить все условия для вычисления регрессии и что только тогда полноценная регрессия может привести к надежным результатам.

Наше облако точек можно описать двумя линиями регрессии – регрессией X на Y и Y на X.

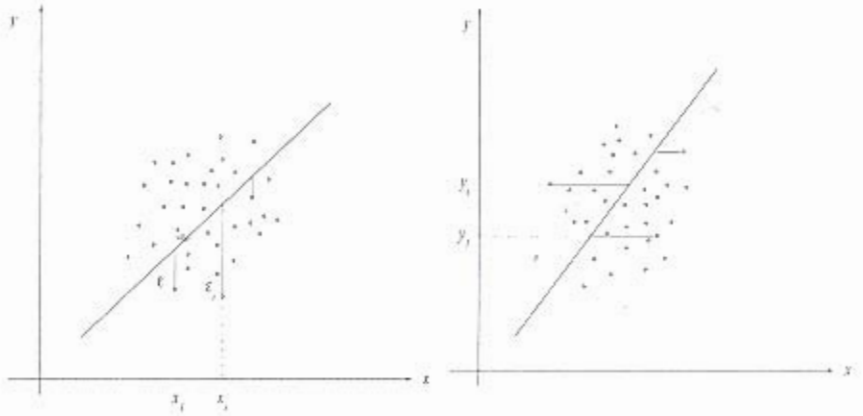


Рис. 42. Облако точек можно описать двумя линиями регрессии – регрессией Y на X (слева) и X на Y (справа)

$Y = b_{0yx} + b_{1yx}x$ - регрессия Y на X.

$X = b_{0xy} + b_{1xy}y$ -регрессия X на Y.

Обозначим через b_{1xy} коэффициент регрессии X на Y, b_{1yx} коэффициент регрессии Y на X. Тогда величина

$$R^2 = b_{1xy} \cdot b_{1yx}$$

является мерой определенности и называется *коэффициент детерминации R-квадрат.*

Чем меньше рассеяние наблюдаемых пар значений относительно прямых регрессии, чем больше точки примыкают к прямым, тем точнее эти прямые определены. Если значение R^2 велико, то это означает, что точки концентрируются около прямой регрессии.

Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем, очевидно, лучше прогноз.

Рассеяние точек относительно прямой регрессии может служить мерой точности, с которой определена прямая. Например, если $R^2 = 0{,}81$, то это означает, что 81% общего рассеяния можно объяснить изменением линейной регрессии при изменении независимой случайной величины, а 19% остаточной изменчивости остаются необъясненными. В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. Значение R-квадрата является индикатором степени подгонки модели к данным (значение R-квадрата близкое к 1.0 показывает, что модель объясняет почти всю изменчивость соответствующих переменных).

Полученную функцию $y = b_0 + b_1x$ можно использовать, чтобы для заданных, а значит, почти безошибочных значений x вычислять предсказанные значения зависимой переменной y.

Итак, задачами регрессионного анализа являются:

1) оценить коэффициент регрессии и свободный член;
2) построить доверительные интервалы для них;
3) проверить гипотезу о значимости регрессии;
4) оценить степень адекватности модели.

Данные могут быть описаны различными регрессионными моделями, приведем некоторые из них:

- линейная $y = b_0 + b_1x$;
- мультипликативная $y = b_1x^2$;
- экспоненциальная $y = e^{b_0+b_1x}$;
- обратная $y = \dfrac{1}{b_0 + b_1x}$.

С помощью линии регрессии мы описываем наши значения, но не можем количественно оценить силу связи между параметрами, для этого служит коэффициент корреляции.

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Линейная корреляция

Как показано выше, облако точек можно описать двумя линиями регрессии – регрессией X на Y и Y на X. Чем меньше угол между этими прямыми, тем сильнее зависимость между параметрами.

Характер и сила связи определяются с помощью коэффициента корреляции r.

По своему характеру корреляционная связь может быть прямой и обратной, а по силе – сильной, средней, слабой. Кроме того, связь может отсутствовать или быть полной.

1. Если $|r|=1$, то Y,X – связаны линейной связью.
2. Если $|r|=0$, то Y,X – не коррелируют.
3. Чем ближе $|r|$ к 1, тем теснее прямолинейная корреляция между величинами Y,X.

Сила и характер связи между параметрами

Сила связи	Характер связи	
	прямая (+)	обратная (-)
Полная	1	–1
Сильная	от 0,7 до 1	от – 0,7 до –1
Средняя	от 0,7 до 0,3	от –0,7 до –0,3
Слабая	от 0,3 до 0	от –0,3 до 0
Связь отсутствует	0	0

Корреляционный анализ экспериментальных данных для двух случайных величин заключает в себе следующие основные приемы:

1. Вычисление выборочных коэффициентов корреляции.
2. Составление корреляционной матрицы.
3. Проверка статистической гипотезы значимости связи.

Если имеется выборка объема n (x_1,y_1), (x_2,y_2),…,(x_n,y_n) с совместным распределением, то величина

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

называемая выборочным коэффициентом корреляции между X и Y, оценивает теоретическую корреляцию и представляет собой эмпирическую меру зависимости между X и Y.

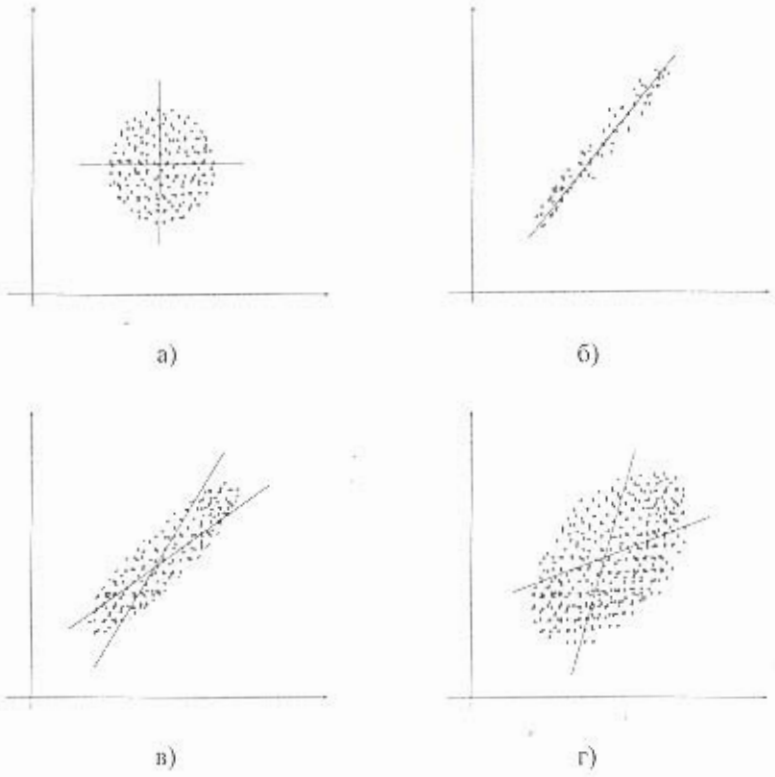


Рис. 43. Зависимость между параметрами: а) зависимости нет; б) полная ; в) слабая; г) сильная.

Коэффициент корреляции обладает следующими свойствами:

1.$-1 \le r \le +1$,

2.при $r = +1$ имеется прямая функциональная зависимость,

3.при $r = -1$ имеется обратная функциональная зависимость,

4.если $r = 0$, то X и Y называют некоррелированными.

Независимые случайные переменные не коррелированы; две случайные переменные тем сильнее коррелированны, чем ближе значение $|r|$ к 1.

Проверка гипотезы о значимости коэффициента корреляции

О статистической взаимосвязи говорят, что она существует или отсутствует, имеет направление (положительна или отрицательна) и характеризуется силой (сильная, слабая). Если в результате исследования нулевая гипотеза не отвергается, то «взаимосвязи нет». В случае, когда нулевая гипотеза отклоняется, говорят о существовании связи исследуемых случайных величин.

Сформулируем гипотезы H_0 и H_1:

H_0: $r = 0$ (т.е. корреляции нет),

H_1: $r \neq 0$.

Зададим уровень значимости α.

Статистикой критерия здесь является следующее выражение:

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2},$$

где t – статистика, имеющая распределение Стьюдента с (n–2) степенями свободы.

При $|t| \ge t_{(n-2),\alpha}$ гипотеза H_0: $r = 0$ отвергается с уровнем значимости α. Это значит, что между параметрами существует значимая корреляция. При $|t| \le t_{(n-2),\alpha}$ у нас нет оснований отвергать H_0: $r = 0$, т.е. можно утверждать, что между параметрами нет значимой корреляции.

Пример.

Вычислим коэффициент корреляции между показателями охвата населения прививками и заболеваемостью брюшным тифом.

Сначала вычислим точечные оценки математических ожиданий для каждого показателя:

$$\overline{X} = 7{,}7 \text{ и } \overline{Y} = 5{,}6$$

Обозначим

$$d_x = \left(x_i - \overline{x}\right),$$
$$d_y = \left(y_i - \overline{y}\right).$$

Районы	охват населения прививками (в%)	Заболеваемость брюшным тифом. (в%)	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
	X	Y					
A	14.7	1.4	7.0	-4.2	49.0	17.64	-29.4
B	13.4	1.4	5.7	-4.2	32.49	17.64	-23.94
C	9.6	2.3	1.9	-3.3	3.61	10.89	-6.27
D	8.1	2.1	0.4	-3.5	0.16	12.25	-1.4
E	5.5	6.2	-2.2	0.6	4.84	0.36	-1.32
F	5.2	6.9	-2.5	1.3	6.25	1.69	-3.25
G	4.4	8.6	-3.3	3.0	10.89	9.0	-9.9
H	4.4	10.8	-3.3	5.2	10.89	27.04	-17.16
I	4.0	11.0	-3.7	5.4	13.69	29.16	-19.98

Тогда формула для подсчета коэффициента корреляции примет следующий вид:

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2}\sqrt{\sum d_y^2}}.$$

Итак,

$$r_{xy} = -0{,}87.$$

Проверим значимость полученного результата

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2} = \frac{-0{,}87}{\sqrt{1 - (-0.87)^2}}\sqrt{9 - 2} = -4{,}68 \cdot$$

	Уровень значимости		
Число степ свободы	0.10	0,05	0,01
7	1,89	2,36	3,5

Допустим мы задали $\alpha = 0{,}01$ тогда при n–2=9–2=7 степенях свободы значение $t_{крит}$=3,5.

Мы получили значение $|t|$=4,68, оно больше критического, следовательно, мы отвергаем гипотезу о незначимости коэффициента корреляции, следовательно, между показателями охвата населения прививками и заболеваемостью брюшным тифом существует значимая корреляционная связь. Причем поскольку $r_{xy} = -0{,}87$ мы можем утверждать, что между этими показателями существует сильная обратная корреляция, т.е., чем больше население охвачено прививками, тем меньше показатель заболеваемости брюшным тифом.

Ранговая корреляция

Ранговая корреляция применяется для обработки данных непараметрическими методами. Если нужно определить взаимозависимость между рядами, распределенными не по нормальному закону, а когда двумерная выборка (x_i,y_i) относится к произвольному непрерывному распределению. В этом случае можно установить зависимость между Y и X с помощью коэффициента ранговой корреляции Спирмена.

Ранг наблюдения – это тот номер, который получит наблюдение в совокупности всех данных – после их упорядочения по определенному правилу (например, от меньших величин к большим).

Процедура перехода от совокупности наблюдений к последовательности их рангов называется *ранжированием*. Результат ранжирования называют *ранжировкой*.

Данные выстраиваются в порядке возрастания (или убывания) и далее им присваивают ранги. Если отдельные показатели ряда встречаются несколько раз, то каждому из них присваивают одинаковый ранг, равный среднему рангу.

Порядковый. номер.	1	2	3	4	5	6	7	8	9
Данные	20	21	22	22	23	24	24	24	25
Ранги	1	2	3,5	3,5	5	7	7	7	9

Итак, коэффициент ранговой корреляции Спирмена вычисляется по следующей формуле:

$$\rho_S = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)},$$

где 6 – постоянный коэффициент, d – разность рангов, n – число наблюдений (объем выборки).

Для проверки гипотезы о значимости коэффициента корреляции применяют следующую статистику:

$$t_S = \frac{\rho_S}{\sqrt{1 - \rho_S^2}} \cdot \sqrt{n - 2},$$

которая имеет распределение Стьюдента с (n–2) степенями свободы.

При $t_S \geq t_{(n-2),\alpha}$ гипотеза H_0: $r = 0$ отвергается с уровнем значимости α.

При $t_S \leq t_{(n-2),\alpha}$ нет оснований отвергать H_0: $r = 0$.

Пример.

Вычислим величину и определим характер связи между содержанием йода в пище и воде и пораженностью населения зобом.

Воспользуемся формулой для коэффициента корреляции Спирмена для рангов $\rho = -0{,}964$.

Вычисленный коэффициент ранговой корреляции показывает, что связь между содержанием йода в пище и воде и пораженностью населения зобом высокая и обратная, т.е. чем больше содержится йода в продуктах питания и воде, тем меньше среди населения доля пораженных зобом.

Кол-во йода в воде и пище (в γ)	Пораженность населения зобом (%)	Порядковые номера (ранги)		Разность рангов	Квадрат разности рангов
		кол-во йода	Пораженности населения зобом		
1	2	3	4	5	6
X	Y	x	y	$d=x-y$	d^2
201	0,2	1	7	– 6	36
178	0,6	2	6	– 4	16
155	1,1	3	4	– 1	1
154	0,8	4	5	– 1	1
126	2,5	5	3	2	4
81	4,4	6	2	4	16
71	16,9	7	1	6	36
n=7	ρ= – 0,964				Σd^2=110

Достоверность корреляций Спирмена оценивается по таблице:

Число пар	Уровень значимости (α)		Число пар	Уровень значимости (α)	
	0,05	0,01		0,05	0,01
4	1,0		16	0,425	0,601
5	0,9	1	18	0,399	0,564
6	0,829	0,943	20	0,377	0,534
7	0,714	0,893	22	0,359	0,508
8	0,643	0,833	24	0,343	0,485
9	0,6	0,783	26	0,329	0,465
10	0,564	0,746	18	0,317	0,448
12	0,506	0,712	30	0,306	0,432

Если вычисленный коэффициент при данном числе сравниваемых пар превышает табличное значение, то связь между признаками признается достоверной. Нецелесообразно вычислять коэффициент связи при числе коррелируемых пар меньше 4-х.

В рассматриваемом примере вычисленный коэффициент ранговой корреляции, равный –0,964, превышает табличное значение при уровне значимости 0,01 и потому должен быть признан значимым с вероятностью ошибки менее 0,01.

Следует добавить, что коэффициент корреляции можно вычислять и тогда, когда данные носят полуколичественный приближенный характер, отражая лишь общий порядок следования величин.

В практических задачах наибольший интерес представляют следующие вопросы:

1) существует значимая корреляционная зависимость Y от X или нет, т.е. отлично ли генеральное корреляционное отношение от нуля или равно нулю;
2) если корреляционная зависимость существут, то какой вид имеет функция регрессии (линейный, нелинейный или иной).

АНАЛИЗ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

Таблица сопряженности признаков

Существует множество признаков, различных явлений и вещей, измерение которых затруднено или вовсе невозможно. Например, как измерить признак «профессия» или «вид патологии», а как сравнить эти признаки для получения статистического представления о профессиональной заболеваемости?

Для начала следует определить, существует ли вообще связь между изучаемыми признаками, или же они ведут себя независимо друг от друга.

Предположим, что у нас есть два качественных признака A и B, пусть прзнак A имеет r градаций, а признак В s градаций. Предположим, что у нас есть выборка объема n из интересующей нас генеральной совокупности. Каждый объект выборки может обладать одним из уровней признака A и одновременно каким-либо уровнем признака B. По этой выборке мы можем определить частоты событий A_i и B_j по отдельности и в любых комбинациях.

Обозначим через n_{ij} частоту события A_iB_j. Число появлений признака A_i (частота события A_i) равно:

$$n_{i.} = \sum_{j=1}^{s} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{is}.$$

Аналогично, частота появления события B_j равна

$$n_{.j} = \sum_{i=1}^{r} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{rj}.$$

Общее число наблюдений, т.е. объем выборки

$$n_{..} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij},$$

замена индекса точкой означает результат суммирования по этому индексу. Полученные частоты представляют в виде таблицы сопряженности признаков или просто таблицы сопряженности.

	B_1	B_2	…	B_j	…	B_S	
A_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
…	…	…	…	…	…	…	…
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i.}$
…	…	…	…	…	…	…	…
A_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	$n_{..}$

Введем гипотезу, отрицающую связь между признаками. H_0: связи нет – это гипотеза о независимости признаков.

Признаки A и B называют независимыми, если оказываются независимыми события: «признак A принимает значение A_i» и «признак B принимает значение B_j», притом для всех пар i, j.

Т.е. признаки A и B называются независимыми, если

$$p(A_i B_j) = p(A_i) \cdot p(B_j),$$

для всех A_i и B_j Вероятности событий A_i и B_j подсчитываем из полученных данных:

$$p(A_i) = \frac{n_{i.}}{n_{..}},$$

$$p(B_j) = \frac{n_{.j}}{n_{..}}.$$

Выразим вероятности события A_iB_j

$$p(A_i B_j) = \frac{\overline{n}_{ij}}{n_{..}},$$

где n_{ij} -частота события A_iB_j при условии, что нулевая гипотеза верна. Посчитаем эту частоту

$$\frac{\overline{n}_{ij}}{n_{..}} = \frac{n_{i.}}{n_{..}} \cdot \frac{n_{.j}}{n_{..}},$$

$$\overline{n}_{ij} = n_{i.} \cdot \frac{n_{.j}}{n_{..}}.$$

Величины $\overline{n}_{ij} = n_{i.} \cdot \frac{n_{.j}}{n_{..}}$ называются ожидаемыми частотами при выполнении нулевой гипотезы.

	B_1	B_2	…	B_j	…	B_S	
A_1	$\overline{n}_{11}$	$\overline{n}_{12}$	…	$\overline{n}_{1j}$	…	$\overline{n}_{1s}$	$\overline{n}_{1.} = n_{1.}$
A_2	$\overline{n}_{21}$	$\overline{n}_{22}$	…	$\overline{n}_{2j}$	…	$\overline{n}_{2s}$	$\overline{n}_{2.} = n_{2.}$
…	…	…	…	…	…	…	…
A_i	$\overline{n}_{i1}$	$\overline{n}_{i2}$	…	$\overline{n}_{ij}$	…	$\overline{n}_{is}$	$\overline{n}_{i.} = n_{i.}$
…	…	…	…	…	…	…	…
A_r	$\overline{n}_{r1}$	$\overline{n}_{r2}$	…	$\overline{n}_{rj}$	…	$\overline{n}_{rs}$	$\overline{n}_{r.} = n_{r.}$
	$\overline{n}_{.1} = n_{.1}$	$\overline{n}_{.2} = n_{.2}$	…	$\overline{n}_{.j} = n_{.j}$	…	$\overline{n}_{.s} = n_{.s}$	$\overline{n}_{..} = n_{..}$

При выполнении гипотезы ожидаемые частоты не должны сильно отличаться от наблюдаемых частот n_{ij} .

	B_1	B_2	…	B_j	…	B_S	
A_1	$n_{11} - \overline{n}_{11}$	$n_{12} - \overline{n}_{12}$	…	$n_{1j} - \overline{n}_{1j}$	…	$n_{1s} - \overline{n}_{1s}$	$\overline{n}_{1.} = n_{1.}$
A_2	$n_{21} - \overline{n}_{21}$	$n_{22} - \overline{n}_{22}$	…	$n_{2j} - \overline{n}_{2j}$	…	$n_{2s} - \overline{n}_{2s}$	$\overline{n}_{2.} = n_{2.}$
…	…	…	…	…	…	…	…
A_i	$n_{i1} - \overline{n}_{i1}$	$n_{i2} - \overline{n}_{i2}$	…	$n_{ij} - \overline{n}_{ij}$	…	$n_{is} - \overline{n}_{is}$	$\overline{n}_{i.} = n_{i.}$
…	…	…	…	…	…	…	…
A_r	$n_{r1} - \overline{n}_{r1}$	$n_{r2} - \overline{n}_{r2}$	…	$n_{rj} - \overline{n}_{rj}$	…	$n_{rs} - \overline{n}_{rs}$	$\overline{n}_{r.} = n_{r.}$
	$\overline{n}_{.1} = n_{.1}$	$\overline{n}_{.2} = n_{.2}$	…	$\overline{n}_{.j} = n_{.j}$	…	$\overline{n}_{.s} = n_{.s}$	$\overline{n}_{..} = n_{..}$

Если видимые различия между наблюдаемыми частотами и частотами, рассчитанными на основании гипотезы о независимости признаков, можно объяснить случайными колебаниями, то отвергать гипотезу независимости нет оснований.

Осталось условиться, как сопоставить два ряда частот, как измерить различие между ними.

Наиболее распространенным является метод Пирсона–Фишера. Для формулировки критерия Пирсона–Фишера обозначим наблюдаемые частоты через n_{ij}, а ожидаемые, теоретические час-

тоты через n_{ij}. Если статистическая модель правильно описывает наблюдения, то числа n_{ij} и n_{ij} должны быть близкими друг к другу, а сумма квадратов отклонений $(n_{ij} - n_{ij})^2$ не должна быть большой.

В качестве меры близости наблюдаемых и ожидаемых частот рассмотрим статистику:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij} - \overline{n}_{ij}\right)^2}{\overline{n}_{ij}},$$

где сумма берется по всем ячейкам таблицы сопряженности. В данном случае χ^2 есть мера согласия опытных данных с теоретической моделью.

Если верна модель, по которой рассчитаны теоретические частоты n_{ij}, то при неограниченном росте числа наблюдений распределение случайной величины χ^2 стремится к распределению хи-квадрат с (r–1)·(s–1) степенями свободы. Для зависимых статистика χ^2 неограниченно возрастает при увеличении n. Поэтому большие (неправдоподобно большие для хи-квадрат) значения указывают на взаимную зависимость.

Какие же значения χ^2 надо считать настолько большими, что они не совместимы с нулевой гипотезой? Очевидно те, появление которых при гипотезе маловероятны, т.е. те, которые превосходят критические значения хи-квадрат, соответствующие выбранному уровню значимости.

Применение критерия χ^2 правомерно, если ожидаемое число в любой из клеток больше или равно 5.

В частном случае таблиц сопряженности, когда признаки A и B принимают только по 2 значения (обычно первое из них – наличие признака, а второе – его отсутствие), статистика χ^2 упрощается

$$\chi^2 = \frac{n\left(n_{11}n_{22} - n_{12}n_{21}\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

В этой ситуации статистика χ^2 имеет распределение хи-квадрат с одной степенью свободы (2–1)·(2–1) = 1.

Приведенная ранее формула для χ^2 в случае таблицы 2×2 (т.е. при 1 степени свободы) дает несколько завышенные значения. Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных χ^2 дискретен. На практике это приведет к тому, что нулевая гипотеза будет отвергаться слишком часто. Чтобы компенсировать этот эффект, в формулу вводят поправку Йейтса (Yates):

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(\left|n_{ij} - \overline{n}_{ij}\right| - \frac{1}{2}\right)^2}{\overline{n}_{ij}}.$$

Заметим, поправка Йейтса, применяется только при 1 степени свободы, т.е. для таблиц 2×2.

Если признаки оказались взаимосвязаны (гипотеза об их независимости проверена и отвергнута), исследователя интересует сила их связи. Было предложено много различных коэффициентов, называемых мерами связи. Рассмотрим меру связи на примере таблицы 2×2. Коэффициент Юла:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

Пример.

Тромбоз шунта у больных на гемодиализе.

Гемодиализ позволяет сохранить жизнь людям, страдающим хронической почечной недостаточностью. При гемодиализе кровь больного пропускают через искусственную почку – аппарат, удаляющий из крови продукты обмена веществ. Искусственная почка подсоединяется к артерии и вене больного: кровь из артерии поступает в аппарат и оттуда, уже очищенная, – в вену. Так как гемодиализ проводится регулярно, больному устанавливают артериовенозный шунт. В артерию и вену на предплечье вводят тефлоновые трубки; их концы выводят наружу и соединяют друг с другом. При очередной процедуре гемодиализа трубки разъединяют между собой и присоединяют к аппарату. После диализа трубки вновь соединяют, и кровь течет по шунту из артерии в вену. Завихрения тока крови в местах соединения трубок и сосудов приводят к тому, что шунт часто тромбируется.

Тромбы приходится регулярно удалять, а в тяжелых случаях даже менять шунт.

Руководствуясь тем, что аспирин препятствует образованию тромбов, Г. Харатер решил проверить, нельзя ли снизить риск тромбоза назначением небольших доз аспирина (160 мг/сут.). Было проведено контролируемое испытание. Все больные, согласившиеся принять участие в испытании и не имевшие противопоказаний к аспирину, были случайным образом разделены на две группы: 1-я получала плацебо, 2-я – аспирин. Ни врач, дававший больному препарат, ни больной не знали, был это аспирин или плацебо. Такой способ проведения испытания (он называется двойным слепым) исключает "подсуживание" со стороны врача или больного и, хотя он технически сложен, дает наиболее надежные результаты. Исследование проводилось до тех пор, пока общее число больных с тромбозом шунта не достигло 24. Группы практически не различались по возрасту, полу и продолжительности лечения гемодиализом.

В 1-й группе тромбоз шунта произошел у 18 из 25 больных, во 2-й – у 6 из 19. Можно ли говорить о статистически значимом различии доли больных тромбозом, а тем самым об эффективности аспирина?

Занесем результаты испытания в таблицу.

	Тромбоз есть	Тромбоза нет	
Плацебо	18	7	25
Аспирин	6	13	19
	24	20	44

(Во всех клетках больше 5-ти значений, следовательно, можно применять критерий хи-квадрат.)

Посмотрим на клетки, расположенные на диагонали, идущей из верхнего левого в нижний правый угол. Числа в них заметно больше, чем числа в других клетках таблицы.

Это наводит на мысль о связи между приемом аспирина и риском тромбоза.

Теперь вычислим ожидаемые числа, которые мы получили бы, если бы аспирин не влиял на риск тромбоза.

	Тромбоз есть	Тромбоза нет	
Плацебо	13,64	11,36	25
Аспирин	10,36	8,64	19
	24	20	44

Сравним таблицы. Числа в клетках довольно сильно различаются. Следовательно, реальная картина отличается от той, которая наблюдалась бы, если бы аспирин не оказывал влияния на риск тромбоза. Теперь посчитаем значение критерия хи-квадрат, которое будет характеризовать эти различия одним числом.

	Тромбоз есть	Тромбоза нет	
Плацебо	18 – 13,64= 4,36	7 – 11,36= – 4,36	25
Аспирин	6 – 10,36= – 4,36	13 – 8,64= 4,36	19
	24	20	44

$$\chi^2 = \frac{4{,}36^2}{13{,}64} + \frac{4{,}36^2}{11{,}36} + \frac{4{,}36^2}{10{,}36} + \frac{4{,}36^2}{8{,}64} =$$

$$= 4{,}36^2\,(0{,}0733 + 0{,}0880 + 0{,}0965 + 0{,}1157) = 7{,}1000\,.$$

Применим поправки Йейтса

$$\chi^2 = \frac{(4{,}36-0{,}5)^2}{13{,}64} + \frac{(4{,}36-0{,}5)^2}{11{,}36} + \frac{(4{,}36-0{,}5)^2}{10{,}36} + \frac{(4{,}36-0{,}5)^2}{8{,}64} =$$

$$= (3{,}86)^2\,(0{,}0733 + 0{,}0880 + 0{,}0965 + 0{,}1157) = 5{,}5650$$

Значения хи-квадрат для одной степени свободы представлены ниже.

	α							
ν	0,5	0,25	0,1	0,05	0,025	0,01	0,005	0,001
1	0,46	1,32	2,71	3,84	5,02	6,63	7,88	10,83

Пусть мы задали 5% -ный уровень значимости $\alpha = 0,05$. Тогда критическое значение = 3,84. Полученное значение $\chi^2=5,56$ больше, чем критическое, следовательно, мы отвергаем гипотезу о том, что аспирин не влияет на проявление тромбоза шунта. Следовательно, мы можем утверждать (с достоверностью 5%), что использование аспирина эффективно снижает риск тромбоза.

ДИСПЕРСИОННЫЙ АНАЛИЗ

Понятие о дисперсионном анализе

Дисперсионный анализ был разработан в 20-х годах XX-го столетия английским математиком и генетиком Рональдом Фишером.

То, что оказывает влияние на конечный результат, называется *фактором* или *факторами*, если их несколько. Конкретную реализацию фактора называют *уровнем* фактора. Значение измеряемого признака называют *откликом*.

Для сравнения влияния факторов на результат необходим определенный статистический материал. Обычно его получают следующим образом: каждый из k способов обработки применяют несколько раз (не обязательно одно и тоже число раз) к исследуемому объекту и регистрируют результаты. Итогом подобных испытаний являются k выборок, вообще говоря, разных объемов (численностей).

Одной из главных конечных целей в задачах однофакторного анализа является оценка величины влияния конкретного способа обработки на изучаемый отклик. Эта задача также может быть сформулирована в форме сравнения влияния двух или нескольких способов обработки между собой, т.е. оценки различия действий между уровнями фактора.

Но прежде чем судить о количественном влиянии фактора на измеряемый признак, полезно спросить себя, есть ли такое влияние вообще. Нельзя ли объяснить расхождения наблюдаемых в опыте значений для разных уровней одного фактора действием чистой случайности. На статистическом языке это предположение означает, что все данные в таблице принадлежат одному и тому же распределению. Это предположение обычно именуют нулевой гипотезой. Для проверки нулевой гипотезы могут быть использованы различные критерии: как параметрические, опирающиеся на предположение о нормальности распределения дан-

ных (F-отношение), так и непараметрические, не требующие подобных допущений (ранговые критерии Краскела–Уоллиса).

Если нулевая гипотеза об отсутствии эффектов обработки отвергается, то проводится оценка действия этих эффектов или контрастов между ними и строятся доверительные интервалы для этих характеристик.

Если же критерий не позволяет отвергнуть нулевую гипотезу об отсутствии эффектов обработки, что обычно на этом анализ может быть завершен. Но иногда вывод об отсутствии эффектов обработки нас не может устроить, так как он противоречит теоретическим предпосылкам или результатам предыдущих исследований. Тогда следует выяснить, нет ли каких-либо еще факторов, влияющих на имеющиеся наблюдения. Ниже будет рассмотрен метод двухфакторного анализа, используемые для решения задач, в которых на конечный результат влияют не один, а два фактора.

Дисперсионный анализ предназначен для исследования двух и более выборок путем сравнения выборочных дисперсий. В общем случае простейшая задача для дисперсионного анализа может выглядеть следующим образом.

Пусть имеется несколько (l) независимых выборок:

$$X_{11},\ldots, X_{1n_1};$$

$$\ldots$$

$$X_{l1},\ldots, X_{ln_l};$$

произведенных из нормальных генеральных совокупностей с неизвестными средними $m_1,\ldots, m_l$ и неизвестными одинаковыми дисперсиями σ^2.

Зададимся вопросом, что заставляет нас, взглянув на несколько выборок, думать, что различия между ними случайны.

Рассмотрим два расположения выборок.

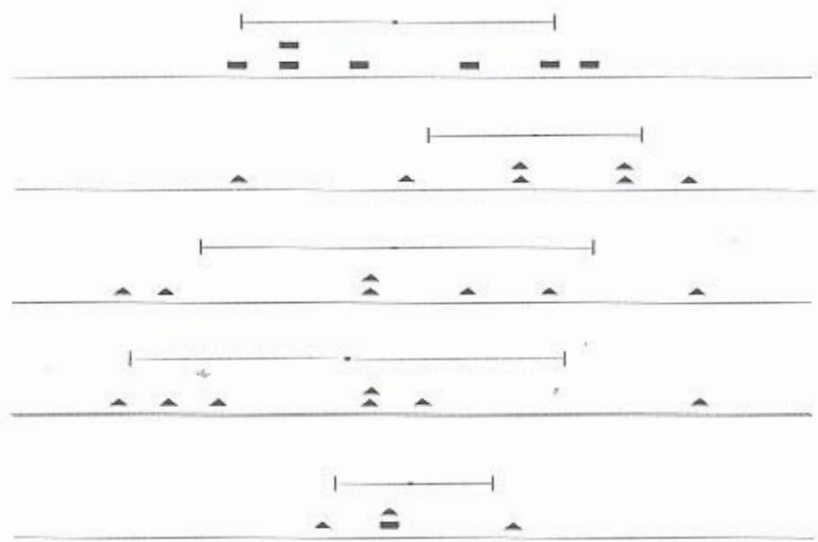


Рис.44. Разброс выборочных средних (нижняя линия) меньше разброса в каждой из выборок

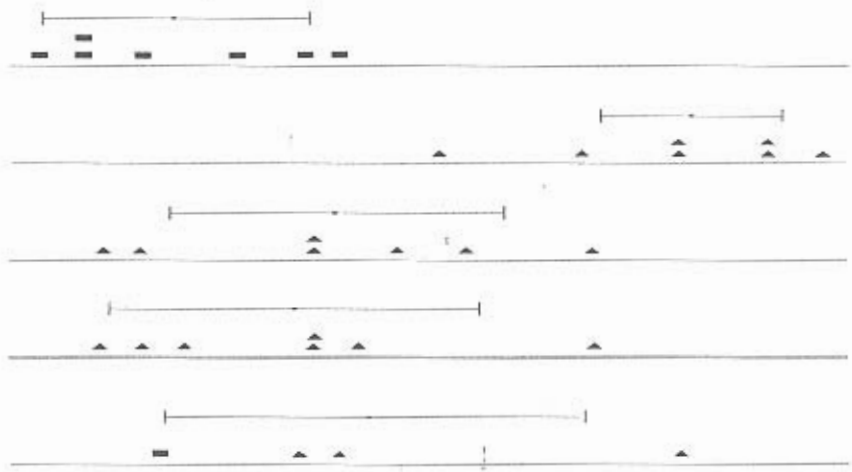


Рис.45. Разброс выборочных средних (нижняя линия) превышает разброс в каждой из выборок

Сравнив рисунки, всякий скажет, что выборки на рис. 44 не различаются, а на рис. 45 различаются. Почему? Сравним разброс значений внутри выборок с разбросом выборочных средних. Разброс выборочных средних на рис. 44 значительно меньше разброса в каждой из выборок. На рис. 45 картина обратная – разброс выборочных средних превышает разброс в каждой из выборок.

Итак, чтобы оценить величину различий, нужно каким-то образом сравнить разброс выборочных средних с разбросом значений внутри групп.

Дисперсионный анализ можно использовать для определения значимости различия средних значений какого-либо параметра жизнедеятельности у животных в нескольких группах, подвергавшихся воздействию препаратов разной дозировки. В последнем случае каждый из препаратов является неким фактором, который может оказывать существенное влияние на изучаемый параметр, а может и не оказывать такого влияния. Дисперсионный анализ обычно применяют для изучения влияния факторов, характеризующихся несколькими уровнями (в рассмотренном примере – дозами используемых препаратов).

В зависимости от количества изучаемых факторов различают *однофакторный* и *многофакторный дисперсионные анализы*.

Однофакторный дисперсионный анализ

Пусть исследуется влияние некоторого фактора A, имеющего m постоянных уровней, на формирование значений некоторой нормально распределенной величины X, причем на всех уровнях распределение значений величины X является нормальным, а генеральные дисперсии неизвестны, но одинаковы.

Пусть также количество проведенных наблюдений при действии фактора на каждом из его уровней одинаково и равно n, полученные результаты представлены в таблице, приведенной ниже.

Номер испытания	Уровень фактора			
	A_1	A_2	…	A_m
1	x_{11}	x_{21}	…	x_{m1}
2	x_{12}	x_{22}	…	x_{m2}
…			…	
n	x_{1n}	x_{2n}	…	X_{mn}
Групповое среднее	$\overline{x}_{1.}$		…	$\overline{x}_{m.}$
Общее среднее.	$\overline{x}_{..}$			

Все значения величины x_{ij}, (где i=1, 2, ..., n; j=1, 2, ..., m), наблюдаемые при каждом фиксированном уровне действия фактора A_i,., составляют *группу*, и в последней строке таблицы представлены соответствующие выборочные групповые средние, вычисленные по формуле

$$\overline{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

и в общем случае различающиеся между собой. Однако самого факта этих различий еще недостаточно для того, чтобы сделать вывод о существенном влиянии изучаемого фактора на величину X, поскольку необходимо убедиться в том, что данные различия вызваны именно изучаемым фактором, а не случайными причинами. С этой целью в подобных случаях и применяют дисперсионный анализ.

Итак, всего измерений

$$N = n_1 + n_2 + \ldots + n_j + \ldots + n_m .$$

Условимся называть групповым средним величину среднего значения, вычисленного по столбцу данной таблицы. Общее среднее по всей выборке складывается из средних значений данных по столбцам:

$$\overline{x}_{..} = \frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ij} ,$$

где N – суммарное число всех измерений по всем факторам (градациям фактора).

Среднее для каждого уровня:

$$\overline{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} .$$

Назовем общей дисперсией величину, зависящую от суммы квадратов разностей каждого значения и общего среднего:

$$S^2_{\text{общ}} = \frac{1}{N-1} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_{..})^2 .$$

Дисперсия групповых средних или межгрупповая дисперсия (факторная дисперсия):

$$S^2_{\text{гр}} = \frac{1}{m-1} \sum_{j=1}^{m} n_j (\overline{x}_{.j} - \overline{x}_{..})^2 .$$

Остаточная дисперсия:

$$S^2_{\text{ост}} = \frac{1}{N-m} \left[(N-1) S^2_{\text{общ}} - (m-1) S^2_{\text{гр}} \right].$$

В основе однофакторного дисперсионного анализа лежит сравнение межгрупповой и остаточной дисперсий.

Условимся, что фактор, воздействующий на измеряемые величины, существенно влияет на среднее значение в том случае, если генеральная межгрупповая дисперсия больше генеральной остаточной дисперсии. В противном случае делают заключение об отсутствии значимого влияния фактора на генеральные средние. Сравнение межгрупповой и остаточной дисперсий проводится по критерию Фишера.

Последовательность проверки гипотезы такова:

1. Формулируем нулевую и альтернативную гипотезы.

Нулевая гипотеза H_0: групповые генеральные средние равны ($\mu_1 = \mu_2 = \ldots = \mu_m$), а также различие выборочных средних получилось случайно, реального влияния фактор не оказывает.

Альтернативная гипотеза предполагает, что различие между выборочными средними не случайно и обусловлено влиянием фактора.

2. Задается уровень значимости α.
3. Вычисляются $S^2_{гр.}$ и $S^2_{ост.}$.

Если $S^2_{гр.} \le S^2_{ост.}$, то признается нулевая гипотеза.

Если $S^2_{гр.} > S^2_{ост.}$, то вычисляется функция:

$$F = \frac{S^2_{гр}}{S^2_{ост}} .$$

4. После вычисления $F_{набл.}$ находится $F_{критич.}$ по таблицам критических значений распределения Фишера. $F_{критич.}$ должно соответствовать числам степеней свободы, $k_{гр.} = m-1$ и $k_{ост.} = N - m$.
5. Сравниваются $F_{набл.}$ и $F_{критич.}$ Если $F_{набл.} < F_{критич.}$ то принимается гипотеза H_o о существенном влиянии фактора на средние значения. Если $F_{набл.} > F_{критич.}$ то нулевая гипотеза отвергается и делают вывод, что фактор не влияет существенно на средние значения.

Математическая модель, на которой основано вычисление критических значений F, предполагает следующее:

1. Каждая выборка независима от остальных выборок.
2. Каждая выборка случайным образом извлечена из исследуемой совокупности.
3. Совокупность нормально распределена.
4. Дисперсии всех выборок равны.

При существенном нарушении хотя бы одного из этих условий нельзя пользоваться дисперсионным анализом. В этом случае надо использовать его непараметрический аналог.

Основная идея дисперсионного анализа состоит в сравнении групповой дисперсии, порождаемой воздействием фактора и остаточной дисперсии, обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на измеряемую величину. В этом случае средние наблюдаемых значений на каждом уровне также значимо различаются.

Гипотеза:

H_0: фактор НЕ влияет,

H_1: фактор влияет.

Критерий $F = \frac{S^2_{ср}}{S^2_{ост}}$ имеет распределение Фишера с $(m-1),(N-m)$ степенями свободы.

Если $F_{эксп} > F_{крит}$, то нулевая гипотеза отвергается, следовательно, есть влияние фактора.

Если $F_{эксп} < F_{крит}$, то нет оснований отвергать нулевую гипотезу, следовательно, нет влияния фактора.

Базовая идея дисперсионного анализа заключается:

в разложении общей дисперсии изучаемых признаков на составляющие в соответствии с возможными источниками вариации;

вычислении F-отношений в качестве тестовой статистики;

проверки значимости нулевой гипотезы (об отсутствии существенного влияния данного фактора на общий разброс данных).

Чтобы определить величину различий средних значений нескольких независимых выборок, нужно попытаться сравнить разброс самих выборочных средних с разбросом значений признака вокруг соответствующего группового среднего внутри групп. Чем больше разброс средних и меньше разброс значений внутри групп, тем меньше вероятность того, что данные группы представляют собой случайные выборки из одной и той же генеральной совокупности.

Модель однофакторного дисперсионного анализа.

Фактор – это качество или свойство, в соответствии с которым в нашей модели производится классификация.

$x_{ij} = \mu + \varsigma_i + z_{ij}$,

x_{ij} – измеренная величина,

μ – общее среднее,

ς_i – индивидуальный вклад i-го уровня фактора (случайная величина),

z_{ij} – индивидуальный вклад остаточного эффекта (т.е. вклад факторов, которые мы не в состоянии описать с помощью известных в данный момент качественных или количественных параметров системы). Тогда гипотеза примет следующий вид:

H_0: $\xi_1 = \xi_2 = \ldots = \varsigma_k = 0$ (т.е. уровни фактора не вносят свой вклад),

H_1: хотя бы один уровень фактора вносит вклад.

Далее используется F-критерий.

Итак, схема порядка операций для однофакторного дисперсионного анализа:

исходные данные группируются в виде комбинационной таблицы таким образом, чтобы градации регулируемого фактора располагались по горизонтали в верхней части таблицы, образуя ее графы или столбцы, а значения результативного признака (варианты) группировались соответственно по градациям фактора;

рассчитывают вспомогательные величины: объем выборки, общую среднюю, групповые средние, общую сумму квадратов, межгрупповую сумму квадратов, внутригрупповую (остаточную) сумму квадратов;

определяют числа степеней свободы;

определяют средние квадраты отклонений или дисперсии по отношениям сумм квадратов отклонений к соответствующим числам степеней свободы;

определяют эффективность действия фактора на результативный признак. Для этого служит дисперсионное отношение, или критерий Фишера;

результат дисперсионного анализа сводят в таблицу.

Источник вариации, дисперсии	Сумма квадратов (отклонений)	Число степеней свободы	Средние квадраты	Дисперсионное отношение, (F)
Межгрупповая	$SS_{межгр} = \sum_{j=1}^{m} n_j \left(\bar{x}_{\cdot j} - \bar{x}\right)^2$	$M-1$	$s^2_{ср} = \frac{SS_{межгр}}{m-1}$	$F = \frac{S^2_{ср}}{S^2_{внутригр}}$
Внутригрупповая (остаточная)	$SS_{внутригр} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{i\cdot}\right)^2$	$N-m$	$s^2_{внутригр} = \frac{SS_{внутригр}}{N-m}$	
Общая	$SS_{общ} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2$	$N-1$	$s^2_{общ} = \frac{SS_{общ}}{N-1}$	F следует распределению Фишера с (m-1),(N-m) степенями свободы.

Двухфакторный дисперсионный анализ. Иерархическая модель

В иерархической модели подразумевается, что один фактор – основной, а внутри основного фактора каждый уровень может быть разделен на подуровни главного фактора.

Пример.

Пусть есть четыре препарата, лечащие одно и то же: A_1, A_2, A_3, A_4.

При клинических испытаниях изучается некоторая величина X. Каждый из препаратов выпускается различными фирмами B_1, B_2, B_3, ... B_{10}. Мы должны выбрать препараты от конкретной фирмы, чтобы препараты были качественными. Здесь – иерархическая структура: главный фактор – препарат, подуровни – фирмы-поставщики.

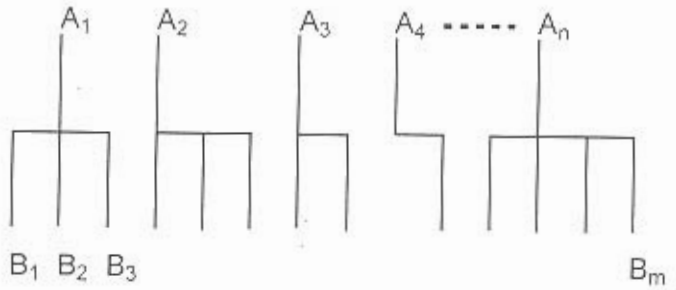


Рис. 46. Иерархическая модель

Рассмотрим математическую модель иерархического дисперсионного анализа.

Пусть главный фактор имеет k-уровней, а подчиненный фактор - m_i уровней в i-ом уровне главного фактора, n_{ij} – количество измерений в j-ом уровне подчиненного фактора i-го уровня главного фактора, тогда

$$X_{ijt} = \mu + \xi_i + \delta_{ij} + z_{ijt},$$

где $1 \le i \le k$; $1 \le j \le m_i$; $1 \le t \le n_{ij}$,

X_{ijt} – значение изучаемой величины,

μ – общее среднее,

ξ_i – случайная величина, характеризующая влияние i-го уровня главного фактора,

δ_{ij} – случайная величина, характеризующая влияние j-го подуровня i-го уровня главного фактора,

z_{ijt} – индивидуальный вклад остаточного эффекта (т.е. фактора, который мы не в состоянии описать с помощью известных в данный момент качественных или количественных параметров системы), т.е. случайные причины.

Сделаем вспомогательные вычисления:

$N_{i.} = \sum_{j=1}^{m_i} n_{ij}$ – количество измерений в i-ом уровне главного фактора,

$N_{..} = \sum_{i=1}^{k} N_{i.}$ – количество всех измерений,

$\sum_{i=1}^{k} m_i$ – количество подуровней подчиненного фактора,

$\overline{x}_{i..}$ – оценка среднего значения i-го уровня главного фактора,

$\overline{x}_{...}$ – оценка общего среднего,

$\overline{x}_{ij.}$ – оценка среднего значения j-го уровня подчиненного фактора i-го уровня главного фактора.

$$s^2_{гл.фак.} = \frac{1}{k-1} \sum_{i=1}^{k} N_{i.} \left(\overline{x}_{i..} - \overline{x}_{...}\right)^2,$$

$$s^2_{подч.фак} = \frac{1}{\sum_{i=1}^{k} m_i - k} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(\overline{x}_{ij.} - \overline{x}_{i..}\right)^2,$$

$$s^2_{ост} = \frac{1}{N_{..} - \sum_{i=1}^{k} m_i} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \sum_{t=1}^{n_{ij}} \left(\overline{x}_{ijt} - \overline{x}_{i..} \right)^2 .$$

Здесь проверяются следующие гипотезы:

H^1_0: главный фактор не влияет на результат,

H^1_1 : главный фактор влияет на результ.

Критерий $F^1 = \frac{s^2_{гл.фак.}}{s^2_{ост}}$ имеет распределение Фишера с (k–1), $\left(N_{..} - \sum_{i=1}^{k} m_i \right)$ степенями свободы.

H^2_0: подчиненный фактор не влияет на результат.

H^2_1: подчиненный фактор влияет на результат.

Критерий $F^2 = \frac{s^2_{подч.фак.}}{s^2_{ост}}$ имеет распределение Фишера с $\left(\sum_{i=1}^{k} m_i - k \right), \left(N_{..} - \sum_{i=1}^{k} m_i \right)$ степенями свободы.

Перекрестная модель

Пример.

1-й фактор – лекарство (A,B,C,D).

2-й фактор способ применения (мы хотим выяснить каким путем лучше вводить препарат) (I,II,III).

Для каждого препарата исследуются все три способа.

	A	B	C	D
I	xxx	xxxxxx	xxx	xxxxxx
II	xxxxx	xxx	xxxxxx	xxx
III	xxx	xxx	xxx	xx

В каждой ячейке должно быть как минимум 2 измерения.

Математическая модель:

Пусть I фактор имеет r-уровней, а II фактор c-уровней. $n_{ij}=n$ – количество измерений в i-ом уровне I фактора и j-м уровне II фактора (пусть для простоты будет одинаковое количество),

$$x_{ijt} = \mu + \xi_i + \eta_j + (\xi\eta)_{ij} + z_{ijt},$$

где $1 \le i \le r$; $1 \le j \le c$; $1 \le t \le n_{ij}$

x_{ijt} – результат измерения,

μ – общее среднее,

ξ_i – случайная величина, характеризующая вклад I фактора,

η_j – случайная величина, характеризующая вклад II фактора,

$(\xi\eta)_{ij}$ – случайная величина, характеризующая вклад межфакторного взаимодействия,

z_{ijt} – случайная величина, характеризующая вклад неучтенных факторов.

Сделаем вспомогательные вычисления.

$\overline{x}_{i..}$ – оценка среднего значения i-го уровня I фактора,

$\overline{x}_{.j.}$ – оценка среднего значения j-го уровня II фактора,

$\overline{x}_{...}$ – оценка общего среднего,

$\overline{x}_{ij.}$ – оценка среднего значения i-го уровня I фактора j-го уровня II фактора (среднее в ячейке),

$$s^2_{Iфакт.} = \frac{1}{r-1} nc \sum_{i=1}^{r} \left(\overline{x}_{i..} - \overline{x}_{...} \right)^2 ,$$

$$s^2_{IIфакт.} = \frac{1}{r-1} nr \sum_{j=1}^{c} \left(\overline{x}_{.j.} - \overline{x}_{...} \right)^2 ,$$

$$s^2_{межс.фак.} = \frac{1}{(r-1)(c-1)} n \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\overline{x}_{ij.} - \overline{x}_{i..} - \overline{x}_{.j.} + \overline{x}_{...} \right)^2 ,$$

$$s^2_{ост.} = \frac{1}{rc(n-1)} \sum_{j=1}^{r} \sum_{j=1}^{c} \sum_{i=1}^{n} \left(x_{ijt} - \overline{x}_{...} \right)^2 ,$$

Здесь проверяются следующие гипотезы:

H^1_0: I фактор не влияет на результат.

H^1_1 : I фактор влияет на результат

Критерий $F^1 = \frac{S^2_{I\,факт.}}{S^2_{ост}}$ имеет распределение Фишера с $(r-1),rc(n-1)$ степенями свободы.

H^2_0: II фактор не влияет на результат.

H^2_1 : II фактор влияет на результат.

Критерий $F^2 = \frac{S^2_{II\,факт.}}{S^2_{ост}}$ имеет распределение Фишера с $(c-1),rc(n-1)$ степенями свободы.

H^{12}_0: II фактор не влияет на результат.

H^{12}_1 : II фактор влияет на результат.

Критерий $F^{12} = \frac{S^2_{межс.\,факт}}{S^2_{ост}}$ имеет распределение Фишера с $(r-1)(c-1),rc(n-1)$ степенями свободы.

Результат дисперсионного анализа сводят в таблицу.

Источник вариации (дисперсии)	Число степеней свободы	Средние квадраты	Дисперсионное отношение. (F)
I фактор	$r-1$	$s^2_{Iфакт.} = \frac{1}{r-1} nc \sum_{i=1}^{r} (\bar{x}_{i..} - \bar{x}_{...})^2$	F следуейт распределению Фишера с $(m-1),(N-m)$ степенями свободы.
II - фактор	$c-1$	$s^2_{IIфакт.} = \frac{1}{r-1} nr \sum_{i=1}^{c} (\bar{x}_{.j.} - \bar{x}_{...})^2$	
Межфакторное взаимодействие	(r-1)(c-1)	$s^2_{межс.фак.} = \frac{1}{(r-1)(c-1)} n \sum_{i=1}^{r} \sum_{j=1}^{c} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2$	
Остаточная дисперсия	rc(n-1)	$s^2_{ост.} = \frac{1}{rc(n-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{l=1}^{n} (x_{ijl} - \bar{x}_{ij.})^2$	
Общая	rcn-1		

Пример.

На одной из опытных станций испытывалась урожайность шести местных сортов пшеницы. Опыт проводился в четырехкратной повторности по каждому сорту. Результаты испытания приведены в таблице.

Цейтнер	Сорт					
	1	2	3	4	5	6
1	26,1	25,0	27,2	23,6	30,0	23,0
2	29,2	24,3	26,4	27,2	33,0	26,0
3	30,0	28,5	31,0	25,2	36,0	26,0
4	27,3	29,0	26,4	24,8	29,8	24,8
Среднее	28,2	26,7	27,8	25,2	32,2	25,0
общ среднее	27,5					

Из данных таблицы видно, что на одни и теже условия выращивания сорта пшеницы реагируют по-разному. Подвергнем эти данные дисперсионному анализу.

H_0: нет влияния фактора (сорт пшеницы), т.е. варьирование носит случайный характер.

	SS	ст. свободы	S2	F	$F_{0,05;0,05}$	H_0
Фактор	140,0	5	28,01	6,4	2,8	Reject
Остаток	79,6	18	4,4			
Общее	219,6	23	9,55			

Можно заключить, что разница в урожайности между сортами пшеницы не случайна. Т.е. есть влияние фактора.

Пример.

На учебно-опытном участке изучалось влияние различных способов внесения в почву органических удобрений на урожай зеленой массы кукурузы. Каждый вариант опыта имел трехкратную повторность. Результаты опыта оказались следующими:

кг	Способ внесения удобрения			
	1	2	3	4
1	21,2	23,6	24	29,2
2	28	22,6	30	28
3	31,2	28	29,2	27
Среднее	26,8	24,7	27,7	28,1
Общее ср.	26,8			

Видно, что результаты опыта варьируют как по вариантам, так и по повторностям. Чтобы установить случайны или не случайны различия между средними групп, подвергнем эти данные дисперсионному анализу.

H_0: нет влияния фактора (способ удобрения), т.е. варьирование носит случайный характер.

	SS	Ст. св	Дисп	F	$F_{8,3;0.05}$	H_0
Фактор	20,23	3	6,74	1,71	8,84	Нет осн отв.
Остаток	92,32	8	11,54	Здесь отношение большего к меньшему		
Общ	112,55	11	10,23			

Нулевая гипотеза остается в силе, т.е. варьирование носит случайный характер. Т.е. нет влияния фактора.

АНАЛИЗ ВЫЖИВАЕМОСТИ

Особенность методов *анализа выживаемости* состоит в том, что они применяются к неполным данным. Отметим также, что более часто, чем обычная функция распределения, в этих методах используется так называемая функция выживания, представляющая собой вероятность того, что объект проживет время больше *t*. Построение *таблиц времен жизни*, оценивание функции выживания с помощью процедуры *Каплана–Мейера* являются описательными методами исследования цензурированных данных. Некоторые из предложенных методов позволяют сравнивать выживаемость в двух и более группах.

Цензурированные и нецензурированные данные

До сих пор мы имели дело только с полными данными: мы знали исход лечения у каждого больного, срок наблюдения всех пациентов был одинаков, и никто из них не выбыл из-под наблюдения до завершения исследования. Однако ситуация, когда исследование должно быть завершено до наступления исхода у всех больных для клинических испытаний, скорее правило, чем исключение.

Наиболее типичный пример исследования такого рода – это изучение выживаемости, когда пациентов наблюдают от начала болезни до смерти. Обычно больных включают в исследование на всем его протяжении, поэтому оно всегда заканчивается до смерти последнего больного. Истинная продолжительность болезни остается неизвестной. Кроме того, исследование может потерять больного из виду до завершения исследования, если тот, к примеру, переехал в другой город. Наконец, больной может умереть по причине, не связанной с изучаемым заболеванием, например, погибнуть в автокатастрофе. Во всех этих случаях длительность заболевания остается неизвестной, мы знаем только,

что она превышает некоторый срок. Такие данные называются цензурированными.

Время жизни – это время до появления некоторого (заранее определенного) события. Например, таким событием может быть развитие заболевания, реакция на лечение, рецидив (повторение) или смерть. Таким образом, временем жизни может быть время до начала развития заболевания, время от начала лечения до реакции на него, время ремиссии (от начала улучшения здоровья до рецидива), время до смерти.

Мы могли бы применять к данным известные нам параметрические или непараметрические методы. Но особенность этих данных в том, что время жизни объектов бывает не известно. Например, пациент может быть жив или находиться в ремиссии после окончания исследования. В этих случаях точное время жизни не известно. Такие данные называются цензурированными (цензурированное время).

Применение такого типа анализа обусловлено тем, что исследователи должны анализировать данные, не дожидаясь пока у всех обследуемых наступит изучаемое событие, а пациенты входят в исследование не одновременно, а на всем протяжении исследования.

Разберем три типа, при которых возникают цензурированные данные.

В экспериментах над животными начинают эксперимент с фиксированным числом животных, но из-за ограниченности во времени или дороговизны исследователи не могут ждать смерти (появления события) у всех животных. По истечении времени эксперимента животные забиваются. Время жизни умерших животных – это время от начала эксперимента до момента смерти (момента появления события), т.е. точные или нецензурированные наблюдения. Время жизни забитых животных точно не известно, им приписывают время равное продолжительности эксперимента, т.е. цензурированные данные. Некоторые животные могли внезапно погибнуть или исчезнуть из поля зрения наблюдателя - это тоже цензурированные наблюдения, им приписывают время от начала эксперимента до их потери. В таких экспериментах, если нет внезапно выбывших, все цензурированные обследования имеют время, равное времени эксперимента.

Пример.

Шести крысам были введены опухолевые клетки. Исследовали время развития опухоли определенного размера (т.е. опухоль определенного размера является событием). Исследователи определили продолжительность эксперимента – 30 недель. На рисунке показаны результаты. У крыс A, B, D опухоль развилась на 10, 15, 25-й неделях соответственно. У крыс C и E опухоль не развилась до конца исследования, их время – 30+ недель. Крыса F внезапно погибла без опухоли после 19-й недели, ее время 19+. Итак данные для анализа выживаемости следующие: 10, 15, 30+, 25, 30+, 19+. Знак плюс означает цензурированные данные.

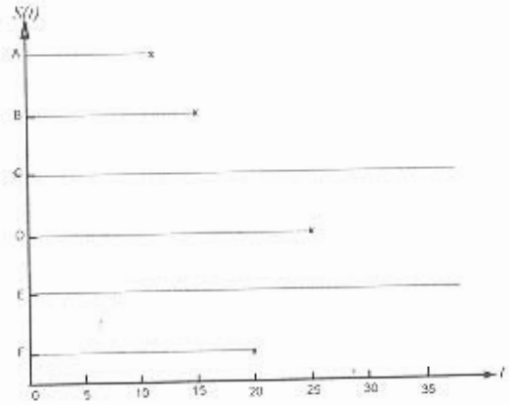


Рис. 47. Цензурированные данные (1-й тип)

Другой случай в экспериментах над животными – это когда дожидаются появления события у определенного числа животных, а остальных забивают. В таких случаях (если нет внезапно выбывших) время жизни цензурированных наблюдений равно наибольшему из времен нецензурированных наблюдений.

Пример.

В эксперименте с шелтью крысами экспериментаторы могут решить остановить эксперимент после развития

опухоли у четырех крыс. Тогда данные будут выглядеть следующим образом: 10, 15, 35+, 25, 35, 19+.

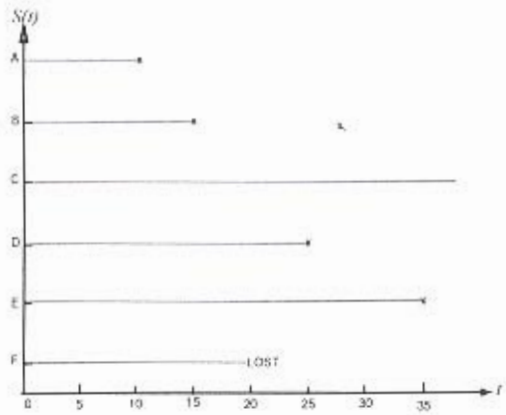


Рис. 48. Цензурированные данные (2-й тип)

В клинических исследованиях период проведения исследования фиксирован, а пациенты входят в него в разное время. Некоторые могут умереть до завершения эксперимента, их время жизни точно известно. Другие могут исчезнуть из поля зрения до завершения эксперимента, но они могут быть живы и после завершения эксперимента. Для таких пациентов время жизни, равно времени от их появления в исследовании до их последнего визита к врачу. Для тех пациентов, которые живы после завершения исследования, время жизни это время от их появления в исследовании до конца исследования. Два последних примера – это цензурированные данные.

Пример.

Предположим, что в течение одного года проводилось исследование, событием в нем являлось наступление рецидива после выхода в ремиссию (улучшение), шестеро пациентов с лейкемией вошли в клиническое исследование на всем протяжении. Пациенты A, C, E достигли ремиссии на 2, 4, 9-м месяцах лечения и дали рецидив через 4, 6 и 3 месяца соответственно. Пациент B достиг ремиссии на третьем месяце лечения и исчез из под контроля (например, уехал в

другой город или просто выписался) через четыре месяца, период его ремиссии считается как 4+. Пациенты D и F достигли ремиссии через 5 и 10 месяцев соответственно и до окончания исследования не дали рецидива, поэтому время их ремиссии считается 8+ и 3+ месяцев соответственно. Итак, получились следующие данные 4, 4+, 6, 8+, 3, 3+ месяцев.

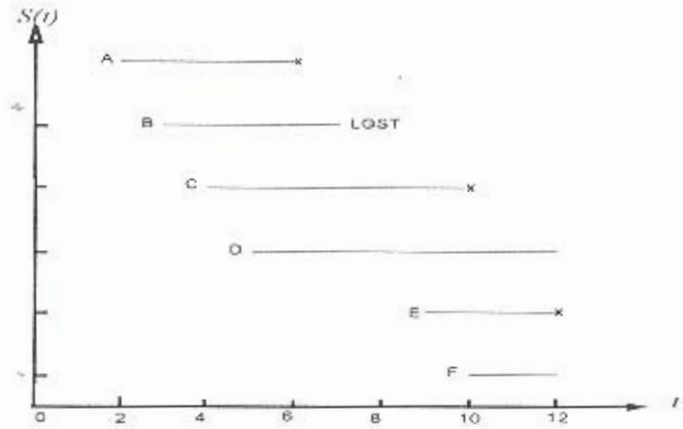


Рис. 49 Цензурированные данные (3-й тип)

Типы 1) и 2) – однократно цензурированные данные, тип 3) – последовательно цензурированные данные или случайно цензурированные. Все эти типы называются цензурированными справа. Когда нет цензурированных данных, исследование называется полным.

Определим требования, которым должны удовлетворять все исследования выживаемости:

1. Для всех исследуемых известно время начала наблюдения.
2. Для всех исследуемых известно время окончания наблюдения, а также – наступило событие или исследуемый выбыл.
3. Выбор наблюдаемых произведен случайно.

Кривая выживаемости

Полученные данные сводят в таблицы (таблица времени жизни) и по ним строятся графики функции (кривые выживаемости).

Поскольку время жизни заранее не известно, можно сказать, что оно является случайной величиной и для нее существует функция распределения.

Выживаемость S(t) – это вероятность прожить время большее t с момента начала наблюдения

$$S(t) = P(T>t).$$

Как правило, вместо этой формулы используют другую. Заменяют $P(T>t) = 1 - P(T<t)$,

$$S(t) = 1 - P(T<t),$$

где $P(T<t)$ – это вероятность наступления события (гибели) до времени t.

Свойства функции $S(t)$:

1. $S(t) = 1$ t = 0;
2. $S(t) = 0$ t = ∞.

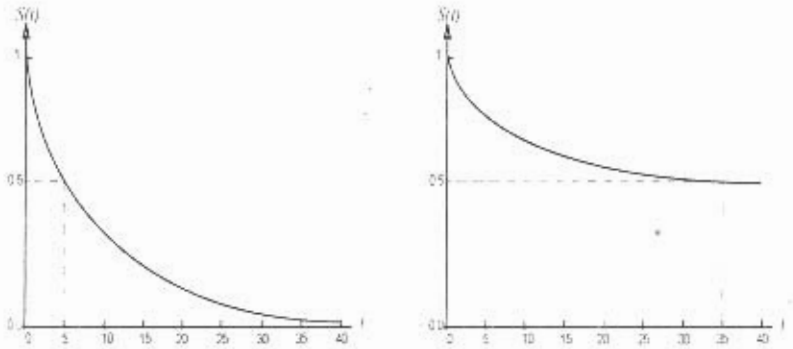


Рис. 50. Функция выживаемости: а) выживаемость низкая; б) выживаемость высокая

График функции $S(t)$ называется кривой выживаемости. Крутой график свидетельствует о низкой выживаемости, пологий– о высокой. Кривая выживаемости используется для нахождения медианы выживаемости и других процентилей времени жизни и для сравнения распределений двух и нескольких групп.

Если у нас исследование полное, т.е. нет цензурированных данных, то функция выживаемости оценивается отношением:

$$S(t) = \frac{\text{число переживших момент } t}{\text{объем совокупности}}$$

В тех случаях, когда есть цензурированные данные, числитель не всегда может быть определен, и эта формула дает ошибочный результат. В таких случаях применяют моментный метод Каплана–Мейера

$$S(t) = \prod \left(1 - \frac{d_t}{n_t}\right),$$

где d_t – число событий (умерших) в момент t, n_t – число наблюдавшихся в момент t. П (большая греческая «пи») – символ произведения. В данном случае она обозначает, что надо перемножить все значения $(1-d_t/n_t)$ для всех моментов, когда произошло хотя бы одно событие. В принципе можно перемножить и по остальным моментам, но при $d_t = 0$, $(1-d_t/n_t)=1$, а умножение на 1 на результатах не сказывается.

Наиболее полная характеристика выживаемости – это кривая выживаемости. Обобщенным же показателем является медиана выживаемости – это наименьшее время, для которого выживаемость меньше 0,5. Однако если число умерших меньше половины числа наблюдаемых, медиану определить нельзя.

Таблицы времени жизни

Есть несколько способов составления таблицы времени жизни.

Рассмотрим два способа, первый более применим для больших наборов данных, второй – для малого числа обследуемых.

Способ Катлера-Эдерера (Cutler-Ederer).

Интервалы времени	Количество наблюдаемых объектов к началу интервала	Количество событий произошедших в данном интервале времени	Количество выбываний, произошедших в данном интервале времени	доля наступления события в интервале	доля выживших в интервале	Кумулятивная доля выживших.
i	n_i	d_i	w_i	q_i	$p_i=1-q_i$	$S_i=p_ip_{i-1}\dots p_1$

Количество изучаемых объектов – это число объектов, которые были "живы" в начале рассматриваемого временного интервала.

Доля исследуемых, для которых событие наступило в *i*-ом интервале, это отношение числа объектов, для которых событие наступило в *i*-ом интервале, к числу объектов, изучаемых на этом интервале

$$q_i = \frac{d_i}{n_i - \frac{1}{2}w_i},$$

где w_i количество выбываний, произошедших в данном интервале.

Доля выживших, т.е. тех, для которых событие не наступило в *i*-ом интервале, равна единице минус *доля исследуемых*, для которых событие наступило в *i*-ом интервале,

$$p_i = 1 - q_i.$$

Кумулятивная доля выживших (функция выживания) – это кумулятивная доля выживших к началу соответствующего временного интервала. Поскольку вероятности выживания считают-

ся независимыми на разных интервалах, эта доля равна произведению долей выживших объектов по всем предыдущим интервалам. Полученная доля, как функция от времени, называется также *выживаемостью* или *функцией выживания,* точнее, это оценка функции выживания

$$\overline{S_i} = p_1 \cdot p_2 \cdot \dots \cdot p_{i-1} \cdot p_i.$$

Способ Каплана–Мейера (Kaplan–Meier).

Момент времени	Количество наблюдаемых объектов к моменту времени	Количество событий произошедших в данный момент времени	Количество выбываний, произошедших в данный момент времени	доля наступления события	доля выживших	Кумулятивная доля выживших.
i	n_i	d_i	w_i	q_i	$p_i=1-q_i$	$S_i=p_ip_{i-1}\dots p_1$

Во втором способе первый столбец не разбивается на интервалы, а в нем записываются моменты, в которые произошло хотя бы одно событие.

$$q_i = \frac{d_i}{n_i},$$

$$\overline{S}(t) = \prod\left(1 - \frac{d_i}{n_i}\right),$$

где d_i – число умерших в момент времени i, n_i – число наблюдавшихся к моменту i (это произведение по всем моментам

времени, когда произошла хотя бы одна смерть, за период от 0 до t).

Полученные результаты расчетов представляются в виде таблицы, строки которой соответствуют моментам времени, в которые происходила хотя бы одна смерть, а также в виде графика. Точки на графике тоже соответствуют моментам, когда умер хотя бы один из наблюдавшихся. Точки соединяются ступенчатой линией, этот график и будет выборочной оценкой кривой выживаемости. Кроме того, построенную кривую можно охарактеризовать и обобщенным показателем, например, медианой. Для этого надо найти точку, в которой кривая выживаемости впервые опускается ниже 0,5. Если в исследовании число умерших было меньше половины, найти медиану невозможно. При этом обобщенным показателем может быть любой другой процентиль (>50). По таблицам выживаемости более точно, чем при использовании общепринятого расчета среднего значения, может быть определен и показатель, называемый средней продолжительностью жизни.

Результаты, полученные в таблице, представляют в виде графика (кривой выживаемости). На самом деле эта кривая представляет собой оценку кривой выживаемости. Оценку точности приближения дает стандартная ошибка выживаемости; ее можно рассчитать по формуле Гринвуда (Greenwood)

$$SE(\overline{S}_i) = \overline{S}_i \sqrt{\sum \left(\frac{q_i}{n_i - d_i - \frac{1}{2} w_i} \right)},$$

где сумма берется по всем интервалам (моментам).

Доверительные границы для функции выживаемости рассчитываются следующим образом:

$$\overline{S}(t) - z_\alpha \cdot SE(\overline{S}(t)) < S(t) < \overline{S}(t) + z_\alpha \cdot SE(\overline{S}(t)).$$

Обычно определяют 95%-ный доверительный интервал. Тогда $\alpha = 1 - 0{,}95 = 0{,}05$. Соответствующее значение $z_\alpha = 1{,}96$.

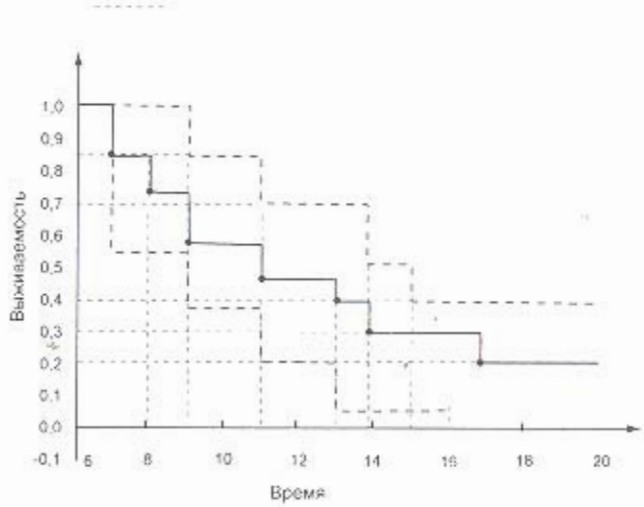


Рис. 51. Кривая выживаемости и доверительный интервал

Отложив на графике доверительные границы, мы увидим расширяющийся «рукав» – доверительную область для выживаемости. Причина расширения доверительной области понятна: чем меньше остается наблюдаемых, тем больше ошибка.

Приведенная выше формула дает симметричную оценку, которая может выйти за граничные значения 1 и 0. Простейший способ подправить такую оценку состоит в том, чтобы значения большие 1 заменить на 1, а значения меньшие 0 – заменить на 0.

Пример.

Рассмотрим анализ выживаемости на примере. Предположим, что 10 пациентов с острым приступом некоей болезни вошли в исследуемую группу на протяжении всего исследования (15 месяцев). Пусть все 10 дали отклик на лечение, и у них наступила ремиссия. Мы будем изучать продолжительность времени ремиссии, а событием будет рецидив (т.е. новый приступ).

	Start	End		Censor
A	1	8	7	
B	1	13	12	
C	2	9	7	
D	3	15	12	C
E	3	14	11	C
F	6	14	8	
G	6	15	9	
H	8	14	6	
I	8	15	7	C
J	12	14	2	

Пациенты A,B,C,F,G,H,J достигли ремиссии на 1,1,2,6,6,8,12 месяцах исследования соответственно и у них случился рецидив через 7,12,7,8,9,6,2 месяцев соответственно.

Пациенты D и I достигли ремиссии на 3-м и 8-м месяце исследования, и по окончании исследования оставались в ремиссии и не дали рецидива (эти данные являются цензурированными).

Пациент E достиг ремиссии на 3-м месяце и через 11 месяцев выписался (уехал), про него не известно, был ли рецидив (это тоже цензурированное данное).

Итак, мы имеем следующий набор данных: 7, 12, 7, 12+, 11+, 8, 9, 6, 7+, 2.

Построим таблицу времени жизни (ремиссии), а также кривую выживаемости и доверительные интервалы для нее.

Воспользуемся вторым способом (так как у нас немного данных)

Моменты	наблюдаемые	Давшие рецидив	Доля давших рецидив	Доля переживших момент без рецидива	Кумулятивная доля	$S_i - z_\alpha SE$ α=0,05 z=1,96	$S_i + z_\alpha SE$ α=0,05 z=1,96
1	2	3	4	5	6	7	8
	n_i	d_i	q_i	$p_i=1-q_i$	S_i		
2	10	1	0,10	0,900	0,90	0,715	1,085 (1)
6	9	1	0,11	0,889	0,80	0,557	1,043 (1)
7	8	2	0,25	0,750	0,60	0,314	0,886
8	5	1	0,20	0,800	0,48	0,169	0,791
9	4	1	0,25	0,750	0,36	0,051	0,669
12	2	1	0,5	0,5	0,18	-0,113 (0)	0,473

где $SE = S_i \sqrt{\sum \left(\frac{d_i}{n_i (n_i - d_i)} \right)}$

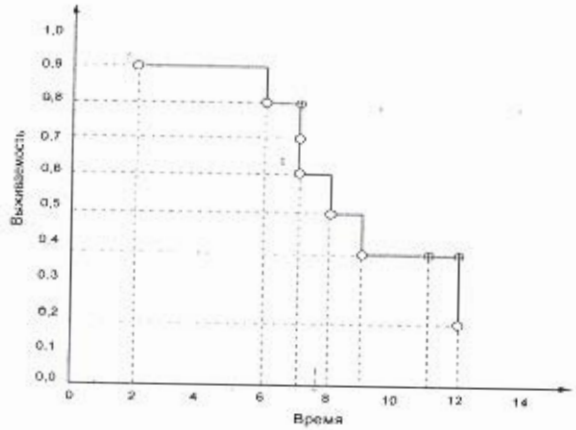


Рис. 52. График кривой выживаемости к примеру

Сравнение двух кривых выживаемости

В клинических исследованиях возникает необходимость сравнить выживаемость разных групп больных. Рассмотрим случай двух групп.

Нулевая гипотеза состоит в том, что в обеих группах выживаемость одинакова. Если бы больные не выбывали, и все больные наблюдались бы равное время, нам бы подошел анализ таблиц сопряженности признаков. Но выбывания неизбежны. Для сравнения кривых выживаемости нужны специальные методы. Рассмотрим два критерия.

Сначала сформулируем нулевую и альтернативную ей гипотезы:

H_0: $S_1(t) = S_2(t)$ методы эквивалентны,
H_1: $S_1(t) \neq S_2(t)$ методы не эквивалентны,
или $S_1(t) > S_2(t)$ 1-й метод эффективней,
или $S_1(t) < S_2(t)$) 2-й метод эффективней.

Логранговый тест. (Log-rank)

Для применения логрангового теста составляют общую таблицу выживаемости для обеих групп, находят ожидаемое число умерших для одной из групп

$$E_{it} = \frac{n_{it} d_{общ\,t}}{n_{общ\,t}},$$

где E_{it} – ожидаемое число умерших в группе i в момент времени t,

n_{it} – число наблюдавшихся в группе к этому моменту.

$d_{общ\,t}$ – общее число смертей в этот момент в обеих группах,

$n_{общ\,t}$ – общее число наблюдавшихся к этому моменту.

Выбывшие учитываются косвенно, влияя на число наблюдавшихся.

Далее суммируют разности наблюдаемого и ожидаемого числа умерших.

$$U_L = \sum (d_{it} - E_{it}).$$

Сумма берется по всем моментам t, когда хотя бы одна смерть наступила в любой из двух групп.

U_L приближенно подчиняется нормальному распределению со стандартным отклонением

$$S_{U_L} = \sqrt{\sum \frac{n_{1t} n_{2t} d_{общ\,t} (n_{общ\,t} - d_{общ\,t})}{n^2_{общ\,t} (n_{общ\,t} - 1)}},$$

где как и раньше сумма берется по всем моментам t, когда наблюдалась хотя бы одна смерть.

Статистика критерия

$$z = \frac{U_L}{S_{U_L}}$$

распределена по стандартному нормальному закону распределения.

Неважно для какой из групп вычисляется U_L, для другой группы она равна по абсолютной величине, но имеет противоположный знак.

Поправка Йейтса

Когда дискретное распределение описывается нормальным распределением, которое по своей сути непрерывно, это приводит к излишней мягкости критерия: мы несколько чаще, чем следовало, отвергаем нулевую гипотезу. Чтобы компенсировать влияние дискретности, применяют поправку Йейтса. В случае логрангового критерия это делается таким образом:

$$z = \frac{|U_L| - \frac{1}{2}}{S_{U_L}}.$$

Теперь попробуем применить логранговый метод.

Пример.

После операции 10 пациентов случайным образом были разделены на две группы: в одной группе проводилась определенная терапия после операции, а в другой – такая терапия не проводилась. Эти пациенты наблюдались два года. Ниже приводится время ремиссии в месяцах.

Группа 1: 23, 16+, 18+, 20+, 24+.

Группа 2: 15, 18, 19, 19, 20.

Надо проверить, есть ли различия между кривыми выживаемости.

Нулевая гипотеза $S_1(t)=S_2(t)$.

Альтернативная гипотеза $S_1(t)\neq S_2(t)$.

	Группа1		Группа2		Объединенная				
Месяц	наблюдений	рецидив	наблюдений	рецидив	наблюдений	рецидив	Ожидаемое число рецидивов в группе1	Слагаемое для U_L	Слагаемое для S^2_{UL}
15	5	1	5	0	10	1	0,50	0,50	0,25
18	4	1	4	0	8	1	0,50	0,50	0,25
19	3	2	3	0	6	2	1,00	1,00	0,4
20	1	1	3	0	4	1	0,25	0,75	0,19
23	0	0	2	1	2	1	0,00	0,00	0

$$S_{U_L} = \sqrt{1{,}09} = 1{,}04\,,\ z = \frac{2{,}75}{1.04} = 2{,}64\ .$$

С поправкой Йейтса z = 2,16, Для α = 0,05, z = 1,96 .

Следовательно, нулевая гипотеза отклоняется и принимаем альтернативную гипотезу.

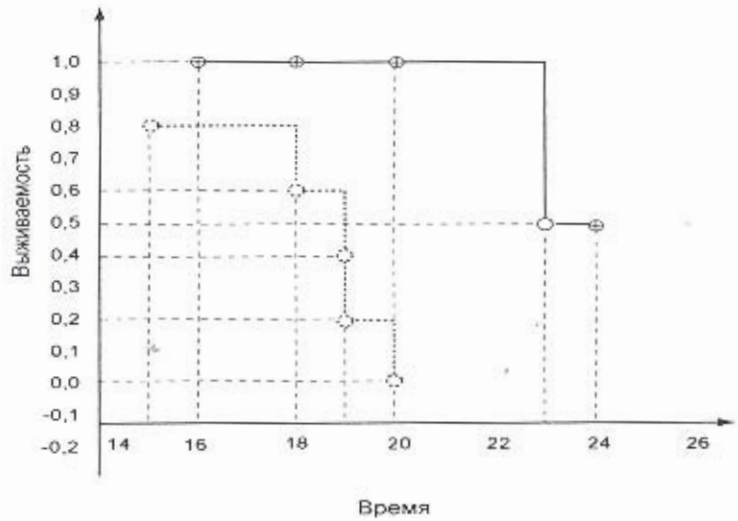


Рис. 53. Сравнение двух кривых выживаемости. Есть значимое различие

Критерий Гехана

Критерий Гехана вычисляют так: каждого пациента из первой группы сравнивают с каждым пациентом из второй группы. Результат сравнения оценивают как +1, если больной из первой группы наверняка прожил дольше; –1, если он наверняка прожил меньше; 0, если невозможно наверняка сказать, кто из них прожил дольше. Последнее возможно в трех случаях: если оба выбыли, если один выбыл до того, как второй умер, и если время наблюдения одинаково. Аналитически это можно выразить формулой:

$$U_{ij} = \begin{cases} +1 \cdot if \cdot x_i > y_j \cdot or \cdot x_i^+ \geq y_j \\ 0 \cdot if \cdot x_i = y_j \cdot or \cdot x_i^+ < y_j \cdot or \cdot y_j^+ < x_i \cdot or \cdot \left(x_i^+, y_j^+\right) \\ -1 \cdot if \cdot x_i < y_j \cdot or \cdot x_i \leq y_j^+ \end{cases}.$$

Результаты сравнения для каждого больного суммируют; эту сумму обозначим U_i.

$$U_{i.} = \sum_{j=1}^{n_2} U_{ij}.$$

В свою очередь сумма всех U_i дает величину W_U,

$$W_U = \sum_{i=1}^{n_2} U_{i.} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij},$$

стандартная ошибка которой определяется по формуле:

$$S_{W_U} = \sqrt{\frac{n_1 n_2 \sum_{i=1}^{n_1} U_{i.}^2}{(n_1 + n_2)(n_1 + n_2 - 1)}}.$$

И, наконец, вычисляют z:

$$z = \frac{W_U}{S_{W_U}},$$

которое распределено по стандартному нормальному закону.

Поправка Йейтса применяется к критерию Гехана точно так же, как и к логранговому критерию.

$$z = \frac{|W_U| - \frac{1}{2}}{S_{U_i}}.$$

Пример.

В эксперименте сравнивали два метода лечения (А и В) опухоли.

Ниже приводится время от начала лечения до смерти в месяцах.

Группа А: 8, 12, 12+, 15, 16, 17.

Группа В: 9, 11, 13, 16, 16+, 20.

Надо проверить, есть ли различия между двумя кривыми выживаемости.

Нулевая гипотеза $S_1=S_2$ (подходы эквивалентны).

Альтернативная гипотеза $S_1 \neq S_2$ (подходы неэквивалентны).

Применим критерий Гехана.

		Group1						
		8	**12**	**15**	**12+**	**16**	**17**	**ΣU_{ij}**
Group2:	**9**	–1	1	1	1	1	1	
	11	–1	1	1	1	1	1	
	13	–1	–1	1	0	1	1	
	16	–1	–1	–1	0	0	1	
	16+	–1	–1	–1	0	–1	0	
	20	–1	–1	–1	0	–1	–1	
	U_i	–6	–2	0	2	1	3	$\Sigma U_{ij} = -2$
	$U^2_{i.}$	36	4	0	4	1	9	$\Sigma U_i^2 = 54$

$$S_{U_W} = \sqrt{\frac{6 \times 6 \times 54}{(6+6) \times (6+6-1)}} = 3{,}83, \quad |z| = \left|\frac{-2}{3{,}83}\right| = 0{,}52 \ (p = 0{,}3).$$

С поправкой Йейтса $|z| = 0{,}39$ ($p = 0{,}34$), для $\alpha = 0{,}05$ $z = 1{,}96$.

Следовательно, нет оснований отвергать гипотезу.

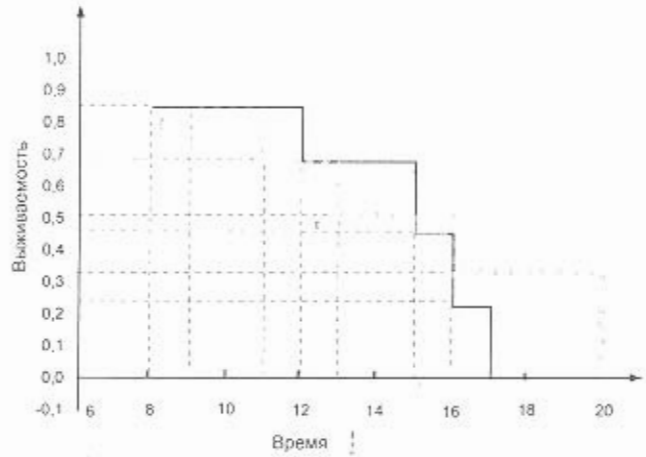


Рис. 54. Сравнение двух кривых выживаемости. Значимого различия нет

ПЛАНИРОВАНИЕ И ПРОВЕДЕНИЕ МЕДИКО-БИОЛОГИЧЕСКОГО ЭКСПЕРИМЕНТА

Медико-биологический эксперимент можно разделить на следующие этапы: планирование и проведение эксперимента, анализ полученных данных, в том числе, статистический анализ данных, описание результатов и выводы.

Под планированием подразумевается определение цели исследования, описание генеральной совокупности, определение выборки: ее объем, случайность отбора, представительность. Проведение включает в себя производство выборок в соответствии с планом, создание файла или таблицы с результатами измерений. В анализ входит проверка корректности собранных данных и статистический анализ. Последний предполагает визуализацию данных (построение гистограмм), проверку гипотез о виде выборочного распределения, оценку параметров распределения (точечное и интервальное оценивание), выявление связей и зависимостей. В выводах подробно описываются результаты эксперимента и статистического анализа.

Рассмотрим эти этапы более подробно.

В ходе планирования необходимо определить цели исследования, сформулировать гипотезы, определить уровни значимости для них, а также чувствительность (мощность) критерия, по которому гипотезы будут проверяться. Напомним, что мощность критерия – это единица минус вероятность ошибки II рода т.е. это вероятность принять альтернативную гипотезу, когда она верна.

На чувствительность критерия влияют:

Уровень значимости – чем меньше уровень значимости, тем ниже чувствительность. Однако будьте осторожны, увеличение уровня значимости может привести к увеличению риска отвергнуть верную нулевую гипотезу, т.е. найти различия там, где их нет.

Величина эффекта, например разность между оценками средних значений двух выборок. Чем больше величина эффекта, тем больше чувствительность критерия. Влиять на величину эффекта невозможно.

Объем выборки – чем больше объем выборки, тем больше чувствительность.

Нетрудно рассчитать чувствительность критерия задним числом, когда уже известна величина эффекта. К сожалению, мы не знаем этого параметра во время планирования исследования. Величину эффекта узнать невозможно (обычно ее оценка и является целью исследования), поэтому при расчете чувствительности нужно указать минимальную величину эффекта, которую мы хотим выявить. Чувствительность редко рассчитывают заранее, между тем делать это необходимо: иначе мы рискуем проводить исследования, заведомо обреченные на неуспех.

Если после проведения исследования эффект обнаружен, то чувствительность уже не важна. В противном случае – если эффекта не выявлено – она приобретает первостепенное значение. В самом деле, если мы не обнаружили статистически значимых различий при чувствительности 80%, то с высокой вероятностью можно утверждать, что различий действительно нет. Иными словами, мы получили отрицательный результат. Если же чувствительность составляла 25%, то мы просто не получили никакого результата.

Чаще всего исследования проводят в крупных клиниках, куда попадают далеко не все больные. Поэтому при всем желании больных в клиниках трудно назвать представительной выборкой. Это несоответствие обязательно нужно иметь в виду, решая, на какую совокупность больных могут быть (и в какой мере) распространены полученные в исследовании результаты.

Любой статистический метод исходит из предположения, что выборка извлечена из совокупности случайно (это обеспечивает представительность выборки). Если это условие не выполняется, т.е. если выборка непредставительна, то статистические методы не дают правильного результата. Если же выборка представительна, то надо учитывать какую совокупность она представляет.

«Извлечены случайно» означает, что вероятность оказаться выбранным одинакова для всех членов совокупности. Например, если групп две (экспериментальная и контрольная) и их размеры

равны, то любой член совокупности может равновероятно попасть в любую из групп.

Обеспечить равную вероятность попадания в любую из групп непросто. Прежде всего надо исключить всякое влияние человека. Предназначенные для этого методы называются *рандомизацией*. Задача рандомизации – обеспечить такой подбор больных, чтобы контрольная группа ни в чем не отличалась от экспериментальной, кроме метода лечения. На этапе оценки результатов необходимо исключить влияние исследователя и исследуемого. Для этого предназначен слепой метод. В идеале нужно проводить двойной слепой метод: ни больной, ни врач не знают, какой из способов лечения был применен. Двойной слепой метод не всегда осуществим, поэтому используют также простой слепой метод (информация есть у врача, но нет у пациента, или наоборот) и частично слепой (и врач, и больной располагают частью информации). В любом случае, информацию, которой располагают участники исследования, следует свести к минимуму. Надо учитывать тот факт, что если у кого-либо из участников исследования есть возможность влиять на построение групп, эта возможность будет использована.

Для рандомизации не достаточно, чтобы выбор не зависел от исследователя. Он должен быть независимым и от самих испытуемых. Приведем пример из области лабораторных исследований. Двадцать крыс, сидящих в клетке, нужно разделить на две группы. Выпустим из клетки 10 крыс и назовем их контрольной группой. Представительна ли она? Скорее всего, нет. Вероятно, из клетки первыми выбегут самые сильные и агрессивные особи.

Для того, чтобы получить случайную выборку, надо воспользоваться генератором или таблицей случайных чисел.

Чем лучше проведено исследование, тем менее вероятно, что его результат смещен в пользу исследуемого метода.

Чтобы правильно выбрать статистический метод обработки данных, необходимо учитывать характер интересующего нас признака (количественный, порядковый или качественный) и тип распределения (нормальное или нет). Выше были разобраны основные статистические методы внутри каждого типа, рассмотренны простейшие модели. С помощью таблицы можно выбрать критерий, которым следует воспользоваться.

	Исследование		
Признак	***Сравнение двух групп***	***Сравнение более двух групп***	***Связь признаков***
Количественный (распределение нормальное)	Критерий Стьюдента	Дисперсионный анализ	Линейная регрессия, коэффициент корреляции Пирсона
Порядковый (ранжированные данные)	Критерий Манна-Уитни	Критерий Крускала-Уоллиса	Коэффициент ранговой корреляции Спирмена
Качественный	Критерий ХИ-квадрат	Критерий ХИ-квадрат	Коэффициент сопряженности Юла.
Выживаемость	**Критерий гехана, Логранговый критерий**		

Выбор статистического критерия в случае числовых признаков требует пояснения. Если известно, что распределение признака в совокупности нормально, можно использовать параметрический метод. Если распределение далеко от нормального, следует воспользоваться непараметрическими аналогами этих методов.

Обнаружив, что нулевая гипотеза об отсутствии эффекта не может быть отвергнута, нужно выяснить почему так происходит. Для этого надо определите чувствительность критерия, если она мала, причиной может быть малый объем выборки. Но если чувствительность велика, то эффект действительно отсутствует. Попытайтесь понять, в самом ли деле процедура получения данных обеспечивает их представительность, в противном случае все последующие выкладки потеряют смысл. Проверте корректность применения методов обработки экспериментальных данных.

ЛИТЕРАТУРА

Агапов Г.И. Задачник по теории вероятностей: Учеб. пос. для вузов. –2-е изд.,доп.– М.:Высшая школа, 1994.–112 с.:ил.

Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ /Пер. с англ. – М.: Мир, 1982. - 488 с.,ил.

Белицкая Е.Я. Учебное пособие по медицинской статистике. – Ленинградское отделение: Медицина,1972. – 178 с.

Большев Л.Н., Смирнов Н. В. Таблицы математической статистики. – М.: Наука, 1965.

Боровиков В.П., Боровиков И.П. STATISTIKA R – Статистический анализ и обработка данных в среде Windows R. – М.: Информационно-издательский дом "Филинъ", 1997. – 608 с.

Варден Б.Л ван дер. Математическая статистика / Пер. с нем. – М.: Изд-во иностр. лит., 1960. – 436 с.

Венцель Е.С. Теория вероятностей. – М.: Наука, 1969.

Венцель Е.С., Овчаров Л.А. Прикладные задачи теории вероятностей. – М.: Радио и связь, 1983. – 416., ил.

Гланц С.. Медико-биологическая статистика /Пер. с англ. – М.: Практика, 1998. – 459 с.

Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пос. для вузов. Изд. 6-е, стер. – М.: Высшая школа, 1997. – 479 с.: ил.

Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. пос. для вузов. 4-е изд., стереотип.-М.: Высшая школа, 1998.– 400 с.,ил.

Громыко Г.Л. Общая теория статистики: Практикум. – М.: ИНФРА-М, 1999. – 139 с.

Данко П.Е., Попов А. Г., Кожевникова Т.Я. Высшая математика в упражнениях и задачах. В 2-х ч. Ч. II: Учеб. пособие для втузов. – изд 5-е, испр. – М.: Высшая школа, 1997. – 416 с.: ил.

Дерффель К. Статистика в аналитической химии /Пер. с нем. – М.: Мир, 1994. – 268 с.: ил.

Дунин-Барковский И.В., Смирнов Н.В. Теория вероятностей и математическая статистика в технике. – М.: Гостехиздат, 1955.

*Емельянов Г.*В., Скитович В.П. Задачник по теории вероятностей и математической статистике. – Ленинград.: Издательство ленинградского университета, 1967.

Закс Л. Статистическое оценивание /Пер. с нем. – М.: Статистика, 1976. – 598 с.: ил.

Калинина В.Н. Математическая статистика: Учеб. для техникумов. – изд. 2-е, стер. – М.: Высшая школа, 1998. – 336 с.: ил.

Колемеев В.А. и др Теория вероятностей и математическая статистика: Учеб. пособие для экон. спец. вузов. – М.: Высшая школа., 1991. – 400с.: ил.

Колмогоров А.Н. Основные понятия теории вероятностей. – М.: Наука, 1974.

Лакин Г.Ф. Биометрия: Учеб.пос. для биологич. спец. вузов. – 3-е изд., перераб. и доп. – М.: Высшая школа, 1980. – 293 с.: ил.

Морозов Ю.В. Основы высшей математики и статистики: Учебник. – М.: Медицина, 1998. – 232 с.: ил.

Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. – М.: ГЭОТАР МЕДИЦИНА, 2000. – 256 с.

Тюрин Ю.Н., Макарова А.А. Статистический анализ данных на компьютере /Под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528с.: ил.

Феллер Вильям. Введение в теорию вероятности и ее приложения Т.1. /Пер. с анг. – М.: Мир, 1967.

Хастингс Н., Пикок Дж. Справочник по статистическим распределениям /Пер. с англ. – М.: Статистика, 1980. – 95 с. ил.

Шмойлова Р.А. Практикум по теории статистики: Учеб.пособие – М.: Финансы и статистика, 2000. – 416 с.: ил.

Campbell R.C. Statistics for biologists. 3rd ed. – Cambridge University Press, 1989.

Dawson-Saunders Beth, Trapp Robert G. Basic & Clinical Biostatistics. – 2nd ed. –Appleton & Lange, 1994.

Lee Elisa T. Statistical methods for Survival Data Analysis 2nd ed. – John Wiley&Sons, Inc. 1992.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1

Таблица 1. Значения функции $\frac{\lambda^k e^{-\lambda}}{k!}$

λ	k								
	0	1	2	3	4	5	6	7	8
0,10	0,9048	0,0905	0,0045	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
0,50	0,6065	0,3033	0,0758	0,0126	0,0016	0,0002	0,0000	0,0000	0,0000
1,00	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0001	0,0000
1,50	0,2231	0,3347	0,2510	0,1255	0,0471	0,0141	0,0035	0,0008	0,0001
2,00	0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034	0,0009
2,50	0,0821	0,2052	0,2565	0,2138	0,1336	0,0668	0,0278	0,0099	0,0031
3,00	0,0498	0,1494	0,2240	0,2240	0,1680	0,1008	0,0504	0,0216	0,0081
3,50	0,0302	0,1057	0,1850	0,2158	0,1888	0,1322	0,0771	0,0385	0,0169
4,00	0,0183	0,0733	0,1465	0,1954	0,1954	0,1563	0,1042	0,0595	0,0298
4,50	0,0111	0,0500	0,1125	0,1687	0,1898	0,1708	0,1281	0,0824	0,0463
5,00	0,0067	0,0337	0,0842	0,1404	0,1755	0,1755	0,1462	0,1044	0,0653
5,50	0,0041	0,0225	0,0618	0,1133	0,1558	0,1714	0,1571	0,1234	0,0849
6,00	0,0025	0,0149	0,0446	0,0892	0,1339	0,1606	0,1606	0,1377	0,1033
6,50	0,0015	0,0098	0,0318	0,0688	0,1118	0,1454	0,1575	0,1462	0,1188
7,00	0,0009	0,0064	0,0223	0,0521	0,0912	0,1277	0,1490	0,1490	0,1304
7,50	0,0006	0,0041	0,0156	0,0389	0,0729	0,1094	0,1367	0,1465	0,1373
8,00	0,0003	0,0027	0,0107	0,0286	0,0573	0,0916	0,1221	0,1396	0,1396
8,50	0,0002	0,0017	0,0074	0,0208	0,0443	0,0752	0,1066	0,1294	0,1375
9,00	0,0001	0,0011	0,0050	0,0150	0,0337	0,0607	0,0911	0,1171	0,1318
9,50	0,0001	0,0007	0,0034	0,0107	0,0254	0,0483	0,0764	0,1037	0,1232
10,00	0,0000	0,0005	0,0023	0,0076	0,0189	0,0378	0,0631	0,0901	0,1126
10,50	0,0000	0,0003	0,0015	0,0053	0,0139	0,0293	0,0513	0,0769	0,1009
11,00	0,0000	0,0002	0,0010	0,0037	0,0102	0,0224	0,0411	0,0646	0,0888
11,50	0,0000	0,0001	0,0007	0,0026	0,0074	0,0170	0,0325	0,0535	0,0769
12,00	0,0000	0,0001	0,0004	0,0018	0,0053	0,0127	0,0255	0,0437	0,0655
13,00	0,0000	0,0000	0,0002	0,0008	0,0027	0,0070	0,0152	0,0281	0,0457
13,50	0,0000	0,0000	0,0001	0,0006	0,0019	0,0051	0,0115	0,0222	0,0375
14,00	0,0000	0,0000	0,0001	0,0004	0,0013	0,0037	0,0087	0,0174	0,0304
15,00	0,0000	0,0000	0,0000	0,0002	0,0006	0,0019	0,0048	0,0104	0,0194
16,00	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0026	0,0060	0,0120

Таблица 2. Значения функции $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$

X	0	1	2	3	4	5	6	7	8	9
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952

Таблица 3. Значения функции $\chi_k^2(x)$

x	k							
	1	2	3	4	5	10	15	20
0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
0,1	0,7518	0,9512	0,9918	0,9988	0,9998	1,0000	1,0000	1,0000
0,2	0,6547	0,9048	0,9776	0,9953	0,9991	1,0000	1,0000	1,0000
0,3	0,5839	0,8607	0,9600	0,9898	0,9976	1,0000	1,0000	1,0000
0,4	0,5271	0,8187	0,9402	0,9825	0,9953	1,0000	1,0000	1,0000
0,5	0,4795	0,7788	0,9189	0,9735	0,9921	1,0000	1,0000	1,0000
0,6	0,4386	0,7408	0,8964	0,9631	0,9880	1,0000	1,0000	1,0000
0,7	0,4028	0,7047	0,8732	0,9513	0,9830	1,0000	1,0000	1,0000
0,8	0,3711	0,6703	0,8495	0,9384	0,9770	0,9999	1,0000	1,0000
0,9	0,3428	0,6376	0,8254	0,9246	0,9702	0,9999	1,0000	1,0000
1	0,3173	0,6065	0,8013	0,9098	0,9626	0,9998	1,0000	1,0000
1,4	0,2367	0,4966	0,7055	0,8442	0,9243	0,9992	1,0000	1,0000
1,8	0,1797	0,4066	0,6149	0,7725	0,8761	0,9977	1,0000	1,0000
2	0,1573	0,3679	0,5724	0,7358	0,8491	0,9963	1,0000	1,0000
2,5	0,1138	0,2865	0,4753	0,6446	0,7765	0,9909	0,9999	1,0000
3	0,0833	0,2231	0,3916	0,5578	0,7000	0,9814	0,9996	1,0000
3,5	0,0614	0,1738	0,3208	0,4779	0,6234	0,9671	0,9990	1,0000
4	0,0455	0,1353	0,2615	0,4060	0,5494	0,9473	0,9977	1,0000
4,2	0,0404	0,1225	0,2407	0,3796	0,5210	0,9379	0,9970	0,9999
4,4	0,0359	0,1108	0,2214	0,3546	0,4934	0,9275	0,9961	0,9999
4,6	0,0320	0,1003	0,2035	0,3309	0,4666	0,9162	0,9950	0,9999
4,8	0,0285	0,0907	0,1870	0,3084	0,4408	0,9041	0,9937	0,9998
5	0,0253	0,0821	0,1718	0,2873	0,4159	0,8912	0,9921	0,9997
6	0,0143	0,0498	0,1116	0,1991	0,3062	0,8153	0,9797	0,9989
7	0,0082	0,0302	0,0719	0,1359	0,2206	0,7254	0,9576	0,9967
8	0,0047	0,0183	0,0460	0,0916	0,1562	0,6288	0,9238	0,9919
9	0,0027	0,0111	0,0293	0,0611	0,1091	0,5321	0,8775	0,9829
10	0,0016	0,0067	0,0186	0,0404	0,0752	0,4405	0,8197	0,9682

Таблица 4. Значения функции $F_{k,l}$

Процентные точки (5%)

l	k							
	1	2	3	4	5	10	15	20
1	161,4462	199,4995	215,7067	224,5833	230,1604	241,8819	245,9492	248,0156
2	18,5128	19,0000	19,1642	19,2467	19,2963	19,3959	19,4291	19,4457
3	10,1280	9,5521	9,2766	9,1172	9,0134	8,7855	8,7028	8,6602
4	7,7086	6,9443	6,5914	6,3882	6,2561	5,9644	5,8578	5,8025
5	6,6079	5,7861	5,4094	5,1922	5,0503	4,7351	4,6188	4,5581
6	5,9874	5,1432	4,7571	4,5337	4,3874	4,0600	3,9381	3,8742
7	5,5915	4,7374	4,3468	4,1203	3,9715	3,6365	3,5107	3,4445
8	5,3176	4,4590	4,0662	3,8379	3,6875	3,3472	3,2184	3,1503
9	5,1174	4,2565	3,8625	3,6331	3,4817	3,1373	3,0061	2,9365
10	4,9646	4,1028	3,7083	3,4780	3,3258	2,9782	2,8450	2,7740
11	4,8443	3,9823	3,5874	3,3567	3,2039	2,8536	2,7186	2,6464
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,7534	2,6169	2,5436
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,6710	2,5331	2,4589
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,6022	2,4630	2,3879
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,5437	2,4034	2,3275
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,4935	2,3522	2,2756
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,4499	2,3077	2,2304
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,4117	2,2686	2,1906
19	4,3808	3,5219	3,1274	2,8951	2,7401	2,3779	2,2341	2,1555
20	4,3513	3,4928	3,0984	2,8661	2,7109	2,3479	2,2033	2,1242
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,3210	2,1757	2,0960
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,2967	2,1508	2,0707
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,2747	2,1282	2,0476
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,2547	2,1077	2,0267
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,2365	2,0889	2,0075
26	4,2252	3,3690	2,9752	2,7426	2,5868	2,2197	2,0716	1,9898
27	4,2100	3,3541	2,9603	2,7278	2,5719	2,2043	2,0558	1,9736
28	4,1960	3,3404	2,9467	2,7141	2,5581	2,1900	2,0411	1,9586
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,1768	2,0275	1,9446
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,1646	2,0148	1,9317
40	4,0847	3,2317	2,8387	2,6060	2,4495	2,0773	1,9245	1,8389

Таблица 5. Значения функции $t_k(x)$

x	k							
	1	2	3	4	5	10	15	20
0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
0,1	0,5317	0,5353	0,5367	0,5374	0,5379	0,5388	0,5392	0,5393
0,2	0,5628	0,5700	0,5729	0,5744	0,5753	0,5773	0,5779	0,5782
0,3	0,5928	0,6038	0,6081	0,6104	0,6119	0,6148	0,6159	0,6164
0,4	0,6211	0,6361	0,6420	0,6452	0,6472	0,6512	0,6526	0,6533
0,5	0,6476	0,6667	0,6743	0,6783	0,6809	0,6861	0,6878	0,6887
0,6	0,6720	0,6953	0,7046	0,7096	0,7127	0,7191	0,7213	0,7224
0,7	0,6944	0,7218	0,7328	0,7387	0,7424	0,7501	0,7527	0,7540
0,8	0,7148	0,7462	0,7589	0,7657	0,7700	0,7788	0,7819	0,7834
0,9	0,7333	0,7684	0,7828	0,7905	0,7953	0,8054	0,8088	0,8106
1	0,7500	0,7887	0,8045	0,8130	0,8184	0,8296	0,8334	0,8354
1,4	0,8026	0,8518	0,8720	0,8829	0,8898	0,9041	0,9091	0,9116
1,8	0,8386	0,8932	0,9152	0,9269	0,9341	0,9490	0,9540	0,9565
2	0,8524	0,9082	0,9303	0,9419	0,9490	0,9633	0,9680	0,9704
2,5	0,8789	0,9352	0,9561	0,9666	0,9728	0,9843	0,9877	0,9894
3	0,8976	0,9523	0,9712	0,9800	0,9850	0,9933	0,9955	0,9965
3,5	0,9114	0,9636	0,9803	0,9876	0,9914	0,9971	0,9984	0,9989
4	0,9220	0,9714	0,9860	0,9919	0,9948	0,9987	0,9994	0,9996
4,2	0,9256	0,9739	0,9877	0,9932	0,9958	0,9991	0,9996	0,9998
4,4	0,9289	0,9760	0,9891	0,9942	0,9965	0,9993	0,9997	0,9999
4,6	0,9319	0,9779	0,9903	0,9950	0,9971	0,9995	0,9998	0,9999
4,8	0,9346	0,9796	0,9914	0,9957	0,9976	0,9996	0,9999	0,9999
5	0,9372	0,9811	0,9923	0,9963	0,9979	0,9997	0,9999	1,0000
5,5	0,9428	0,9842	0,9941	0,9973	0,9986	0,9999	1,0000	
6	0,9474	0,9867	0,9954	0,9981	0,9991	0,9999	1,0000	
6,5	0,9514	0,9886	0,9963	0,9986	0,9994	1,0000		
7	0,9548	0,9901	0,9970	0,9989	0,9995	1,0000		
7,5	0,9578	0,9913	0,9975	0,9992	0,9997	1,0000		
8	0,9604	0,9924	0,9980	0,9993	0,9998	1,0000		
9	0,9648	0,9939	0,9986	0,9996	0,9999	1,0000		
10	0,9683	0,9951	0,9989	0,9997	0,9999	1,0000		

Таблица 6. Равномерно распределенные случайные числа

10 09 73 25 33	76 52 01 35 86	34 67 35 48 76
37 54 20 48 05	64 89 47 42 96	24 80 52 40 17
08 42 26 89 53	19 64 50 93 03	23 20 90 25 60
99 01 90 25 29	09 37 67 07 15	38 31 13 11 65
12 80 79 99 70	80 15 73 61 47	64 03 23 66 53
66 06 57 47 17	34 07 27 68 50	36 69 73 61 70
31 06 01 08 05	45 57 18 24 06	35 30 34 26 14
85 26 97 76 02	02 05 16 56 92	68 66 57 48 18
63 57 33 21 35	05 32 54 70 48	90 55 35 75 48
73 79 64 57 53	03 52 96 47 78	35 80 83 42 82
98 52 01 77 67	14 90 56 86 07	22 10 94 05 58
11 80 50 54 31	39 80 82 77 32	50 72 56 82 48
81 45 29 96 34	06 28 89 80 83	13 74 67 00 78
88 68 54 02 00	86 50 75 84 01	36 76 66 79 51
99 59 46 73 48	87 51 76 49 69	91 82 60 89 28
65 48 11 76 74	17 46 85 09 50	58 04 77 69 74
80 12 43 56 35	17 72 70 80 15	45 31 82 23 74
74 35 09 98 17	77 40 27 72 14	43 23 60 02 10
69 91 62 68 03	66 25 22 91 48	36 93 68 72 03
09 89 32 05 05	14 22 56 85 14	46 42 75 67 88
91 49 91 45 23	68 47 92 76 86	46 16 28 35 54
80 33 69 45 98	26 94 03 68 58	70 29 73 41 35
44 10 48 19 49	85 15 74 79 54	32 97 92 65 75
12 55 07 37 42	11 10 00 20 40	12 86 07 46 97
63 60 64 93 29	16 50 53 44 84	40 21 95 25 63
61 19 69 04 46	26 45 74 77 74	51 92 43 37 29
15 47 44 52 66	95 27 07 99 53	59 36 78 38 48
94 55 72 85 73	67 89 75 43 87	54 62 24 44 31
42 48 11 62 13	97 34 40 87 21	16 86 84 87 67
23 52 37 83 17	73 20 88 98 37	68 93 59 14 16

ПРИЛОЖЕНИЕ 2

Формулы комбинаторики

Перестановки. Перестановками называют комбинации, состоящие из одних и тех же элементов и отличающиеся только порядком их расположения. Число всех возможных перестановок.

$$P_k = k!$$

Задачи

235. Сколькими способами могут восемь человек стать в очередь к театральной кассе?

236. Сколько пятизначных чисел можно составить из цифр 1,2,4,6,7,8, если никакую цифру не использовать более одного раза? Сколько среди этих чисел будет четных? Сколько нечетных?

237. Пять мальчиков и пять девочек рассаживаются в ряд на десять подряд расположенных мест, причем мальчики садятся на нечетные места, а девочки на четные. Сколькими способами они могут это сделать?

238. Сколько упорядоченных пар символов (X,Y) можно образовать, если на место X можно подставлять либо a, либо b, либо c, а на место Y либо 1, либо 2, либо 3, либо 4? Нарисуйте дерево, на котором показано множество всех возможных пар (X,Y).

239. Выходя из вагона, некто обнаружил у себя в кармане никель, дайм, квотер и полдоллара (никель, дайм, квотер и полдоллара – монеты США достоинством в 5,10,25 и 50 центов). Сколькими способами он может дать на чай носильщику?

240. Восемь лабораторных животных нужно упорядочить в соответствии с их способностями выполнять определенные задания. Каково число возможных упорядочений, если допустить, что одинаковых способностей нет?

241. Для лечения заболевания применяют пять лекарств. Полагают, что последовательность, в которой применяют лекарства, оказывает существенное влияние на результаты лечения. Сколько имеется различных порядков назначения этих лекарств?

242. Сколькими способами каждой из 20 аминокислот можно поставить в соответствие однозначным образом некоторую из 20 троек нуклеотидов?

Сочетания. Сочетаниями называют комбинации, составленные из n различных элементов по k, которые отличаются хотя бы одним элементом. Число всех возможных сочетаний:

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

Свойства:

$$C_n^k = C_n^{n-k}.$$

$$C_{n+1}^{k+1} = C_n^{k+1} + C_n^k, k < n.$$

$$C_n^0 = C_n^n = 1.$$

Задачи

243. Найдите n из следующих уравнений: $C_n^2 = 45$, $C_n^8 = C_n^{12}$.

244. Подрядчику нужны 4 плотника, а к нему с предложением своих услуг обратились 10. Сколькими способами он может выбрать среди них четырех?

245. Сколько пятибуквенных слов, каждое из которых состоит из 3-х согласных и 2-х гласных, можно образовать из букв слова УРАВНЕНИЕ?

246. Колода карт содержит 36 различных листов. Сдача карт одному игроку состоит из 6-ти карт, порядок раздачи которых не важен. Каково число всех возможных сдач?

247. Имеются 10 белых и 5 черных шаров. Сколькими способами можно выбрать 7 шаров, чтобы среди них были три черных?

248. Сколькими способами можно группу из 12 человек разбить на две подгруппы, в одной из которых должно быть не более 5, а во второй не более 9 человек?

249. Необходимо выбрать в подарок 4 из 10 имеющихся различных книг. Сколькими способами можно это сделать?

Размещения. Размещениями называют комбинации, составленные из n различных элементов по k, которые отличаются либо составом элементов, либо их порядком. Число всех возможных размещений:

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

Задачи

250. Сколько можно составить семизначных телефонных номеров?

251. Сколько различных билетов с указанием станций отправления и станций назначения можно отпечатать для ж/д, на которой 50 станций?

Связь комбинаторных формул

$$A_n^k = P_k C_n^k.$$

Производные

- $C' = 0$.
- $x' = 1$.
- $(x^m)' = mx^{m-1}$.
- $(\ln x)' = \frac{1}{x}$.
- $(a^x)' = a^x \ln a$ (a>0).
- $(e^x)' = e^x$.
- $(\cos x)' = -\sin x$.
- $(\sin x)' = \cos x$.
- $(\mathrm{tg}x)' = \frac{1}{\cos^2 x}$.
- $(\mathrm{ctg}x)' = \frac{-1}{\sin^2 x}$.
- $(\mathrm{arctg}x)' = \frac{1}{1+x^2}$.
- $(\mathrm{arcctg}x)' = \frac{-1}{1+x^2}$.

Производная сложной функции

- $\{f[g(x)]\}' = f'(g(x))g'(x)$.

Задачи

Применяя формулы и правила дифференцирования, найти производные следующих функций:

252. $y = 2x^3 - 5x^2 + 7x + 4$.

253. $y = x^2 e^x$.

254. $y = x^4 \operatorname{arctg} x$.

255. $y = \dfrac{\arcsin x}{x}$.

256. $y = \cos^2 x$.

257. $y = \sin(2x+3)$.

258. $y = \dfrac{7}{x^3}$.

259. $y = 3x^3 \ln x - x^3$.

260. $y = x^2 \sin x + 2x\cos x - 2\sin x$.

261. $y = e^{-x} - \sin e^{-x} \cos e^{-x}$.

Интегралы

- $\int cf(x)dx = c\int f(x)dx$.
- $\int x^m dx = \dfrac{x^{m+1}}{m+1} + c$ (m≠–1).
- $\int \dfrac{1}{x} dx = \ln|x| + c$.
- $\int a^x dx = \dfrac{a^x}{\ln a} + c$ (a>0).
- $\int e^x dx = e^x + c$.
- $\int \sin x dx = -\cos x + c$.
- $\int \cos x dx = \sin x + c$.

- $\int \operatorname{tg} x dx = -\ln|\cos x| + c$.
- $\int \operatorname{ctg} x dx = \ln|\sin x| + c$.
- $\int \dfrac{1}{\cos^2 x} dx = \operatorname{tg} x + c$.
- $\int \dfrac{1}{\sin^2 x} dx = -\operatorname{ctg} x + c$.
- $\int \dfrac{1}{1+x^2} dx = \operatorname{arctg} x + c$.

Задачи

Применяя формулы и правила интегрирования, найти:

262. $\int (2x^3 - 5x^2 + 7x - 3)dx$.

263. $\int \left(\sqrt{x} + \dfrac{1}{\sqrt[3]{x}}\right)^2 dx$.

264. $\int 2^x 3^{2x} 5^{3x} dx$.

265. $\int e^{3\cos x} \sin x dx$ ***.

266. $\int (2\sin x + 3\cos x)dx$.

267. $\int x\sqrt{x} dx$.

268. $\int e^{3x} 3^x dx$.

269. $\int x\cos(x^2) dx$.

270. $\int\limits_{\pi/6}^{\pi/4} \dfrac{dx}{\cos^2 x}$.

271. $\int\limits_0^1 xe^{-x} dx$.

ПРИЛОЖЕНИЕ 3

Греческий алфавит

Прописные	Строчные	Название
Α	α	альфа
Β	β	бета
Γ	γ	гамма
Δ	δ	дельта
Ε	ε	э псилон
Ζ	ζ	дзета
Η	η	эта
Θ	θ	тета
Ι	ι	йота
Κ	κ	каппа
Λ	λ	ламбда
Μ	μ	мю
Ν	ν	ню
Ξ	ξ	кси
Ο	ο	о микрон
Π	π	пи
Ρ	ρ	ро
Σ	σ	сигма
Τ	τ	тау
Υ	υ	и псилон
Φ	φ	фи
Χ	χ	хи
Ψ	ψ	пси
Ω	ω	омега

Латинский алфавит

Прописные	Строчные	Название
A	a	а
B	b	бе
C	c	це
D	d	де
E	e	э
F	f	эф
G	g	ге
H	h	га
I	i	и
J	j	йот
K	k	ка
L	l	эль
M	m	эм
N	n	эн
O	o	о
P	p	пе
Q	q	ку
R	r	эр
S	s	эс
T	t	те
U	u	у
V	v	ве
X	x	икс
Y	y	ипсилон
Z	z	зета

ОТВЕТЫ

Случайные события

1. а) Ω = {ГГГ, ГГР, ГРГ, ГРР, РГГ, РГР, РРГ, РРР}; б) {ГГГ, ГГР, ГРГ, РГГ}. **2.** A∪B=A; AB=B. **3.** $\overline{A}$ – попадание в область, лежащую вне круга A; $\overline{B}$ – попадание в область, лежащую вне круга B; $\overline{A \cup B}$ – попадание в круг A или в круг B; $A \cup B$ – попадание в область вне обоих кругов A и B; AB попадание в общую часть кругов; $\overline{AB}$ попадение в область, лежащую вне общей части кругов A и B. **4.** а) $A \subset BC$, т.е. BC происходит всякий раз, как происходит A; б) $B \subset A$ и $C \subset A$, т.е. каждый раз, как только происходит B или C, происходит также и A. **5.**а) $A\overline{B}\overline{C}$; б) $AB\overline{C}$; в) ABC; г) $A \cup B \cup C$; д) $AB \cup AC \cup BC$; е) $A\overline{B}\overline{C} \cup \overline{A}B\overline{C} \cup \overline{A}\overline{B}C$; ж) $\overline{A}BC \cup A\overline{B}C \cup AB\overline{C}$ з) $\overline{A}\,\overline{B}\,\overline{C}$; и) $\overline{ABC}$. **6.** а) да; б) нет; в)да. **7.** а) да; б) нет; в)нет. **8.** 1) $A+C=E$; 2) $AC=K$; 3) $EF=G$; 4) $G+E=E$; 5) $GE=G$; 6) $BD=H$; 7) $E+K=E$.

Вероятность случайного события.

9. 5/36. **10.** а) 1/90; б) 1/81. **12.** а) 1/6; б) 1/18; в) 1/2; г) 1/18. **13.** 3/4. **14.** 1/720. **15.** 1/2. **16.** а) 3/10; б) 7/10. **17.** 2/6; 1/2. **18.** 3/8. **19.** 1/6; 1/3. **20.** 5/18; 11/36. **21.** 7/9. **22.** 1/60. **23.** 7/15. **24.** 1/720. **25.** 1/20. **26.** $1-\frac{C_{n-m}^{k}}{C_{n}^{k}}$. **27.** 0,0938. **28.** 29 $\frac{12!}{12^{12}}$. **29.** $\frac{C_S^K C_{N-S}^{L-K}}{C_N^L}$. **30.** 14/55. **31.** 0,1. **32.** 0,6. **33.** 24/91. **34.** 0,65. **35.** 1) да, 2) нет 3) да 4) нет. **36.** a/(a+b). **37.** (a-1)/(a+b-1). **38.** (a-1)/(a+b-1). **39.** a/(a+b). **40.** a/(a+b). **41.** a(a-1)/[(a+b)(a+b-1)]. **44.** 1/n!. **45.** $1/n^n$. **46.** 0,01765; 0,000969; 0,00001845; $0{,}715\times10^{-7}$. **47.** 5/9. **48.** 1/2. **49.** 2/(N-1). **50.** а) 1; б) 1/5; в) 3/5. **52.** 0,25. **53.** 81/1001. **54.** $\frac{1}{6^n}$; $\frac{n}{6^n}$. **55.** 5/12. **56.** 1/4.

57. Событие A представить как сумму двух непересекающихся событий AB и $A\overline{B}$. **59.** Независимы. **60.** Зависимы. **63.** 2/3. **64.** 48/95. **65.** 7/9. **66.** 0,857375. **67.** 1/24; 0; 1/4; 1/3; 3/8. **68.** 67/91. **69.** 5/6. **71.** $p_1(1-p_2)+(1-p_1)p_2$. **72.** $p_1+p_2-2p_1p_2$. **73.** 0,14. **74.** 0,38. **75.** 0,2. **76.** 1/495. **77.** 0,7. **78.** 0,18. **79.** 0,384. **80.** 57/115. **81.** 0,9. **82.** $c-b$. **83.** 1 – c. **85.** 13/30. **86.** 2/3. **87.** $\frac{n+2}{2(n+1)}$. **88.** 0,85. **89.** 0,52. **90.** Безразлично. **91.** 0,87. **92.** $\frac{a+b+c/3}{a+b+c}$ **95.** 20/21. **96.** 6/7. **97.** 3/7. **99.** 0,1. **100.** 4/29. **101.** 5/11. **102.** 0,3. **104.** 0,279. **105.** 0,22. **107.** $\frac{2}{3}p_1+\frac{1}{3}p_2$. **111.** отлично 0,58; плохо 0,002.

Повторные испытания

112. 0,2787. **113.** 2 из 4. **114.** 3 из 4. **115.** 3/16. **116.** 0,3723. **117.** а) 0,77; б) 0,02. **118.** 0,73728. **119.** а) 0,31; б) 0,48. **120.** 5. **121.** 24 или 25. **122.** 0,9596. **123.** 0,0916. **124.** 0,2385. **125.** 0,0375. **126.** а) 0,0613; б) 0,9197; с) 0,019. **127.** а) 0,224; б) 0,1992; в) 0,5768; г) 0,95. **128.** а) 0,0532; б) 0,0219. **129.** 0,051. **130.** 0,0231. **131.** а) 0,8882; б) 0,8944; в) 0,1056. **132.** 0,9737. **133.** 0,05. **134.** $\{3{,}4 \le \mu \le 23{,}2\} = \{4 \le \mu \le 23\}$. **135.** 0,0041. **136.** 122 0,0782. **137.** 177. **138.** а) 1 из 2; б) не менее 2-х из 4-х. **139.** а) 0,1792; б) 0,74. **140.** а) 0,52; б) 0,62. **141.** 0,1563. **142.** 0,965. **143.** а) 0,4236; б) 0,5; в) 0,5. **146.** 0,95945. **147.** 0,6826. **148.** 100. **149.** 0,1813. **150.** 0,6826. **151.** 80/243. **152.** а) 0,774; б) 0,0021. **153.** $n \ge 22$. **154.** $n \ge 300$. **155.** 10. **156.** 5. **157.** 0,22. **158.** 0,11. **159.** 0,09. **160.** а) 0,195; б) 0,969. **161.** а) 0; б) 0,99534; 0,5; 0,00466.

Случайные величины.

168. в) 0,8. **172.** б) 0,25. **173.** а) A=0,5, B=$\frac{1}{\pi}$; б) $\frac{1}{\pi(1+x^2)}$; в) 0,5. **175.** $\frac{\sqrt{2}}{4}$. **177.** 1. **178.** б) 0,5858. **180.** 1/2. **181.** а) 2/9; в) 13/27. **182.** 0,5; 0,25. **183.** 0,25; 0,75.

Числовые характеристики случайной величины

184. 6. **185.** а) 11; б) 30. **186.** $x_3 = 21$, $p_3 = 0{,}2$. **187.** 3/5. **188.** λ. **189.** 69. **190.** 15,21; 3,9. **191.** а) 8,545, 2,923; б) 248,95, 15,78. **192.** 2. **193.** 0,8. **194.** 0,48. **195.** λ. **196.** 2/3. **197.** 0. **198.** а) 3/4; б) 11/16. **199.** 2. **200.** $\frac{c^2}{2}$. **201.** 0. **202.** 9; 40. **203.** 61. **204.** 7/2; 35/12. **205.** 0,495. **206.** 4/3. **207.** а) 4,5. **208.** $\frac{(\pi^2-8)}{4}$. **209.** 2. **210.** p_1=0,4, p_2=0,1, p_3=0,5. **211.** p_1=0,2, p_2=0,3, p_3=0,5. **218.** 1/2. **219.** $D[X]=\frac{(b-a)^2}{12}$, $\sigma[X]=\frac{(b-a)}{2\sqrt{3}}$. **220.** a, $\frac{l^2}{3}$. **224.** 1; 25. **225.** 5. **226.** 3; $\sqrt{3}$

СОДЕРЖАНИЕ

Елена Анатольевна Лукьянова

МЕДИЦИНСКАЯ СТАТИСТИКА

Учебное пособие

Редактор *Ж.В. Медведева*
Технический редактор *Ю.В. Чванова*
Корректор *О. Бельтран-Легас*
Компьютерная верстка *Е.А. Лукьянова*
Дизайн обложки *И.Ю. Проценко*

Тематический план 2001 г., № 9