ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА С ИСПОЛЬЗОВАНИЕМ КОМПЬЮТЕРА

Ю.М.Барабашева, Г.Н. Девяткова, В.Н. Тутубалин, Е.Г. Угер

Предисловие

Данная брошюра ставит перед собой скромную задачу - представить справочное пособие, необходимое для преподавателей и студентов в связи с изучением части курса "Математические методы в биологии" на Биологическом факультете МГУ. Эта часть курса включает вероятностно-

статистические методы с ориентацией на обработку тех наблюдений, которые могут встретиться в биологии. Пособие состоит из трех частей.
Первая часть - "Задачи" - представляет собой список задач, взятых из различных источников и собранных вместе для удобства (С.Гроссман, Дж.Тернер «Математика для биологов», М.: «Высшая школа», 1983; Мешалкин Л.Д. «Сборник задач по теории вероятностей», М.: Из-во МГУ, 1963; Д.И.Золотаревская «Теория вероятностей. Задачи с решениями», М.: УРСС, 2003; Х.М.Андрухаев «Сборник задач по теории вероятностей», М.: «Просвещение», 1985; Н.Я.Сотникова «Первоапрельский задачник по теории вероятностей для студентов нематематиков» (http://www.astro.spbu.ru/staff/nsot/Teaching/tver/zadachi.html); и др.). Эта часть предназначена для освоения того языка, на котором разговаривают теория вероятностей и математическая статистика: случайное событие, вероятность, условная вероятность, случайная величина, плотность распределения вероятности и т.д. Некоторые задачи повторяются в очень близкой формулировке, что должно облегчить преподавателю составление вариантов контрольных работ. Для решения ряда задач предлагается пользоваться компьютерными пакетами вместо того, чтобы обращаться к громоздким справочникам.

Применение компьютера является принципиально важным для образования еще и в другом отношении. Приемы математической статистики (в простейшем случае, например, доверительные интервалы или уровни значимости при проверке гипотез) имеют гарантированную в некотором смысле надежность. Но гарантии эти формулируются в вероятностных терминах. Например: доверительный интервал ловит неизвестное значение определяемого параметра не хуже, чем с *вероятностью* 95%; при проверке гипотезы *вероятность* ошибки первого рода не более 5%. К чему относятся эти *вероятности*? А они относятся не к единственной серии наблюдений, с помощью которой вычисляется доверительный интервал или проверяется гипотеза, а к *ансамблю* многих серий таких наблюдений, который является лишь мыслимым, но реально не существует. Важно, что использование компьютера позволяет имитировать и предъявить студенту этот ансамбль, не существующий в практике обычных научных исследований. Студенту нужно понимать некоторую ограниченность и условность вероятностного подхода. Подобные компьютерные упражнения (и некоторые вводные соображения к ним) сведены во второй раздел, который так и называется "Упражнения". Исходные данные, необходимые для выполнения упражнений, находятся по адресу: http://www.math.msu.su/department/probab//index-k.html или http://ecology.genebee.msu.ru/3_SOTR/CV_Barabasheva.htm.

Наконец, третья часть данного пособия именно и представляет собой справочник по основным понятиям, формулам и фактам теории вероятностей

и математической статистики. Частью содержание этого раздела представляет собой резюме того языка, который необходимо знать при практическом применении вероятностных методов. Но знание наиболее сложных формул, как правило, не требуется для приложений, если есть возможность воспользоваться компьютером.

Таким образом, предлагается для студентов нематематических специальностей, в частности для биологов, сократить объем материала, содержащего более сложные математические формулы и теоремы, заменив их некоторыми компьютерными упражнениями.

Раздел I. Задачи

Условия представленных задач частично отредактированы. Например, не следует говорить "вес коровы есть нормально распределенная случайная величина" (Золотаревская, задача 131), потому что в любом стаде коров лишь конечное число, так что при случайном выборе коровы из стада функция распределения веса есть ступенчатая функция. Но можно сказать "распределение по весу коров в некотором стаде приблизительно описывается нормальным законом" (задача 99). Кстати, если явно выписать плотность нормального распределения (например, задача 101; Золотаревская: задача 128), то мы столкнемся вот с какой проблемой.

В математике все рассматриваемые величины, по умолчанию, безразмерны. Но в любых приложениях так или иначе возникают единицы измерения. Например, в математике вполне возможна функция $\sin t$, где t - время, но в физике бывает только $\sin \omega t$, где ω - частота (синус от размерной величины невозможен). Если вес коровы выражать в килограммах, то его математическое ожидание будет тоже в кг, но дисперсия - в $кг^2$. А нормальная плотность для веса будет иметь размерность $кг^{-1}$ (чтобы интеграл от нее по какому-то отрезку значений веса мог дать безразмерную величину - вероятность попадания в этот отрезок веса наудачу взятой коровы). Или можно перейти к безразмерным величинам, разделив все возможные веса на единицу, т.е. 1кг, но по умолчанию это не так и это нужно явно оговорить. Вот и проводилась редакция условий задач, в частности, такого рода.

1. В рассматриваемой популяции людей M обозначает множество всех мужчин, T – множество людей, больных туберкулезом. Описать в содержательных терминах состав множеств: $\overline{M}; \overline{T}; MT; T \setminus M$.

2. Каждый из 4-х потоков студентов выбирает по 1 представителю в комитет. Сколькими способами можно выбрать состав представителей, если потоки насчитывают 47, 51, 54, 55 студентов?

3. В трех пробирках, поставленных в штатив для пробирок (с 3-мя отверстиями) содержатся разные препараты: C_1, C_2, C_3. Посчитать число всех возможных расположений этих препаратов в штативе по порядку слева направо.

4. В распоряжении агрохимика есть шесть различных минеральных удобрений. Ему необходимо провести эксперименты по изучению совместного влияния каждой возможной тройки из этих удобрений. Сколько всего экспериментов ему придется провести, если:
   a) порядок внесения удобрений несущественен;
   b) порядок внесения удобрений существенен?

5. Из группы в 9 крыс нужно выбрать 3 и посадить их в клетки C_1, C_2, C_3 . Сколькими способами это можно сделать?

6. Лабораторная крыса помешена в лабиринт и должна избрать один из пяти возможных путей. Лишь один из них ведет к поощрению в виде пищи. В предположении, что крыса с одинаковой вероятностью избирает любой путь, какова вероятность выбора пути, ведущего к пище?

7. Сколькими способами можно разместить двенадцать мышей, занумерованных от 1 до 12, в четырех клетках *A*, *B*, *C*, *D* по три мыши в каждой? Какова вероятность того, что в первую клетку попадут мыши с №№ 1,2,3? (Какие предположения Вы при этом делаете?)

8. Некоторая популяция растений состоит из особей трех генотипов, помеченных *AA, Aa* и *aa.* Численности каждого типа составляют соответственно 200, 600 и 50. Допустим, что из этой популяции случайно выбирают одно растение. Какова вероятность того, что:
   a) это растение типа *AA;*
   b) это растение типа *AA* или *Aa?*

9. В большой популяции плодовой мушки 25% мух имеют мутацию глаз, 50% мух имеют мутацию крыльев, а 40%, имеющих мутацию глаз имеют мутацию крыльев. Какова вероятность того, что у мухи, наудачу выбранной из этой популяции, есть:
   a) хотя бы одна мутация;
   b) есть мутация глаз, но нет мутации крыльев?

10. Бросаются 2 кости. Пусть событие *A* – сумма очков нечетная; *B* – хотя бы на одной кости выпала 1. Описать в содержательных терминах события *AB, AUB,* $AU\overline{B}$. Найти их вероятности. (Какие предположения Вы при этом делаете?)

11. Профессор выставляет 20 различных оценок за контрольные работы 20 студентам (100-балльная система) и заносит их в компьютер. В результате компьютерного сбоя все оценки случайно переставились. Какова вероятность того, что:
    a) каждый студент получит свою прежнюю оценку;
    b) ровно 19 студентов получат свои прежние оценки?

12. Для лечения некоторой хронической болезни применяются пять лекарств *a*, *b*, *c*, *d*, *e*. Врач хочет провести сравнительное исследование трех из этих пяти лекарств. Три исследуемых лекарства отбираются из данных пяти случайным образом. Чему равна вероятность того, что:
    a) лекарство *a* будет исследовано;
    b) будут исследованы лекарства *a* и *b*;
    c) будет исследовано, по крайней мере, одно из лекарств *a* и *b*?

13. Из колоды карт (52 карты) наудачу выбирают 3 карты. Найти вероятность того, что это тройка, семерка, туз.

14. Водопроводчик Вася поздно вечером возвращается домой. У него в руках связка из пяти ключей, причем только один подходит к дверям квартиры. По причинам, о которых можно только догадываться, Вася пробует ключи наугад так, что при каждой попытке любой ключ, включая нужный, выбирается с одинаковой вероятностью. За этим захватывающим зрелищем через глазок дверей соседней квартиры внимательно следят Иван Кузьмич и Пелагея Марковна. Иван Кузьмич готов биться об заклад, что Васька и с третьей попытки в дом не попадет. Сердобольная же Пелагея Марковна утверждает, что, по крайней мере, на третий раз дверь поддастся. У кого больше шансов победить в споре?

15. На полке в почвенной лаборатории случайно расставлены бюксы с различными образцами почвы: 8 бюксов с влажной почвой, 6 – с сухой. Найти вероятность того, что из пяти выбранных наудачу бюксов:
    a) 4 будут с сухой почвой;
    b) по крайней мере 4 будут с сухой почвой.

16. На плодовой опытной станции для посадки в теплице подготовили 20 черенков: 8 черенков алычи, остальные черенки сливы. Случайным образом отобрано 3 черенка. Найти вероятность того, что:
    a) ровно 1 из трех отобранных является черенком алычи;
    b) хотя бы 1 является черенком алычи.

17. В клетке содержат 6 белых и 4 серых мыши. Рассмотрим эксперимент, состоящий в случайном извлечении из клетки трех мышей.

a) Опишите пространство элементарных исходов этого эксперимента в двух случаях: когда учитывается и когда не учитывается порядок извлечений.
b) Вычислите вероятности для следующих возможных комбинаций цвета мышей (3 белых; 2 белых и I серая; 1 белая и 2 серых; 3 серых).
c) Найдите вероятность того, что выбрано по крайней мере 2 белых мыши.

18. Для посадки на садовом участке подготовлены саженцы 2-х сортов черной смородины: 6 саженцев сорта Селеченская и 8 – сорта Вологда. По небрежности они были случайно смешаны. Случайно выбираются 3 саженца для посадки. Какова вероятность того, что будут посажены:
    a) 3 саженца сорта Селеченская;
    b) хотя бы 2 саженца сорта Селеченская;
    c) хотя бы 1 саженец сорта Селеченская;
    d) ни одного саженца сорта Селеченская?

19. Из 10 лотерейных билетов 2 – выигрышных. Определить вероятность того, что среди наудачу взятых 5-и билетов:
    a) один выигрышный;
    b) два выигрышных;
    c) хотя бы один выигрышный.

20. В ящике 5 красных и 4 зеленых яблока. Сколькими способами можно выбрать 3 яблока из ящика (без учета порядка извлечений)?
    Какова вероятность того, что:
    a) выбраны 3 красных яблока;
    b) 3 зеленых яблока;
    c) 1 красное яблоко;
    d) яблоки разных цветов?

21. Из 20 человек, одновременно заболевших гриппом, 15 выздоровели полностью за 3 дня. Из этих 20 человек случайно выбираются 5. Какова вероятность того, что:
    a) все 5 выздоравливают за 3 дня;
    b) выздоравливают, по крайней мере, 4 человека;
    c) ни один не выздоравливает?

22. Пустые горшочки с медом Винни-Пух ставит на полочку вместе с полными для того, чтобы вид уменьшающегося числа горшков не слишком портил ему настроение. В настоящий момент в Пуховом буфете вперемежку стоят 5 горшочков с медом и 6 абсолютно пустых. Какова вероятность того, что:

a) в двух взятых наудачу горшочках окажется мед;
b) мед окажется хотя бы в одном из двух?

23. В классе 20 учеников, из которых 8 отличников. Класс наудачу разделен на две равные части. Какова вероятность того, что:
    a) в каждой части одинаковое число отличников;
    b) все отличники оказались в одной половине?

24. Требуется выбрать наудачу 6 человек из группы в 6 мужчин и 6 женщин. Какова вероятность того, что:
    a) будет выбрано ровно трое мужчин;
    b) будет выбрано, по крайней мере, 5 мужчин;
    c) будет выбрано больше мужчин, чем женщин?

25. В команде из 20 спортсменов 5 мастеров спорта. По жеребьевке выбирают трех спортсменов. Найти вероятность того, что:
    a) все трое из выбранных – мастера спорта;
    b) выбрано, по крайней мере, два мастера;
    c) не выбрано ни одного мастера спорта.

26. В точке C полимерной цепочки длиной L произошел «обрыв» во время реакции. Допуская, что положение C равновозможно, определить вероятность того, что точка C удалена от левого конца на расстояние меньше l.

27. Шкала секундомера имеет цену деления 0,2 секунды. Какова вероятность получить результат с ошибкой не более 0,05 сек., если отсчет округляется до целого деления в ближайшую сторону?

28. Капля, размеры которой пренебрежимо малы, попадает наудачу в область, имеющую форму квадрата со стороной 2 см. Вероятность попадания в любую замкнутую область пропорциональна площади этой области. Найти вероятность того, что расстояние от капли до левого нижнего угла квадрата не более 1 см.

29. Иван и Петр договорились встретиться между часом и двумя часами дня. Каждый приходит в назначенное место в случайный момент данного промежутка времени и ждет другого в течение четверти часа, но не позже 2 часов дня, после чего уходит. Обозначая x – момент прихода Ивана, y – момент прихода Петра, построить на плоскости Oxy пространство элементарных событий, а также области, отвечающие событиям:
    A – встреча состоялась;
    B – Иван ушел, не дождавшись Петра;
    C – Петр пришел раньше Ивана и дождался его;

D – встреча произошла до 1 час. 20 мин.
Найти вероятности всех четырех событий.

30. Брошены две игральные кости. Событие A = {выпадение шестерки на первой кости}. Событие B = {сумма выпавших очков равна 7}. Являются ли события A и B независимыми?

31. Два стрелка независимо друг от друга стреляют по мишени. Вероятность попадания в мишень первого стрелка равна 0,3; вероятность попадания второго стрелка – 0,8. Найти вероятности следующих событий:
    a) первый стрелок промахнулся, а второй попал;
    b) хотя бы один стрелок попал.

32. Докажите, что если два события A и B с положительными вероятностями несовместны, то они зависимы.

33. Какова вероятность, что в семье с двумя детьми (без близнецов) оба ребенка - мальчики, если известно, что в семье есть мальчик? (Считать приближенно, что вероятность рождения мальчика равна 1/2 независимо от любой другой информации.)

34. Известно, что при трех бросаниях игральной кости цифра 6 выпала хотя бы один раз. Какова вероятность того, что она выпала два раза?

35. Предположим, что в семье с 3-мя детьми (без близнецов) все возможные распределения детей по полу и возрасту равновероятны. Определим событие A - “в семье имеются дети обоих полов” и событие B - “в семье имеется не более одной девочки”. Установить:
    a) независимы ли события A и B;
    b) независимы ли события A и B для семей с 2-мя детьми.

36. Игральная кость бросается два раза. X_1 и X_2 – числа выпавших очков. Рассматриваются события: A={X_1 делится на 2; X_2 делится на 3} и B={X_1 делится на X_2}. Установить, являются ли случайные события A и B независимыми, используя понятие условной вероятности.

37. На фабрике, изготавливающей лекарства, лаборатории №1, №2, №3 производят, соответственно, 25%, 35% и 40% всей продукции. Брак составляет, соответственно, 5%, 4% и 2%. Какова вероятность того, что:
    a) случайно выбранное лекарство – доброкачественное;
    b) случайно выбранное лекарство оказалось бракованным;
    c) случайно выбранное лекарство изготовлено лабораторией № 1, если известно, что оно бракованное?

38. В лаборатории 3 клетки. В клетке №1 содержатся 2 коричневые и 3 белые мыши; в клетке №2 – 4 коричневые и 2 белые мыши; в клетке № 3 – 5 коричневых и 5 белых мышей. Наудачу выбирается клетка и из нее наудачу – мышь. Какова вероятность, что она белая?

39. В магазин поступают телевизоры от 3 фирм. На долю 1-ой фирмы приходится 50% от общего числа поставок, на долю 2-ой фирмы – 20%, а на долю 3-ей фирмы – 30%. Из практики известно, что бракованными (т.е. попадающими в гарантийный ремонт) оказываются 4% телевизоров, поставляемых 1-ой фирмой, 3% поставляемых 2-ой фирмой и 5% поставляемых 3-ей фирмой. Найти вероятность того, что:
    a) купленный в данном магазине телевизор окажется бракованным;
    b) оказавшийся бракованным телевизор был произведён 1-ой фирмой.

40. Среди 25 экзаменационных билетов 20 – «счастливых» и 5 «несчастливых». Два студента по очереди наудачу вытягивают билеты. У кого вероятность вытащить «счастливый» билет больше: у первого или у второго?

41. 5% новорожденных мальчиков и 0,25% новорожденных девочек страдают дальтонизмом. 50% новорожденных – мальчики, и 50% - девочки. Какова вероятность того, что наудачу выбранный новорожденный – дальтоник?

42. Студент пришел на зачет, зная из 30 вопросов только 20. Зачет ставится за один правильный ответ. Какова вероятность сдать зачет, если после первого неудачного ответа преподаватель задает еще только один вопрос? Вопросы выбираются наудачу.

43. В семье 3 дочери Вера, Надежда, Любовь. В соответствии с возрастом обязанности по мытью посуды распределены в соотношении 4:3:3. Вероятность разбить посуду при мытье у каждой из дочерей равна соответственно 0,01; 0,02; 0,03. Родители, находясь в комнате, слышат звон разбитой посуды. Какова вероятность того, что посуду мыла Вера? Надежда? Любовь?

44. В двух коробках лежат цветные карандаши. В первой 3 красных и 2 синих, во второй 5 красных и 1 синий. Вася из каждой коробки вынул наудачу по одному карандашу и один из этих двух карандашей (тоже выбранный наудачу) подарил Маше. Какова вероятность, что у Маши оказался красный карандаш?

45. Семена для посева поступают в агрофирму из трех семеноводческих хозяйств. Причем первое и второе хозяйства присылают по 40% всех

семян. Всхожесть семян из первого хозяйства равна 90%, второго – 85%, третьего – 95%.

a) Определить вероятность того, что наудачу взятое семя не взойдет.
b) Пусть наудачу взятое семя не взошло. Какова вероятность, что оно получено от второго хозяйства?

46. В популяции цветных мышей 2/3 особей имеют гетерозиготный генотип (*Aa*), а 1/3 – гомозиготный тип (*AA*). Предположим, что альбинос (*aa*) и цветная мышь, наудачу выбранная из популяции, скрещиваются и дают потомство численностью 4 мышонка. (Известно, что признак альбинизма рецессивен.) Какова вероятность того, что все 4 мышонка будут цветными?

47. В первой коробке 25 ампул, из них 13 стандартных, во второй 15 ампул, из них 6 стандартных. Из первой коробки взята наудачу ампула и переложена во вторую коробку. Найти вероятность того, что ампула, извлеченная наудачу из второй коробки, будет стандартной.

48. Часы, изготовленные на 3-х заводах, поступают в магазин. Первый завод изготавливает 40% всей продукции; второй 45%; третий 15%. В продукции первого завода спешат 80% часов; в продукции второго спешат 70%; третьего - 90%. Найти вероятность того, что:
    a) наудачу выбранные часы не спешат;
    b) наудачу выбранные часы спешат;
    c) если купленные часы спешат, какова вероятность того, что они произведены на втором заводе?

49. На летней практике 70% первокурсников и 30% второкурсников. Среди первокурсников 60% девушек, среди второкурсников – 50% девушек. Все девушки по очереди дежурят на кухне. Найти вероятность того, что в случайно выбранный день на кухне дежурит первокурсница.

50. Популяция людей случайным образом разбита на две группы одинаковой численности. Первая группа в течение 10 лет придерживалась диеты с высоким содержанием ненасыщенных жиров; вторая (контрольная группа) питалась по обычной диете, богатой насыщенными жирами. Через 10 лет никто не умер, но вероятность возникновения сердечнососудистых заболеваний в этих группах составляла соответственно 31% и 48%. Какова вероятность того, что:
    a) случайно выбранный человек из этой популяции – здоров;
    b) случайно выбранный человек имеет сердечнососудистое заболевание;

c) если случайно выбранный человек имеет сердечнососудистое заболевание, то какова вероятность того, что этот человек из контрольной группы?

51. 5% новорожденных мальчиков и 0,25% новорожденных девочек страдают дальтонизмом. 50% новорожденных – мальчики, и 50% - девочки. Наудачу выбранный новорожденный – дальтоник. Какова вероятность того, что это мальчик?

52. На летней практике 2 группы первокурсников и 1 группа второкурсников. У первокурсников в каждой группе 5 девушек и 3 юношей. У второкурсников 4 девушки и 4 юноши. По жеребьевке выбрали одну группу и из нее наудачу студента для поездки в город.
    a) Какова вероятность того, что выбрана девушка?
    b) Выбранный оказался юношей. Какова вероятность того, что он первокурсник?

53. Пусть в коробке есть три новых и три уже использованных теннисных мяча. Для первой игры наудачу берут из коробки два мяча и затем их возвращают в коробку. Какова вероятность для второй игры из этой коробки наудачу вынуть два новых мяча?

54. При некотором обследовании на туберкулез вероятность обнаружить заболевание у больного туберкулезом равна 0,9. Вероятность принять здорового человека за больного равна 0,01. Пусть доля больных туберкулезом по отношению ко всему населению равна 0,001. Найти вероятность того, что человек здоров, если он был признан больным при обследовании.

55. Среди 10 000 000 монет одна фальшивая (с двух сторон – герб). Наудачу выбирается монета и бросается 10 раз. 10 раз выпал герб. Какова вероятность того, что эта монета – фальшивая?

56. В лабораторию для экспериментов поступают партии животных по 4 штуки. Отбор этих 4-х производят случайно из большого количества животных. С точки зрения эксперимента, проводимого в лаборатории, 40% животных имеет дефект. При получении партии из 4 животных в лаборатории производят выборочную проверку, а именно, партию из четырех животных принимают, если при проверке трех из них, выбранных наугад, они оказываются годными. Найти вероятность принятия партии.

57. На садовом участке посажены три дерева: вишня, слива и яблоня. Вероятность того, что приживется вишня, равна 0,7; для сливы и для ябло-

ни вероятности прижиться соответственно равны 0,8 и 0,9. Какова вероятность того, что:

a) приживутся ровно два дерева;
b) приживутся не менее двух деревьев;
c) приживется хотя бы одно дерево?

58. У пользователя имеются три дискеты для компьютера, изготовленные на фирмах *K*, *L* и *M*, по одной дискете от каждой из этих фирм, причем штампы фирм на дискетах отсутствуют. Две из имеющихся трех дискет оказались бракованными. Какова вероятность того, что бракованными являются дискеты фирм *L* и *M*, если брак в продукции фирмы *K* составляет 10%, а в продукции фирм *L* и *M* - соответственно 20% и 13%?

59. Три крысы обучаются выполнению трех различных заданий (по одной крысе на каждое задание). Вероятности того, что крысы после обучения выполняют свои задания за 1 мин. составляют соответственно 2/3, 1/2 и 1/3. Какова вероятность того, что:

a) все три крысы выполнят свои задания за 1 мин?
b) только две крысы выполнят свои задания за 1 мин?

60. При обследовании заболевания легких проверялось 10 000 человек в возрасте свыше 60 лет. Оказалось, что 4000 человек из этой группы являются постоянными курильщиками. У 1800 из курящих обнаружились серьезные нарушения в легких. Среди некурящих серьезные нарушения в легких имели 1500 человек. Являются ли (по этим данным) курение и наличие нарушений в легких независимыми событиями?

61. Предполагается, что все возможные распределения детей по полу и возрасту равновероятны и близнецов нет.

a) Какова вероятность того, что у девочки, о которой известно, что она растет в семье, где четыре ребенка, есть старший брат?
b) Сколько детей должна планировать семейная пара, чтобы вероятность наличия хотя бы одного мальчика была выше 70%?
c) Сколько детей должна планировать семейная пара, чтобы вероятность наличия хотя бы одного мальчика и одной девочки была выше 70%?
d) Какова вероятность того, что в семье из шестерых детей три мальчика и три девочки?
e) Какова вероятность того, что в семье из шестерых детей все дети одного и того же пола?

62. Представим, что в одной семье есть восемь детей — четыре мальчика и четыре девочки. Предполагается, что все возможные распределения детей по полу и возрасту равновероятны и близнецов нет.

a) Какова вероятность того, что старший ребенок — мальчик?
b) Какова вероятность того, что все четыре мальчика старше четырех девочек?

63. В семье три ребенка и близнецов нет. Предполагается, что все возможные распределения детей по полу и возрасту равновероятны.
Определим события: *A* (первый ребенок — девочка), *B* (второй ребенок — мальчик), *C* (третий ребенок — мальчик), *D* (первые два ребенка — мальчики) и *E* (хотя бы один ребенок — мальчик).
a) Вычислить вероятности этих пяти событий.
b) Являются ли независимыми события: *A* и *B*; *A* и *E*; *B* и *E*?

64. Двое бросают монету по очереди. Выигрывает тот, у кого раньше появится герб. Найти вероятности выигрыша каждого из игроков.

65. Сколько раз нужно подбросить два игральных кубика, чтобы вероятность выпадения хотя бы один раз двух шестерок была больше 0,5?

66. Известно, что 10% семян огурцов не всходят при посеве. Какова вероятность, что из 4 посеянных семян взойдут:
a) ровно два;
b) от 1 до 3 семян?

67. Студенты выполняют контрольную работу. Для получения положительной оценки достаточно решить две задачи из трех, написанных на карточке. Для каждой задачи предлагается 5 различных ответов, из которых только один правильный. Студент плохо знает материал и поэтому выбирает ответы наугад. Какова вероятность того что он получит положительную оценку?

68. Что вероятнее: выиграть у равносильного противника три партии из четырех или пять из восьми? (Ничейный вариант партии исключен.)

69. Сколько раз надо подбрасывать монету, чтобы вероятность появления герба три раза равнялась 0,25?

70. По каналу связи передается кодовая комбинация из 5 символов. Вероятность искажения одного символа при приёме равна 0,1. Найти вероятность того, что хотя бы один символ будет искажен.

71. В тестовом задании 8 вопросов, на каждый дано 4 варианта ответов, среди которых 1 правильный. Какое наиболее вероятное число правильных ответов даст отвечающий наудачу?

72. Что вероятнее произойдет при бросании правильной монеты – герб выпадет один раз при двух бросаниях или герб выпадет три раза при шести бросаниях?

73. «Осторожный фальшивомонетчик». Дворцовый чеканщик кладет в каждый ящик вместимостью в сто монет одну фальшивую. Король подозревает чеканщика и подвергает проверке монеты, взятые наудачу по одной в каждом из 50 ящиков. Какова вероятность того, что чеканщик не будет разоблачен?

74. Магазин получил 1000 бутылок минеральной воды. Вероятность, того, что при перевозке бутылка окажется разбитой, равна 0,003. Найти вероятность того, что магазин получит хотя бы одну разбитую бутылку. Сравнить точное и приближенные значения вероятностей (воспользоваться ***Excel*** или ***Statistica***).

75. Среди коконов некоторой партии 40% цветных. Какова вероятность того, что из 10 коконов, случайно отобранных из партии:
    a) 3 цветных;
    b) не более 3-х цветных?

76. Вероятность всхожести семян некоторого сорта растений 0,8. Какая вероятность больше: из десяти наугад выбранных семян взойдет хотя бы одно или из двенадцати семян взойдут хотя бы два?

77. Из поступивших в магазин телефонов третья часть белого цвета, однако цвет становится виден только после распаковки. Найти вероятность того, что из шести не распакованных телефонов:
    a) ровно 2 белых;
    b) есть хотя бы один белый.

78. В некоторой популяции насекомых 30% всех насекомых инфицированы. Что вероятнее: найти хотя бы одну инфицированную особь из 10 наугад выбранных насекомых или более одной из 12?

79. Обреченным на смерть пациентам в качестве последнего шанса можно предложить опасную операцию, в результате которой выживают 80% всех оперированных. Какова вероятность того, что ровно 80% из 5 оперированных выживут?

80. При одном обороте антенны радиолокационной станции, следящей за космическим объектом, объект будет обнаружен с вероятностью 0,5. Сколько оборотов должна совершить антенна, чтобы вероятность обнаружения объекта была больше 0,99?

81. Слушатели вводного курса по количественному химическому анализу достигают приемлемого результата в 80% титрований. Один студент добился приемлемого результата лишь однажды в шести титрованиях. Какова вероятность наступления такого события?

82. На станциях отправления поездов находятся 1000 автоматов для продажи билетов. Вероятность выхода из строя одного автомата в течение часа равна 0,004. Какова вероятность того, что в течение часа из строя выйдет два, три или пять автоматов? (Решать с помощью ***Excel*** или ***Statistica***).

83. Вероятность выигрыша в лотерею равна 0,3. Куплено 5 билетов. Найти наивероятнейшее число выигравших билетов и соответствующую вероятность. (Решать с помощью ***Excel*** или ***Statistica***).

84. Лечение некоторого заболевания приводит к выздоровлению в 75% случаев. Лечилось 6 больных. Какова вероятность того, что:
    a) выздоровеют все шестеро;
    b) не выздоровеет ни один;
    c) выздоровеет по крайней мере один?

85. Вероятность того, что саженец елки прижился и будет успешно расти равна 0,8. Посажено 400 елочных саженцев. Какова вероятность того, что успешно вырастут не более 325 деревьев. Сравнить точное и приближенные значения вероятностей. (Решать с помощью ***Excel*** или ***Statistica***).

86. На некотором поле повреждены гербицидами 15% растений мяты. Найти наивероятнейшее число поврежденных гербицидами растений мяты среди 20 растений, отобранных с этого поля случайным образом. (Решать с помощью ***Excel*** или ***Statistica***).

87. Склады семенного картофеля перед посадкой проверяют на отсутствие очагов гниения. В проверенном складе оказалось 20% клубней с пятнами. Найти наивероятнейшее число клубней без пятен из 9 клубней, отобранных случайным образом и вероятность наивероятнейшего числа клубней без пятен. Сравнить точное и приближенные значения вероятностей. (Решать с помощью ***Excel*** или ***Statistica***).

88. У клевера определенного сорта бывает в среднем 84% позднеспелых растений. Какова вероятность того, что 52 растения из 60 растений клевера, отобранных случайным образом, являются позднеспелыми? Сравнить точное и приближенные значения вероятностей. (Решать с помощью ***Excel*** или ***Statistica***).

89. При скрещивании двух кормовых сортов люпина во втором поколении теоретически ожидаемым отношением алкалоидных растений к безалкалоидным является отношение 9:7. Найти вероятность того, что среди полученных 150 гибридных растений половина растений будут алкалоидными. Сравнить точное и приближенные значения вероятностей. (Решать с помощью ***Excel*** или ***Statistica***).

90. Опытный участок засеян семенами костра безостого. На одной из делянок этого участка в травостое содержится 4% сорных растений - клевера белого и разнотравья. Какова вероятность того, что среди 125 растений этой делянки, отобранных случайным образом, имеются:
    a) ровно 3 сорных;
    b) не более трех сорных?

    Сравнить точное и приближенные значения вероятностей. (Решать с помощью ***Excel*** или ***Statistica***).

91. При механизированной уборке картофеля повреждается в среднем 10% клубней. Найти вероятность того, что в случайной выборке из 200 клубней картофеля повреждено от 15 до 50 клубней. Найти точное и приближенное значение и обосновать выбор вида приближения. (Решать с помощью ***Excel*** или ***Statistica***).

92. По гипотезе Менделя в опытах по скрещиванию желтого гибридного гороха вероятность появления зеленого гороха равна ¼. При 34153 опытах скрещивания (сделанных несколькими разными исследователями) в 8506 случаях был получен зеленый горох. Допуская, что во всех опытах вероятность получения зеленого гороха была постоянной и равной ¼, найдите:
    a) вероятность неравенства $0.245<\nu<0.255$, где ν- частота появления зеленого гороха;
    b) вероятность того, что при повторении опытов (34153 опыта) отклонение относительной частоты ν от ¼ по абсолютной величине будет больше полученного в описанном опыте;
    c) сколько аналогичных опытов нужно сделать, чтобы с вероятностью 0,99 можно было бы утверждать, что отклонение относительной частоты от ¼ не превзойдет 0,01?

93. Монета подброшена 3 раза. Найти распределение вероятностей числа выпадений герба. Нарисовать функцию распределения.

94. Вероятность того, что лотерейный билет окажется выигрышным, равна 0,1. Покупатель купил 5 билетов. Найти распределение вероятностей числа выигрышей у владельца этих билетов. Найти их математическое ожидание и дисперсию.

95. Распределение случайной величины ξ задано таблицей

a_i	1	2	3	4	5
$P\{\xi = a_i\}$	$0{,}4 \cdot b$	$0{,}2 \cdot b$	$0{,}1 \cdot b$	$0{,}2 \cdot b$	$0{,}1 \cdot b$

Найти b, $M\xi$, $D\xi$, $P\{\xi>3\}$, $P\{3<\xi<8\}$.

96. Случайная величина ξ имеет следующее распределение:

a_i	-2	-1	0	1	2
$P\{\xi = a_i\}$	c	$2 \cdot c$	$2 \cdot c$	$4 \cdot c$	c

Найти c, $M\xi$, $D\xi$, $P\{0<\xi<3\}$, $P\{\xi \geq 0\}$.

97. В большой популяции дрозофилы у 25% мух имеется мутация крыльев. ξ - число мух с этой мутацией в случайной выборке из 5 особей. Каково пространство значений, распределение вероятностей и функция распределения случайной величины ξ? Чему равны математическое ожидание и дисперсия этой случайной величины?

98. Хозяин булочной обслуживает небольшое количество людей. Вероятность продать k батонов в день задается соотношением:

	k=1, 2,…,25	k=26, 27,…, 49
$P(k)$=	$a \cdot k$	$a \cdot (50-k)$

a) При каком a мы имеем дело с распределением случайной величины?
b) Какова вероятность того, что в некоторый день лавочник продаст менее 26 батонов?
c) Какова вероятность того, что в некоторый день лавочник продаст более 22 и менее 28 батонов?
d) Какое минимальное число батонов должен иметь лавочник, чтобы удовлетворять спрос по крайней мере в 95% случаев?

99. Распределение по весу коров в некотором стаде приблизительно описывается нормальным законом с математическим ожиданием a=470кг и средним квадратическим отклонением σ = 30кг.
a) Какая доля коров имеет вес, не превосходящий 500кг?
b) Указать вес, который не превосходит десятая часть коров.
c) Какова вероятность того, что две из трех коров, отобранных случайным образом, будут (каждая) иметь вес более 470 и менее 530кг?

(Решать с помощью ***Excel*** или ***Statistica***).

100. По результатам опытов выявлено, что всхожесть семян в разных партиях ярового ячменя приблизительно описывается нормальным законом с математическим ожиданием равным 65%. Доля 0,9398 от

общего числа этих партий имеет всхожесть семян в пределах 65 ± 18,8%. Найти интервал, симметричный относительно математического ожидания, в который с вероятностью 0,9973 попадут возможные значения всхожести. (Решать с помощью ***Excel*** или ***Statistica***).

101. У яровой пшеницы определенного сорта длина главного колоса в сантиметрах приблизительно описывается законом распределения с плотностью

$$f(x) = \frac{1}{1{,}2\sqrt{2\pi}} e^{-(x-6{,}6)^2/2{,}88} \quad см^{-1}$$

Найти интервал, в который с вероятностью 0,9973 попадут возможные значения длины этого колоса. (Решать с помощью ***Excel*** или ***Statistica***).

102. Настриг шерсти у овец определенной породы приблизительно описывается нормальным законом. С вероятностью 0,9973 значения этой величины принадлежат симметричному относительно математического ожидания интервалу (7;10,6) кг. Найти интервал, симметричный относительно математического ожидания, в котором с вероятностью 0,95 заключены возможные значения настрига шерсти. (Решать с помощью ***Excel*** или ***Statistica***).

103. Длина початка кукурузы определенного сорта приблизительно описывается нормальным законом с математическим ожиданием a=12,6см. У 68,26% початков длина принимает значение, принадлежащее интервалу (11,4; 13,8) см. Какой процент початков имеет длину более 14,1см? (Решать с помощью ***Excel*** или ***Statistica***).

104. Процентное содержание гумуса в пахотном горизонте исследуемого участка почвы приблизительно описывается равномерным распределением на отрезке [1,8; 4,0]. Найти плотность распределения и функцию распределения этой случайной величины. Вычислить среднее значение и дисперсию процентного содержания гумуса, а также вероятность того, что в пахотном горизонте содержится от 2 до 3% гумуса. (Решать с помощью ***Excel*** или ***Statistica***).

105. Рацион с пониженным содержанием йода вызывает увеличение щитовидной железы у 60% животных данной популяции. Для эксперимента случайно выбирают подряд животных, пока не попадется животное с увеличенной щитовидной железой.
   a) Найти распределение числа ξ извлеченных животных (считая и больное).
   b) Найти значение функции распределения в точке 3,5.

106. Случайная величина распределена равномерно на отрезке (0;1).

a) Нарисовать функцию распределения.
b) Определить Mξ, Dξ.
c) Найти $P\{\xi>0,7\}$.
d) Найти такое x_p, что $F_\xi(x_p)=p$ для p=0,3; 0,6. Как называются эти величины?

107. Случайная величина ξ распределена равномерно на отрезке (a,b).
a) Какое распределение имеет случайная величина $\eta=(\xi-a)/(b-a)$?
b) Найти ее математическое ожидание, дисперсию и квантили порядка 0,3 и 0,9.

108. Плотность вероятности задается формулой:

$$p(x)=\begin{cases} 0 & x\le 0 \\ a\cdot x & 0<x\le 1 \\ 0 & x>1 \end{cases}$$

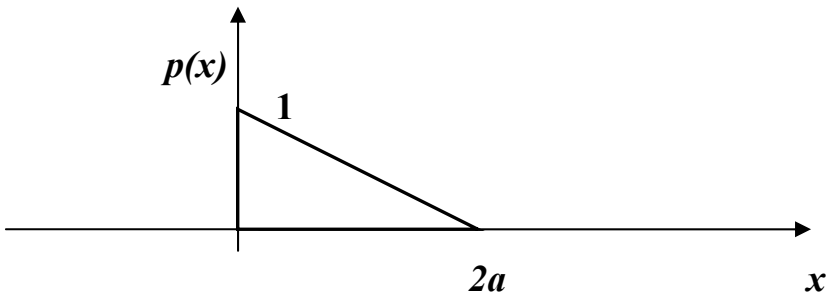
Найти:
a) величину a;
b) функцию распределения $F(x)$;
c) Mξ, Dξ;
d) $P(\xi>0,5)$, $P(0,3<\xi<0,7)$;
e) квантили $x_{0,2}$; $x_{0,8}$.

109. Плотность вероятности задается формулой:

$$p(x)=\begin{cases} 0 & x\le 0 \\ x & 0<x\le a \\ 0 & x>a \end{cases}$$

Найти:
a) величину a;
b) функцию распределения $F(x)$;
c) Mξ, Dξ;
d) $P(\xi>0,5)$, $P(0,3<\xi<0,7)$;
e) квантили $x_{0,1}$; $x_{0,5}$.

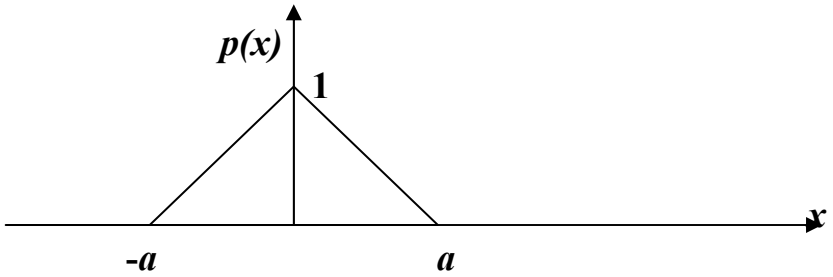
110. Плотность случайной величины задана графически



Найти:

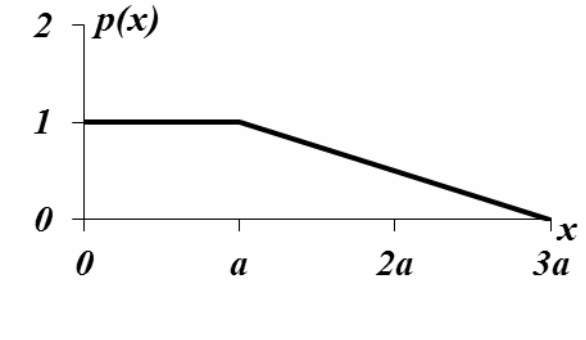
a) величину *a;*
b) функцию распределения $F(x)$;
c) $M\xi$, $D\xi$;
d) $P(\xi>-1)$, $P(0<\xi<0,7)$;
e) квантили $x_{0,2}$, $x_{0,9}$.

111. Плотность случайной величины задана графически



Найти:
a) величину *a;*
b) функцию распределения $F(x)$;
c) $M\xi$, $D\xi$;
d) $P(\xi>-1)$, $P(0<\xi<0,7)$;
e) квантили $x_{0,3}$; $x_{0,8}$.

112. Плотность случайной величины задана графически



Найти:
a) величину *a;*
b) функцию распределения $F(x)$;
c) $M\xi$;
d) $P(\xi>-1)$, $P(0<\xi<0,7)$;
e) квантили $x_{0,3}$, $x_{0,8}$.

113. Случайная величина ξ имеет следующее распределение:

a_i	-2	-1	0	1	2
$P\{\xi= a_i\}$	$2\cdot c$	$3\cdot c$	$4\cdot c$	$5\cdot c$	c

Найти:
a) величину c;
b) $M\xi$, $D\xi$;

c) $P\{0<\xi<3\}$, $P\{\xi>1\}$.

114. Случайная величина принимает значения:

-1	0	2	3	4

С вероятностями

a	$2a$	$3a$	a	$2a$

Найти:

a) величину a;
b) $M\xi$, $D\xi$;
c) $P\{\xi>3\}$, $P\{0<\xi<8\}$.

115. Из кладовой биологического факультета доставляют пробирки в пачках по 50 штук. Потребность в пробирках в данный день может с равной вероятностью быть 0,1,…,10 пачек.

a) Каково ожидаемое число пробирок, доставляемых в данный день?
b) Каково стандартное отклонение?

116. Известно, что для здорового человека значения pH крови примерно описываются нормально распределенной случайной величиной со средним 7,4 и стандартным отклонением 0,02. Какова вероятность того, что:

a) уровень pH превосходит 7,43;
b) находится между 7,39 и 7,45;
c) какой уровень не превосходит pH с вероятностью 0,95?

(Решать с помощью ***Excel*** или ***Statistica***).

117. Продолжительность жизни растений данного вида в определенной среде представляет собой непрерывную случайную величину ξ с плотностью распределения $f(x)=\frac{1}{120}e^{-x/120}$ $сут^{-1}$. Найти:

a) функцию распределения ξ;
b) долю растений, погибающих в течение 100 дней;
c) среднюю продолжительность жизни растений.
d) Если растение прожило 100 дней, какова вероятность того, что оно проживет еще 100 дней?

118. Длительность жизненного цикла определенного вида бактерий является экспоненциально распределенной случайной величиной со средней длительностью жизни 12 часов. Вычислить:

a) вероятность того, что данная бактерия закончит свое существование за 12 часов;
b) вероятность того, что бактерия, прожившая сутки, погибнет в течение следующих суток.

119. Установите, какие из приведенных ниже функций представляют собой функции плотности вероятности для непрерывной случайной величины:

а) $f(x)=\begin{cases}1, & -1/2<x<1/2\\ 0, & x\le -1/2;\quad x\ge 1/2\end{cases}$ б) $f(x)=\begin{cases}(1-x^2), & -1\le x\le 1\\ 0, & x\notin[-1,1]\end{cases}$

в) $f(x)=\begin{cases}1/2*\sin x, & 0\le x\le \pi\\ 0, & x\notin[0,\pi]\end{cases}$ г) $f(x)=\begin{cases}e^x, & -\infty<x<0\\ 0, & x>0\end{cases}$

д) $f(x)=\begin{cases}\sin x, & -\pi/2\le x\le \pi\\ 0, & x\notin[-\pi/2;\pi]\end{cases}$

- Найдите соответствующие функции распределения.
- Найдите вероятности того, что случайные величины принимают значения между 0 и 1.
- Найдите медиану и квантили порядка 0,1 и 0,8.

120. Фермер содержит 15 коров, 5 из которых дают удои более чем по 4500л молока в год. Случайным образом отобраны 3 принадлежащие этому фермеру коровы. Найти закон распределения случайной величины X - числа коров среди отобранных, дающих указанные высокие удои.

121. Три терморегулятора работают независимо друг от друга. Вероятность бесперебойной работы в течение смены первого терморегулятора равна 0,6. Для второго и третьего терморегуляторов эти вероятности соответственно равны 0,8 и 0,9.

a) Найти закон распределения случайной величины X - числа терморегуляторов, бесперебойно работающих в течение смены.
b) Вычислить математическое ожидание, дисперсию и среднее квадратическое отклонение величины X.

122. В некоторой популяции ржи наряду с зеленозерными растениями содержатся желтозерные, причем желтозерными являются 25% всех растений. Из этой популяции случайным образом отобраны четыре растения ржи.

a) Найти закон распределения случайной величины ξ - числа зеленозерных растений ржи среди четырех отобранных;
b) построить функцию распределения величины ξ.

Раздел II. Упражнения

Задачи курса теории вероятностей и математической статистики в основном можно свести к следующим: вероятностно-статистическая обработка наблюдений (тема общая и обширная, но имеющая для биологии не фундаментальное, а методическое значение) и рассмотрение несложных вероятностных моделей для решения прикладных задач.

1.Связь теории вероятностей с обработкой экспериментальных данных

Рассмотрим в простейшей форме одну из прикладных задач медицинской генетики. Пусть известно, что некоторое заболевание может возникнуть по причине отсутствия в организме синтеза определенного белка, причем известно, какой именно ген отвечает за синтез этого белка и какая именно мутация может произойти в этом гене, которая делает синтез белка невозможным. Мы планируем объем медицинской помощи лицам, которые больны данным заболеванием, для чего необходимо прежде всего знать, какова доля этих лиц среди всего населения. В простейшем (но достаточно реальном) случае независимой комбинации любых аллелей дело сводится к оценке аллельной частоты мутации p.

В самом деле, обозначим, как обычно в генетике, через A нормальный ген и через a мутацию. Возможны генотипы AA, Aa и aa. Если нормальный ген присутствует хотя бы в одной из родительских хромосом, то синтез белка возможен, так что аллель a в данном случае может считаться рецессивной. При чисто случайном сочетании аллелей, достающихся от каждого из родителей, вероятность генотипа AA составляет q^2, где $q = 1 – p$ – аллельная частота отсутствия мутации. Генотип Aa может получиться двумя способами (аллели A и a могут достаться от каждого из родителей), следовательно, его вероятность есть $2pq$, а вероятность генотипа aa (который и соответствует заболеванию) равна p^2. (Эти вероятности p^2, $2pq$ и q^2 представляют собой известный в генетике закон Харди-Вайнберга.)

Собственно говоря, при планировании объема медицинской помощи нас интересует доля больных среди всего населения, т.е. как раз p^2. Но единственный способ определения подобных вероятностей – экспериментальный. Можно было бы, в принципе, обследовать достаточно большую выборку из всего населения и оценить в ней долю больных. Но методы современной молекулярной биологии позволяют выяснить структуру генома человека (в интересующем месте) и тем самым оценить аллельную частоту p. Поскольку $p>p^2$, речь в таком случае пойдет об оценке большей величины. Но чем больше частота события, тем легче её оценивать в смысле необходимого количества наблюдений (см. ниже соответствующие математические выкладки). Таким образом, нужно взять выборку людей из интересующей группы населения, секвенировать соответствующие участки их геномов и оценить по этой выборке вероятность p как частоту встречаемости в геномах рецессивной аллели. Понятно, что в идеале выборка должна быть случайной, но такое осуществить невозможно. Практически речь идет обычно об исследовании выборки образцов крови доноров, которая сдается с целью переливания. Следовательно, мы должны предположить, что аллельная частота для доноров такова же, как и для всего населения. Если принять это

допущение, то возникает вопрос о том, каков должен быть объем выборки для достижения некоторой заданной точности определения p. (Не забудем, что число исследованных аллелей равно удвоенному числу образцов крови доноров, попавших в выборку.)

На какую степень точности следует ориентироваться? Планируя какие-то новые виды медицинской помощи, мы понимаем, что (по крайней мере, в ближайшее время) мы вряд ли сможем её оказать более, чем 1% всего населения. То есть разумно ориентироваться на значение p^2 порядка 0,01, т.е. значение p порядка 0,1. Ошибка в оценке объема помощи связана не с абсолютной ошибкой в оценке значения p (или p^2), а с относительной ошибкой: вряд ли допустимо ошибиться вдвое, но на 10% ошибиться можно. Если мы ошибемся в оценке p на относительную величину δ, т.е. вместо истинной вероятности p примем вероятность $p(1+\delta)$, то вместо p^2 мы примем величину $p^2(1+\delta)^2 = p^2(1+2\delta+\delta^2) \approx p^2(1+2\delta)$. Иными словами, относительная ошибка в значении p примерно удвоится при оценке объема медицинской помощи. Теперь немного теории.

Принципиально подбор супружеских пар мог бы быть связан с наличием или отсутствием тех или иных аллелей, но исследование подобных вопросов находится за гранью возможного. Поэтому обычно считается, что комбинации аллелей, достающихся от родителей, происходят независимо. В таком случае выявление интересующей аллели может быть уподоблено последовательности независимых испытаний, каждое из которых имеет два исхода – успех, т.е. наличие аллели (вероятность которого есть p), и неудачу – ее отсутствие (вероятность которого есть q) . Такие испытания называются *испытаниями Бернулли*. В теорию вероятностей они вошли с незапамятных времен в связи с азартными играми. В частности, если искомая аллельная частота на самом деле равна ½, то ситуация аналогична бросанию монеты. Вот и рассмотрим, насколько точно может быть определена эта частота успеха с помощью того или иного числа бросаний монеты.[1]

Основными параметрами, характеризующими случайную величину в теории вероятностей, считаются ее математическое ожидание (обозначается M) и дисперсия (обозначается D), а также стандартное отклонение, обозначаемое σ, которое равно квадратному корню из дисперсии. Обозначим через $\mu(n)$ число успехов в n испытаниях Бернулли (с вероятностью p успеха в отдельном испытании). Оказывается, что

$$\mathbf{M}\,\mu(n) = np;\ \mathbf{D}\,\mu(n) = npq;\ \sigma_{\mu(n)} = \sqrt{npq}\ .$$

В теории вероятностей в виде математических теорем устанавливается следующее правило. Во многих случаях (в частности, и для испытаний Бернулли) отклонение случайной величины от математического ожидания на

[1] Реально аллельные частоты могут быть гораздо меньше 1/2, скажем, составлять величину порядка 1/10. В этом случае для их оценки требуется гораздо больше наблюдений, так что на первых порах даже компьютерное моделирование этого процесса несколько громоздко. Поэтому мы начинаем со случая p=1/2.

σ (или еще больше) вполне вероятно: такая вероятность составляет около 1/3. Но отклонение на 2σ (или больше) гораздо менее вероятно: примерно 1/20, т.е. 5%. Отклонение на 2,5σ (или больше) имеет вероятность около 1%, а на 3σ – доли процента. То есть, лишь в одном случае из 20 можно вместо ожидаемого значения $\mathbf{M}\mu(n)=np$ получить значение μ, такое, что $|\mu(n) - np| > 2\sqrt{npq}$.

Понятно, что если вероятность *p* нам не известна, то за ее оценку на основании опыта следует взять *частоту* успеха в *n* испытаниях Бернулли, т.е. величину $\hat{p} = \mu(n)/n$. Если бы оказалось, что $\mu(n) = \mathbf{M}\mu(n) = np$, то частота в точности равнялась бы вероятности *p*. Но поскольку $\mu(n)$ может отклоняться от *np* на величину порядка $\sqrt{npq}$, частота может отклоняться от вероятности на величину порядка $\sqrt{\frac{pq}{n}}$. При этом относительная ошибка составит величину порядка $\frac{1}{p}\sqrt{\frac{pq}{n}} = \sqrt{\frac{q}{np}}$, т.е. будет тем больше, чем меньше *p*.

Вместо доказательства математических теорем мы сейчас займемся тем, что проверим эти утверждения о достигаемой точности экспериментально.

В настоящее время принято считать, что случайные эксперименты можно имитировать с помощью компьютера. С этой целью употребляются так называемые *псевдослучайные числа.* Они создаются небольшими и быстро работающими компьютерными программами, называемыми датчиками случайных чисел. Такой датчик имитирует (например) эксперимент, состоящий в случайном бросании точки на единичный отрезок [0;1]. Любые другие случайные величины моделируются как функции от этих случайных чисел. В частности, испытания Бернулли можно моделировать по следующему правилу. Успех (обозначаемый 1) происходит в случае, если полученное случайное число $\leq p$, а если оно $> p$, то происходит неудача (обозначаемая 0).

Кроме быстроты выполнения случайных экспериментов, компьютер имеет еще то преимущество, что позволяет быстро строить графики, и в частности график числа успехов $\mu(n)$ как функции *n*. Интересно, что глазомерный анализ таких графиков позволяет выявить некоторые подробности, ускользающие при доказательстве математических теорем.

Такого рода задача предлагается в упражнении 1.

2.Проблема проверки статистических гипотез

Теперь мы рассмотрим другой вариант применения статистических методов с целью лучшего понимания результатов эксперимента. Речь идет о проблеме выбора между двумя "вариантами действий": можно ли сколько-нибудь обоснованно утверждать, что один вариант лучше, чем другой.

Возьмем в качестве примера эксперимент по конкурентному взаимодействию в двухвидовом сообществе инфузорий (виды *A* и *B*). В

начале эксперимента в некоторую среду помещается определенное количество особей того и другого вида. Затем они начинают размножаться, конкурируя, например, за пищевой ресурс, и, в конце концов, после смены ряда поколений может получиться так, что в живых останутся только особи одного вида, скажем, вида *A* (то есть можно сказать, что вид *A* победил в конкурентной борьбе).

Другим примером может быть сравнение урожайности двух сортов. На двух соседних делянках высеваются сорта *A* и *B*, и если в конце концов выясняется, что сорт *A* дал больший урожай, то это называется «победой» сорта *A*.

В тех случаях, когда подобное соревнование, проведенное в нескольких экспериментах, всегда кончалось победой одного и того же участника (например вида *A*), никакого вероятностного рассмотрения не требуется: ясно, что *A* всегда оказывался сильнее, чем вид *B*, и можно предположить, что и при проведении аналогичных экспериментов в будущем вид *A* всегда будет одерживать победу. Но ситуация существенно изменяется в том случае, когда исход опыта не вполне однозначен: допустим, что при проведении нескольких опытов вид *A* одержал больше побед, чем вид *B*, но все-таки иногда побеждал *B*.

Понятно, что если *A* побеждал чаще, то правильно сделать вывод о том, что этот вид в каком-то смысле сильнее, чем вид *B*. Этот вывод должен обладать предсказательной силой: на самом деле, имеется в виду, что, если в будущем сделать много аналогичных экспериментов, то *A* опять-таки будет побеждать чаще. Однако такой вывод нельзя сделать иначе, как в вероятностных терминах. Нужно предположить, что отдельные повторения эксперимента образуют *статистический ансамбль* в том смысле, что можно говорить о *вероятности* победы *A* в каждом отдельном эксперименте, причем эта вероятность *p* (в простейшем случае, на котором мы и остановимся) считается одинаковой во всех экспериментах (в том числе и в мыслимых будущих экспериментах). Кроме того, отдельные эксперименты считаются независимыми. В теории вероятностей есть свое математическое определение независимости, но в данный момент знать его не обязательно: достаточно предположить, что в содержательном смысле эксперименты устроены так, что ход одного не влияет на ход другого. Независимость будет обеспечена, если различные эксперименты проводятся в различных сосудах с питательной средой. При помещении ситуации в вероятностный контекст утверждение о том, что вид *A* сильнее вида *B* сводится к тому, что *p* - вероятность победы *A* - больше, чем вероятность победы *B*, или (что то же самое) к утверждению $p>1/2$.

Очевидно, что для того, чтобы можно было утверждать, что $p>1/2$, надо уметь обоснованно отвергнуть утверждение о том, что $p=1/2$. Исследование возможностей подобного обоснования и называется проблемой статистической проверки гипотез.

Рассмотрим сначала обычное понятие логического вывода в условиях, когда нет никакой случайности и все утверждения носят детерминированный характер. Применительно к проверке научных гипотез один из частных видов логического вывода может выглядеть следующим образом:

1-ая посылка: если гипотеза *H* верна, то наступление события *S* невозможно;

2-я посылка: произведен опыт и событие *S* произошло.

Заключение: гипотеза *H* неверна.

Правомерность такой логики никаких сомнений не вызывает.

Однако в вероятностном контексте происходит нечто другое.

При рассмотрении вероятностной модели мы имеем пространство исходов эксперимента $\Omega=\{\omega\}$, выдвигаемая гипотеза *H* задает вероятности исходов, а различные подмножества множества Ω являются событиями.

Аргументом против гипотезы *H* является не наступление события *S*, вовсе невозможного при верной гипотезе *H*, а наступление события маловероятного (если *H* верна). Таким образом, в первой посылке мы заменяем «невозможно» на «маловероятно». Хотелось бы в заключении тоже сказать, что «*H* маловероятна». Но такое заключение в большинстве случаев невозможно, потому что выразимую числом вероятность имеют массовые явления, а научные гипотезы обычно к таким не относятся. Мы должны как-то выразить малую веру в гипотезу *H*.

Надо сказать, что эта проблема обсуждается в науке уже, по крайней мере, лет триста. И простого и естественного её решения наука так и не нашла. Выработана не вполне естественная процедура так называемой "статистической проверки гипотез".

Формально это выглядит так. Заранее зададим положительное, близкое к нулю число α, называемое *уровнем значимости*. Затем, руководствуясь соображениями о том, как проверяемая гипотеза может быть нарушена, выделяем во множестве $\Omega=\{\omega\}$ всех исходов эксперимента некоторое подмножество *S* (событие *S)*, которое называется *критическим подмножеством*. Это подмножество обладает следующим свойством: если гипотеза *H верна*, вероятность наступления события *S* не превосходит α. После того, как выбраны α и *S,* производим опыт и получаем результат ω. Если $\omega \in S$ (т.е. наступает событие *S*), то можно сказать, что гипотеза *H* скорее всего неверна. Но принята более осторожная формулировка: *гипотеза H отклоняется на уровне значимости* α. Если же $\omega \notin S$ (событие *S* не наступило), то говорят: *на уровне значимости* α *гипотеза H не отклоняется.*

Задача этого рода предлагается в упражнении 2.

3. Практические задания

Упражнение 1.

1. Используя компьютерную имитацию бросания монеты (файл «*PR1*») вычислите частоты выпадения герба $\hat{p}$ для последовательности из N=1000

наблюдений и нарисуйте график частоты выпадения герба $\hat{p}(n)$ в зависимости от числа бросаний n.

- Повторите 10 раз последовательность 1000-кратного бросания монеты. Сколько раз из этих 10 частота выпадения герба не попала в интервал $0{,}5 \pm 2\sqrt{\frac{pq}{N}} = 0{,}5 \pm \sqrt{\frac{1}{1000}}$? Для определения доли таких случаев объедините полученные результаты в вашей группе студентов.
- Стабилизируется ли частота выпадения герба при увеличении числа испытаний? Если да, то к какому значению? Заметьте, что приближение к этому значению обычно происходит «с одной стороны» (сверху или снизу), а не путем колебаний «вокруг» этого значения.

2. Рассмотрите число выпадений герба $\mu(n)$ при n бросаниях монеты. Постройте график, состоящий из точек с координатами (n,μ), для n=1÷100, и на том же графике нарисуйте прямую $y=x/2$, состоящую из точек $(n, n/2)$.

На этом же рисунке постройте прямые $y=x/2 \pm \frac{\sqrt{x}}{2}$; $y=x/2 \pm \frac{2\sqrt{x}}{2}$; $y=x/2 \pm \frac{3\sqrt{x}}{2}$.

Получите различные реализации процесса бросания монеты.

Попадают ли реализации в границы, заданные построенными прямыми?

Указания.

В файле «PR1» приведены необходимые последовательности, в которых нуль соответствует стороне монеты с цифрой, а единица - стороне с гербом. Для подсчета количества гербов достаточно найти сумму чисел в этой последовательности.

Обратите внимание, что при любом действии на этом листе Вы получаете новую реализацию. В частности, получение новой реализации может быть осуществлено нажатием клавиши Del в любой пустой ячейке.

К п.1. Для получения частоты выпадения герба в зависимости от числа бросаний необходимо воспользоваться функцией СРЗНАЧ. Например, если реализация последовательности бросаний расположена в диапазоне (B3:B1002), то в ячейке C3 запишите формулу: = СРЗНАЧ (C$3:C3) и скопируйте (протяните) ее до C1002. В столбце C Вы получите значения частоты выпадения герба для каждого n.

К п.2. Для получения величины $\mu(n)$ воспользуйтесь таким же приемом, как и для получения $\hat{p}(n)$, только вместо функции СРЗНАЧ, воспользуйтесь функцией СУММ.

Построение графика в Excel.

Главное меню Вставка->Диаграмма->Точечная, выберите с правой стороны вид графика (например, точки, соединенные линиями)->Далее->закладка РЯД->кнопка ДОБАВИТЬ->справа в открывшихся окнах:

в окно "имя" занесите название кривой , которую Вы изображаете (по умолчанию ей присваивается имя "Ряд1");

в окне "Значения X" укажите диапазон ячеек, в которых находятся значения абсцисс изображаемой кривой. Для этого нажмите красную стрелку в квадратике, который находится в правом конце окошка, выделите требуемый диапазон ячеек, снова нажмите красную стрелку (если абсциссы равны1,2,...,можно не заполнять это окно);

в окне "Значения Y" таким же образом укажите диапазон ячеек, в которых находятся значения ординат.

->Далее (этот пункт можно пропустить и сразу нажать "Готово") выберите необходимые закладки и внесите заголовки и др. ->Готово

Упражнение 2.

Вернемся к примеру с конкуренцией 2-х видов в 10 пробирках (см. Раздел II, п.2), где Ω - набор 10-позиционных последовательностей букв *A* и *B*. Гипотеза *H* состоит в том, что все эти последовательности равновероятны (с вероятностями, равными 2^{-10}). Наша задача выбрать подходящее *S* для заданного α. (Заметим, что аналогичное множество Ω мы получаем и при 10-кратном бросании правильной монеты).

Формально мы можем выбрать из этого множества Ω любое событие *S*, вероятность которого при верной гипотезе *H* не превосходит α. Но тут должны вступать неформальные соображения об альтернативе. В данном случае при неверной гипотезе *H* (мы предполагаем, что возможное нарушение состоит в том, что *A* «сильнее» *B*, т.е. $p>1/2$) естественно выбрать в качестве *S* ту совокупность, где содержится наибольшее количество букв *A*.

Обозначим через $\mu(10)$ число букв *A* в 10-позиционной последовательности. Ниже приводится таблица вероятностей для различных значений, которые может принимать $\mu(10)$ (повторим, что они вычислены, исходя из соображения равновероятности всех возможных 10-позиционных последовательностей из двух букв).

$\mu(10)$	вероятность
0	0,001
1	0,010
2	0,044
3	0,117
4	0,205
5	0,246
6	0,205
7	0,117
8	0,044
9	0,010
10	0,001

Далее следует задать α, выбрать множество S, такое, что сумма вероятностей для элементов, входящих в него, не превышает α, и отвергать гипотезу о равноправии видов A и B в случае попадания во множество S (наступления события S).

1. Задайте α=0,055. Определите по таблице множество S. В файле «*PR2*» (лист «данные1») представлены результаты смоделированных 100 серий экспериментов по конкуренции видов A и B в предположении, что виды равноправны. Используя в качестве экспериментальных эти данные, проверьте гипотезу H. (При такой постановке задачи гипотеза H отвергается, когда в результате эксперимента мы попадаем в S). Посчитайте среднюю относительную частоту попадания в область S.
2. На листе «данные2» аналогичным образом представлены результаты экспериментов, в которых предположение о равноправии A и B нарушено (и в этом случае проверяемую гипотезу хотелось бы отвергать всегда). Проделайте такую же проверку гипотезы H, что и в п.1 данного упражнения, подсчитайте среднюю частоту попаданий в S и сравните ее с результатом, полученным в п.1.

Указания.

В файле «PR2» на листе «данные1» в ячейках B2:CW11 представлены результаты смоделированных 100 серий экспериментов по конкуренции видов A и B (в предположении, что виды равноправны). В 12-ой строке для каждой серии подсчитайте количество букв «A» (например с помощью функции СЧЁТЕСЛИ(B2:B11;"=A")). В любой свободной ячейке (например, B15) подсчитайте долю серий экспериментов (из 100), в которых мы попали в критическую область S (используя функцию СЧЁТЕСЛИ(B12:CW12;">7"). Запишите этот результат (долю попаданий в S) в другую свободную ячейку. При этом в ячейках B2:CW11, в 12-ой строке и в ячейке B15 появятся результаты для новых 100 серий экспериментов. Запишите новый результат под предыдущим. Проделайте это несколько раз (но не менее 15). Вычислите среднюю долю попаданий в S и сравните его с α.

Упражнение 3.

Пусть в испытаниях Бернулли число испытаний n=50,

1. Изобразите на одном графике зависимости биномиальных вероятностей P_k (формула (4.3.1)) от k для p = 0,3; 0,5; 0,7.
   - Опишите поведение P_k.
   - Для каких значений k вероятности P_k принимают максимальное значение?
   - Как связаны вероятности P_k для p=0,3 и p=0,7?
2. Для каждого значения p=0,01; 0,05; 0,5 постройте на одном листе три графика зависимости от k: биномиальных вероятностей (формула (4.3.1)),

нормальное приближение этих вероятностей (формула (4.3.3)) и Пуассоновское приближение (формула (4.3.2)).

- В каком случае биномиальные вероятности лучше описываются пуассоновским приближением, а когда – нормальным?

Указания.

Для вычислений воспользуйтесь функциями БИНОМРАСП, НОРМРАСП, ПУАССОН в ***Excel*** *или Binom, Normal, Poisson в* ***Statistica.***

Упражнение 4.

1. На выпечку 1000 булочек с изюмом полагается 10000 изюмин. Найти вероятности P_k , того, что в купленной нами булочке ровно k изюмин (k= 0, 1, …, 10000).
   Какие предположения позволяют нам воспользоваться схемой испытаний Бернулли?
2. Нарисуйте вероятности P_k для k=0,…,20 (точные значения с помощью формулы биномиальной вероятности, а также с помощью Пуассоновского и нормального приближений).
3. Какое число изюмин в одной булочке мы можем считать указателем того, что среднее число изюмин в одной булочке меньше 10, рискуя при этом ошибиться с вероятностью не более 0,05.

Указания.

1. *«Испытания Бернулли» состоят в том, что мы (естественно, мысленно) проверяем, попала ли каждая из 10000 изюмин в нашу булочку. Поскольку всего булочек 1000, то для каждой изюмины вероятность попасть в нашу булочку равна 1/1000=0,001. Т.е. n=10000, p=0,001* и для пуассоновского приближения $\lambda=10$.
2. *При построении графиков воспользуйтесь функциями, предложенными в указаниях к упражнению 3.*
3. *Рассуждения здесь такие же, как в упражнении 2 (выдвигается основное предположение-гипотеза: $\lambda=10$ на уровне значимости 0,05 затем ищется «критическое» множество S).*

Упражнение 5.

Известно, что вероятность рождения мальчика ≅ 0,515.

- Какова вероятность того, что среди 10000 новорожденных мальчиков будет не больше, чем девочек?
- Какое точное распределение имеет число новорожденных мальчиков?
- Какое приближение для этого распределения целесообразно применить в данном случае? Обоснуйте выбор графически.

Указания.

При построении графиков по оси абсцисс выберите область, где значения вероятности превосходят 0,001.

Упражнение 6.

С 1881 по 1900 гг. В Швейцарии родилось 1 359 671 мальчиков и 1 285 086 девочек. Считая, что мы можем ошибиться с вероятностью 0,001, определить, совместимы ли эти данные с предположением о том, что вероятность рождения мальчика равна 0,5? 0,515?

Указания.

Для решения можно воспользоваться нормальным приближением. Далее рассуждения о совместимости выдвинутого предположения с реальными данными аналогичны рассуждениям в упражнениях 2 и 4.

Упражнение 7.

1. Постройте на одном графике плотности распределения для трех нормальных случайных величин: $\mathbf{M}\xi_1=-2$, $\mathbf{M}\xi_2=0$, $\mathbf{M}\xi_3=2$; $\mathbf{D}\xi_1=\mathbf{D}\xi_2=\mathbf{D}\xi_3=1$.
   - Как изменяется вид плотности при изменении математического ожидания?
2. Постройте на одном графике плотности распределения для трех нормальных случайных величин: $\mathbf{M}\xi_1=\mathbf{M}\xi_2=\mathbf{M}\xi_3=0$; $\mathbf{D}\xi_1=1$, $\mathbf{D}\xi_2=4$, $\mathbf{D}\xi_3=9$.
   - Как изменяется вид плотности при увеличении дисперсии?

Упражнение 8.

Изобразите на одном графике функции распределения Стьюдента для числа степеней свободы n=3,5,10,30. На том же графике начертите функцию стандартного нормального распределения.

- Начиная с какого n (числа степеней свободы) можно пользоваться нормальным приближением для распределения Стьюдента?

Указания.

Для получения соответствующих функций распределения воспользуйтесь функциями СТЬЮДРАСП и НОРМСТРАСП в ***Excel*** *или Student и Normal в* ***Statistica.***

Упражнение 9.

1. Постройте на одном графике плотности распределения для трех χ^2 распределенных случайных величин с числом степеней свободы n=1,5,10.
   - Как изменяется вид плотности при увеличении числа степеней свободы?
2. Постройте на одном графике функцию распределения χ^2 с 30 степенями свободы и функцию распределения нормальной случайной величины с математическим ожиданием 30 и дисперсией 60.
   - Можно ли пользоваться нормальным приближением при нахождении квантилей распределения χ^2?

Указания.
Для получения значений плотности и функции распределения воспользуйтесь функциями Chi2, iChi2, Normal и iNormal в ***Statistica.***

Упражнение 10.

В файле «*PR10*» приведены 100 значений случайных чисел с нулевым математическим ожиданием и дисперсией 1.

- Подсчитайте долю значений, вышедших за σ, 2σ, 3σ *(*т.е. отклоняющихся от среднего значения на одно, два, три стандартных отклонения). Согласуется ли результат с неравенством Чебышева?
- Вычислите точное значение вероятности таких отклонений, зная, что $\xi \sim N(0,1)$.

Упражнение 11.

В файле «*PR11*» приведены значения 1000 случайных величин имеющих нормальное распределение с математическим ожиданием, равным 2 и дисперсией, равной 9.

- Постройте график зависимости среднего арифметического от числа наблюдений;
- рассмотрите колебания среднего арифметического вокруг значения математического ожидания.

Указания.
Данное упражнение рекомендуется выполнять в ***Excel*** *(см. указания к упражнению 1).*

Упражнение 12.

В файле «*papilio»* приведены данные из работы по видообразованию бабочек рода *Papilio.* Переменные:
TERG - восьмикратная длина *tergit* (мм);
UNCUS – восьмикратная длина *uncus* (мм);
S – название вида в следующих обозначениях: 1- *Papilio multicaudatus*; 2 - *Papilio rutulus*; 3 - *Papilio glaucus*; 4 - *Papilio eurymedon*.

Пользуясь данными из файла «*papilio»* выполните следующие задания:

1. Постройте накопленную выборку для переменных *TERG* и *UNCUS*. Постройте график накопленной суммы.
2. Проведите прямую, соединяющую конечные точки на графике, и вычислите отклонения накопленной суммы от этой прямой. Постройте график отклонений.
   - Выявляются ли неоднородности выборок для каждой из переменных на графике для накопленных параметров?

- Выявляются ли неоднородности выборок для каждой из переменных на графике отклонений?
- Насколько неоднородность поведения графиков соответствуют разделению на виды?

Упражнение 13 .

Сгенерировать 10 выборок объемом 100 каждая из распределения χ^2 с двумя степенями свободы (переменные v1,v2,.,v20).

Начертить в нормальном масштабе эмпирические функции распределения для переменных: v1; (v1+v2)/2; (v1+v2+…+v20)/20 и объяснить поведение полученных графиков с помощью ЦПТ.

Упражнение 14.

Пользуясь данными из файла «*papilio»* (см. упражнение 12), выполните следующие задания:

1. Для переменной *TERG* постройте на одном графике эмпирические функции распределения в нормальном масштабе для каждого вида.
2. То же самое проделайте для переменной *UNCUS.*
   - Как меняются от вида к виду значения переменных *TERG* и *UNCUS* и их разбросы.
   - Можно ли по этим переменным объединить какие-либо два вида?

Упражнение 15.

В файле «*eggs*» приведены данные по размеру и весу куриных яиц категорий «0», «1», «2». Переменные: *weight* – вес яйца в г; *length* – размер по высоте в мм; *width* – размер в самой широкой части в мм.

Пользуясь данными из файла «*eggs*» выполните следующие задания:

1. Постройте эмпирические функции распределения в нормальном масштабе для переменных *length, width, weight* отдельно для каждой категории.
2. На каждом из этих графиков проведите прямую и на глаз оцените математическое ожидание, стандартную ошибку и квантили порядка 0,3; 0,6; 0,75.

Упражнение 16.

Пользуясь данными из файла «*eggs*» выполните следующие задания:

- С помощью компьютерных программ оцените для всех переменных из файла (для каждой категории отдельно) математическое ожидание, стандартное отклонение и квантили, перечисленные в упражнении 15.2.
- Сравните полученные результаты с соответствующими оценками в упражнении 15.2.
- Постройте доверительные интервалы для математического ожидания веса для каждой категории.

Упражнение 17.

В файле «*PR17*» представлены 10 выборок из стандартного нормального распределения по 250 наблюдений каждая.

- Постройте для каждой выборки доверительные интервалы для математического ожидания с доверительной вероятностью 0,95.
- Определите, сколько построенных интервалов накрывают математическое ожидание.
- Получите аналогичные выборки еще 10 раз и определите общую долю интервалов, накрывающих математическое ожидание.

Указания.

Это упражнение проще делать с помощью ***Excel****.*

Для каждой из представленных выборок в ячейках M2:V3 вычислите нижнюю и верхнюю границы 95% доверительных интервалов для математического ожидания. На графике эти границы изобразятся точками. При этом видно, какие интервалы накрывают математическое ожидание, а какие – нет. Для получения новых выборок достаточно поставить курсор в любую свободную ячейку и нажать «Del».

Для нахождения оценок:

математического ожидания - воспользуйтесь функцией СРЗНАЧ,

стандартного отклонения – СТАНДОТКЛОН.

α-квантиль распределения Стьюдента с n степенями свободы равна

t_α =СТЬЮДРАСПОБР(1-α;n) для α>0,5 и t_α =-СТЬЮДРАСПОБР(α;n) для α<0,5;

корень квадратный находится с помощью функции КОРЕНЬ.

Упражнение 18.

Пользуясь данными по определению заряда электрона, полученными Р.Милликеном в 1913 году (файл «*Milliken*», заряд электрона в 10^{-10} ед. СГС), выполните следующие задания:

- Постройте эмпирическую функцию распределения в нормальном масштабе.
- Проверьте нормальность распределения.
- Постройте 95% доверительный интервал для заряда электрона.
- Найдите в интернете современное значение заряда электрона. Попадает ли оно в доверительный интервал, полученный Вами?

Упражнение 19.

Пользуясь данными из файла «*eggs*» выполните следующие задания:

1. Проверьте нормальность распределения графически и с помощью процедур проверки гипотез для переменной «*weight*» категория «1».
2. Проверьте гипотезу о равенстве весов I и II категорий.

Упражнение 20.

В учебнике «Статистический анализ данных на компьютере» (Ю.Н.Тюрин, А.А. Макаров. 1998) приводятся данные о времени реакции на звук (x) и на свет (y) у группы испытуемых. И, хотя единицы измерения не приведены, следует считать, что это миллисекунды (файл «*PR20*»). Предполагается, что условия проведения эксперимента обеспечивают независимость данных, полученных на одном испытуемом от аналогичных данных для других испытуемых.

В учебнике предлагается проверить с помощью критерия знаков для разности (x-y) гипотезу о том, что время реакции на звук и на свет можно считать одинаковым.

В нашем упражнении мы предлагаем более тщательно исследовать приведенные данные:

- постройте диаграмму рассеяния (x,y);
- оцените коэффициент корреляции между переменными x и y;
- постройте эмпирическую функцию распределения для разности (x-y);
- на основании визуального анализа диаграммы рассеяния и графика эмпирической функции распределения удалите резко выделяющееся наблюдение и проделайте еще раз задания предыдущих пунктов;
- возможно ли применение t-критерия для проверки гипотезы о равенстве нулю математического ожидания переменной (x-y)? Если да, то проведите эту проверку;
- проверьте с помощью критерия знаков гипотезу о том, что время реакции на свет и на звук в группе испытуемых можно считать одинаковой.

Упражнение 21.

По данным из файла «*eggs*» (см. Упражнение 15) найдите зависимость веса яйца от его линейных размеров для категорий «0», «1», используя модуль регрессионного анализа *"Multiple Regression"* пакета ***Statistica*** .

1. Объясните следующие полученные величины:
   - оценки B_0, B_1, B_2;
   - multiple R; R^2; F; p;
   - Standard error of estimate;
   - Std.Err. of B.
2. Исследуйте остатки:
   - постройте графики зависимости остатков $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i})$ от значений переменных y, x_1, x_2 и от номера наблюдения;
   - по этим графикам проведите визуальный анализ зависимости разброса остатков от величин переменных;
   - постройте график эмпирической функции распределения остатков в нормальном масштабе для визуальной проверки нормальности.

3. Постройте графики остатков $\hat{\varepsilon}$ для категории «2». Сравните их с остатками $\hat{\varepsilon}$ для категорий «0», «1» (п.2).
4. Исследуйте возможность упростить полученную модель (для категорий «0», «1»).
   - С помощью *t*- критерия проверьте гипотезы: $b_1=1$ и $b_2=2$.
   - Если не отвергнуты гипотезы предыдущего пункта, вычислите $\hat{\hat{y}} = b_0 + x_1 + 2 \cdot x_2$.
   - Получите остатки $\hat{\hat{\varepsilon}} = y - \hat{\hat{y}}$. Сравните свойства остатков $\hat{\varepsilon}$ и $\hat{\hat{\varepsilon}}$ (графически и их среднеквадратичные отклонения).

Указания.

Предполагается, что вес яйца пропорционален его объему. Если считать, что яйцо является телом вращения, то его объем пропорционален $(length)^{\alpha}$ *и* $(width)^{\beta}$ *(где* α *и* β *неизвестны, но близки к 1 и 2 соответственно), поэтому можно предположить, что* $weight \approx K^* (length)^{\alpha} * (width)^{\beta}$. *Для того, чтобы неизвестные параметры* α *и* β *входили линейно в это равенство, прологарифмируем его (введя для простоты следующие обозначения:* $y=ln(weight)$; $x_1=ln(length)$; $x_2=ln(width)$). *Тогда* $y \approx b_0+b_1x_1+b_2x_2$ *или*

$$y=b_0+b_1x_1+b_2x_2+\varepsilon \qquad (1),$$

что соответствует модели регрессии (в предположении $My= b_0+b_1x_1+b_2x_2$; ε *- нормально распределенная случайная величина с нулевым математическим ожиданием и постоянной дисперсией).*

Данные приведены для трех категорий, определяемых размером. По этому признаку разделите данные на 2 части: одну часть (категории «0» и «1») используйте для получения оценок параметров регрессии; вторую (категория «2») – для оценки качества модели. Другими словами, для проверки качества приближения следует получить оценки неизвестных параметров по категориям «0» и «1» (п.п.1-2), а затем «испытать» модель (1) на категории «2» (п.3).

Раздел III. Справочные материалы

4. Основы теории вероятностей

4.1. Пространство элементарных исходов, события. Задание вероятностей. Операции над событиями

Случай дискретного (т.е. конечного или счетного) пространства элементарных событий

Множеством (*или* ***пространством) элементарных событий*** *(или элементарных* ***исходов****) называется конечное или счетное множество* $\Omega=\{\omega\}$. *Считается, что каждому его элементу* ω *приписано*

неотрицательное число P(ω), называемое вероятностью элементарного события ω. Сумма вероятностей всех элементарных исходов равна единице, или в формульной записи $\sum_{\omega} P(\omega_i) = 1$.

Событием A *называется любое подмножество множества элементарных событий, или в формульной записи* $A \subseteq \Omega$.

Говорят, что событие *A* в опыте наступило, если опыт закончился элементарным событием ω, являющимся одним из элементов *A* (в формульной записи $\omega \in A$). Элементарные события ω, которые входят в подмножество *A*, называются еще *благоприятными* для события *A*.

Вероятностью *P(A) любого события A называется сумма вероятностей элементарных событий, входящих в подмножество A, или в формульной записи* $P(A) = \sum_{\omega_i \in A} P(\omega_i)$.

Пустое подмножество множества элементарных событий ∅ *называется* ***невозможным событием****; его вероятность считается равной 0. Подмножество, совпадающее со всем множеством элементарных событий* Ω*, называется* ***достоверным событием. Его вероятность равна 1.*** Впрочем, иногда в множество Ω включается несколько элементарных событий с нулевыми вероятностями, и в этом случае *достоверным* событием можно называть любое событие вероятности 1.

Дополнением *события A (обозначается* $\bar{A}$ *или* $\Omega \setminus A$*) называется событие, состоящее из таких элементарных событий, которые не входят в событие A.* В теории вероятностей дополнение называется также ***отрицанием****,* а иногда употребляется более старый термин ***противоположное событие.*** Имеем равенство

$$P(A) + P(\bar{A}) = \sum_{\omega \in A} P(\omega) + \sum_{\omega \notin A} P(\omega) = \sum_{\omega \in \Omega} P(\omega) = 1.$$

Это равенство выражается словами в виде простенькой теоремы: *сумма вероятностей события и его отрицания равна единице.*

По определению, ***сумма событий*** *есть их теоретико-множественное объединение, а именно: событие* $A \cup B$ (иногда обозначаемое $A+B$) *состоит из таких элементарных событий, которые входят хотя бы в одно из событий A или B.* ***Пересечение событий*** *есть их теоретико-множественное пересечение, т.е. событие* $A \cap B$ (обычно обозначаемое AB) *состоит из таких элементарных событий, которые входят в каждое из событий A и B.*

Вероятность суммы событий *равна сумме их вероятностей минус вероятность пересечения:*

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

Частный случай этой теоремы возникает когда события *A* и *B* не имеют общих элементов, т.е. $AB = \varnothing$. (Такие события *A* и *B* называются еще *несовместными,* поскольку ни в каком эксперименте они не могут наступить одновременно.) В этом случае $P(AB) = 0$**,** и получается формула

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

т.е. вероятность суммы несовместных событий равна сумме их вероятностей. Очевидно, это правило справедливо для суммы не только двух, но и любого числа несовместных событий.

В случае, когда все элементарные исходы равновозможны, ***вероятность события равна*** *отношению числа элементарных событий, благоприятствующих этому событию, к общему числу элементарных событий* (случай классической вероятности).

Случай общего пространства элементарных событий.

Вообще говоря, множество элементарных событий не обязано быть конечным или счетным. Однако считается, что для достаточно широкого класса его подмножеств может быть определена вероятность, обладающая свойством *вероятность счетной суммы несовместных событий равна сумме их вероятностей.* Например, обобщением классической вероятности является геометрическая вероятность. ***Геометрическая вероятность*** определяется как *вероятность попадания в область* $g \subseteq G$ *при бросании наудачу точки в область G и равна* $P(g) = \frac{mes\ g}{mes\ G}$, *где mes означает меру области (длина, площадь, объем).*

4.2. Условная вероятность. Вероятность произведения событий. Независимость. Формула полной вероятности. Формула Байеса

Условной вероятностью *события B при условии, что событие A наступило, называется величина* $P(B/A) = P(AB)/P(A)$. Отсюда:

$$P(AB) = P(A)P(B/A),$$

т.е. вероятность совместного наступления двух событий равна вероятности одного из них, умноженной на условную вероятность второго при условии, что первое наступило.

Два события A и B называются ***независимыми,*** *если выполняется равенство* $P(AB) = P(A)P(B)$.

Любое событие B может быть представлено в виде

$$B = B\Omega = BH_1 + B\,H_2 + \ldots + BH_n + \ldots,$$

где $H_1, H_2, \ldots, H_n, \ldots$ конечное или счетное множество непересекающихся событий (т.е. $H_iH_j = \varnothing$ при $i \neq j$), каждое из которых имеет положительную вероятность, и выполняется равенство: $\Omega = H_1 + H_2 + \ldots + H_n + \ldots$, т.е. $H_1, H_2, \ldots, H_n, \ldots$ образуют полную группу событий. Тогда $P(B) = \sum_i P(BH_i) = \sum_i P(H_i)P(B/H_i)$.

Эта формула называется *формулой полной вероятности.*

В тех же обозначениях *формула Байеса* определяет условные вероятности $P(H_i/B)$:

$$P(H_i/B) = \frac{P(H_iB)}{P(B)} = \frac{P(H_i)P(B/H_i)}{\sum_i P(H_i)P(B/H_i)}.$$

4.3. Испытания Бернулли. Поведение биномиальных вероятностей

В научной и практической деятельности постоянно приходится проводить многократно повторяющиеся испытания в сходных условиях. Предположим, что при этом результаты предшествующих испытаний никак не сказываются на последующих. Очень важен простейший тип таких испытаний, когда в каждом из испытаний некоторое событие A может появиться с одной и той же вероятностью p, и эта вероятность остается неизменной независимо от результатов предшествующих или последующих испытаний. Этот тип испытаний был впервые исследован знаменитым швейцарским ученым Якобом Бернулли (1654—1705): отсюда название *испытания Бернулли.*

Схема следующая. Имеется n независимых испытаний, в каждом из которых возможны 2 исхода: событие A и противоположное ему $\overline{A}$ (обычно их называют успехом У= A и неудачей Н= $\overline{A}$). Вероятности этих событий соответственно обозначают $p=P($У$)$ *и* $q=P($Н$)$, где $p+q=1$. Эти вероятности не меняются от испытания к испытанию.

В испытаниях Бернулли нас интересует величина μ - количество успехов при n испытаниях, которая может принимать значения 0, 1,…, n. Вероятность события $\{\mu=k\}$ (обозначим его P_k) вычисляется по формуле:

$$P_k = C_n^k p^k q^{n-k} = \frac{n!}{(n-k)!k!} p^k q^{n-k} \quad . \qquad (4.3.1)$$

В настоящее время вычисления по последней формуле без труда выполняются с помощью компьютерных пакетов. Однако следует иметь в виду ряд приближенных выражений для этих вероятностей. Это так называемые *пуассоновское* и *нормальное* приближения. Они важны и в компьютерную эпоху, поскольку действуют также и в некоторых таких случаях, когда исходная схема испытаний отличается от схемы Бернулли. Степень этого отличия в практических ситуациях часто уточнить невозможно. Поэтому область применимости указанных приближений сознательно обрисовывается расплывчато. В общем, речь идет о больших значениях n (несколько десятков и более). Если при этом $p << ½$, для (4.3.1) рекомендуется применять пуассоновское приближение:

$$P_k \approx \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{где } \lambda = np \, . \qquad (4.3.2)$$

Если, наоборот, p близко к единице, то вместо успехов следует рассматривать неудачи и пуассоновское приближение применять к

вероятностям неудач. Если же n велико, а p не близко к 0 или 1, то применяют нормальное приближение :

$$\sqrt{npq}\,P_k \approx \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(k-np)^2}{2npq}\};\qquad (4.3.3)$$

На практике обычно требуются не столько вероятности $P_к$ отдельных значений $к$ числа успехов, сколько вероятности того, что эти значения попадают в тот или иной интервал от $к_1$ до $к_2$.

Вероятность события $\{k_1 \le \mu \le k_2\}$ можно вычислить с помощью функции Лапласа $\Phi(x)$:

$$P\{k_1 \le \mu \le k_2\} \approx \Phi(x_2) - \Phi(x_1), \text{ где } x_1 = \frac{k_1 - np}{\sqrt{npq}},\quad x_2 = \frac{k_2 - np}{\sqrt{npq}}.$$

4.4. Понятие случайной величины. Функция распределения случайной величины. Независимость случайных величин

Случайная величина ξ *это числовая функция, определенная на пространстве элементарных исходов* $\xi=\xi(\omega)$. Мы рассматриваем дискретные и непрерывные случайные величины.

F(x) - функция распределения *случайной величины* ξ *- определяется по формуле* $F(x)=P(\xi \le x)$. Это неубывающая функция, принимающая значения от 0 до 1.

Дискретная случайная величина может быть задана таблицей вида:

Значения	a_1	a_2	...	a_k	...
$p_i=P(\xi=a_i)$	p_1	p_2	...	p_k	...

Сумма вероятностей отдельных значений дискретной случайной величины всегда равна 1, т.е. $\sum_k p_k = 1$, $p_k>0$. Тогда функция распределения принимает значения $F(a_i)=\sum_{a_k \le a_i} p_k$.

Для непрерывной случайной величины функция распределения задается следующим образом:

$$F(x) = \int_{-\infty}^{x} f(t)dt\,,$$

где ***f(x) - плотность распределения случайной величины*** ξ. Здесь $F(x)$ удовлетворяет условиям: $\lim_{x\to-\infty} F(x) = 0$; $\lim_{x\to\infty} F(x) = 1$.

Плотность распределения $f(x)$ – неотрицательная функция и $\int_{-\infty}^{\infty} f(x)dx = 1$, т.е. площадь под кривой, изображающей плотность, равна 1.

Вероятность попадания случайной величины в заданный интервал может быть вычислена двумя способами:

1) через функцию распределения $P(a < \xi \le b) = F(b) - F(a)$;

2) через плотность распределения $P(a < \xi \le b) = \int_a^b f(x)dx.$

Квантилью x_p (*p-квантилью, квантилью уровня p, квантилью порядка p) непрерывной случайной величины* ξ, *имеющей функцию распределения* $F_\xi(x)$, *называют решение* x_p *уравнения* $F_\xi(x) = p$, $p \in (0, 1)$. Квантили, наиболее часто встречающиеся в практических задачах, имеют свои названия:

медиана - *квантиль уровня 0,5;*

нижняя квартиль - *квантиль уровня 0,25;*

верхняя квартиль - *квантиль уровня 0,75;*

децили - *квантили уровней 0,1; 0,2; ... ; 0,9;*

процентили - *квантили уровней 0,01; 0,02; ... ; 0,99 (часто эти числа выражают в процентах).*

Совместное рассмотрение двух или нескольких случайных величин приводит к понятию системы случайных величин. Такая система называется также ***многомерной случайной величиной***.[2] В частности ***двумерной случайной величиной*** *называют систему из двух случайных величин* (ξ_1, ξ_2), *для которых определена вероятность* $P(\xi_1 \le x; \xi_2 \le y)$ *совместного выполнения неравенств* $\xi_1 \le x$ *и* $\xi_2 \le y$, *где x и y - любые действительные числа.*

Случайные величины ξ_1 ***и*** ξ_2 ***независимы***, *если независимы события* $\xi_1 \le x$ *и* $\xi_2 \le y$ *для любых x и y.*

Распределение дискретной двумерной случайной величины задается набором вероятностей $p_{ij}=P(\xi_1=a_i;\ \xi_2=bj)$, где a_i - значения, которые принимает ξ_1, bj - значения, которые принимает ξ_2 ($\sum_{i,j} p_{ij} = 1$). Для независимых ξ_1 и ξ_2 $p_{ij} = p_i p_j$, *где* $p_i=P(\xi_1=a_i)$; $p_j=P(\xi_2=bj)$.

Для двух непрерывных случайных величин $F(x,y) = P(\xi_1 \le x; \xi_2 \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$. В случае независимости ξ_1, ξ_2 выполняются соотношения $f(x,y)=f_1(x)f_2(y)$ и $F(x,y) = F_1(x) \cdot F_2(y)$.

4.5. Основные параметры случайной величины

Математическое ожидание *случайной величины* $\mathbf{M\xi}$***:***

1) Для дискретной случайной величины ξ, *заданной рядом распределения:*

$$M\xi = \sum_k a_k p_k;$$

2) Для непрерывной случайной величины ξ *заданной плотностью распределения:*

$$M\xi = \int_{-\infty}^{\infty} x \cdot f(x)dx;$$

Дисперсия *случайной величины* $\mathbf{D\xi} = M(\xi - M\xi)^2 = M\xi^2 - (M\xi)^2$.

[2] Более подробно будет рассмотрена в параграфе 4.6.

Вычисление $\mathbf{M}(\xi-\mathbf{M}\xi)^2$ и $\mathbf{M}\xi^2$ производят, пользуясь свойством: $\mathbf{M}\varphi(\xi)=\sum_k \varphi(a_k)p_k$ для дискретных случайных величин и $\mathbf{M}\varphi(\xi)=\int_{-\infty}^{\infty}\varphi(x)f(x)dx$ для непрерывного случая.

Среднее квадратическое отклонение случайной величины $\sigma=\sqrt{\mathbf{D}\xi}$

Математическое ожидание и среднее квадратичное отклонение имеют ту же размерность, что и случайная величина ξ.

Свойства математического ожидания:

1. Математическое ожидание постоянной величины равно самой постоянной, т.е. $\mathbf{M}c = \boldsymbol{c}$.
2. Постоянный множитель можно выносить за знак математического ожидания, т.е. $\mathbf{M}(a\xi) = a\mathbf{M}\xi$.
3. Математическое ожидание суммы двух случайных величин равно сумме математических ожиданий слагаемых**,** т.е. $\mathbf{M}(\xi+\eta) = \mathbf{M}\xi+ \mathbf{M}\eta$**.**
4. Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий, т.е. $\mathbf{M}(\xi\eta) = \mathbf{M}\xi\cdot\mathbf{M}\eta$.

Свойства дисперсии:

1. Дисперсия постоянной равна нулю, т.е. $\mathbf{D}c = \mathbf{0}$.
2. Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат $\mathbf{D}(a\xi) = a^2\mathbf{D}\xi$.
3. Дисперсия суммы независимых случайных величин равна сумме их дисперсий $\mathbf{D}(\xi+\eta) = \mathbf{D}\xi+ \mathbf{D}\eta$.

Ковариацией называется выражение:

$$cov(\xi,\eta)=\mathbf{M}[(\xi-\mathbf{M}\xi)\cdot(\eta-\mathbf{M}\eta)]=\mathbf{M}\xi\eta-\mathbf{M}\xi\cdot\mathbf{M}\eta.$$

Если случайные величины ξ и η независимы, то их коэффициент ковариации равен нулю; обратное в общем случае неверно.

Коэффициентом корреляции случайных величин ξ и η называется число:

$$\rho_{\xi\eta} = corr(\boldsymbol{\xi,\eta}) = \frac{\mathbf{M}(\boldsymbol{\xi}-\mathbf{M}\boldsymbol{\xi})(\boldsymbol{\eta}-\mathbf{M}\boldsymbol{\eta})}{\sqrt{\mathbf{D}\boldsymbol{\xi}\cdot\mathbf{D}\boldsymbol{\eta}}} = \frac{\mathrm{cov}(\boldsymbol{\xi},\boldsymbol{\eta})}{\sqrt{\mathbf{D}\boldsymbol{\xi}\cdot\mathbf{D}\boldsymbol{\eta}}},\ |\rho_{\xi\eta}|\le 1.$$

Приведем примеры часто встречающихся законов распределения случайных величин.

Дискретные случайные величины.

1. Случайная величина, распределенная по ***закону Бернулли,*** принимает 2 значения: {0; 1} с вероятностями (1-p) и p соответственно. Закон распределения определяется одним параметром p. $\mathbf{M}\xi=p;$ $\mathbf{D}\xi=p\,(1-p)$.

2. ***Биномиальная случайная величина*** (число успехов в n испытаниях Бернулли) принимает значения $\{0,1,\ldots, k,\ldots, n\}$ с вероятностями $P(\xi=k)= C_n^k p^k q^{n-k}$. Закон распределения определяется параметрами p и n. $\mathbf{M}\xi= np;$ $\mathbf{D}\xi= np\ (1-p)$.
3. ***Пуассоновская случайная величина*** принимает значения $\{0,1,\ldots, k,\ldots\}$ с вероятностями $P(\xi =k)= e^{-\lambda}\frac{\lambda^k}{k!}$. У распределения Пуассона один параметр λ. $\mathbf{M}\xi=\lambda$, $\mathbf{D}\xi=\lambda$.

Непрерывные случайные величины.

1. ***Равномерно распределенная*** на отрезке $[a,b]$ случайная величина ξ принимает значения $b\geq\xi\geq a$. Ее плотность на отрезке $[a,b]$ равна константе, вне этого отрезка равна нулю.
2. ***Экспоненциально*** *(или* ***показательно) распределенная случайная величина*** ξ может принимать любые неотрицательные действительные значения. Ее плотность равна $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$, а функция распределения $F(x)=\begin{cases}0, & x<0\\ 1-e^{-\lambda x}, & x\geq 0\end{cases}$. Математическое ожидание и дисперсия вычисляются по формулам: $\mathbf{M}\xi=\frac{1}{\lambda}$, $\mathbf{D}\xi=\frac{1}{\lambda^2}$.
3. ***Нормальная случайная величина*** может принимать любые действительные значения. Ее плотность равна $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\{\frac{-(x-a)^2}{2\sigma^2}\}$.

   Это распределение имеет два параметра a и σ. $\mathbf{M}\xi=a;$ $\mathbf{D}\xi=\sigma^2$. (Краткая запись: $\xi\sim N(a,\sigma^2)$). Нормальное распределение с параметрами $a=0$; $\sigma=1$, т.е. $N(0,1)$ называется стандартным нормальным распределением. Нормальный закон распределения обладает следующими свойствами:

- случайная величина η, полученная из нормальной случайной величины ξ линейным преобразованием $\eta=c_1\xi+c_2$ тоже подчиняется нормальному закону распределения;
- сумма двух случайных величин $\xi_1\sim N(a_1,\sigma_1^2)$ и $\xi_2\sim N(a_2,\sigma_2^2)$ имеет нормальное распределение с параметрами $\mathbf{M}(\xi_1+\xi_2)=a_1+a_2;$ $\mathbf{D}(\xi_1+\xi_2)=\sigma_1^2+\sigma_2^2+cov(\xi_1,\xi_2)$.

С нормальным распределением связаны другие распределения.

4. Случайная величина имеет ***логнормальное распределение***, если её логарифм имеет нормальное распределение. Распределение такой случайной величины ξ задаётся плотностью вероятности, имеющей вид:

$f_\xi(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\{\dfrac{-(\ln x - a)^2}{2\sigma^2}\}, & x > 0 \\ 0, & x \le 0 \end{cases}$, где $\sigma > 0$, $a \in R$. Тогда говорят, что ξ имеет логнормальное распределение с параметрами a и σ^2. В краткой записи: $\xi \sim \mathrm{Log}N(a,\sigma^2)$, и тогда $\eta = \ln\xi \sim N(a,\sigma^2)$.

5. ***Распределение χ^2 (хи-квадрат) с k степенями свободы*** — это распределение суммы квадратов k независимых стандартных нормальных случайных величин. В формульной записи: $\xi_i \sim N(0,1)$, $\eta = \xi_1^2 + \xi_2^2 + \ldots + \xi_k^2$, $\eta \sim \chi_k^2$; $\mathbf{M}\eta = k$, $\mathbf{D}\eta = 2k$.
6. Пусть $\xi_0, \xi_1, \xi_2, \ldots, \xi_n$ – независимые стандартные нормальные случайные величины. Тогда распределение случайной величины η, заданной формулой $\eta = \dfrac{\xi_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \xi_i^2}}$ называется ***распределением Стьюдента (***или ***t-распределением*****) с n степенями свободы:** $\eta \sim t_n$.

4.6. Совместное распределение нескольких случайных величин

На пространстве элементарных исходов можно задать несколько случайных величин $\xi_1, \xi_2, \ldots, \xi_n$. Такой набор случайных величин, рассматриваемых совместно, называется n-мерным случайным вектором или n-мерной случайной величиной. Для такой случайной величины определяется совместная функция распределения: $F(x_1,x_2,\ldots,x_n) = P\{\xi_1 \le x_1, \xi_2 \le x_2 \ldots, \xi_n \le x_n,\}$.

Случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ независимы (в совокупности), если для любых $x_1, x_2, \ldots, x_n$ имеет место равенство

$$F_{\xi_1,\xi_2,\ldots,\xi_n}(x_1, x_2, \ldots, x_n) = F_{\xi_1}(x_1) \cdot F_{\xi_2}(x_2) \cdot \ldots \cdot F_{\xi_n}(x_n)$$

Перечислим свойства функции совместного распределения. Для простоты обозначений ограничимся вектором (ξ,η) из двух величин.

- Для любых x,y верно неравенство: $0 \le F_{\xi\eta}(x,y) \le 1$.
- $F_{\xi\eta}(x,y)$ не убывает по каждой координате вектора (x,y).
- $\lim\limits_{x\to-\infty} F_{\xi\eta}(x,y) = 0$ и $\lim\limits_{y\to-\infty} F_{\xi\eta}(x,y) = 0$.
- Двойной предел существует и равен 1, т.е. $\lim\limits_{x\to\infty}\lim\limits_{y\to\infty} F_{\xi\eta}(x,y) = 1$.
- Чтобы по функции совместного распределения получить функцию распределения величины ξ следует устремить $y \to \infty$ т.е. $\lim\limits_{y\to\infty} F_{\xi\eta}(x,y) = F_\xi(x)$; для получения функции распределения η следует устремить $x \to \infty$ т.е. $\lim\limits_{x\to\infty} F_{\xi\eta}(x,y) = F_\eta(y)$.

1. Для дискретных случайных величин ξ, η их совместное распределение, как и в одномерном случае, проще задавать набором вероятностей $p_{k,l} = P(\xi = a_k, \eta = b_l)$ в виде таблицы:

ξ \ η	b_1	b_2	…	b_l	…
a_1	p_{11}	p_{12}	…	p_{1l}	…
a_2	p_{21}	p_{22}	…	p_{2l}	…
…	…	…	…	…	…
a_k	p_{k1}	p_{k2}	…	p_{kl}	…
…	…	…	…	…	…

где $\sum_i \sum_j p_{ij} = 1$.

Тогда совместная функция распределения равна $F_{\xi\eta}(x,y) = \sum_{i:a_i \le x} \sum_{j:b_j \le y} p_{ij}$.

Частный закон распределения ξ можно получить по формуле:

$$P\{\xi = a_i\} = p_{\xi i} = \sum_j P\{\xi = a_i, \eta = b_j\} = \sum_j p_{ij}.$$

Аналогично получается закон распределения η :

$$P\{\eta = b_j\} = p_{\eta j} = \sum_i P\{\xi = a_i, \eta = b_j\} = \sum_i p_{ij}.$$

Необходимым и достаточным условием независимости ξ и η является условие: $p_{ij}=p_{\xi i} \cdot p_{\eta j}$. А из свойства коэффициента ковариации следует, что для независимых случайных величин $corr(\xi,\eta)=0$. Обратное не всегда верно, т.е. из равенства коэффициента корреляции нулю не следует независимость случайных величин.

2. Для непрерывных случайных величин ξ,η функция распределения задается следующим образом: $F_{\xi\eta}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{\xi\eta}(s,t)dsdt$, где $f_{\xi\eta}(x,y)$-плотность совместного распределения.

Свойства плотности:

- $f_{\xi\eta}(x,y) \ge 0$ для любых x,y.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\xi\eta}(x,y)dxdy = 1.$

Необходимым и достаточным условием независимости ξ и η является условие: $f_{\xi\eta}(x,y) = f_{\xi}(x) \cdot f_{\eta}(y).$

4.7. Неравенство Чебышева. Закон больших чисел. Центральная предельная теорема

Неравенство Чебышева.

Пусть у случайной величины ξ математическое ожидание равное a и дисперсия σ^2 конечны. Тогда для любого $\varepsilon > 0$

$$P\{|\xi - a| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}.$$

Закон больших чисел.

Пусть $\xi_1, \xi_2, \ldots, \xi_n$ - последовательность попарно независимых случайных величин, имеющих ограниченные в совокупности дисперсии, т. е. $D\xi_i \leq C$ для любого i. Тогда, каково бы ни было $\varepsilon>0$, справедливо соотношение

$$\lim_{n\to\infty} P\left\{\left|\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} - \frac{M\xi_1 + M\xi_2 + \ldots + M\xi_n}{n}\right| < \varepsilon\right\}.$$

Закон больших чисел устанавливает факт приближения среднего значения большого числа случайных величин к некоторым постоянным. Представление о распределении суммы случайных величин дает центральная предельная теорема (ЦПТ). ЦПТ представляет собой группу теорем, утверждающих, что сумма достаточно большого числа случайных величин распределена приближённо по нормальному закону.

Центральная предельная теорема.

Пусть $\xi_1, \xi_2, \ldots, \xi_n$ — независимые случайные величины с математическими ожиданиями $M\xi_1, M\xi_2, \ldots, M\xi_n$ и дисперсиями $D\xi_1, D\xi_2, \ldots, D\xi_n$, соответственно. Пусть

$$U_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n - (M\xi_1 + M\xi_2 + \ldots + M\xi_n)}{\sqrt{D\xi_1 + D\xi_2 + \ldots + D\xi_n}},$$

тогда при справедливости некоторых условий, обеспечивающих малость вклада любого из слагаемых в U_n, для любого x выполняется равенство:

$$\lim_{n\to\infty} P(U_n < x) = \Phi(x),$$

где $\Phi(x)$- функция распределения стандартного нормального закона распределения.[3]

5. Математическая статистика

5.1. Понятие выборки

При решении многих научных и практических вопросов мы сталкиваемся с наборами чисел $x_1, x_2, \ldots, x_n$, которые получаются при повторениях одного и того же эксперимента. Например, это может быть n измерений одной и той же биологической характеристики у различных объектов исследования, записанных в том порядке, как они получены.

Чтобы использовать вероятностно-статистические приемы обработки таких данных, необходимо построить вероятностную модель. Она заключается в том, что мыслится некий (ненаблюдаемый) набор независимых

[3] Условия, о которых идёт речь, не будем здесь формулировать. Их можно найти в специальной литературе (см., например, Гнеденко Б. В. Курс теории вероятностей: Учебник. 7-е изд., исправл. — М.: Эдиториал УРСС, 2001).

одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$, реализацией которого являются полученные в эксперименте числа $x_1, x_2, \ldots, x_n$.

Таким образом, определим ***выборку*** *как совокупность реально полученных значений $x_1, x_2, \ldots, x_n$, за которыми стоят независимые одинаково распределенные случайные величины* $\xi_1, \xi_2, \ldots, \xi_n$.

При проведении экспериментов или полевых наблюдений предположения, содержащиеся в определении выборки, могут нарушаться. Например, может появиться систематическое смещение результатов в зависимости от номера наблюдения. Поэтому рекомендуется рассмотреть график зависимости результата наблюдения от его номера (в порядке получения отдельных наблюдений), чтобы убедиться в отсутствии заметных временных трендов в самих наблюдениях, либо в размахе их колебаний (которых не должно быть в выборке). Также следует рассмотреть график накопленных наблюдений, выявляющий неоднородность выборки. Если же есть основания для выделения отдельных групп наблюдений (например, наблюдения, сделанные в различных лабораториях), то надо по мере возможности сравнивать результаты для этих групп.

5.2. Эмпирическая функция распределения

Существует удобный способ графического представления данных выборки $x_1, x_2, \ldots, x_n$, называемый ***эмпирической функцией распределения*** $\boldsymbol{F_n(x)}$. $F_n(x) = \frac{1}{n} \cdot \{\textit{число } x_i \le x\}$, т.е. это доля наблюдений, лежащих левее x, включая точку x.

Графически эмпирическая функция представляет собой ступенчатую функцию со скачками в точках $x_1, x_2, \ldots, x_n$. Величина скачка равна $\frac{1}{n}$, если в точке x_i нет совпадающих наблюдений; если же совпадающие наблюдения есть, то величина скачка равна $\frac{k}{n}$, где k – число совпадающих наблюдений в точке x_i.

При построении графика эмпирической функции распределения с помощью пакетов программ ***Excel*** или ***Statistica*** удобно изображать середины ступенек $i/n-1/2n=(2i-1)/2n$. Тогда построение графика сводится к следующему.

Упорядочим выборку в порядке возрастания: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ (последовательность, которая получается после такого переупорядочения, называется *вариационным рядом*). Точки с координатами $(x_{(i)}, (2i-1)/2n)$ изображают середины ступенек эмпирической функции распределения. Для наглядности их можно соединить.

Эмпирическая функция распределения дает наглядное изображение выборочных данных без потери информации (если мы действительно имеем дело с моделью выборки, где порядок x_i не важен). Менее совершенным (и устаревшим в компьютерную эпоху) способом графического представления

данных является так называемая *гистограмма.* Для построения гистограммы берут числовой интервал $[a,b]$, содержащий выборочные данные, и делят его на несколько равных или неравных интервалов с помощью точек $a=a_1<a_2<...<a_{k+1}=b$. Вычисляют числа n_i, равные количеству тех выборочных значений, которые попали в i -ый интервал. Частоты $h_i=n_i/n$ попадания выборочных значений в различные интервалы изображают графически с помощью прямоугольников с основаниями $[a_i,a_{i+1})$ и такими высотами, чтобы *площадь* прямоугольника равнялась (в некотором масштабе) частоте h_i. В этом случае наблюдения огрубляются в соответствии с шириной интервалов (так как рассматривается не точное значение каждого отдельного наблюдения, а лишь интервал, в который оно попало). Говорят, что при этом получаются *группированные данные.*

Если выдвигается гипотеза о том, что наблюдения образуют выборку из нормального распределения, то для проверки этой гипотезы применяется так называемый *нормальный масштаб.*

Поскольку теоретическая функция распределения нормального закона имеет вид $F(x) = \Phi\{(x-a)/\sigma\}$, где Φ - *функция Лапласа,* то, сделав обратное преобразование $\Phi^{-1}[F(x)]$, получим функцию $(x-a)/\sigma$, графиком которой является прямая линия. Но так как эмпирическая функция $F_n(x)$ более или менее близка к теоретической, то при таком же преобразовании значений эмпирической функции должна получиться ломаная линия более или менее близкая к прямой. Эта близость обычно оценивается глазомерно. При практическом построении эмпирической функции в нормальном масштабе следует изображать точки с координатами $(x_{(i)}, \Phi^{-1}([2i-1]/2n))$.

Теперь, когда нам известно, как графически можно проверить нормальность распределения, вернемся к центральной предельной теореме.

5.3. Оценка параметров

Задача оценивания параметров распределения – одна из основных задач математической статистики. На содержательном уровне задача оценивания параметров распределения формулируется так: располагая выборкой $x_1, x_2,..., x_n$ случайной величины ξ, необходимо получить оценку $\hat{a}$ неизвестного параметра a генеральной совокупности и ее статистические свойства. Различают два основных типа оценок: ***точечные оценки*** и ***интервальные оценки (***или ***доверительные интервалы)***.

5.3.1. Точечные оценки

Точечная оценка неизвестного параметра a - функция от выборочных значений (или, как говорят, *статистика*) $\hat{a} = \hat{a}(x_1,...,x_n)$, т.е. сама является случайной величиной, но для конкретной выборки $x_1, x_2,..., x_n$ она принимает числовое значение. Желательно, чтобы это значение было близко (в вероятностном смысле) к величине оцениваемого параметра.

Свойства оценок.

1.***Несмещенность.***

Статистика $\hat{a}(x_1,\ldots,x_n)$ называется ***несмещенной оценкой*** параметра a, если математическое ожидание оценки равняется оцениваемому параметру: $M\hat{a}(x_1,\ldots,x_n)=a$.

Если $|M\hat{a}(x_1,\ldots,x_n)-a| \xrightarrow[n\to\infty]{} 0$, (т.е. смещение стремится к нулю при увеличении размера выборки), то такую оценку называют ***асимптотически несмещенной.***

2.***Состоятельность.***

Статистика $\hat{a}(x_1,\ldots,x_n)$ называется ***состоятельной оценкой*** параметра a, если с ростом размера выборки оценка стремится по вероятности к оцениваемому параметру: $\lim\limits_{n\to\infty} P\{|\hat{a}(x_1,\ldots,x_n)-a|<\varepsilon\}=1$ при любом сколь угодно малом ε.

Примеры точечных оценок.

Оцениваемый параметр	**Статистика**	**Свойства**
Математическое ожидание	$\bar{x}=\sum_{i=1}^{n} x_i$	Несмещённость, состоятельность
Дисперсия	$\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2$	Асимптотическая несмещённость, состоятельность
Дисперсия	$\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2$	Несмещённость, состоятельность
Параметр p биномиального распределения	μ/n (μ - наблюденное число успехов)	Несмещённость, состоятельность

5.3.2. Интервальные оценки

Другим типом оценок статистических параметров являются доверительные интервалы.

Доверительный интервал для неизвестного параметра a - это интервал (a_1,a_2), построенный по выборке, который содержит (*накрывает*) истинное значение параметра a с вероятностью, не меньшей заданного значения γ (верхняя a_2 и нижняя a_1 границы этого интервала являются статистиками: $a_1=a_1(x_1, x_2,\ldots, x_n)$; $a_2= a_2(x_1, x_2,\ldots, x_n)$). Здесь γ называется доверительной вероятностью.

Доверительные интервалы используются, когда нам нужны надежные границы, в которые попадает значение оцениваемого параметра.

Примеры доверительных интервалов

Исходные предпосылки	**Оценка**	**Границы интервала (ε=1-γ; γ - доверительная вероятность)**	**Пояснения**
$x_1,x_2,\ldots x_n$ - выборка из нормального распределения $N(a,\sigma^2)$ с неизвестным мат.ожиданием a и известной дисперсией σ^2	математическое ожидание a	$\left(\overline{x}\pm\frac{\sigma}{\sqrt{n}}\cdot u_{1-\frac{\varepsilon}{2}}\right)$	u_p - квантиль порядка p стандартного нормального распределения
$x_1,x_2,\ldots x_n$ - выборка из нормального распределения $N(a,\sigma^2)$ с неизвестным мат.ожиданием a и неизвестной дисперсией σ^2	математическое ожидание a	$\left(\overline{x}\pm\frac{s}{\sqrt{n}}\cdot t_{n-1,1-\frac{\varepsilon}{2}}\right)$	t_p - квантиль порядка p распределения Стьюдента с (n-1) степенями свободы
$x_1,x_2,\ldots x_n$ - выборка из нормального распределения $N(a,\sigma^2)$ с неизвестным мат.ожиданием a и неизвестной дисперсией σ^2	дисперсия σ^2	$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\varepsilon}{2}}},\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\varepsilon}{2}}}\right)$	$\chi^2_{n-1,p}$ - квантиль порядка p распределения «хи-квадрат» с (n-1) степенями свободы
Рассматривается биномиальное распределение числа успехов μ в n испытаниях Бернулли. Частота успехов $w=\mu/n$	параметр p биномиального распределения	$w-\sqrt{\frac{w(1-w)}{n}}u_{\varepsilon/2}$, $w+\sqrt{\frac{w(1-w)}{n}}u_{1-\varepsilon/2}$,	u_p - квантиль порядка p стандартного нормального распределения. n-несколько десятков

5.4. Проверка статистических гипотез

5.4.1. Постановка задачи

Пусть в эксперименте доступна наблюдению случайная величина ξ (т.е. у нас есть выборка $x_1, x_2,\ldots, x_n$), ее функция распределения $F(x)$ неизвестна полностью или неизвестны ее параметры. Тогда любое утверждение, касающееся распределения ξ называется ***статистической гипотезой***.

В параграфе 1.2 мы уже рассматривали такую постановку задачи проверки гипотез (для частного случая проверки равенства p=1/2 биномиального распределения). Возможны и другие гипотезы: о равенстве параметров двух или нескольких распределений, о независимости выборок и проч. Обычно наряду с проверяемой гипотезой H_o (*нулевой*, или *основной*) формулируется и противоречащая ей гипотеза H_a (*конкурирующая*, или *альтернативная*), которая принимается, если отвергнута нулевая гипотеза.

Для проверки статистической гипотезы используется специально подобранная случайная величина g, зависящая от выборочных значений $g=g(x_1,x_2,\ldots,x_n)$ с известным законом распределения, называемая

статистическим критерием.[4] Множество возможных значений критерия разбивается на два непересекающихся подмножества: одно из них W_α (*критическая область*) содержит значения критерия, при которых нулевая гипотеза отклоняется, второе (*область принятия гипотезы*) — значения g, при которых она не отклоняется. Поскольку статистический критерий характеризует величину отклонения от нулевой гипотезы, то критическая область часто имеет вид: либо $g<k_p$ (*левосторонняя область)*, либо $g>k_p$ (*правосторонняя область)*, либо $|g|>k_p$ (*двусторонняя область)*. Значения k_p, отделяющие критическую область от области принятия гипотезы, называются *критическими точками*.

При проверке гипотезы могут быть допущены ошибки двух видов: *ошибка первого рода*, если отклонена верная нулевая гипотеза, и *ошибка второго рода*, если принята неверная нулевая гипотеза.

		Верная гипотеза	
		H_0	H_a
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка *второго* рода)
	H_a	H_0 неверно отвергнута (Ошибка *первого* рода)	H_0 верно отвергнута

Вероятность ошибки первого рода, называемая *уровнем значимости критерия,* обычно обозначается α; вероятность ошибки второго рода обычно обозначается β. Вероятность отвергнуть H_0, когда она неверна, называется мощностью критерия (правило проверки гипотезы называют *критерием*); мощность критерия равна $1-\beta$.

Ошибки первого и второго рода связаны между собой. При уменьшении вероятности ошибки первого рода для заданного критерия g и выбранной критической области увеличивается вероятность ошибки второго рода (уменьшается мощность критерия). Статистику критерия g и критическую область стараются выбрать так, чтобы при заданной α максимизировать мощность критерия.

Методика проверки гипотез состоит в следующем.

1. Формулируется ***нулевая гипотеза*** H_0 о распределении вероятностей $F(x)$. Гипотеза формулируется исходя из требований прикладной задачи. Чаще всего рассматриваются две гипотезы —

[4] Вопрос о том, какую функцию g надо взять для проверки той или иной гипотезы, часто не имеет однозначного ответа. Есть целый ряд требований, которым должна удовлетворять «хорошая» статистика g. Неформальное правило состоит в том, чтобы выбирать в качестве критерия величину, характеризующую степень отклонения от нулевой гипотезы.

основная H_0 и ***альтернативная*** H_a . Иногда альтернатива не формулируется в явном виде.[5] Задаётся *статистический критерий* $g(x_1,x_2,...x_n)$, для которого в условиях справедливости гипотезы H_0 выводится функция распределения $\Phi(g)$[6]. Фиксируется ***уровень значимости*** — допустимая для данной задачи вероятность ***ошибки первого рода***, то есть того, что гипотеза на самом деле верна, но будет отвергнута процедурой проверки. Это должно быть достаточно малое число $\alpha \in [0,1]$. На практике часто полагают α=0,01 или 0,05.

2. На множестве допустимых значений статистики g выделяется ***критическое множество*** W_α значений статистики g, такое, что $P\{g \in W_\alpha\} = \alpha$. [7]

3. Собственно ***статистический тест (статистический критерий)*** заключается в проверке условия: попадает ли $g(x_1,x_2,...x_n)$ в критическую область W_α.

Если $g(x_1,x_2,...x_n) \in W_\alpha$, то делается вывод: «данные противоречат нулевой гипотезе при уровне значимости α». Гипотеза H_0 отвергается.

Если $g(x_1,x_2,...x_n) \notin W_\alpha$, то делается вывод: «данные не противоречат нулевой гипотезе при уровне значимости α». Гипотеза H_0 не отвергается.

Итак, ***статистический критерий*** определяется статистикой g и критическим множеством W_α, которое зависит от уровня значимости α.

Альтернативная методика на основе *p*-значения

P-значение (*p-value*) — это наименьшая величина уровня значимости, при которой H_0 отвергается для данного значения статистики критерия g. В формульной записи: $p\text{-value}=\min\{\alpha: g \in W_\alpha\}$

Если достигаемый уровень значимости достаточно мал (близок к нулю), то нулевая гипотеза отвергается. В частности, его можно сравнивать с фиксированным уровнем значимости α; тогда альтернативная методика будет эквивалентна классической (при $p<\alpha$ имеем $g(x_1,x_2,...x_n) \in W_\alpha$ и гипотеза H_0 отвергается, при $p>\alpha$ имеем $g(x_1,x_2,...x_n) \notin W_\alpha$, и делается противоположный вывод).

Рассмотрим далее конкретные задачи проверки статистических гипотез.

---

[5] В зависимости от формулируемых предположений нулевая гипотеза может быть *простой* или *сложной*.

Простой гипотезой называют предположение, состоящее в том, что неизвестная функция распределения $F(x)$ отвечает некоторому совершенно конкретному вероятностному распределению (например, данные являются выборкой из равномерного распределения на отрезке [-1,1], математическое ожидание равно известной константе).

Сложной гипотезой называют предположение о том, что неизвестная функция распределения $F(x)$ принадлежит некоторому множеству распределений, состоящему из более чем одного элемента (например, $F(x)$- функция нормального распределения).

[6] Для многих задач вид функций $g(x_1,x_2,...x_n)$ и $\Phi(g)$ известен.

[7] Определение границы критического множества как функции от уровня значимости α является строгой математической задачей, которая в большинстве практических случаев имеет готовое простое решение.

5.4.2. Статистические критерии о параметрах

Одновыборочные критерии.

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α-уровень значимости)*	*Пояснения*
$x_1,x_2,\dots x_n$ выборка из нормального распределения $N(a,\sigma^2)$ с известной дисперсией σ^2	Математическое ожидание a равно a_0	$a\neq a_0$	$u=\dfrac{\bar{x}-a_0}{\sigma/\sqrt{n}}$	$\lvert u\rvert > u_{1-\alpha/2}$	u_p- p-квантиль $N(0,1)$
		$a>a_0$		$u>u_{1-\alpha}$	
		$a<a_0$		$u<u_{\alpha}$	
$x_1,x_2,\dots x_n$ выборка из нормального распределения с неизвестной дисперсией (одновыборочный t-критерий)	Математическое ожидание a равно a_0	$a\neq a_0$	$t=\dfrac{\bar{x}-a_0}{s/\sqrt{n}}$	$\lvert t\rvert > t_{1-\alpha/2}$	t_p- p-квантиль распределения Стьюдента с n-1 степенями свободы.
		$a>a_0$		$t>t_{1-\alpha}$	
		$a<a_0$		$t<t_{\alpha}$	
$x_1,x_2,\dots x_n$ - выборка из нормального распределения с неизвестными параметрами (одновыборочный χ^2–критерий)	Дисперсия σ^2 равна σ_0^2	$\sigma^2\neq\sigma_0^2$	$\chi^2=(n-1)\dfrac{s^2}{\sigma_0^2}$	$[0,\chi^2_{n-1,\,\alpha/2}]\cup[\chi^2_{n-1,1-\alpha/2},\infty]$	$\chi^2_{n-1,p}$- p-квантиль распределения χ^2 с n-1 степенями свободы.
		$\sigma^2>\sigma_0^2$		$\chi^2>\chi^2_{n-1,1-\alpha/2}$	
		$\sigma^2<\sigma_0^2$		$\chi^2<\chi^2_{n-1,\alpha/2}$	
k - число успехов в n испытаниях Бернулли	$p=p_0$	$p\neq p_0$	$u=\dfrac{w-np_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\lvert u\rvert > u_{1-\alpha/2}$	(приближенный критерий) u_p- p-квантиль $N(0,1)$
		$p>p_0$		$u>u_{1-\alpha}$	
		$p<p_0$		$u<u_{\alpha}$	

Двухвыборочные критерии.

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α-уровень значимости)*	*Пояснения*
$x_1,x_2,\dots,x_n$ и $y_1,y_2,\dots,y_m$ - выборки из двух независимых нормальных распределений $N(a_1,\sigma^2)$ и $N(a_2,\sigma^2)$ соответственно (с одинаковыми неизвестными дисперсиями σ^2) (Двухвыборочный t-критерий)	Математические ожидания равны, т.е. $a_1=a_2$	$a_1\neq a_2$	$t=\dfrac{\bar{x}-\bar{y}}{s}\cdot\sqrt{\dfrac{nm}{n+m}}$ $S^2=\dfrac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$	$\lvert t\rvert > t_{1-\alpha/2}$	t_p- p-квантиль распределения Стьюдента с (n+m-2) степенями свободы
		$a_1>a_2$		$t>t_{1-\alpha}$	
		$a_1<a_2$		$t<t_{\alpha}$	

$x_1,x_2,\ldots,x_n$ и $y_1,y_2,\ldots,y_m$ - выборки из двух независимых нормальных распределений $N(a_1,\sigma_1^2)$ и $N(a_2,\sigma_2^2)$ соответственно (с неравными неизвестными дисперсиями σ_1^2 и σ_2^2) (Двухвыборочный *t*-критерий)	Математические ожидания равны, т.е. $a_1=a_2$	$a_1 \neq a_2$	$t=\dfrac{\bar{x}-\bar{y}}{\sqrt{(\frac{S_x^2}{n}+\frac{S_y^2}{m})}}$ $S_x^2=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ $S_y^2=\dfrac{\sum_{i=1}^{m}(y_i-\bar{y})^2}{m-1}$	$\lvert t\rvert > t_{1-\alpha/2}$	t_p- p-квантиль распределения Стьюдента с l степенями свободы, где $l=[1/d]$, а $d=\dfrac{\left(\frac{\frac{S_x^2}{n}}{\frac{S_x^2}{n}+\frac{S_y^2}{m}}\right)^2}{n-1}+\dfrac{\left(\frac{\frac{S_y^2}{m}}{\frac{S_x^2}{n}+\frac{S_y^2}{m}}\right)^2}{m-1}$
		$a_1 > a_2$		$t>t_{1-\alpha}$	
		$a_1 < a_2$		$t<t_{\alpha}$	
$x_1,x_2,\ldots,x_n$ и $y_1,y_2,\ldots,y_n$ - связанные выборки из нормальных распределений $N(a_1,\sigma_1^2)$ и $N(a_2,\sigma_2^2)$ соответственно.[8]	Математические ожидания равны, т.е. $a_1=a_2$	$a_1 \neq a_2$	$t=\dfrac{(\bar{x}-\bar{y})\sqrt{n}}{\sqrt{S_x^2+S_y^2-2S_{xy}}}$ $S_{xy}=\dfrac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$	$\lvert t\rvert > t_{1-\alpha/2}$	t_p- p-квантиль распределения Стьюдента с $(n-1)$ степенями свободы
		$a_1 > a_2$		$t>t_{1-\alpha}$	
		$a_1 < a_2$		$t<t_{\alpha}$	
$x_1,x_2,\ldots,x_n$ и $y_1,y_2,\ldots,y_m$ - выборки из двух независимых нормальных распределений $N(a_1,\sigma_1^2)$ и $N(a_2,\sigma_2^2)$ (Двухвыборочный *F*-критерий)	Дисперсии равны, т.е. $\sigma_1^2=\sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F=\dfrac{S_x^2}{S_y^2}$	$(0,F_{n-1,m-1,\alpha/2}]$ $\cup[F_{n-1,m-1,1-\alpha/2},\infty)$	F_p – p- квантили F-распределения с $(n-1)$ и $(m-1)$ степенями свободы.
		$\sigma_1^2 > \sigma_2^2$		$F>F_{n-1,m-1,1-\alpha}$	
		$\sigma_1^2 < \sigma_2^2$	$S_x^2=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ $S_y^2=\dfrac{\sum_{i=1}^{m}(y_i-\bar{y})^2}{m-1}$	$0<F<F_{n-1,m-1,\alpha}$	
k и l- значения двух независимых биномиальных		$p_\xi \neq p_\eta$	$u=$	$\lvert u\rvert > u_{1-\alpha/2}$	u_p – p- квантиль стандартного
		$p_\xi > p_\eta$		$u>u_{1-\alpha}$	

[8] Т.е. элементы x_i, y_i соответствуют одному и тому же объекту, но измерения сделаны в разные моменты времени (например, измерение давления до и после приема лекарств). В этом случае процедура сравнения выборочных средних аналогична проверке равенства нулю среднего выборочного разности $\{x_i-y_i\}$.

случайных величин ξ и η с параметрами n, p_ξ и m, p_η соответственно.	$p_\xi = p_\eta$	$p_\xi < p_\eta$	$\frac{(\nu_\xi - \frac{1}{2n}) - (\nu_\eta + \frac{1}{2m})}{\sqrt{\nu(1-\nu)(\frac{1}{n}+\frac{1}{m})}}$ $\nu_\xi = k/n;\ \nu_\eta = l/m;$ $\nu = (k+l)/(n+m)$	$u < u_\alpha$	нормального распределения (критерий приближенный; его рекомендуется применять, когда наблюдений несколько десятков).
k и l- значения двух независимых пуассоновских случайных величин ξ и η с параметрами λ_ξ и λ_η соответственно.	$\lambda_\xi = \lambda_\eta$	$\lambda_\xi \neq \lambda_\eta$ $\lambda_\xi > \lambda_\eta$ $\lambda_\xi < \lambda_\eta$	$u = \frac{k-l}{\sqrt{k+l}}$	$\lvert u \rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_\alpha$	u_p – p- квантиль стандартного нормального распределения.
$x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$ - выборки из нормальных распределений $N(a_1, \sigma_1^2)$ и $N(a_2, \sigma_2^2)$ соответственно.	Коэффициент корреляции ρ равен числу ρ_0, т.е. $\rho = \rho_0$	$\rho \neq \rho_0$ $\rho > \rho_0$ $\rho < \rho_0$	$u = \frac{\sqrt{n-3}}{2} \cdot \left(\ln\frac{1+r}{1-r} - \ln\frac{1+\rho_0}{1-\rho_0} \right)$	$\lvert u \rvert > u_{1-\alpha/2}$ $u > u_{1-\alpha}$ $u < u_\alpha$	u_p – p- квантиль стандартного нормального распределения

5.4.3. Критерии согласия

Для проверки гипотезы о предполагаемом виде закона распределения случайной величины применяют критерии согласия. Основные критерии согласия – критерий хи-квадрат и критерий Колмогорова.

Критерий «хи-квадрат»

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α-уровень значимости)*	*Пояснения*
Простая нулевая гипотеза (распределение $F_0(x)$ задано полностью)					
$x_1, x_2, \ldots, x_n$ - выборка значений случайной величины ξ с неизвестной функцией распределения $F(x)$, n>50.	Случайная величина ξ имеет заданное распределение $F(x) = F_0(x)$	$F(x) \neq F_0(x)$	$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$ Подробности вычислений см. ниже	$\chi^2 > \chi^2_{k-1,1-\alpha}$	$\chi^2_{k-1,p}$ – p-квантиль распределения хи-квадрат с k-1 степенями свободы

Сложная нулевая гипотеза (распределение $F_0(x)$ задано с точностью до параметров)					
$x_1,x_2,\ldots,x_n$ - выборка значений случайной величины ξ с неизвестной функцией распределения $F(x)$, n>50.	Случайная величина ξ имеет заданное распределение $F(x) = F_0(x,\theta_1,\ldots,\theta_r)$, где $\theta_1,\ldots,\theta_r$ – *неизвестные* параметры	$F(x) \neq F_0(x)$	$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$ $\hat{p}_i$ -оценка вероятности p_i	$\chi^2 > \chi^2_{k-r-1,1-\alpha}$	$\chi^2_{k-r-1,p}$ -p-квантиль распределения хи-квадрат с k-r-1 степенями свободы

Для вычисления статистики χ^2 необходимо область значений выборки $[x_{\min}, x_{\max}]$ разбить на k интервалов так, чтобы число наблюдений, попавших в i-тый интервал (n_i) было не меньше 5-ти. При этом p_i – вероятность попадания в i-тый интервал случайной величины с функцией распределения $F_0(x)$. В случае, когда параметры $\theta_1,\ldots,\theta_r$ распределения $F_0(x,\theta_1,\ldots,\theta_r)$ неизвестны, их оценивают по выборке $\hat{\theta}_1,\ldots,\hat{\theta}_r$, а величины $\hat{p}_i$ вычисляют с помощью $F_0(x,\hat{\theta}_1,\ldots,\hat{\theta}_r)$.

Критерий Колмогорова

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α- уровень значимости)*	*Пояснения*
$x_1,x_2,\ldots,x_n$ - выборка значений случайной величины ξ с неизвестной функцией распределения $F(x)$.	Случайная величина ξ имеет заданное распределение $F(x) = F_0(x,\theta_1,\ldots,\theta_r)$, где $\theta_1,\ldots,\theta_r$ – *известные*	$F(x) \neq F_0(x)$	$\lambda = \sqrt{n} \cdot D_n$, где $D_n = \sup_{-\infty<x<\infty} \lvert F_0(x) - F_n(x)\rvert$ ($F_n(x)$ - эмпирическая функция распределения).	$\lambda > \lambda_{крит}$ или при малых n $D_n > D_{крит}$	При n>20 и верной H_0 λ имеет функцию распределения Колмогорова $K(x)$. $\lambda_{крит}$ - (1-α) квантиль распределения Колмогорова.
	Случайная величина ξ имеет распределение $F(x)=N(\mu,\sigma^2)$ с неизвестными μ и σ (параметры оценивают по выборке)	$F(x) \neq N(\mu,\sigma^2)$	$\tilde{\lambda} = D_n(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}})$	$\tilde{\lambda} > \tilde{\lambda}_{крит}$	Критические точки приведены ниже.

Функция распределения Колмогорова $K(x)$ имеет вид:

$$K(x) = 1 + 2 \cdot \sum_{i=1}^{\infty} (-1)^i \exp(-2k^2x^2)$$

При малых n для нахождения критических точек (т.е. границ критических областей) для величины D_n можно воспользоваться таблицами (например, Л.Н.Большев, Н.В.Смирнов Таблицы математической статистики М. Наука. 1983). Приведем здесь фрагмент таблицы (α - уровень значимости)

n \ α	0,2	0,1	0,05	0,02	0,01
4	0,49	0,56	0,62	0,69	0,73
6	0,41	0,47	0,52	0,58	0,62
8	0,36	0,41	0,45	0,51	0,54
10	0,32	0,37	0,41	0,46	0,49
12	0,30	0,34	0,37	0,42	0,45
14	0,27	0,31	0,35	0,39	0,42
16	0,26	0,29	0,33	0,36	0,39
18	0,24	0,28	0,31	0,34	0,37
20	0,23	0,26	0,29	0,33	0,35

Критические точки (границы критической области) для статистики $\tilde{\lambda}$ найдены Х.Лилиефорсом и приводятся в таблицах, например, в книге Ю.Н.Тюрина «Непараметрические методы статистики» М. Знание. 1978. Приведем здесь отрывок из таблицы процентных точек $\tilde{\lambda}_{крит}$ для проверки нормальности распределения при неизвестных параметрах:

Статистика $\tilde{\lambda}$	α (уровень значимости)				
	0.15	*0.10*	*0.05*	*0.025*	*0.01*
$D_n(\sqrt{n}-0.01+\frac{0.85}{\sqrt{n}})$	0.775	0.819	0.895	0.955	1.035

5.4.4. Непараметрические методы статистики

Наряду с рассмотренными ранее, существуют методы математической статистики, не предполагающие знания функционального вида генеральных распределений. Название "непараметрические методы" подчеркивает их отличие от классических - параметрических - методов, в которых предполагается, что генеральное распределение известно с точностью до конечного числа параметров, и которые позволяют по результатам наблюдений оценивать неизвестные значения этих параметров и проверять гипотезы относительно их значений. В частности, непараметрическим является приведенный выше критерий Колмогорова, но имеется и ряд других. Приведем здесь наиболее часто используемые: критерий знаков; критерий знаковых рангов; критерий Вилкоксона-Манна-Уитни; критерий Ван-дер-Вардена.

Критерий знаков.

Исходные предпосылки	H_0	H_a	*Критерий*	*Пояснения*
$x_1,x_2,...x_n$ выборка из неизвестного распределения $F(x,Me)$ с неизвестной медианой.	Медиана *Me* равна заданному числу *m*. $Me=m$	$Me\neq m$ $Me>m$ $Me< m$	Статистика критерия *S*=число положительных разностей (x_i-m)	Статистика *S* – число успехов в испытаниях Бернулли. Проверка гипотез сводится к проверке гипотез о вероятности *p* в испытаниях Бернулли

Критерий знаковых рангов.

Ранг R_i – порядковый номер i-го наблюдения в вариационном ряду. При совпадении наблюдений можно использовать: метод случайного ранга, когда номера совпадающих рангов выбираются по жребию; метод среднего ранга, когда каждому из совпадающих наблюдений приписывается ранг, равный среднему арифметическому их порядковых номеров; а можно просто отбросить совпадающие наблюдения.

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α- уровень значимости)*
$x_1,x_2,\dots x_n$ выборка из неизвестного непрерывного распределения $F(x,Me)$ с неизвестной медианой.	Медиана Me равна заданному числу m. $Me=m$	$Me\neq m$	Статистика критерия $T^+=\sum_{(x_i-m)>0} R_i\lvert x_i-m\rvert$	$[0,T^+_{\frac{\alpha}{2}}]\cup[T^+_{1-\frac{\alpha}{2}},n(n+1)/2]$
		$Me> m$		$[T^+_{1-\alpha},n(n+1)/2]$
		$Me< m$		$[0,T^+_{\alpha}]$

При малых n критические точки для статистики T^+ можно найти в справочниках (см., например, М.Холлендер, Д.А.Вульф. «Непараметрические методы статистики.» М: Финансы и статистика. 1983). T^+_p при $n\to\infty$ ($n>25$) является p-квантилью нормального распределения с параметрами $\left(\frac{n(n+1)}{4},\frac{n(n+1)(2n+1)}{24}\right)$.

Критерий Вилкоксона-Манна-Уитни

Постановка задачи для критерия Манна-Уитни совпадает с постановкой задачи для критерия Вилкоксона. Статистика U Манна-Уитни определяется как число пар (x_i, y_j) таких, что $x_i < y_j$, среди всех mn пар, в которых первый элемент - из первой выборки, а второй - из второй. Можно показать, что $U= mn+m(m+1)/2- W$. Поскольку W и U линейно связаны, то часто говорят не о двух критериях - Вилкоксона и Манна-Уитни, а об одном - критерии Вилкоксона (Манна-Уитни).

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α- уровень значимости)*
$x_1,x_2,\dots,x_m$ и $y_1,y_2,\dots,y_n$ - выборки значений двух независимых случайных величин с неизвестными функциями распределения $F(x)$ и $F(x-\Delta)$.	$\Delta=0$	$\Delta\neq 0$	Статистика критерия W. О вычислении W см. ниже.	$W< W_{\alpha/2}$ и $W> W_{1-\alpha/2}$
		$\Delta> 0$		$W> W_{1-\alpha}$
		$\Delta< 0$		$W< W_{\alpha}$

Статистика W двухвыборочного критерия Вилкоксона определяется следующим образом. Все элементы объединенной выборки $x_1,x_2,\dots,x_m$ $y_1,y_2,\dots,y_n$ упорядочиваются в порядке возрастания. Элементы первой

выборки $x_1,x_2,\ldots,x_m$ имеют в общем вариационном ряду ранги $R_1, R_2, \ldots, R_m$. Тогда статистика Вилкоксона - это сумма рангов элементов первой выборки $W = R_1 + R_2 + \ldots + R_m$.

Для $n \geq m \geq 5$ и верной H_0 распределение W имеет вид:

$P\{W \leq \omega\} \approx \Phi(x)+\varphi(x)\cdot(x^3-3x)\cdot (m^2+n^2+mn+m+n)/[20mn\cdot(m+n+1)]$, где $\Phi(x)$ и $\varphi(x)$- функция и плотность нормального распределения;

$x=(\omega-\mathbf{M}W+0.5)/ (\mathbf{D}W)1/2$; $\mathbf{M}W=n(m+n+1)/2$; $\mathbf{D}W=mn(m+n+1)/12$.

При малых n критические точки для статистики W можно найти в справочнике (см., например, Л.Н.Большев, Н.В.Смирнов. «Таблицы математической статистики.» М.Наука.1983).

Критерий Ван-дер-Вардена

Исходные предпосылки	H_0	H_a	*Критерий*	*Критическая область (α- уровень значимости)*
$x_1,x_2,\ldots,x_n$ и $y_1,y_2,\ldots,y_m$ - выборки значений двух независимых случайных величин с неизвестными функциями распределения $F(x)$ и $F(x-\Delta)$.	$\Delta=0$	$\Delta \neq 0$	$X=\sum_{i=1}^{n} u\left(\frac{R_i}{n+m+1}\right)$, где u$(p)$ - p-квантиль стандартного нормального распределения, R_i - i-ый ранг	$X<X_{\alpha/2}$ и $X>X_{1-\alpha/2}$
		$\Delta > 0$		$X>X_{1-\alpha}$
		$\Delta < 0$		$X<X_{\alpha}$

Для нахождения критической области используются значения X_p - p-квантили распределения Ван-дер-Вардена.

При малых n критические точки для статистики X можно найти в справочнике (см., например, Л.Н.Большев, Н.В.Смирнов «Таблицы математической статистики» М.Наука.1983).

При $n+m \geq 20$ распределение статистики Ван дер Вардена асимптотически нормально с нулевым математическим ожиданием и дисперсией

$$DX = \frac{mn}{(m+n)(m+n+1)} \cdot \sum_{i=1}^{n+m} u^2\left(\frac{i}{m+n+1}\right)$$

5.5. Регрессионный анализ

Все рассмотренные ранее методы исследования относятся к одномерным случайным величинам. В реальных исследованиях чаще всего имеют дело с *многомерными данными*. Многочисленные математико-статистические методы, предназначенные для обработки таких данных, исходят из предположения статистической однородности и нормального распределения случайных величин, входящих в модель.

Одним из самых распространенных методов анализа многомерных данных является регрессионный анализ. Он рассматривает задачи, связанные с построением функциональных зависимостей между различными количественными переменными. Например, изучение зависимости массы тела животного от его возраста; первичной продукции фитопланктона от концентрации биогенных элементов; урожайности от количества внесенных удобрений и др.

Формально задача ставится так. Имеется набор из $k+1$ переменной $Y, X_1,\ldots,X_k$ и предполагается, что они связаны равенством $Y=f(X_1,\ldots,X_k, \beta_0,\beta_1,\ldots,\beta_k)+\varepsilon$, где вид функции f известен с точностью до параметров $\beta_0,\beta_1,\ldots, \beta_k$, а ε - случайная ошибка, характеризующая изменчивость переменной Y, а также влияние факторов, не учтенных в модели. Переменная Y называется зависимой переменной (или объясняемой переменной, или откликом); переменные $X_1,\ldots,X_k$ - независимыми переменными (или объясняющими переменными, или факторами, или предикторами). Целью регрессионного анализа является оценка неизвестных параметров по наблюденным в ряде экспериментов значениями $Y,X_1,\ldots,X_k$.

Обычно задача разбивается на три этапа:

- выбор модели регрессии, что заключает в себе предположения о виде зависимости Y от $X_1,\ldots,X_k$ и $\beta_0,\beta_1,\ldots, \beta_k$;
- оценка параметров $\beta_0,\beta_1,\ldots, \beta_k$ в выбранной модели;
- проверка статистических гипотез о регрессии.

Если $f(X_1,\ldots,X_k, \beta_0,\beta_1,\ldots, \beta_k)$ линейна по неизвестным параметрам, то такая модель называется *линейной моделью регрессионного анализа.* Она имеет вид:

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_kX_k+\varepsilon \text{ [9]}. \qquad (5.5.1)$$

В случае $k=1$ мы говорим о простой регрессии, при $k>1$ о множественной регрессии.

Рассмотрим подробнее линейную модель, к которой обычно стараются свести дело в приложениях.

В результате эксперимента имеются матрица наблюдений вида

$$\begin{pmatrix} y_1 & x_{11} & x_{21} & \ldots & x_{k1} \\ y_2 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ y_n & x_{1n} & x_{2n} & \ldots & x_{kn} \end{pmatrix},$$

где i-ая строка соответствует i-му наблюдению, а модель имеет вид $y_i=\beta_0+\beta_1(x_{1i})+\ldots+\beta_k(x_{ki})+\varepsilon_i$. Случайные ошибки ε_i реально не наблюдаются, но предполагается, что это - независимые, нормально распределенные

[9] В общем случае можно было бы записать модель в виде $Y=\beta_0+\beta_1g_1(X_1,\ldots,X_k)+\ldots+\beta_k g_k(X_1,\ldots,X_k)+\varepsilon$, тогда, перейдя к новым переменным $\widetilde{X}_i=g_i(X_1,\ldots,X_k)$, получим ту же модель (2.5.1). В частности, $g_1,\ldots g_k$ могут быть степенями одной переменной и тогда Y приближается полиномом.

случайные величины с одинаковой неизвестной дисперсией σ^2. Необходимо оценить неизвестные параметры $\beta_0,\beta_1,\ldots,\beta_k$.

Значения $b_0,b_1,\ldots,b_k$ - оценки параметров $\beta_0,\beta_1,\ldots,\beta_k$ - получают минимизируя сумму $S=\sum_{i=1}^{n}(y_i - b_0 - b_1x_{1i} - \ldots - b_kx_{ki})^2$ (метод наименьших квадратов - МНК).[10]

Оценка функции регрессии получается по формуле:

$$\hat{Y} = b_0+b_1X_1+\ldots+b_kX_k, \qquad (5.5.2)$$

а для i-го наблюдения $\hat{y}_i = b_0 + b_1x_{1i} + \ldots + b_kx_{ki}$.

Оценки $b_0,b_1,\ldots,b_k$, получаемые МНК, являются линейными функциями от наблюдений $y_1,\ldots,y_n$.

Нахождение оценок $b_0,b_1,\ldots,b_k$ реализовано во многих статистических пакетах, в частности в пакете ***Statistica*** и в ***Excel.***

Анализ полученных результатов.

Единственным сколько-нибудь надежным методом оценки работоспособности результатов, получаемых с помощью регрессионного анализа, является проверка их на новых данных (т.е. на данных аналогичных наблюдений, но не использованных в процессе оценки коэффициентов регрессии). Следует предостеречь от распространенной ошибки, когда имеющийся массив данных делят на две (или несколько) частей *путем случайной выборки,* а затем оценивают коэффициенты на одной части, а проверяют результат на остальных. Такая проверка ни в чем не убеждает по следующей причине.

В приложениях вероятностных методов более всего надо опасаться нарушения статистической однородности, т.е. постоянства во времени вероятностных свойств случайных величин, входящих в модель. Но при проверке с помощью случайного отбора мы таких нарушений не заметим, поскольку при случайном отборе статистические свойства получаемых частей экспериментального материала автоматически будут одинаковы. Например, если мы рассмотрим простейшую ситуацию, когда во всех наблюдениях определялась одна и та же константа, но появилась ошибка в виде тренда во времени, то мы заметим это, если просто нарисуем график наблюдений в зависимости от времени. Но мы ничего не заметим, если возьмем две случайные выборки из наших наблюдений. Деление на части нужно производить по содержательным признакам (например, по времени или месту получения наблюдений). Следует имитировать ту цель, для которой вычислялась регрессия. Например, если это возможно более точное

[10] Можно отказаться от равенства дисперсий во всех точках наблюдений, если известно, что дисперсии ошибок $\sigma^2_i = \sigma^2/w_i$, где «веса» w_i известны. Если все наблюдения умножить на $\sqrt{w_i}$, то мы придем к задаче с равными дисперсиями. В этом случае минимизируется сумма:
$S=\sum_{i=1}^{n} w_i(y_i - b_0 - b_1x_{1i} - \ldots - b_kx_{ki})^2$. Часто отказываются и от нормальности распределения, оставляя предположение о статистической независимости ошибок ε_i.

определение значений объясняемой переменной с помощью значений объясняющих переменных, то нужно проверить, какая точность оказалась достигнутой для новых данных.

Однако применяются и некоторые формальные приемы, которые позволяют составить примерное представление о качестве полученной регрессионной модели без привлечения новых данных.

В различных пакетах статистических программ, кроме непосредственно оценок $b_0, b_1, \ldots, b_k$ и их стандартных отклонений,[11] рассчитываются величины, необходимые для указанной цели, а именно: $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$, где T в нижнем индексе означает *TOTAL*, $\bar{y}$ - выборочное среднее. Эта сумма представляется в виде

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{или} \quad SS_T = SS_{res} + SS_{Reg}.$$

SS_T характеризует разброс данных, SS_{res} - разброс данных около предсказанных регрессией значений. Отношение $R^2 = SS_{Reg}/SS_T$ есть доля вариации Y, объясненная регрессией (5.5.2). Это отношение называется *коэффициентом детерминации*, оно характеризует качество приближения $\hat{Y}$ к зависимой переменной Y. R^2 принимает значения между 0 и 1.[12] Чем ближе эта величина к 1, тем лучшей считается полученная регрессионная формула. Величина $SS_{res}/(n\text{-}k\text{-}1)$ дает несмещенную оценку дисперсии σ^2, а величину $\sqrt{SS_{res}/(n - k - 1)}$ называют стандартной ошибкой.

Кроме перечисленных выше величин обычно рассчитывается значение статистики F (статистика F-критерия Фишера) и ее p-значение для проверки значимости отклонения выбранной модели регрессии от модели $Y = const + \varepsilon$.[13]

---

[11] Зная величину оценки b_i и ее стандартное отклонение s_{bi}, можно построить доверительный интервал для β_i: $[b_i - s_{bi} \cdot t_{1-\alpha/2},\ b_i + s_{bi} \cdot t_{1-\alpha/2}]$, где $t_{1-\alpha/2}$ –квантиль t-распределения с n-k-1 степенями свободы, $\alpha = 1-\gamma$, γ-доверительная вероятность. Также с помощью t-критерия можно проверить равенство этого коэффициента заданному числу, в частности обычно проверяется равенство β_i нулю, или, как говорят, значимость соответствующего коэффициента.

[12] Значение коэффициента детерминации R^2, возрастает с ростом числа переменных в регрессии, что не означает улучшения качества предсказания. Поэтому для оценки качества подгонки регрессионной модели к наблюдаемым значениям y_i вводится скорректированный (*adjusted*) коэффициент детерминации

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

[13] Критерий Фишера применяется и для более широкой задачи подбора модели, а именно, для проверки гипотезы H_0 о том, что q коэффициентов регрессии равны 0, т.е. H_0: $\beta_k = \beta_{k-1} = \beta_{k-q+1} = 0$ *(редуцированная модель)*; H_a: все коэффициенты отличны от 0. Обозначим остаточные суммы SS_{res} для полной и SS_{resH0} для редуцированной модели. Тогда статистика $F = \dfrac{(SS_{resH0} - SS_{res})/q}{SS_{res}/(n-k-1)}$ при верной H_0 имеет F-распределение с q и n-k-1 степенями свободы. Этим критерием пользуются также в пошаговом регрессионном анализе, когда последовательно добавляют или удаляют переменные в модели.

Все эти показатели, характеризующие качество описания выбранной моделью имеющихся данных, скорее позволяют отвергнуть совсем неудачную модель, чем подтвердить правильность выбора функциональной зависимости. Более обоснованное решение можно получить, сравнивая наблюдаемые значения y_i с оценками $\hat{y}_i$.

Анализ остатков $\hat{\varepsilon}_i = y_i - \hat{y}_i$ дает некоторое представление как о качестве приближения, так и о выполнении предпосылок регрессионного анализа. Непосредственно ошибки $\varepsilon_i = y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki}$ не наблюдаются, поэтому рассматриваются «кажущиеся» ошибки или остатки – разности между наблюдаемыми значениями y_i и оценками $\hat{y}_i$, т.е. $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Эти величины должны вести себя как независимые нормально распределенные случайные величины. В пакетах статистического анализа обычно реализуется построение графика зависимости $\hat{\varepsilon}_i$ от номера наблюдения. Такой график позволяет показать: наличие зависимости, не учтенной в модели; неравенство дисперсий наблюдений; статистическую зависимость наблюдений в соседних точках и др.

Для проверки нормальности обычно используется график эмпирической функции распределения в нормальном масштабе, а также критерии согласия.