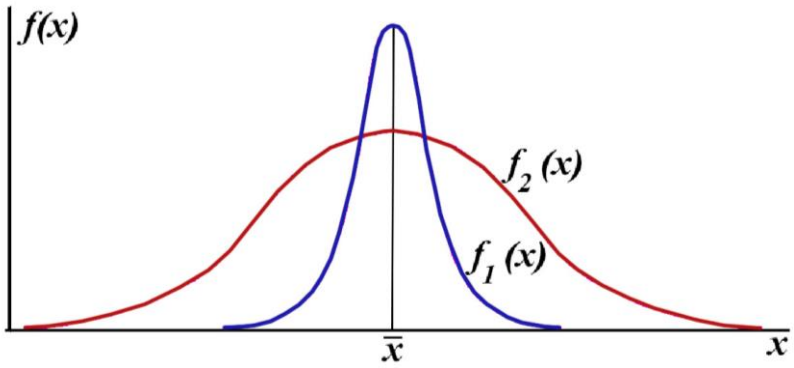
О.А. Подрезов

Методы
статистической обработки
и анализа
гидрометеорологических наблюдений



Бишкек ◊ 2020

КЫРГЫЗСКО-РОССИЙСКИЙ СЛАВЯНСКИЙ УНИВЕРСИТЕТ
ЕСТЕСТВЕННО-ТЕХНИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра метеорологии, экологии и охраны окружающей среды

О.А. Подрезов

Методы статистической обработки и анализа гидрометеорологических наблюдений

(Методы анализа с использованием статистик, аппроксимации распределений, регрессии, корреляции и проверки гипотез)

(учебник для бакалавров - гидрометеорологов)

Бишкек ◊ 2020

УДК
ББК
П 44

Рецензенты:

Рекомендовано к печати Ученым Советом ГОУВПО КРСУ

Подрезов О.А.

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ И АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКИХ НАБЛЮДЕНИЙ (Методы анализа с использованием статистик, аппроксимации распределений, регрессии, корреляции и проверки гипотез). Учебник для бакалавров-гидрометеорологов/Кыргызско-Российский Славянский университет. – Бишкек: Издательство КРСУ, 2020. – 100 с.

В 5 главах учебника по дисциплине «Методы статистической обработки и анализа гидрометеорологических наблюдений», включенной в федеральный компонент учебного плана направления – гидрометеорология, рассматриваются пять основных тем:

- анализ эмпирических распределений на основе использования их статистик;
- нормальный закон, t-распределение Стьюдента, χ^2-Пирсона и их некоторые практические использования;
- некоторые теоретические законы распределений, применяемые в метеорологии для аппроксимации одномерных случайных величин
- эмпирические связи и зависимости случайных величин;
- проверка статистических гипотез.

Учебник содержат классические и современные методы анализа метеорологических наблюдений, изложенные автором с учетом 25-летнего опыта преподавания дисциплины и многочисленных расчетных примеров, полученных преимущественно на основе собственных климатических исследований, с доведением каждого решения до «числа», т.е. получения количественных оценок и выводов гидрометеорологического характера.

Рекомендуется студентам естественно-технического факультета КРСУ – гидрометеорологам, энергетикам и физикам, а также студентам других специальностей наук о Земле.

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Статистические методы являются важнейшим инструментом анализа явлений природы и человеческой деятельности. Эти методы широко используются при исследовании гидрометеорологических процессов, поскольку они являются вероятностными по своей сути и во многих аспектах наиболее полно могут быть изучены и описаны только с позиций статистического подхода.

В соответствии с учебным планом подготовки бакалавров - гидрометеорологов, предусмотрено чтение дисциплины «Методы статистической обработки и анализа гидрометеорологических наблюдений». Подобные по направленности курсы читались автором на протяжении последних 25 лет студентам специальности метеорология сначала Кыргызского государственного университета, а затем Кыргызско-Россйиского Славянского университета. Это позволило не только накопить необходимый опыт чтения дисциплины, но и выработать наиболее целесообразный подход к методологии изложения и составу материала. Автор отказался от доказательной стороны изложения, полагая, что это задачи дисциплины теории вероятностей и математической статистики, которые читаются в составе курса высшей математики. В то же время основной упор сделан на *понятийность* как предметной идеологии, так и материала курса, что позволило достигнуть:

- максимальной простоты математического аппарата, необходимой описательной полноты приводимых рабочих формул с обоснованием возможностей их применения на практике;
- доведения использования всех методов «до числа», т.е. с технической демонстрацией их использования на конкретных примерах и выводах, получаемых из статистического анализа;

В содержание книги вошли следующие 5 тем:

1. Анализ эмпирических распределений на основе использования их статистик.
2. Нормальный закон, t-распределение Стьюдента, χ^2-Пирсона
и их некоторые практические использования.
3. Некоторые теоретические законы распределений, применяемые в метеорологии для аппроксимации одномерных случайных величин.
4. Эмпирические связи и зависимости случайных величин.
5. Проверка статистических гипотез.

Разумеется, этими темами далеко не исчерпывается содержание классических статистических методов. Однако автор сознательно пошел на выделение именно этих тем, по-

лагая, что глубокие знания всегда есть результат постоянного самообразования исследователя, а цель настоящего учебного пособия – быть только «запалом» этого процесса.

Из этих же соображений, численные решения ориентированы, прежде всего, на общедоступную программу Excel. Более сложные и узкоспециальные программы могут быть без труда освоены самостоятельно, если будет достигнута главная цель, поставленная при написании настоящих конспектов лекций.

Настоящий учебник базируется на аналогичом курсе 2009 г., изданном в более широком объеме для специальности, метеорология. Учебник предназначен, прежде всего, для бакалавров-гидро-метеорологов, однако автор надеется, что он будет также полезен географам, экологам и студентам других специальностей наук о Земле, где велика потребность применения статистических методов с использованием математического аппарата разумной сложности применительно к этим направлениям и специальностям.

О.А.Подрезов.

Бишкек, 2020 г.

Вводная глава:

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1. Специфика статистических законов и методов

По характеру проявления причинных связей все законы природы делятся на два класса: детерминированные и статистические. Классический пример детерминированных законов – небесная механика. Зная положение планет Солнечной системы, можно практически с большой точностью рассчитать их положение, как в прошлом, так и в далеком будущем и предсказать, например, время и место солнечных и лунных затмений.

Однако предсказание погоды в пределах даже нескольких суток, а тем более предсказание изменений климата в пределах всего одного столетия, несмотря на большие усилия науки в этом направлении, с такой же степенью успешности просто невозможны. Причина в том, что на главные закономерные или детерминированные зависимости в процессах развития погоды и климата накладываются многие случайные составляющие, имеющие существенное влияние, но трудно поддающиеся учету. Их роль может быть столь велика, что они часто полностью изменяют течение процесса на какой-либо стадии и делают его конечный результат трудно предсказуемым или, по крайней мере, всегда придают его исходу элемент неопределенности. С теоретической точки зрения, какие факторы считать основными или закономерными, а какие второстепенными или случайными *принципиально безразлично*. Можно поставить задачу, повышая неограниченно точность решения, включать в нее все новые и новые факторы, от самых существенных, до самых ничтожных. Однако такой подход привел бы к непомерной сложности и практической невозможности получения решения. Например, любая горная страна имеют сложную орографию с большим хитросплетением передовых и внутренних хребтов с их различной высотой, ориентацией и расчлененностью, а также межгорными впадинами и долинами разного размера, формы и высоты. Все это создает очень сложные местные климатические условия внутри горных стран, когда на главную причину изменения местных климатов – высотный фактор – накладываются множественные влияния орографии. Но количественно описать связи орография–климат (кроме высоты) очень трудно. Поэтому условно многие из них приходится считать не учитываемыми в отдельности или случайными, именно за счет их многообразия и множественности влияния.

Поэтому должна существовать *принципиальная разница* в методах учета *основных факторов*, влияющих в главных чертах на течение метеорологических и иных случайных процессов, и *второстепенных факторов*, которым условно отводится роль *возмущений или погрешностей*. Элемент неопределенности, сложности, многопричинности, присущий случайным явлениям и процессам, требует создания специальных методов для их изучения. Эти методы базируются на теории вероятностей и представлены ее прикладной частью – математической статистикой.

Таким образом, математическая статистика изучает и описывает массовые явления природы и общества, имеющие случайный характер, с одновременным действием множества причин их обуславливающих, когда результаты каждой отдельной реализации процесса точно предсказать невозможно, но можно предсказать область всего спектра реализаций с определенной долей вероятности. Математические законы и методы, лежащие в основе такого подхода, называются *статистическими*. Цель их применения состоит в том, чтобы, минуя слишком сложное и практически невозможное изучение отдельного явления, встать на путь описания массовых случайных явлений, выделив в них основные тенденции развития и случайные возмущения, что позволит осуществить объективный научный прогноз.

Например, согласно этим законам, будущее состояние климатической системы, т.е. погода и климат определяются неоднозначно, а лишь с некоторой долей вероятности, являющейся объективной мерой возможности реализации выявленных в прошлом тенденций изменения системы. Именно такое решение этих задач является наиболее целесообразным на сегодняшний день.

В настоящее время теория вероятностей и математическая статистика очень глубокие и чрезвычайно широкие по своим приложениям области науки. Отдельные их части постоянно отпочковываются в самостоятельные научные направления. Очень широко и плодотворно математическая статистика используется в метеорологии и других науках о Земле, поэтому включение в учебные планы *специальности метеорология* дисциплин, изучающих применение статистических методов, стало обязательным.

2. Понятие случайного события и его вероятности. Непосредственный подсчет вероятностей

Для изучения явлений природы и общества производят *опыты или наблюдения*. Опыты и наблюдения могут быть поставлены самой природой или обществом, а также человеком. Например, Вы бросаете монету, чтобы определить, как часто будет наблюдаться

выпадение герба или решки. Это простейший опыт, организованный человеком. Напротив, можно наблюдать сколько раз за год в Бишкеке будет отмечаться туман. Хотя наблюдателем является человек, опыт «воспроизводит» природа. Интуитивно ясно, для того чтобы результаты опытов можно было объективно сравнивать между собой, необходимо обеспечить, насколько это возможно, одинаковость комплекса условий, в которых осуществляется опыт. Например, наблюдение за туманом надо проводить в одном и том же месте (число дней с туманом в году будет различно в нижней северной и верхней южной частях г. Бишкек). Кроме того, надо использовать один и тот же критерий тумана и метод его оценки (видимость менее 1000 м). Наблюдения должны лежать внутри интервала в 30–40 лет, согласно современных понятий о климате, т.к. в противном случае можно попасть при одном опыте в одни климатические условия, а при другом – в уже заметно изменившиеся условия.

Результатом каждого опыта, наблюдения или испытания является *событие*. Оно будет носить случайный характер, т.к. может произойти (наблюдение тумана в заданные сутки) или не произойти. В последнем случае произойдет событие *противоположное* первому, т.е. отсутствие тумана в эти сутки. Обычно используют обозначения: A – появление события в опыте (день с туманом); $B = \overline{A}$ – появление противоположного события в опыте (день без тумана). При этом одно из них (А или $\overline{A}$) обязательно произойдет.

Чтобы количественно сравнивать между собой события по степени их возможности вводится *специальная численная мера – вероятность события p*. Событиям, которые произойдут всегда – *достоверные события* – приписывается вероятность p, равная 1. Противоположным, т.е. *невозможным* событиям приписывается вероятность p=0. Вероятность всех собственно случайных событий заключена между 0 и 1, чем она больше, тем больше шансов на то, что событие при опыте произойдет.

В некоторых случаях (азартные игры в кости, карты, лотерея и др.) вероятность события $p(A)$ можно подсчитать непосредственно используя очевидную формулу:

$$p(A) = \frac{m}{n}, \qquad (1)$$

где m – число случаев, благоприятных появлению события A, n – общее число случаев (опытов).

Например, вероятность $p(A)$ того, что из колоды в 36 карт наугад будет вытянут туз (событие A) равна $p(A)$=4/36=1/9. Это пример простого искусственно организованного опыта, когда вероятность вытянуть любую карту равна 1/36. Про такие опыты говорят, что они сводятся к «схеме случаев» или «схеме урн». В другом нашем примере с туманом

m будет равно числу суток в году, когда на метеостанции отмечался туман, а n=365 – общее число суток в году. Если туман наблюдался в 15 сутках, то $p(A)$=15/365=0,041 или 4,1%, если p выразить в процентах.

В формуле (1) всегда $0 \le p(A) \le 1$, т.к. m=0 ..., n. Это, так называемая *классическая формула* для вычисления вероятностей. Она долгое время фигурировала в литературе как определение самой вероятности. Сейчас она используется только как формула для *непосредственного* подсчета вероятностей, когда опыт сводится к *схеме случаев*.

Используя (1) и теоремы теории вероятностей можно рассчитать вероятности наступления и более сложных событий, чем рассмотренные выше. Например, вероятность выпадения герба (решки) при одном бросании равна 0,5, а вероятность более сложного события A, состоящего в том, что герб (решка) выпадут три раза подряд равна $p(A)=0{,}5^3=0{,}125$. Аналогично, если, например, в каком-либо месте в отдельно взятый день января вероятность появления тумана одна и та же и равна 0,1, то вероятность того, что туман будет наблюдаться два дня подряд $p(A)=0{,}1^2=0{,}01$, т.е. в 100 раз меньше.

Мы не будем далее развивать эту тему, отсылая читателя к рекомендуемой литературе, где эти вопросы хорошо изложены. Заметим только, что для использования формулы (1) всегда надо тем или иным способом определить m и n, в чем и состоит основная трудность непосредственного подсчета $p(A)$.

3. Частость, или статистическая вероятность, события

Если опыт обладает симметрией возможных исходов (монета, игральные кости, карты и др.), т.е. сводится к схеме случаев, то для непосредственного подсчета вероятности $p(A)$ применима формула (1). В абсолютном большинстве случаев (например, погодно–климатические явления) схема случаев, а, следовательно, и (1), не применимы. Когда мы рассматривали пример с туманом, то предполагалось, что вероятность его появления в каждый отдельный год (или день) одна и та же. Но это обычно не так, т.к. погодные условия непрерывно меняются и вместе с ними меняются вероятности появления таких явлений как туман и других.

Однако формула типа (1) может быть с успехом использована для непосредственного подсчета вероятностей, если произвести не один, а серию из n опытов, в каждом из которых могло произойти или не произойти событие А. Назовем частостью события A в данной серии опытов n величину отношения $p_*(A)$, равную

$$p_*(A) = \frac{m}{n}, \qquad (2)$$

где m – число опытов, в которых появилось A, а n – общее число опытов.

Казалось бы, (1) и (2) совершенно идентичны. Но это не так. Теперь не надо предполагать, что в каждом отдельном опыте $p=const$, что очень важно. Величину $p_*(A)$ называют также *статистической вероятностью*, в отличие от $p(A)$ – математической вероятности. При небольшом числе опытов $p_*(A)$ сама в значительной мере носит случайный характер, меняясь от одной группы опытов к другой. Но при увеличении n она обладает замечательным свойством стабилизироваться, приближаясь к некоторой средней величине $\bar{p}_*(A)$. Например, в каждый отдельно взятый год $p_*(A)$ для числа дней с туманом может сильно меняться, но если брать $p_*(A)$ не за один год, а за 5, 10, 20, 30 лет, то $p_*(A)$ будет все устойчивее, мало колеблясь около некоторого среднего $\bar{p}_*(A)$. Это среднее носит в метеорологии название климатической нормы, если оно определено за 30–40 лет и более. Свойство устойчивости относительных частот $p_*(A)$ с увеличением n многократно проверено в самых различных областях природы и человеческой деятельности и является одной из наиболее характерных закономерностей, наблюдаемых в случайных явлениях. Поэтому между частостью события $p_*(A)$ и его вероятностью $p(A)$ имеет место глубокая органическая связь.

Математическую формулировку связи $p(A)$ и $p_*(A)$ впервые дал Я.Бернулли, показав, что при неограниченном увеличении числа однородных независимых опытов средняя частость события $\bar{p}_*(A)$ будет сколь угодно мало отличаться от его вероятности в отдельном опыте. Только опираясь на эту теорему во многих случаях можно оценить $p(A) \approx \bar{p}_*(A)$.

Например, за 55 лет на метеостанциях Бишкек, Чон-Арык и Байтык наблюдалось число дней с туманом соответственно равное 1595, 2805 и 2530. Тогда, вероятность наблюдения тумана в произвольно взятый день на этих станциях будет приближенно равна:

Бишкек – $\bar{p}_* = \dfrac{1595}{55 \cdot 365{,}25} = 0{,}0794$ или 7,94%;

Чон-Арык – $\bar{p}_* = \dfrac{2805}{55 \cdot 365{,}25} = 0{,}1396$ или 13,96%;

Байтык – $\bar{p}_* = \dfrac{2530}{55 \cdot 365{,}25} = 0{,}1259$ или 12,59%.

Этот пример показывает лишь, как можно оценить вероятность появления тумана, используя данные наблюдений. Но полученного решения явно недостаточно для успешного прогноза тумана на завтра.

Следует отметить, что есть различие в характере приближения частости $p_*(A)$ к вероятности $p(A)$ от стремления *к пределу в математическом смысле*, т.к. при неограничен-

ном увеличении n всегда остается пусть ничтожно малая вероятность того, что $p_*(A)$ в отдельном опыте существенно уклонится от $p(A)$. Поэтому говорят, что $p_*(A)$ *сходится по вероятности* к $p(A)$.

Далее мы увидим, что для определения $p(A)$ события, не сводящегося к схеме случаев, далеко не всегда необходимо из опыта определять его среднюю частоту $\overline{p}_*(A) \approx p(A)$. Это можно сделать косвенно путем различных расчетов, поскольку опыт во многих случаях не возможен. Технология таких косвенных расчетов и составляет главное содержание теории вероятностей и математической статистики. Однако в ее основе всегда лежат те или иные опытные данные, на базе которых выполняются расчеты. Качество и количество таких исходных данных, в конечном счете, определяют надежность получаемых статистических результатов.

В дальнейшем изложении и использовании формулы для статистической вероятности (2) метка * будет опускаться, т.к. из контекста ясно, о чем идет речь.

4. Случайная величина. Практически невозможные и практически достоверные события. Принцип практической уверенности

Случайное событие – это величина качественно-количественного порядка (выпадение герба или решки, наблюдение грозы, тумана, града в заданный день, рождение ребенка с голубыми глазами и др.). Более широким понятием является случайная величина. Введем для нее специальное сокращение – *СВ*.

Случайной величиной называется величина, которая в результате опыта, наблюдения, испытания может принять то или иное определенное численное значение, но заранее неизвестно, какое именно.

Примеры случайных величин:

- число дней с туманом в заданном произвольно году (дискретная СВ),
- значение температуры точно в 12 ч в заданный день (непрерывная СВ, если предположить возможность ее измерения с любой точностью),
- значение максимальной скорости ветра при очередном прохождении атмосферного фронта (непрерывная СВ).

Случайные величины могут быть *дискретными* (число дней с явлениями), *непрерывными* (значения таких метеорологических величин как температура, скорость ветра, влажность, давление и др.), *ограниченными с одного конца* (например, осадки, скорость ветра – не могут быть отрицательными) *ограниченными с обоих концов* (облачность может быть в пределах только от 0 до 10 баллов) или *неограниченными* (например, нельзя для

любой местности указать предел максимальной скорости ветра, который не мог бы быть превышен пусть с самой малой вероятностью). Случайное событие есть частный случай СВ, когда наступлению события можно приписать ее значение 1, а не наступлению – 0. На практике также легко перейти от непрерывной СВ к дискретной, заменив множество ее значений из некоторого узкого диапазона средним значением в диапазоне. Например, все значения скоростей ветра в диапазоне 1,5 ..., 2,5 м/с мы можем округленно отнести к скорости 2 м/с. При статистической обработке метеорологических данных так и приходится поступать в абсолютном большинстве случаев, когда весь диапазон СВ разбивается на ряд интервалов–классов и заменяется численными значениями середин классов.

Современная теория вероятностей и математическая статистика оперируют *преимущественно со случайными величинами*, переходя всюду от случайных событий к СВ, что позволяет эффективно использовать весь их математический аппарат.

В шкале значений вероятностей p принято считать, что при p=0 событие (т.е. численное значение СВ) невозможно, а при p=1 достоверно. Но на практике обычно приходится иметь дело с событиями, про которые можно сказать, что они лишь *практически не возможны* или *практически достоверны*. Практически достоверным событием назовем такое, вероятность которого не в точности равна 1, а близка к ней. Насколько она должна быть близка, теория вероятностей *ответить не может*. Это дело опыта, т.е. характера той задачи, которая решается. Точно также практически невозможным событием назовем такое, вероятность которого не в точности ровна нулю, а близка к нему. Причем необходимая степень близости также определяется исходя из поставленных целей.

Например, Вы планируете на завтра прогулку в лес. Вероятность хорошей погоды на завтра равна 0,90. Значит 10% за то, что погода будет плохая и прогулку планировать не стоит. Однако в данном случае вероятность хорошей погоды p=0,90 можно признать высокой и соответствующей практически достоверному событию, а вероятность плохой погоды (p=0,1) – практически невозможному событию. Напротив, Вы производите тренировочный прыжок с парашютом. Если ваш парашют имеет вероятность отказа работы p=0,01 (это в 10 раз меньше, чем p=0,1 для плохой погоды), то эту вероятность нельзя признать малой, а событие отказа парашюта – практически невозможным, т.к. характер риска здесь стал принципиально иным –это ваша жизнь.

Эти обстоятельства накладывают некоторые ограничения на все решения, имеющие вероятностный характер. Они всегда полностью (т.е. на 100%) не достоверны, а имеют определенную вероятность риска, т.к. всегда остается пусть самый малый шанс того, что, не допустив никакой ошибки в применении теории на практике, вы приняли неверное решение. Однако это не недостаток теории и следствий из нее, а внутренняя суть, т.е. при-

рода массовых случайных явлений, которые можно исследовать только вероятностными методами. Напротив, применив теорию, вы в большинстве случаев будете принимать верные решения. Так, задав грань допустимого риска в 1%, вы в 99% случаев обеспечите принятие правильных решений и лишь в 1% случаев принятие ошибочного решения. Методы теории вероятностей и математической статистики могут и и должны обеспечить решение практических задач именно в этом плане.

Тема 1. АНАЛИЗ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ ИХ СТАТИСТИК

Первичная систематизация и простые методы обработки эмпирических рядов наблюдений метеостанций (эмпирических распределений или выборок), являются необходимыми и важными этапами климато-станистического анализа. Более того, во многих случаях они позволяют получить существенные и часто достаточные по требуемой глубине анализа климатические результаты, не прибегая к более сложным методам статистической обработки. К таким простым методам относится ранжирование метеорологических рядов по возрастанию или убыванию, их группировка и расчет для них эмпирических функций распределения. К этим методам непосредственно примыкает последующий расчет и использование в климатическом анализе основных статистик эмпирических распределений – среднего, дисперсии (среднего квадратического отклонения), коэффициентов вариации, асимметрии и экссцесса. Именно этим первичным, простым, но необходимым на практике статистическим методам обработки и анализа метеорологичесих рядов посвящены тема 1.

Глава 1.1. ЗАКОН РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ И ЕГО ОБЩИЕ СВОЙСТВА

1.1.1. Генеральная совокупность и выборка

Случайная величина (СВ) в п.4 вводной лекции была определена как величина, которая принимает в результате опыта, наблюдения или испытания то или иное численное значение, но заранее неизвестно, какое именно. Проведя n испытаний над СВ, мы получим n ее значений. Представим теперь, что число испытаний неограниченно возрастает, так что оно охватывает весь возможный диапазон ее изменения. *Генеральной совокупностью*

называется все мыслимое множество значений СВ, которое охватывает весь диапазон ее изменений (при данном комплексе условий наблюдений или опыта). Генеральная совокупность может быть *конечной* или *бесконечной*. Например, колода в 36 игральных карт содержит конечный и полный набор переменной. Возможные значения температуры за 100 лет (например, в XX веке) в каком-либо пункте содержат полное множество значений переменной, бесконечное хотя бы потому, что температура непрерывная СВ[1]. Однако от бесконечности здесь легко избавиться, если представить все температуры их часовыми или восьмисрочными значениями. В этом случае будет получена, хотя и большая, но конечная генеральная совокупность, которая полно характеризует температурные условия за 100 лет. Однако если мы хотим получить генеральную совокупность не за 100, а последние 1000 лет, то температурный ряд потребуется расширить на все последнее 1000-летие. Так выглядит соблюдение оговорки, приведенной в определение генеральной совокупности в скобках – «при данном комплексе условий наблюдений или опыта».

Выборочной совокупностью или *выборкой* называется любая выборка (подмножество), взятая из генеральной совокупности. Выборку используют для того, чтобы изучить *ее статистические свойства* и сделать по ним выводы о *статистических свойствах генеральной совокупности*. Обычно по тем или иным причинам выборка намного меньше по объему генеральной совокупности. Под объемом в статистике понимается число членов n выборки или генеральной совокупности.

Чтобы по свойствам выборки можно было объективно судить о свойствах генеральной совокупности, выборка должна быть *репрезентативна* или *показательна*. К сожалению, нет каких-то законченных рецептов, как сделать выборку показательной. Можно только указать на два основных правила, соблюдение которых повышает репрезентативность выборки: 1) выборка должна иметь достаточно большой объем (обычно не менее 30-50 значений СВ); 2) она должна иметь случайный характер (когда исключается возможная тенденциозность подхода к проведению выборки). Кроме того, ее информативность повышается, если элементы или члены выборки статистически независимы (не связны между собой).

В задачах метеорологии вопрос о показательности выборок имеет большое значение. Однако здесь ситуация иная, чем, например, в экономических и общественных науках. Метеорология и климатология оперирует данными фактических наблюдений – метеорологическими рядами – за имеющийся период наблюдений от нескольких лет до 100 и редко

[1] В этом смысле любой конечный интервал изменений температуры, например, за 1 час содержит теоретически бесконечное множество ее значений. Однако нас не должно это беспокоить, т.к. в практическом плане мы всегда сумеем разумно перейти к необходимому конечному числу ее значений.

150–250 лет. Считается, что ансамбль погодных условий, описывающих климат заданного периода времени, остается практически постоянным в течение 20–50 лет (в среднем около 30 лет), а между смежными 30-летиями может уже существенно отличаться, обуславливая временную изменчивость климата. Поэтому различного рода многолетние средние, называемые *климатическими нормами*, обычно рассчитываются за период 30 лет и более. Практически в большинстве задач используется весь доступный период наблюдений метеостанций. Чем более длинный метеорологический ряд использован, тем надежнее получаемые по нему выводы.

Надо особо подчеркнуть, что при различного рода специальных или экспедиционных наблюдениях с неизбежностью приходится оперировать небольшими выборками. Например, чтобы решить вопрос о максимально возможных скоростях ветра на гребнях хребтов Тянь-Шаня и Памиро-Алая, автору в течение 20 лет пришлось организовывать трудные анемосъемки в зимнее время длительностью от одного месяца до 1–5 лет (в последнем случае приборы устанавливались на радиорелейных станциях). Ясно, что расчеты максимальных скоростей, вероятных 1 раз в год, 5, 10 и даже 25 лет, что требовалось при решении задач по оценке нагрузок на ЛЭП, по выборкам длительностью всего один – несколько месяцев сопряжены с возможными значительными погрешностями. Тем не менее, такие расчеты были сделаны, и они дали по-существу единственное решение задачи, основанное на прямых измерениях. Главная погрешность в этом случае проистекала от случайности погодно-ветровых условий зимнего периода в момент проведения анемосъемки (выборки). Если период анемосъемки примерно соответствует средним условиям, то показательность выборки повышается, при экстремально низком и высоком ветровом режиме результаты расчетов будут соответственно занижены или завышены, т.е. непоказательны.

Что касается обеспечения случайности выборки, то в метеорологических наблюдениях она обеспечивается автоматически случайным ходом самих метеорологических процессов. Здесь важно понимать, что выборочные результаты объективно характеризуют только тот период, в котором проводились наблюдения. Их всегда следует осторожно распространять даже на смежный период будущего, помня, что такая временная экстраполяция правомочна при условии неизменности климатических условий, чего никогда нет, т.к. климат всегда, хотя и относительно медленно, изменяется, причем эти изменения носят сложный колебательный и поступательный характер.

1.1.2. Виды эмпирических выборок

Первичная метеорологическая выборка есть результат регистрации наблюдений, носящий название метеорологического ряда. Такой ряд может быть представлен:

1) записью самопишущего прибора или цифрового регистратора;

2) табличной хронологической записью наблюдений (например, 8-срочные данные, данные, осредненные за сутки, месяц, год и т.д.);

3) ранжированной в возрастающем или убывающем порядке выборкой, кс

по классам исходными данными.

Например, фрагменты выборок типа 2 и 3 для средней годовой температуры в Бишкеке имеют вид:

2.	Год	1928	1929	1930	1931	…...	1997	1998	1999	2000
	$T_{сг}$	10,1	9,7	9,9	9,2	…...	12,0	11,4	11,5	12,0
3.	$T_{сг}$ по возрастанию:	8,7	8,8	8,9	9,0	…...	12,0	12,0	12,0	12,0
	$T_{сг}$ по убыванию:	12,0	12,0	12,0	12,0	…...	9,0	8,9	8,8	8,7

Пример сгруппированной выборки по данным о максимальных скоростях 211 бурь (скорость $V \geq 15$ м/с), зарегистрированных на метеостанции Фрунзе за 19 лет, выглядит следующим образом:

Класс V_i, м/с	15–17	17–19	19–21	21–23	23–25	25–27	27–29	29–31
Середина $\overline{V_i}$, м/с	16	18	20	22	24	26	28	30
Частота, n_i	134	59	15	1	1	0	1	0

Заметим, что число случаев попадания СВ в тот или иной класс (n_i) носит название *частоты класса*, в отличие от его *частости*, равной $n_i/\left(n=\sum n_i\right)$. Частость, выраженная в процентах, в метеорологии носит название повторяемости. Ее также называют относительной частотой. Численно она равна статистической вероятности $p_*(x)$ по формуле (2) п. 3 вводной лекции. Так, частота первого класса в последнем примере n_i=134, а его частость или статистическая (эмпирическая) вероятность равна 100%·134/211=63,5%.

Какой вид выборки использовать при анализе зависит от цели анализа и задач наглядности представления его результатов. Разумеется, конечные результаты расчетов во всех случаях будут одинаковы, но вот наглядность, а значит и эффективность анализа, будут различны. Например, в типично рядовом случае, когда рассматривается выборка максимальных скоростей ветра, в хронологически расположенных случаях 211 бурь получаемая степень наглядности невысока. Но стоит только ранжировать выборку или дать ее в сгруппированном виде (как в примере выше), то сразу видно, что, как правило, скорость ветра в бурях не превышает 21 м/с (из 211 случаев только в трех скорости ветра были больше). Видно также, что градации классов 15–17, 17–19 и 19–21 относительно хорошо «наполнены», т.к. их частоты боее 5-10 и соответственно равны 134, 59 и 15 случаев. Во всех остальных классах частоты неустойчивы (0 или 1).

Для группировки выборки существуют ориентировочные практические правила: 1. Число классов k можно оценить по объему выборки n по уравнению:

$$k \approx 5 \lg n, \qquad (1.1)$$

что дает:

объем выборки n	50	100	500	1000	10000
число классов k	8	10	13	15	20

Ширина класса Δx тогда приближенно равна

$$\Delta x = \frac{x_{max} - x_{min}}{k}, \qquad (1.2)$$

где x_{max} и x_{min} – значения максимального и минимального члена выборки.

Однако оценки k и Δx по (1.1) и (1.2) нужно всегда корректировать, исходя из генетических свойств выборки. Всегда высокое качество группировки есть признак квалификации исследователя. Например, в выборке из n=211 значений максимальных скоростей в бурях по (1.1) следует иметь 10–11 классов и $\Delta x \approx 1{,}3$ м/с. Но совершенно ясно, что для удобства ширину классов надо взять целочисленной – 2 м/с. Тогда, откорректированное значение k будет около 7. Дополнительный 8 класс в рассмотенном примере взят для «запаса», необходимого при теоретических расчетах. Еще один вопрос: куда относить случаи, точно попадающие на границу классов при подсчете объемов классов n_i? Рекомендуется использовать один из трех вариантов: 1) делить их пополам и включать поровну в оба смежные класса; 2) относить все такие значения в нижний класс; 3) относить их все в верхний класс. Какой из вариантов целесообразнее должен решить исследователь, исходя из специфики выборки, помня, что, как правило, это не игает особо большой роли.

Как будет показано далее, по выборкам, в первую очередь рассчитываются их основные статистические характеристики, которые кратко называются статистиками. Это среднее значение СВ, ее дисперсия, среднее квадратическое отклонение, коэффициенты вариации, асимметрии и эксцесса. Все результаты расчетов этих характеристик по выборкам носят названия *оценок*, тем самым подчеркивается приближенность получаемых значений статистик. Напротив, результаты тех же расчетов по генеральной совокупности носят название *параметров*, чем подчеркивается точность их значений. Эти термины будут широко использоваться в указанном смысле во всем последующем изложении.

1.1.3. Понятие закона распределения случайной величины и его общие математические свойства

Обозначим СВ через X (например, это температура), а все возможные ее численные значения через x. Пусть СВ непрерывна. Рассмотрим как меняется с изменением x вероятность p события $X<x$, т.е. вероятность попадания СВ X левее точки x, которая может пробегать все множество значений от $-\infty$ до $+\infty$. Очевидно, что $p(X<x)$ есть некоторая функция $F(x)$. Таким образом, введем

$$F(x)=p(X<x). \qquad (1.3)$$

Функция F(x) называется *интегральной функцией распределения или интегральным законом распределения СВ*. Она полностью описывает все свойства как непрерывной, так и дискретной СВ. Ее общий вид показан на рис.1.1. Эта функция ограничена, она меняется в пределах от 0 до 1 и безразмерна, т.к. представляет собой вероятность выполнения неравенства $(X<x)$. Задавая любое x_i (рис.1.1) можно найти F_i, равное вероятности p_i того, что $X<x_i$. Двигаясь по оси x вправо можно определить все эти вероятности p, представленные в общем виде формулой (3).

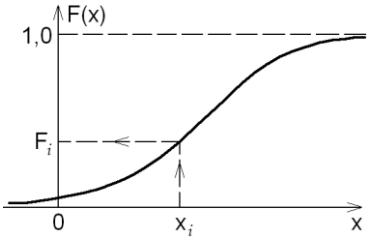


Рис.1.1. Общий вид функции $F(x)=p(X<x)$.

Любое правило, которое устанавливает соответствие между значением случайной величины x и вероятностью ее появлений p (т.е. события $X<x$) называется законом распределения СВ. Это может быть график, таблица, функция, словесное правило. Однако событие $X=x$ можно приписать только дискретной СВ, т.к. для непрерывной СВ любой конечный промежуток содержит несчетное множество ее значений, для каждого из которых $p=0$. Этим вызвана необходимость перехода от события $X=x$ к событию $X<x$. Только на первый взгляд кажется, что сделан переход к менее наглядному и менее удобному на практике событию. Никакой потери информативности и наглядности, как увидим далее,

не произошло. Зато все формулировки приобрели универсальность, т.к. стали справедливыми для всех СВ – дискретных и непрерывных. Введение *функции F(x)* в качестве закона распределения СВ играет решающую роль с точки зрения привлечения для анализа всех возможностей математического аппарата.

Для непрерывной СВ и дифференцируемой *F(x)* существует производная *f(x)*:

$$f(x) = F'(x) = \frac{dF(x)}{dx}. \qquad (1.4)$$

Она характеризет скорость возрастания *F(x)* в каждой точке *x*. Общий вид *f(x)* показан на рис.1.2. Функция принимает только положительные значения *(f(x)≥0)*, может иметь не только колоколообразную

форму, как на рис.1.2, и функционально связана с *F(x)*. Функцию *f(x)* называют *функцией распределения плотности вероятности* (кратко – функцией плотности вероятности) или *дифференциальным законом распределения*, т.к. если известно *f(x)*, то всегда можно перейти к *F(x)* и наоборот. В отличие от *F(x)*, функция *f(x)* имеет размерность x^{-1}.

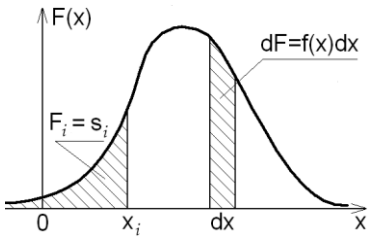


Рис.1.2. Общий вид функции плотности ***f(x)***.

Функции *F(x)* и *f(x)* имеют следующие общие математические свойства:

$$0 \le F(x) \le 1, \; f(x) \ge 0; \qquad (1.5)$$

$$f(x) = dF/dx, \quad F(x) = \int_{-\infty}^{x} f(x)dx; \qquad (1.6)$$

$$dF = f(x)dx; \qquad (1.7)$$

$$\int_{-\infty}^{\infty} f(x)dx = 1. \qquad (1.8)$$

Каждому значению $F_i(x_i)$ на графике рис.1.1 по свойству (1.6) соответствует площадь S_i на графике рис. 1.2, ограниченная снизу отрезком $]-\infty; x_i]$, вертикалью x_i, а сверху – участком кривой *f(x)*.

По свойству (1.7) вероятность попадания СВ на участок *dx* равна произведению *f(x)dx*. Это произведение *dF=f(x)dx* носит название элемента вероятности и численно равно заштрихованной площадке, показанной на рис. 1.2.

Формально любая функция, удовлетворяющая условию $0 \le F(x) \le 1$, может служить законом распределения. Однако это чисто математическое условие. Обычно кроме него требуется выполнение не менее важного физического условия: каждая *F(x)* должна описывать распределение вероятности в том или ином вероятностном процессе, т.е. вытекать из этого процесса.

Одной из важных задач математической статистики является нахождение оценок конкретного вида *F(x)* или *f(x)* по выборке. Эта задача носит название аппроксимации (приближенного представления) генеральной совокупности выборочной *F(x)*, т.е. ее оценкой, найденной по опытным данным.

Зная закон распределения (в виде формулы, таблицы, графика) можно исчерпывающим образом описать случайную величину, решая относительно нее следующие возможные задачи:

1. Задавая значение переменной x_i найти соответствующие им вероятности p_i событий ($X<x_i$), а, следовательно, и вероятности $\overline{p_i}$ противоположных событий ($X \geq x_i$), т.к. $\overline{p_i} = 1 - p_i$. На рис.1.1 для этого надо по заданному x_i определить на оси *y* значение F_i.

2. Задавая значение $F_i=p_i(X<x_i)$ найти соответствующие им x_{p_i}, которые называются квантильными значениями (или сокращенно квантилем). Каждому квантилю x_{p_i} соответствует вероятность p_i того, что СВ *X* не превзойдет его значение. Вероятность p_i называется уровнем квантиля. Эта задача обратная первой и для ее решения надо задать $F_i=p_i$ на оси *y*.

3. Определить вероятностный диапазон Δp попадания в него СВ *X*, используя либо подход 1, либо подход 2. Например, необходимо с вероятностью Δp=0,98 определить вероятные границы размаха переменной *X*. Этот диапазон $\Delta p=(p=0,99)-(p=0,01)$. Найдем квантили $x_{0,99}$ и $x_{0,01}$, их разность Δx и есть искомый диапазон, в которой с вероятностью Δp=0,98 попадает СВ *X*, т.е. 98% всех значений *X* будет лежать внутри интервала $x_{0,01}$ …, $x_{0,99}$.

4. Определить вероятность события $X=x_i$, понимая под x_i некоторый узкий конечный интервал переменной, в котором x_i есть центр интервала. Например, надо найти вероятность появления скорости ветра 20 м/с. Для этого выделим интервал скоростей 19,5 ..., 20,5 м/с и будем считать, что все скорости, которые попадают в этот интервал, равны 20 м/с. Тогда, решение сводится к задаче 3 нахождения двух вероятностей $p_{19,5}$ и $p_{20,5}$. Разность $\Delta p = p_{19,5} - p_{20,5}$ и есть вероятность скорости 20 м/с, т.е. события (19,5≤*V*≤20,5) м/с.

Решив эти задачи, мы полностью опишем интересуемые нас на практике изменения случайной величины, определив все необходимые вероятности, которые позволяет рассчитать функция распределения СВ.

1.1.4. Табличное и графическое представление эмпирических законов распределения

На этапе предварительного анализа выборки всегда полезно проанализировать получающийся по нему эмпирический закон распределения. Это можно сделать несколькими графическими способами.

Ранжирование и анализ не сгруппированной выборки.

Ранжированием выборки называется ее построение в возростающий или убывающий ряд. Это длается в Excelc помощью программы «Сортировка и фильтр». Для ранжировки по возростанию надо выделить исходную выборку, выбрать в этой программе опцию «сортировка от минимального к максимальному» и щелкнуть на ней ЛКМ. Выборка будет построена в возрастающий ряд. Теперь каждому члену этого ряда надо присвоить ранг r_i (от 1 до n), равный номеру члена ряда. Для примера в табл. 1.1 в первых двух столбцах слева показан исходный ряд средних годовых температур T_i на станции Байтык за 86 лет (1915 – 2000 гг.), а в двух последующих столбцах приведен его ранжированный вид и ранги r_i каждого члена ряда.

Для каждого x_i, расположенного в порядке возрастания в выборке, можно рассчитать эмпирические значения $F_i(x_i)$ по одной из формул:

$$F_i(x_i) = \frac{r_i}{n}, \tag{1.9}$$

$$F_i(x_i) = \frac{r_i}{n+1}, \tag{1.10}$$

$$F_i(x_i) = \frac{r_i - 0{,}3}{n + 0{,}4}, \tag{1.11}$$

где r_i – номер члена x_i в ранжированной выборке; n - объем выборки.

Формулу (1.9) целесообразно использовать при очень длинных рядах (больших n≥100), а формулы (1.10) и (1.11) – при более коротких рядах. Затем по оси x отложить натуральную переменную x_i, а по оси y – соответствующие ей значения $F_i(x_i)$ (их можно выразить так же в процентах). Полученная кривая $F_i(x_i)$ носит название *эмпирической функции распределения, эмпирической кривой обеспеченности* или *кумулятивной кривой*. Иногда строятся кривые для выборки, ранжированной в убывающем порядке, т.е. $F_*(x) = p(X \geq x)$. Между $F(x)$ и $F(x)$ справедливо соотношение $F(x) = 1 - F_*(x)$. Полученные кривые дают наглядное представление о виде и свойствах $F(x)$ или $F_*(x)$.

В табл. 1 приведен пример расчета функции обеспеченности $F_i(x_i)$ для средних годовых температур воздуха на МС Байтык за 86-летний период наблюдений 1915-2000 гг. На рис. 1.3 показан построенный по табл. 1.1 точечный график этой функции, который в совокупности с таблицей позволяет решать задачи 1-4, пречисленные в предыдущем п. 1.1.3. Например, найдем минимальные значения средних годовых темератур с обеспеченностью

F, равной 0,01, 0,05 и 0,10. Для этого используем непосредственно таблицу, так как по графику это делается менее точно. Задавая указанные значения F_i определяем эти температуры, которые будут соответственно равны 4,0, 5,2 и 5,6°С. Точно также определим максимальные значения средних годовых температур с обеспеченностью F_i, равной 0,9, 0,95 и 0,99, которые будут равны 7,2, 7,6 и 8,1°С. Именно эти значения обычно нужны потребителю климатической информации, тогда как их определение (расчет табл. 1.1 и построение графика рис. 1.3) и есть основная задача климатолога.

Таблица 1.1

Расчет обеспеченностей средних годовых температур *F (T_i°C)* на МС Байтык за 1915-2000 гг. по формуле (1.10)

Исходный ряд		Ранжированный ряд			Исходный ряд		Ранжированный ряд		
год	T_i°C	T_i°C	r_i	F_i	год	T_i°C	T_i°C	r_i	F_i
1915	6.9	4.0	1	0.0115	1958	6.6	6.6	44	0.5057
1916	7.1	4.9	2	0.0230	1959	6.4	6.6	45	0.5172
1917	6.2	5.0	3	0.0345	1960	5.9	6.6	46	0.5287
1918	5.9	5.2	4	0.0460	1961	6.2	6.6	47	0.5402
1919	6.1	5.2	5	0.0575	1962	5.6	6.6	48	0.5517
1920	7.8	5.3	6	0.0690	1963	7.0	6.7	49	0.5632
1921	5.8	5.5	7	0.0805	1964	6.8	6.7	50	0.5747
1922	5.2	5.5	8	0.0920	1965	6.7	6.7	51	0.5862
1923	5.6	5.6	9	0.1034	1966	6.6	6.7	52	0.5977
1924	7.0	5.6	10	0.1149	1967	5.5	6.7	53	0.6092
1925	7.1	5.6	11	0.1264	1968	6.2	6.8	54	0.6207
1926	6.4	5.6	12	0.1379	1969	6.0	6.8	55	0.6322
1927	7.7	5.6	13	0.1494	1970	7.2	6.8	56	0.6437
1928	7.2	5.7	14	0.1609	1971	6.7	6.8	57	0.6552
1929	6.6	5.7	15	0.1724	1972	6.2	6.9	58	0.6667
1930	7.7	5.7	16	0.1839	1973	6.9	6.9	59	0.6782
1931	5.0	5.8	17	0.1954	1974	5.8	6.9	60	0.6897
1932	6.1	5.8	18	0.2069	1975	5.9	6.9	61	0.7011
1933	4.9	5.8	19	0.2184	1976	6.5	7.0	62	0.7126
1934	7.0	5.9	20	0.2299	1977	6.5	7.0	63	0.7241
1935	6.0	5.9	21	0.2414	1978	7.4	7.0	64	0.7356
1936	5.6	5.9	22	0.2529	1979	5.7	7.0	65	0.7471

1937	5.2	5.9	23	0.2644	1980	6.8	7.0	66	0.7586
1938	7.2	6.0	24	0.2759	1981	7.2	7.0	67	0.7701
1939	5.9	6.0	25	0.2874	1982	7.0	7.0	68	0.7816
1940	7.2	6.0	26	0.2989	1983	5.6	7.0	69	0.7931
1941	7.0	6.1	27	0.3103	1984	4.0	7.1	70	0.8046
1942	6.1	6.1	28	0.3218	1985	6.5	7.1	71	0.8161
1943	5.8	6.1	29	0.3333	1986	6.4	7.1	72	0.8276
1944	5.3	6.2	30	0.3448	1987	6.7	7.1	73	0.8391
1945	8.1	6.2	31	0.3563	1988	6.6	7.2	74	0.8506
1946	7.0	6.2	32	0.3678	1989	5.5	7.2	75	0.8621
1947	6.8	6.2	33	0.3793	1990	6.7	7.2	76	0.8736
1948	6.8	6.2	34	0.3908	1991	7.1	7.2	77	0.8851
1949	6.2	6.4	35	0.4023	1992	5.6	7.2	78	0.8966
1950	6.9	6.4	36	0.4138	1993	6.0	7.2	79	0.9080
1951	6.9	6.4	37	0.4253	1994	6.5	7.3	80	0.9195
1952	7.1	6.4	38	0.4368	1995	7.3	7.4	81	0.9310
1953	5.7	6.5	39	0.4483	1996	6.7	7.5	82	0.9425
1954	6.4	6.5	40	0.4598	1997	5.7	7.7	83	0.9540
1955	7.0	6.5	41	0.4713	1998	6.5	7.7	84	0.9655
1956	7.0	6.5	42	0.4828	1999	6.6	7.8	85	0.9770
1957	7.2	6.5	43	0.4943	2000	7.5	8.1	86	0.9885

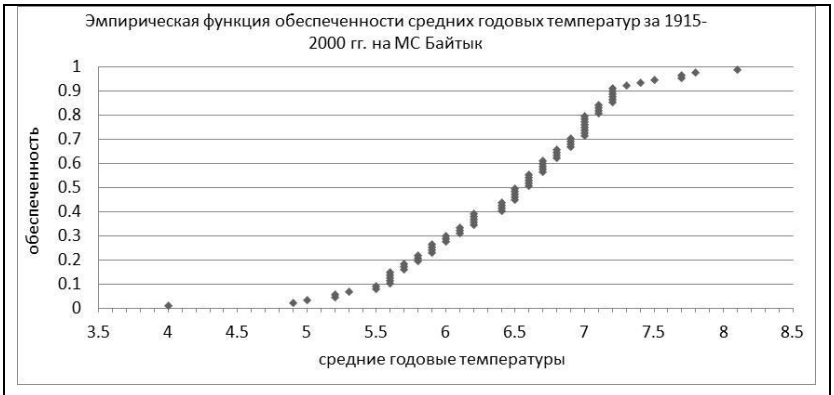


Рис. 1.3 График эмпирической функция обеспеченности средних годовых темпераур воздуха на МС Байтык за 1915-2000 гг.

Группировка исходной выборки и ее анализ.

Определение числа и границ классов группирования. Выборка температур в табл. 1.1 содержит 86 членов, и трудно обозрима для визуального анализа. Произведем ее группировку, исползовав оценочные формулы (1.1) и (1.2) для приближенного определения числа классов группирования k и их ширины ΔT°C. При объеме выборки n=86 по формуле (1.1) найдем ориентировочное число классов k=5*log(86)=9,7 (т.е. примерно 10), а ширина класса ΔT=(8,1-4)/9,7=0,42°C. Для практического примем округленную ширину класса, равную 0,5°C. Так как самым низким значением температуры в выборке является 4,0°C, то в качестве первого класса целесообразно взять 3,5-4,0°C. Тогда, остальные границы классов определяться автоматически и будут соответствовать значениям, показанным в трех первых столбцах табл. 1.2.

Определение частот классов n_i,. По программе Excel «частота» рассчитаем частоты n_i попадания температур исходной выборки, приведенной в табл. 1.1, в каждый класс табл. 1.2 (для этого можно использовать как сгруппированную, так и не сгруппированную выборку).

Для этого сначала в табл. 1.2 надо выделить для записи частот n_i столбец, равный заданному числу классов сгруппированной выборки. Затем выбрать программу «частота» и в ее окне ввести в «массив данных» - массив исходной выборки (т.е. ввесь столбец с выборкой по табл. 1.1), а в «массив интервалов» - массив верхних границ классов группирования по табл. 1.2. После этого щелкнуть ЛКМ на ОК, и щелкнуть курсором мыши в строке формул вверху, где показан ввод данных программы «частота». Теперь надо нажать три клавиши одновременно: Shift-Ctrl-Enter (лучше сначала нажать Shift-Ctrl, а затем дожать Enter).

В табл. 1.2 показан пример группировки выборки средних годовых температур по станции Байтык за 86 лет из табл. 1.1. В ней приведены нижние (НГ) и верхние (ВГ) границы классов группирования, значения середин классов (СК), а также рассчитанные по программе «частота» значения частот классов n_i.

Анлиз сгруппированной выборки. Видно, что полученные частоты класссов n_i образуют массив, который начинается и заканчивается нулевыми классами (n_i=0). Всегда целесообразно при группировке выборки задавать общий маасив классов с запасом таким образом, чтобы начальный и конечный классы были нулевыми. Это объясняется тем, что выборка имеет ограниченный период наблюдений. При его увеличении (т.е. возможном увеличении объема выборки) следует ожидать и возможного расширения массива данных с выходом за пределы предыдущих исходных значений. Как увидим далее, при статисти-

ческом выравнивании выборки теоретическими законами распределений теоретические частоты классов, как правило, выйдут за пределы выборочных, так что иногда надо добавлять слева и справа не по одному нулевому классу, а по два или даже по три.

Полученное распределение частот классов n_i будет отражать все основные свойства распределения средних годовых температур на станции при условии репрезентативности выборки. На границах выборки частоты классов обычно малы и меняются недостаточно закономерно. Говорят, что классы здесь «плохо заполнены», так как имеют частоты менее 5. Но в ее основной средней части закономерность следования частот, как правило, выражена хорошо, если выборка достаточно велика по объему.

Зная частоты классов n_i легко рассчитать их статистические вероятности (или частости) $p_{i=} / \sum n_i$ (в данном случае $\sum n_i=86$), которые приведены в столбце 5 табл. 1.2. В климатологии p_i называются также повторяемостями и обычно выражают в процентах. Поэтому $\sum p_i$ всегда равна 1 или 100%. Значения интегральной функции распределения F_i (функции обеспеченностей или накопленных частостей) находят последовательным суммированием $p_{i,}$ которые приведены в последней столбце табл. 1.2. Разумеетс, полученные массивы n_i, p_i *и* F_i *совершенно однозначно* характеризуют распределение средних годовых температур на МС Байтык, разница только в формате итоговых данных для анализа.

Таблица 1.2

Результат группировки исходной выборки средних годовых температур воздуха на МС Байтык за 1915-2000 гг. (обозначения: *НГ, ВГ и СК* – нижняя, верхняя границы классов температуры и их середина)

НГ, °С	ВГ, °С	СК,°С	Частота n_i	Вероят. p_i	Обеспеч. F_i
3.0	3.5	3.25	0	0	0
3.5	4.0	3.75	1	0.011628	0.011628
4.0	4.5	4.25	0	0	0.011628
4.5	5.0	4.75	2	0.023256	0.034884
5.0	5.5	5.25	5	0.05814	0.093023
5.5	6.0	5.75	18	0.209302	0.302326
6.0	6.5	6.25	17	0.197674	0.500000
6.5	7.0	6.75	26	0.302326	0.802326
7.0	7.5	7.25	13	0.151163	0.953488
7.5	8.0	7.75	3	0.034884	0.988372

8.0	8.5	8.25	1	0.011628	1
8.5	9.0	8.75	0	0	1

Теперь можно построить две гистограммы (столбчатые графики - диаграммы) для распределений значений p_i, и F_i, которые показаны на рис. 1.4 и 1.5. На рис. 1.4 по горизонтальной оси отложены *середины классов* темтературы, а по вертикальной - *вероятности классов* p_i. Вместо середин классов можно также отложить сами *границы классов* (3,0-3,5, 4,0-4,5 и т.д.), что является равносильным. На рис. 1.5 по горизонтальной оси отложены *верхние границы* классов температуры, а по вертикальной оси соответствующие им *значения обеспеченнотей классов* F_i. Этот порядок построения вытекает из понятий p_i и F_i и его надо выполнять обязательно.

Табл. 1.2 совместно с рис. 1.4 обладают высокой визуальной наглядностью, так как позволяют видеть непосредственно основные статистические свойства выборки распределения средних годовых температур на МС Байтык. Так, вместо перечня из 86 значений температур не сгруппированной выборки табл. 1.2 теперь анализу подлежат всего 10 классов табл. 1.3 (10, а не 12 потому, что два концевых класса пустые). Из этих данных, прежде всего, видно, что наиболее часто (в 86% случаев) температуры попадают в 4 центральных класса, имеющих общий диапазон значений от 5,5 до 7,5°С, а на все остальные приходится только 14%. При этом начальным не нулевым классом является класс 3,5-4,0°С, а аналогичным конечным 8,0-8,5°С. В этом диапазоне сосредоточены все выборочные средние годовые температуры на станции Байтык.

Рис. 1.4. Гистограмма распределения статистических вероятностей классов средней годовой температыры на МС Байтык.

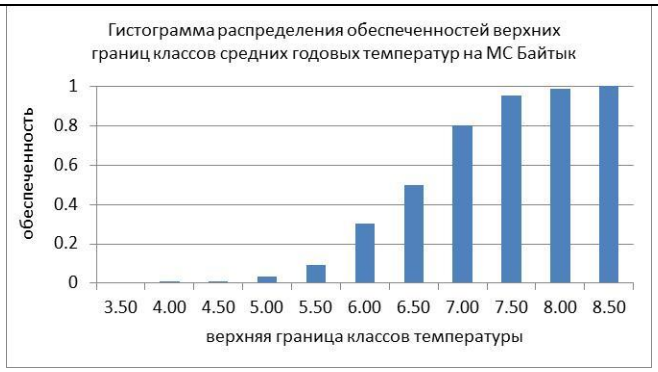


Рис. 1.5. Гистограмма распределения статистических обеспеченностей верхних границ классов средней годовой температуры на МС Байтык

При этом наиболее часто наблюдаемый класс, модальный класс, имеет границы 6,5-7,0°С, вероятность попадания в который равна 30%. Второй модальный класс (5,5-6,0°С) выражен гораздо слабее, имея вероятность около 20% (возможная много модальность как свойство распределений будет рассмотрена в следующей главе 1.2). В целом распределение вероятностей классов на рис. 1.4 лево асимметричего, так как хорошо видно, что его левая ветвь длиннее правой (это свойство распределений также будет рассмотрено в следующей главе 1.2).

Если через середины верхних частей столбиков на рис. 1.5 провести плавную огибающую кривую, то будет получена кривая обеспеченностей F_i, показанная на рис. 1.1, только теперь она представлена для сгруппированной выборки. Используя и численные значения F_i для кждого класса и соответствующие им верхние границы классов температуры из табл. 1.2 (или более грубо по графику рис. 1.5) можно решить все 4 статистические задачи, вытекающие из общих свойств функции распределения, перечисленные в конце п. 1.1.3.

Глава 1.2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ЭМПИРИЧЕСКИХ ЗАКОНОВ РАСПРЕДЕЛЕНИЙ

1.2.1. Понятие начальных, центральных и смешанных моментов. Требования состоятельности, несмещенности и эффективности, предъявляемые к оценкам статистик

Исчерпывающей характеристикой СВ является закон распределения. Но существуют и другие характеристики, которые описывают ее хотя и не полно, но определяют те или иные важные свойства распределений. Обобщенно их часто называют *основными статистиками распределений*. Статистики намного легче получить по выборке, чем закон распределения, часто их бывает достаточно для практического анализа. Кроме того, именно по ним затем может быть найден и сам закон распределения. К таким статистикам относятся: среднее или математическое ожидание, медиана, мода, дисперсия (а также среднее квадратическое отклонение, стандартное отклонение, стандарт), коэффициенты вариации, асимметрии, эксцесса и коэффициент корреляции. Все они вычисляются через так называемые начальные, центральные и смешанные статистические моменты случайной величины (термин «момент» заимствован из механики). Рассмотрим настоящем пункте сначала определения моментов, через которые находятся эти статистики, а затем общие требования, предъявляемые к расчету статисик.

Начальным моментом m_k порядка k случайной величины x называется математическое ожидание (среднее значение) ее k-той степени:

$$m_k = \int_{-\infty}^{\infty} x^k f(x)dx, \qquad (1.12)$$

где $f(x)$ – функция плотности распределения.

В случае выборочного эмпирического распределения

$$m_k = \sum x_i^k p_i = \frac{1}{n}\sum x_i^k, \qquad (1.13)$$

где x_i – i-ое значение случайной величины в выборке; $p_i = \frac{1}{n}$ – статистическая вероятность члена выборки; n – объем выборки (число ее членов).

Центральным моментом μ_k порядка k случайной величины x называется математическое ожидание (среднее значение) ее k-ой степени, рассчитанное относительно центра тяжести распределения, т.е. математического ожидания (*мо*) или среднего значения

$$\mu_k = \int_{-\infty}^{\infty} (x - мо)^k f(x)dx. \qquad (1.14)$$

В случае выборочного (эмпирического) распределения

$$\mu_k = \sum (x_i - \bar{x})^k p_i = \frac{1}{n}\sum (x_i - \bar{x})^k, \qquad (1.15)$$

где $\bar{x} = m_1$ – среднее значение x в выборке.

Обычно рассчитываются и используются начальные и центральные моменты не выше 4 порядка. Из приведенных формул следуют следующие соотношения между начальными и центральными моментами:

$$\begin{aligned} &\mu_0 = 1,\ \mu_1 = 0; \\ &\mu_2 = m_2 - m_1^2; \\ &\mu_3 = m_3 - 3m_1 \cdot m_2 + 2m_1^3; \\ &\mu_4 = m_4 - 4m_1 \cdot m_3 + 6m_1^2 m_2 - 3m_1^4. \end{aligned} \qquad (1.16)$$

Эти формулы используются тогда, когда технически выгодно сначала по выборке рассчитать начальные моменты, а затем по ним центральные моменты или в иных необходимых случаях.

Смешанным начальным моментом m_k^* порядка $k=(k_1+k_2)$ двух случайных величин x и y называется среднее значение произведения $x_i^{k_1}$ и $y_i^{k_2}$:

$$m_k^* = \frac{1}{n}\sum x_i^{k_1} \cdot y_i^{k_2} \ . \qquad (1.17)$$

Смешанным центральным моментом μ_k^* порядка $k=(k_1+k_2)$ двух случайных величин x_i и y_i называется среднее значение их произведений, взятых относительно своих средних $\bar{x}$ и $\bar{y}$, т.е.

$$\mu_k = \frac{1}{n}\sum (x_i - \bar{x})^{k_1} (y_i - \bar{y})^{k_2} . \qquad (1.18)$$

В практике широко используются смешанные моменты второго порядка, когда k=2, k_1=1 и k_2=1. Они, например, необходимы для расчета такого важного показателя как коэффициент корреляции между СВ x и y, который характеризует силу линейной корррелляционной связи между ними.

Оценки статистик, которые вычисляются по приведенным выше формулам моментов различных порядков, называются точечными, так как задаются полученным конкретным числом, т.е. точкой. Точечная оценка статистики обязательно должна сопровождаться оценкой ее погрешности, как это будет показано ниже. Кроме точечных возможно также вычисление интервальных оценок, когда они задаются вероятным интервалом статистики. Например, с заданной вероятностью p найденная оценка математического ожидания (истинное среднее значение) $\bar{x}$ лежит в интервале [a, b], т.е. $\bar{x} \in a..., b$. С интервальными оценками мы познакомимся позже.

К точечным оценкам всех статистик параметров распределений предъявляются общие требования *состоятельности, несмещенности и эффективности*.

Оценка статистики называется *состоятельной,* если по мере роста объема выборки n оценка будет приближаться (сходиться по вероятности) к теоретическому значению статистики или параметра. Состоятельность есть положительное качество оценки (т.е. фор-

мулы, по которой она вычисляется), гарантия, что оценкой можно пользоваться уверенно, по крайней мере, при большом *n*.

Оценка статистики называется *несмещенной*, если при любом числе наблюдений *n* ее математическое ожидание точно равно теоретическому значению оценки. Несмещенность означает, что даже при малом числе наблюдений *n*, формула для оценки статистики исключает возможность возникновения систематической погрешности (смещения) в расчетах оценки. Несмещенность гарантирует отсутствие систематической погрешности при любом *n*.

Оценки параметров по выборке могут быть найдены различными способами, т.е. по формулам, в основе которых могут лежать различные подходы. Например, оценку среднего можно вычислить как обычным путем, используя все выборочные данные, так и найти как полусумму:

$$\bar{x} = \frac{1}{2}\left(x_{макс} + x_{мин}\right).$$

Обе оценки обладают свойством состоятельности и несмещенности. Однако ошибка оценки во втором случае существенно больше и поэтому первый способ ее расчета является предпочтиельным.

Оценка статистики называется *эффективной*, если по сравнению с другими оценками, она обладает наименьшей дисперсией. Таким образом, свойство эффективности является относительным, и его можно использовать, если есть реальная возможность получить оценки различными способами, а затем выбрать из них более эффективную.

В следующем пункте будут рассмотрены основные статистики распределений, основанные на сделанных определениях начальных и центральных статистических моментов.

1.2.2. Статистики положения центра распределения случайной величины - среднее, мода и медиана

Из основных статистик, характеризующих различные свойства распределения СВ, обычно рассматриваются прежде всего те, которые определяют положение центра распределения. Таких статистик три: *среднее значение* (или математическое ожидание), *мода и медиана.* При этом термин математическое ожидание используется обычно для среднего значения СВ, определенного по генеральной совокупности, т.е. когда он выступает как параметр (см. 1.1.2). Напомним, что все статистики, рассчитываемые по выборкам, носят названия *оценок* соответствующих параметров, определяемых по генеральным совокупностям

Среднее значение СВ есть начальный момент первого порядка, который, согласно (1.13), для несгруппированной выборкам может быть рассчитан по формуле

$$\bar{x} = m_1(x) = \frac{1}{n}\sum x_i \ , \qquad (1.19)$$

где x_i – i-тое значение СВ в выборке, n – объем выборки, равный сумме ее членов.

Оцека среднего значения формуле по (1.19) обладает свойствами состоятельности, несмещенности и эффективности. Если она вычислена по ораниченной выборке, то обозначается как $\bar{x}$. Если точное среднее значение найдено по генеральной совокупности, то оно называется математическим ожиданиеми и обозначается как *мо*.

Среднее значение часто называют *центром тяжести* распределения, т.к. вокруг него распределяются (группируются) все остальные значения СВ. В этом смысле $\bar{x}$ может выступать в качестве основной характеристики выборки, т.е. представлять приближенно выборку в целом. Например, в группе из 20 студентов их рост колеблется от 147 до 181 см, при этом среднее значение равно 165 см. В первом приближении можно говорить, что рост студента в данной группе составляет 165 см.

Наглядно геометрическая интерпретация среднего видна в графиках функции плотности *f(x)*. Так, на рис. 1.6, где показаны графики *а*) симметричной; *б*) и *в*) право - и левоасимметричной; *г*) экспоненциальной и *д*) U-образной кривой *f(x)*, при этом положение $\bar{x}$ неизменно характеризует положение «средней области» распределения.

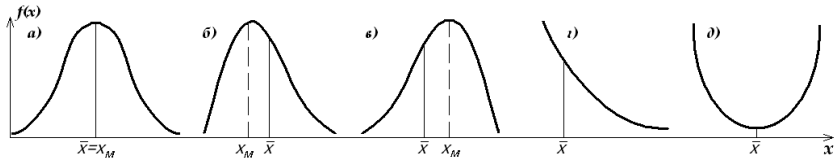


Рис. 1.6. Положение среднего значения $\bar{x}$ и модального значения $x_м$ при различных видах кривой плотности *f(x)*.

Модой называется значение СВ $x_м$, которое соответствует точке максимума на кривой *f(x)* (рис. 1.6*а, б* и *в*). Это наиболее часто встречающееся в выборке значение *x*. Как и среднее, она также характеризует положение центра распределения только с несколько иной позиции. Однако могут быть распределения, не имеющие моды, например, если *f(x)* имеет вид экспоненты или обратной экспоненты (рис. 1.6*г*) или имеющих антимоду (рис. 1.6*д*). В случае симметричного распределения (рис. 1.6*а*) $\bar{x}$ и $x_м$ совпадают. При правоасимметричном распределении (рис. 6*б*) $x_м$ лежит левее, а при левоасиммметричном (рис. 1.6*в*) $x_м$ лежит правее $\bar{x}$.

Когда выборка сгруппирована, то в ней легко определяется модальный класс, имеющий наибольшую частоту или вероятность $p_i=n_i/n$. В этом классе находится мода, значение которой находятся по специальным интерполяционным формулам [23]. Моды может быть две и более, тогда говорят о мультимодальных (много модальных) распределениях. Причиной нескольких мод является то, что выбрка СВ формируется под воздействием не одного, а нескольких главных факторов. В случае появления в одном распределении двух и более мод, прежде всего, ищутся физические причины их появления. Во многих случаях наличие двух мод дает основание разделить выборку на две, соответствующие основным причинам этих мод, и рассматтривать каждую из них отдельно.

Например, в западной части Иссык-Кульской котловины бури возникают под воздействием двух основных причин – это процессы развития гроз в теплый период года и процессы образования местных штормовых ветров улан при вторжениях холода в котловину в холодный период года.Так, в 2017 г. средняя длительность наблюдавшихся 59 бурь (скорости 10 м/с и выше) в международном аэропорту Иссык-Куль (район пос. Тамчи) оказалась равной 3,3 ч, максимальная достигала 27,5 ч, а повторяемости их различных градаций (в часах) составляли:

часы	до 0,5	0,5-1	1-2	2-3	3-5	5-10	10-15	15-20	20-25	25-30
частота	28	7	4	4	3	9	2	1	0	1
повт.,%	47,5	11,9	6,8	6,8	5,1	15,3	3,4	1,7	0,0	1,7

Видно, что большинство бурь кратковременны, почти 50% из них длится в пределах всего 0,5 ч (первый модальный клсасс) - это кратковременные усиления вера при грозах и такие же усиления при быстром прохождении холодных фронтов 2-рода. Но кроме этой первой моды прослеживается вторая, соответствующая модальному классу 5-10 ч (15,3%). Это говорит о том, что режим бурь здесь действительно формируется под действием двух причин, одна из них которых – это кратковременные усиления ветра при грозах и холодных фронтах 2-рода (эти бури абсолютно преобладают по повторяемости), а вторая связана с мощными и интенсивными вторжениями холода в котловину, что приводит к развитию достаточно длительных бор-уланов. Так, в 2017 г. только две из 13 бурь длительностью более 5 ч наблюдались летом, остальные 11 - осенью и весной, когда частота возникновения уланов максимальна

Медианой называется значение СВ $x_{ме}$, которое делит распределение пополам по накопленной вероятности или обеспеченности, т.е. слева и справа от нее лежит 50% объе-

ма распределения, так что для $x_{ме}$ значение F_i =0,5. Медиана стоит точно в центре ранжированного ряда. Отсюда номер члена медианы (ранг) в ранжированном ряду $r_{ме}$ равен

$$r_{ме} = \frac{1}{2}(n+1). \qquad (1.20)$$

Если ряд содержит нечетное число членов n, то $r_{ме}$ – целое число; если n – четное, то $r_{ме}$ – дробное число. Например, в нечетном ряду из 27 членов, $r_{ме}$=14 члену (значение $x_{14}=x_{ме}$, т.е. будет медианным), в четном ряду из 28 членов $r_{ме}$=14,5 (т.е. $x_{ме}$ будет соответствовать полусумме x_{14} и x_{15}). Для сгруппированной выборки сначала по распределению накопленных частот или вероятностей находят медианный интервал, а затем внутри него $x_{ме}$ можно оценить по интерполяционной формуле [23].

В заключение отметим, что среднее, мода и медиана, каждая со своей стороны, характеризует свойства центра распределения СВ (см. например, рис. 1.7, где все три характеристики показаны для правоасимметричного распределения) и взаимно дополняют друг друга. Так, при резко асимметричных распределениях $\bar{x}$ находится в классах, не содержащих наибольшие частоты и здесь весьма полезно наряду с $\bar{x}$ вычислить $x_м$. Точно также при большом разбросе крайних членов выборки среднее может стать ненадежной характеристикой центра распределения и тогда, наряду с $\bar{x}$, следует обязательно вычислть и использовать $x_{ме}$.

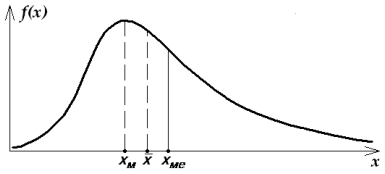


Рис. 1.7. Расположение $\bar{x}$, $x_м$ и $x_{ме}$ при правоасимметричном рас-

1.2.3. Статистики рассеивания распределения случайной величины

Кроме характеристик положения центра распределения СВ – среднего, медианы и моды, – необходимо знать и другие статистики, описывающие различные свойства распределений. Следующими по значимости являются характеристики степени рассеивания СВ около центра распределения. Ими являются дисперсия и определяемые ею средние квадратическое отклонение (СКО) и коэффициент вариации.

Дисперсией СВ (введем на нее обозначение *D(x)*) называется второй центральный момент μ_2, определяемый формулой (1.15) при k=2,

$$D(x) = \mu_2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \qquad (1.21)$$

для несгруппированной выборки,

Кроме того, по (1.16) дисперсия *D(x)* может быть выражена через начальные моменты m_2 и m_1

$$D(x) = \mu_2 = m_2 - m_1^2 . \qquad (1.22)$$

На рис. 1.8 показаны два распределения *f(x)*, имеющие одинаковые $\bar{x}$, но разные дисперсии *D(x)*. Распределение $f_2(x)$, имеющиее больший разброс (больший размах ветвей), имеет и большую дисперсию $D_2(x)$.

Для характеристики рассеивания нельзя было взять первый центральный момент (что, казалось бы, проще), т.к. $\mu_1 \equiv 0$. Подходящим для этой цели является именно μ_2, тем более, что, как увидим далее, он играет ключевую роль в методе наименьших квадратов, являющимся важнейшим в теории вероятностей и математической статистике.

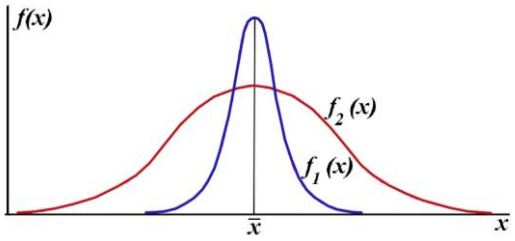


Рис.1.8. Два распределения $f_1(x)$ и $f_2(x)$ с одинаковыми $\bar{x}$, но разными дисперсиями: $D_2(x)>D_1(x)$.

Дисперсия имеет размерность квадрата СВ [x^2], что неудобно на практике. Желательно, чтобы мера рассеивания, как и среднее, имела размерность самой СВ, т.е. [x^1]. Такую величину очень просто получить, если извлечь из дисперсии квадратный корень, приписав ему только знак +. Эта характеристика рассеивания получила название среднего квадратического отклонения (СКО) и обозначается через $\sigma(x)$ и равна:

$$\sigma(x) = \sqrt{D(x)} = \left[\frac{1}{n}\sum (x_i - \bar{x})^2\right]^{0,5} \qquad (1.23)$$

для несгруппированной выборки,

Точно также, исходя из (1.22), имеем:

$$\sigma(x) = \sqrt{D(x)} = \left[m_2 - m_1^2\right]^{0,5} . \qquad (1.24)$$

Например, на двух близко расположенных станциях Бишкек (756 м) и Чон-Арык (1110 м) имеем следующие характеристики распределения годовых сумм осадков по многолетним данным:

Бишкек (756 м) – $\bar{x}_Б$ =422 мм, $\sigma_Б(x)$=97 мм,

Чон-Арык (1110 м) – $\bar{x}_Ч$ =616 мм, $\sigma_Ч(x)$=105 мм.

Наглядно видно, что, несмотря на не очень большое различие высот, в Чон-Арыке осадков выпадает на 46% больше, чем в Бишкеке. С точки зрения среднего годового коли-

чества осадков, Чон-Арык имеет гораздо более влажный климат, чем Бишкек. Но, если оценить изменчивость годовых сумм на обеих станциях, то они разнятся не существенно: $\sigma_Б(x)$=97 мм, а $\sigma_ч(x)$=105 мм. Таким образом, колебания осадков на станциях от года к году около их норм – 422 мм и 616 мм – практически равны, т.е. эта сторона климатических условий одинакова.

Однако сравнивать размах распределений по значениям *σ(x)* можно только для СВ одинаковой размерности. Нельзя, например, сравнить *σ(x)*, вычисленные для распределений температуры (°С) и осадков (мм). Желательно получить какую-то *безразмерную* характеристику изменчивости, которую можно применять одновременно для анализа СВ разной размерности. Такой безразмерной мерой изменчивости является *коэффициент вариации c(x)*, определяемый формулой:

$$c(x) = \frac{\sigma(x)}{\bar{x}}; \qquad c(x)\% = \frac{\sigma(x)}{\bar{x}} \cdot 100\%. \tag{1.25}$$

Коэффициент вариации вычисляется только для существенно положительных СВ ($x_i \geq 0$). В этом его единственный недостаток. Он является относительной долей изменчивости. По определению $c(x) \geq 0$ и, как видно из (1.25) характеризует изменчивость СВ в долях $\bar{x}$ от *σ(x)*. Так, например, если *c(x)*=60%, то это означает, что *σ(x)* составляет 60% от $\bar{x}$.

Так, в предыдущем примере для МС Бишкек получим *c(x)*=23%, а для МС Чон-Арык – *c(x)*=17%. Но если нам надо, например, сравнить изменчивость годовой скорости ветра и осадков (разные СВ) на станции Бишкек, то для этого можно использовать только коэффициент вариации. Так, имеем для Бишкека *c* (осадков)=23%, а *c* (ветра)=16%. Как видно, для обоих СВ различия коэффициентов вариации относительно малы, т.е. они имеют примерно одинаковую изменчивость.

Приведенные выше формулы (1.21)-(1.25) дают состоятельные, но смещенные оценки дисперсии, СКО и коэффициента вариации. Как показывет теория, формулы для несмещенных оценок этих характиристик имеют вид:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2, \tag{1.26}$$

$$s = \sqrt{s^2} = \left[\frac{1}{n-1}\sum_i (x_i - \bar{x})^2\right]^{0,5} \tag{1.27}$$

$$c = \frac{s}{\bar{x}} \quad \text{или} \quad c(\%) = \frac{s}{\bar{x}} 100\%. \tag{1.28}$$

В этих формулах буквы $\sigma^2(x)$ и *σ(x)* для обозначения дисперсии и СКО заменены на $s^2(x)$ *и s(x)*, чтобы специально подчеркнуть, что мы имеем дело с *оценками*, рассчитанны-

ми по ограниченным выборкам, а не точными значениями параметров этих характеристик, найденными по генеральной совокупности. Этот подход в различии обозначений точных *параметров и их оценок* для этих характеристик будем применять и в последующем изложении.

Имеено по формулам (1.26)-(1.28) надо рассчитывать значение оценок дисперсии, СКО и коэффициента вариации при использовании выборочных данных. Казалось бы различие результатов расчетов по (1.21)-(1.24) и (1.26)-(1.28) за счет деления на n или на $(n-1)$ мало и им можно принебречь. Это действительно так при достаточно больших выборках (условно при $n \geq 30$), но в тех случаях, когда n имеет порядок 10 или даже 5 случаев оно составляет для дисперсии соответственно 10 и 20%, что уже существенно и надо учитывать.

Таким образом, формулы (1.21)-(1.24) дают состоятельные, но смешенные на n/(n-1) оценки дисперсии, СКО и вариации, тогда как формулы (1.27)-(1.28) дают их состоятельные и несмещенные оценки.

1.2.4. Статистики асимметрии и эксцесса распределения случайной величины

Большинство распределений не симметричны относительно $\bar{x}$. На графике $f(x)$ для таких распределений левая и правая ветви кривой не одинаковы. Чтобы охарактеризовать степень асимметрии вводится специальная безразмерная характеристика – коэффициент асимметрии или скошенности распределений. Обычно он обозначается через $A(x)$ или A.

Коэффициентом асимметрии A(x) называется безразмерное отношение

$$A(x) = \frac{\mu_3(x)}{\sigma^3(x)}, \qquad (1.29)$$

где μ_3 – третий центральный момент, который может иметь знаки $\pm$.

Если $A(x)>0$ (т.е. $\mu_3>0$), то правая ветвь распределения длиннее левой и говорят, что оно правоасимметрично или имеет положительную асимметрию. Если $A(x)<0$ ($\mu_3<0$), то распределение левоасимметрично или имеет отрицательную асимметрию. Для симметричных распределений $A(x)=0$ ($\mu_3=0$).

Формула (1.29) дает состоятельную, но смещенную оценку асимметрии. Состоятельной и несмещенной оценкой коэффициента асимметрии, которой надо пользоваться на практике, будет формула

$$A = \frac{1}{n-1}\sum_i \frac{(x_i - \bar{x})^3}{s^3}, \qquad (1.30)$$

где *s* – выборочная оценка СКО по формуле (1.27). На практике различают почти симметричные распределения ($A(x)\approx 0$), слабо, умеренно и сильно асимметричные распределения. В эти качественные понятия в зависимости от типа распределения, характера решаемой задачи и объема исходных данных (т.е. точности оценки *A(x)*) вкладываются различные численные значения модулей *A(x)*. Ориентировочно во многих реальных метеорологических задачах можно *условно считать:*

$|A(x)|<0{,}1$ – практическое отсутствие асимметрии,

$|A(x)|=0{,}1\text{-}0{,}2$ – слабая асимметрия,

$|A(x)|=0{,}3\text{-}0{,}5$ – умеренная асимметрия,

$|A(x)|>0{,}6\text{-}1{,}0$ – сильная асимметрия,

$|A(x)|>1{,}0$ – очень сильная асимметрия.

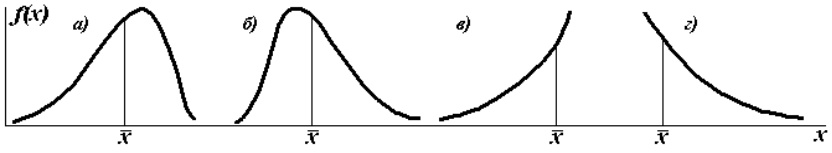


Рис. 1.9. Общий вид *f(x)* имеющих разную асимметрию:
а) и *б*) умеренно лево- и право асимметричные распределения;
в) и *г*) сильно, лево- и право асимметричные распределения.

На рис. 9 показаны графики *f(x)* с умеренной левой (*а*) и правой (*б*) асимметрией, а также *f(x)*, имеющие вид обратных экспоненциальных зависимостей с очень сильной правой (*г*) и очень сильной левой (*в*) асимметрией.

Физически асимметрия распределения метеорологической величины означает, что разброс ее влево и вправо от среднего значения неодинаков и это надо обязательно учитывать при статистико-климатическом анализе. Например, коэффициенты асимметрии суточного количества осадков на МС Бишкек равны:

Месяц:	*Январь*	*Апрель*	*Июль*	*Октябрь*	*Год*
A(x):	2,2	2,0	3,1	1,8	2,0

Это говорит о том, что распределения суточных количеств осадков для всех месяцев и года в целом сильно право асимметричны, т.е. разброс отдельных суточных сумм осадков относительно их многолетних средних (климатических норм) гораздо сильнее в правую сторону от $\bar{x}$, чем в левую. Более того, вправо он не ограничен (что вытекает из природы этой метеорологической величины), т.к. всегда существует пусть очень малая веро-

ятность того, что в следующий раз будет превышено наблюденное до этого максимальное значение осадков.

Имеется еще одна важная характеристика формы распределения – коэффициент эксцесса, который характеризует островершинность функции плотности *f(x) по сравнению с нормальным законом распределения*.

Коэффициентом эксцесса E(x) называется безразмерное выражение

$$E(x) = \frac{\mu_4(x)}{\sigma^4(x)} - 3, \qquad (1.31)$$

где $\mu_4(x)$ – четвертый центральный момент, определяемый (1.15).

Для нормального закона распределения, который имеет самое широкое применение в теории и на практике, $\mu_4/\sigma^4 = 3$. Поэтому, чтобы охарактеризовать эксцесс распределения относительно нормального закона в (1.31) введено вычитаемое – 3. Таким образом, для нормального закона *E(x)=0*, для всех остальных законов распределений *E(x)* может быть нулевым, больше или меньше нуля.

Формула (1.31) дает состоятельную, но смещенную оценку эксцесса. Состоятельной и несмещенной оценкой коэффициента эксцесса, которой надо пользоваться на практике, будет получаемая по формуле (1.32):

$$E = \frac{1}{n-1}\sum_i \frac{(x_i - \bar{x})^4}{s^4} - 3, \qquad (1.32)$$

где s - выборочная оценка СКО по формуле (27).

На рис. 10 показаны графики трех функций плотности *f(x)*, для которых эксцесс равен нулю, больше нуля (островершинные распределения по сравнению с нормальным) или меньше нуля (плосковершинные распределения по сравнению с нормальным).

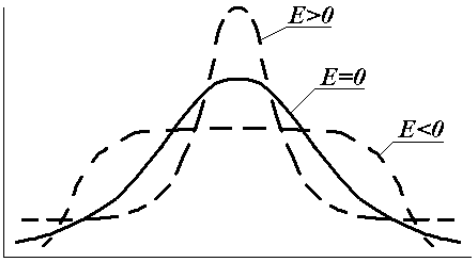


Рис. 1.10. Кривые распределений *f(x)* с различным эксцессом: *E=0* (нормальная кривая), *E>0* (островершинная кривая), *E<0* (плосковершинная кривая).

Точно так же, как и для *A(x)*, в метеорологических задачах условно можно принять следующие градации для характеристики островершинности эмпирических распределений:

$|E|$ <0,2 – практически эксцесс отсутствует,

$|E|$ =0,2-0,3 – слабый эксцесс,

$|E|$=0,3-0,6 – умеренный эксцесс,

$|E|$=0,6-1,0 – сильный эксцесс,

$|E|$>1 – очень сильный эксцесс.

Если $E<-2$, то возникает многовершинность (полимодальность) распределения и, как уже указывалось в п.1.2.2, для анализа выборку часто следует дифференцировать на более простые «однопричинные» выборки и анализировать их по отдельности.

В заключение еще раз подчеркнем, что все рассмотренные основные статистики распределений:

- положение центра (среднее, мода и медиана),
- рассеивание (*СКО* и коэффициент вариации),
- асимметрия (*A(x)*),
- эксцесс (*E(x)*),

легко вычисляются по эмпирическим выборкам. Во многих задачах их знания достаточно, чтобы описать основные свойства СВ и не прибегать к необходимости поиска и более сложных расчетов для нее аппроксимирующего закона распределения *E(x)*.

1.2.5. Средние квадратические ошибки основных статистик

Выше уже подчеркивалось, что все численные значения статистик, получаемых по выборкам (эмпирическим распределениям), называются оценками в отличие от их точных значений или параметров, соответствующих генеральным совокупностям. Поэтому оценка любой статистики *должна непременно сопровождаться оценкой ее ошибки*, а еще лучше – доверительного интервала. Статистическая ошибка возникат не за счет погрешности вычислений, а есть закономерное следствие, вытекающее из ограниченности объема выброрки *n*. В самом общем выводе, чем меньше *n*, тем больше статистическая ошибка определения параметра СВ по выборке. Кроме того, для среднего, дисерсии, СКО и вариации она возрастает с увеличением размаха колебаний СВ.

К сожалению, в общем случае вычисление точных значений ошибок статистик для произвольного распределения далеко не простая задача, которая не решена до настоящего времени. Точные формулы средних квадратических ошибок имеются только для нормального закона:

1. Ошибка среднего значения – $\sigma_{\bar{x}}$

$$\sigma_{\bar{x}} = \frac{\sigma(x)}{\sqrt{n}}. \qquad (1.33)$$

2. Ошибка среднего квадратического отклонения – $\sigma_{\sigma(x)}$

$$\sigma_{\sigma(x)} = \frac{\sigma(x)}{\sqrt{2n-1}}. \qquad (1.34)$$

3. Ошибка коэффициента вариации – $\sigma_{c(x)}$

$$\sigma_{c(x)} = \frac{c(x)}{\sqrt{2n}}\sqrt{1+c^2(x)}. \qquad (1.35)$$

4. Ошибка коэффициента асимметрии – $\sigma_{A(x)}$

$$\sigma_{A(x)} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}. \qquad (1.36)$$

5. Ошибка коэффициента эксцесса – $\sigma_{E(x)}$

$$\sigma_{E(x)} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} \approx \sqrt{\frac{24}{n}} = 2\sigma_{A(x)}. \qquad (1.37)$$

6. Кроме того, средняя квадратическая ошибка статистической вероятности (частости, повторяемости) p_i имеет вид:

$$\sigma_p = \sqrt{\frac{p_i(1-p_i)}{n}} \qquad (1.38)$$

На практике, за неимением лучшего решения, часто приходится пользоваться формулами (1.33)-(1.38) для вычисления средних квадратических значений ошибок для любых распределений. Значения ошибок в этом случае являются приближенными (т.е. оценками), имеющими разную степень точности, в зависимости от типа исследуемого распределения. Однако с этим приходится мириться, т.к. приведение оценки значения статистики без сопровождения ее оценкой ошибки является вообще *статистически не корректным*.

Синтаксис записи оценок статистики с учетом ее ошибки имеет вид: $\bar{x} \pm \sigma_{\bar{x}}$, $\sigma(x) \pm \sigma_{\sigma(x)}$ и т.д. Например, по выборке вычислено среднее значение скорости ветра при бурях на метеостанции $\bar{V}$, равное 19,1 м/с, и ошибка этого среднего $\sigma_{\bar{V}}$ по (1.33), равная ±0,2 м/с. Тогда, запись должна иметь вид: $\bar{V}$ =19,1±0,2 м/с. В этом случае любому, даже неискушенному пользователю, становится ясным - насколько точно оценено среднее значение скорости ветра в бурях. Именно в этом и состоит смысл сопровождения оценок статистик оценками их ошибок.

1.2.6. Технология практического расчета статистик с использованием Excel

Технология практического расчета статистик может быть крайне разнообразна в зависимости от наличия вычислительных средств, программ и задач исследований. Поэтому приведем только перечень основных программ *Excel* с необходимыми краткими пояснениями.

В перечне статистических функций Excel-97 и последующих версий приводятся следующие программы для расчета или определения статистик распределений (в скобках заглавными буквами дается название соответствующей программы):

- ранг (номер) члена выборки в ранжированном ряду (РАНГ),
- минимальное значение члена выборки (МИН),
- максимальное значение члена выборки (МАКС),
- модальное значение (МОДА), но только при наличии в выборке повторяющихся членов, в противном случае выдается значение – ошибка,
- медианное значение (МЕДИАНА),
- частота (число случаев) попадания значения x в заданные классы сгруппированной выборки, если задать значения *верхних границ* классов (ЧАСТОТА),
- оценка среднего значения – $\bar{x}$ (СРЗНАЧ), определяемая по не сгруппированной выборке по формуле (19),
- оценка среднего квадратического отклонения s (СТАНДОТКЛОН.В), определяемая по не сгруппированной выборке по формуле (27),
- генеральное значение среднего квадратического отклонения σ (СТАНДОТКЛОН.Г), определяемого по по формуле (23),
- оценка выборочной дисперсии s^2 (ДИСП), определяемая по формуле (26),
- генеральная дисперсия σ^2 (ДИСПР), определяемая по формуле (21),
- оценка коэффициента асимметрии A (СКОС), определяемого по формуле:

$$A = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3 ,$$

- оценка коэффициента эксцесса E (ЭКСЦЕСС), определяемого по формуле:

$$E = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} .$$

К сожалению, аналогичный расчет средних квадратических ошибок статистик в Excel не предусмотрен. Поэтому их надо рассчитывать по формулам (33)-(38), которые справедливы для нормального закона, в режиме набора формул.

Тема 2. НОРМАЛЬНЫЙ ЗАКОН, t- РАСПРЕДЕЛЕНИЕ СТЬЮДЕНТА, χ^2-ПИРСОНА И НЕКОТОРЫЕ ИХ ПРАКТИЧЕСКИЕ ПРИМЕНЕНИЯ

Нормальный закон распределения является основным законом математической статистики, к нему сходятся по вероятности многие другие законы при неограниченном увеличении числа наблюдений. Он так же играет ключевую роль в построении различных статичтических критериев. В этом плане к нему по практической значимости примыкают три других важных распределения: t – распределение Стьюдента, χ^2 – распределение Пирсона и F – распределение Фишера. В теме 2 будут рассмотрены основные свойства первых трех распределений и некоторые их практические применения при анализе метеорологических рядов. Другие их применения, а также использование F – распределение Фишера, будут рассматриваться далее во всех остальных темах. Следует также отметить, что нормальный закон в некоторых случаях может быть использован и непосредственно для приближенного описания (аппроксимации) метеорологических величин, правда, это в основном касается только их средних значений. Ксожалению, больнинсво метеорологических величин не подчиняется нормальному закону.

Глава 2.1. НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ И ЕГО ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ

2.1.1. Условия возникновения нормального закона, дифференциальная и интегральная функции распределения нормального закона

Нормальный закон распределения (закон Гаусса) играет исключительно важную роль, как в теории статистики, так и в ее практических приложениях. Это наиболее часто

встречающийся в практике закон распределения. Главной его особенностью является то, что он служит *предельным законом*, к которому приближаются многие другие законы при тех или иных допущениях.

Пусть на результате испытания (на проявление метеорологического явления или ситуации) сказываются одновременно множество причин, подчиненных каким угодно законом распределений. Если ни одна из причин не является решающей, и влияние всех их относительно одинаково, то суммарный результат этого влияния будет подчиняться нормальному закону. Например, температура воздуха в любой местности в течение года изменяется очень прихотливо под воздействием самых разнообразных причин. В сумме они формируют среднегодовую температуру, которая является случайной величиной и изменяется от года к году. Распределения средних годовых температур в самых различных местностях хорошо подчиняется нормальному закону.

В течение года реализуется *весь набор возможных* формирующих режим температуры факторов, но в течение одного месяца он ограничен спецификой метеорологического *режима сезона* и, следовательно, составляющие сезонных факторов уже возможно неравнозначны. Среди них могут преобладать отдельные причины, например, в зимние месяцы для Средней Азии – холодные северо-западные вторжения арктического воздуха. Поэтому распределения средних месячных температур по сравнению со средними годовыми температураи хуже подчиняются нормальному закону и могут не подчиняться ему.

Примеры в технике не менее наглядны. Пусть производится контрольная стрельба по мишени по возможности при полном соблюдении одинаковости условий опыта: одно и то же изделие (автомат, винтовка и др.), одна и та же партия патронов, методика прицеливания, внешние условия и т.д. Однако все же на результат каждого выстрела будут накладываться многие, хотя и малые, но не учитываемые причины: различные массы и размеры пуль, массы зарядов в отдельных патронах, колебания плотности воздуха, случайные погрешности прицеливания и др. В результате никогда не удается добиться при большом числе опытов 100% попадания в десятку (почти точку), а величина разброса попаданий относительно центра подчиняется нормальному закону.

К сожалению, абсолютное большинство *исходных* метеорологических величин не подчиняется нормальному закону. Ему подчиняются многие *средние годовые значения* или другие виды средних. Однако и в метеорологии он находит самое широкое применение в различного рода важных методических решениях. Применение математической статистики *на практике без использования этого заеона является просто немыслимым*.

Нормальный закон хорошо изучен и обладает достаточно простыми свойствами. Его функция плотности $f(x)$ описывается выражением:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-мо)^2}{2\sigma^2}}, \qquad (2.1)$$

где x – непрерывная случайная переменная, $x \in]-\infty, \infty[$; σ^2 – дисперсия; *мо* – математическое ожидание.

Соответственно интегральная функция распределения *F(x)* имеет вид:

$$F(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \int_{-\infty}^{x} e^{-\frac{(x-мо)^2}{2\sigma^2}} dx. \qquad (2.2)$$

Интеграл (2.2) в элементарных функциях не выражается, однако это не вызывает каких-либо технических трудностей в его использовании, поскольку он с любой требуемой точностью вычисляется численными способами, а подробные таблицы этого интеграла имеются в нормированной форме.

Кривая распределения (2.1) имеет симметричный колоколообразный вид (рис. 2.1а), ее ветви уходят в бесконечность. Максимальная ордината кривой равна $\frac{1}{\sigma\sqrt{2\pi}}$ и соответствует точке $x=мо=x_м=x_{ме}$, т.е. математическое ожидание (среднее значение x) совпадает с модой и медианой. Вид графика функции *F(x)* показан на рис. 2.1б.

Центральные моменты закона равны:

$$\mu_2=\sigma^2;\ \mu_3=0;\ \mu_4=3\sigma^4.$$

Поэтому асимметрия и эксцесс у нормального закона нулевые

$$A = \frac{\mu_3}{\sigma^3} = 0, \qquad (2.3)$$

$$E = \frac{\mu_4}{\sigma^4} - 3 = 0. \qquad (2.4)$$

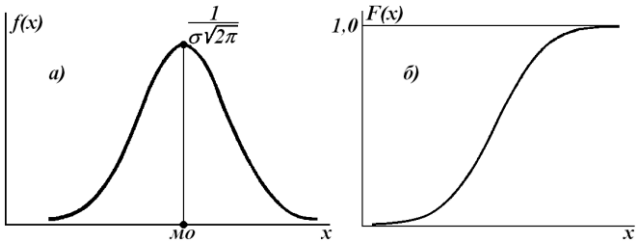


Рис. 2.1. Графики общего вида функций: *а*) *f(x)* и *б*) *F(x)* для нормального закона.

Таким образом, параметрами закона являются два: $мо = \bar{x} = m_1$ и $\sigma^2 = D(x)$, т.е. среднее или математическое ожидание и дисперсия (СКО). Среднее определяет положение центра распределения на оси x, а дисперсия – его размах (рис.2.2).

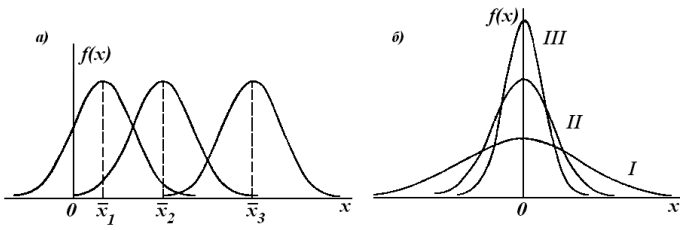


Рис. 2. 2. Влияние параметров: *а*) *мо*$=\bar{x}$ на положение кривой *f(x)* на оси *x* и *б*) параметра σ^2 на размах кривой *f(x)*.

При увеличении σ кривая *f(x)* становится все более плоской, растягиваясь вдоль оси *x*, при уменьшении σ – кривая *f(x)* вытягивается вверх, одновременно сжимаясь с боков и становясь иглообразной (кривые I, II и III на рис. 2.2*б*).

Для обозначения нормального закона в форме (2.2), т.е. для *F(x)* часто используется сокращенная запись: *N(x, мо,* σ*)* или *N(мо,* σ*)*, где *x* – бегущая случайная переменная, а *мо* и σ – численные значения его параметров.

2.1.2. Стандартный нормальный закон. Приближенные критерии нормальности распределения. Компьютерные реализации нормального закона

Итак, нормальный закон *N(x, мо,* σ*)* зависит от двух параметров: *мо* $\approx \bar{x}$ и $\sigma \approx s$. Здесь *мо* и σ – параметры, т.е. точные значения среднего и СКО, когда они определены по генеральной совокупности, а $\bar{x}$ и *s* – их оценки, рассчитанные по выборке. Встает заманчивый вопрос – нельзя ли от переменных параметров *мо* $\approx \bar{x}$ и $\sigma \approx s$ перейти к каким-то таким их постоянным значениям так, чтобы использовать этот переход для унификации вычислений по нормальному закону. Такой переход прост и называется *нормировкой*.

Введем новую *нормированную переменную z*:

$$z = \frac{x - мо}{\sigma(x)}, \tag{2.5}$$

где $z \in]-\infty, \infty[$.

Тогда, можно показать, что

$$\begin{aligned} мо(z) &\approx \bar{z} \equiv 0, \\ \sigma(z) &\approx s(z) \equiv 1. \end{aligned} \tag{2.6}$$

В результате функция плотности и интегральная функция нормального закона для переменной *z* по (2.1) и (2.2) запишутся в виде:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}, \tag{2.7}$$

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz \,. \qquad (2.8)$$

Закон *f(z)* и *F(z)*, выражаемый формулами (2.7) и (2.8), получил название *стандартного* нормального закона (или *нормированного* нормального закона), который часто обозначается как *N(z, 0, 1)* или *N(0, 1)*.

С помощью простой нормировки переменной по (2.5) достигнут очень важный практический результат: стандартный нормаольный закон имея свойство (2.6), зависит тоько от переменной z. Поэтому для него можно составить подробные таблицы функций *f(z)* и *F(z)* для практического использования, как это показано в табл. 2.1 и 2.2 для z_i от 0 до 4,0.

Таблица 2.1

Нормальное стандартное распределение

Плотность вероятности нормированного нормального распределения $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,39894	39892	39886	39876	39862	39644	39822	39797	39767	39733
0,1	39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0,2	39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0,3	38139	38023	37903	37780	37654	37524	37391	37255	37115	36973
0,4	36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0,5	35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0,6	33322	33121	32918	32713	32506	32297	32086	31874	31659	31443
0,7	31225	31006	30785	30563	30339	30114	29887	29659	29430	29200
0,8	28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0,9	26609	26369	26129	25888	25647	25406	25164	24923	24681	24439
1,0	24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1,1	21785	21546	21307	21069	20831	20594	20327	20121	19886	19652
1,2	19419	19186	18954	18724	18494	18265	18037	17810	17585	17360
1,3	17137	16915	16694	16474	16256	16038	15822	15608	15395	15183
1,4	14973	14764	14556	14350	14146	13943	13742	13542	13344	13147
1,5	12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1,6	11092	10915	10741	10567	10396	10226	10059	09893	09728	09566
1,7	09405	09246	09089	08933	08780	08628	08478	08329	08183	08038
1,8	07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1,9	06562	06438	06316	06195	06077	05959	05844	05730	05618	05508

2,0	05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2,1	04398	04307	04217	04128	04041	03955	03871	03788	03706	03626
2,2	03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2,3	02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2,4	02239	02186	02134	02083	02033	01984	01936	01888	01842	01797
2,5	01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2,6	01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2,7	01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2,8	00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2,9	00595	00578	00562	00545	00530	00514	00499	00485	00470	00457
3,0	00443	00430	00417	00405	00393	00381	00370	00358	00348	00337
3,1	00327	00317	00307	00298	00288	00279	00271	00262	00254	00246
3,2	00238	00231	00224	00216	00210	00203	00196	00190	00184	00178
3,3	00172	00167	00161	00156	00151	00146	00141	00136	00132	00127
3,4	00123	00119	00115	00111	00107	00104	00100	00097	00094	00090
3,5	00087	00084	00081	00079	00076	00073	00071	00068	00066	00063
3,6	00061	00059	00057	00055	00053	00051	00049	00047	00046	00044
3,7	00042	00041	00039	00038	00037	00035	00034	00033	00031	00030
3,8	00029	00028	00027	00026	00025	00024	00023	00022	00021	00021
3,9	00020	00019	00018	00018	00017	00016	00016	00015	00014	00014
4,0	00013	00009	00006	00004	00002	00002	00001	00001	00000	00000

Таблица 2.2

Нормальное стандартное распределение

Функция распределения Ф(z) нормированного нормального распределения

z	0	1	2	3	4	5	6	7	8	9
0,0	0,50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0,1	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0,2	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0,3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,4	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0,5	69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0,6	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490

0,7	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0,8	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,9	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1,0	84134	84375	84614	84850	85083	85314	85543	85769	85993	86214
1,1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,3	90320	90490	90658	90824	90988	91149	91308	91466	91621	91774
1,4	91924	92073	92220	92364	92507	92647	92786	92922	93056	93189
1,5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1,6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,7	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,0	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3,1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,5	99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,6	99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3,7	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3,8	99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3,9	99995	99995	99996	99996	99996	99996	99996	99996	99997	99997
4,0	99997	99998	99999	99999	99999	—	—	—	—	—

На протяжении примерно двух-столетий (с начала 19 до конца 20 века, когда началось широкое распространение ПК), вычисление нормального закона, как правило, выполнялось с использованием таблиц вида 2.1 и 2.2 в три следующих этапа: 1) сначала по исходной выборке определялись значения $\bar{x}$ и s (в натуральной переменной x), 2)затем натуральная переменная x_i преобразовывалась по (2.5) в нормированные значения z_i, для которых по таблицам типа 2.1 и 2.2 определялись значения $f_i(z_i)$ или $F_i(z_i)$, 3) после этого найденые значения $f_i(z_i)$ и $F_i(z_i)$ приписывались исходным величинам x_i, в соответствии с обратным преобразованием по (2.5).

Использовать эти таблицы достаточно просто, если учесть четность $f(z)$ и симметричность $F(z)$ относительно $z=0$. Поэтому табуляции даны только для $z \geq 0$, т.е. при пользовании таблицами надо принять

$$f(-z) = f(z) \quad \text{и} \quad F(-z) = 1 - F(z). \qquad (2.9)$$

В заключение приведем весьма полезную табличку вероятностей $p_{\Delta x}$ того, что случайная величина, подчиняющаяся нормальному закону, попадает в интервалы $\Delta x = \bar{x} \pm a\sigma$:

от -1σ до $+1\sigma$, $p_{\Delta x}$=0,6827 ($p_{\Delta x}$≈68,3%),

от -2σ до $+2\sigma$, $p_{\Delta x}$=0,9545 ($p_{\Delta x}$≈95,5%),

от -3σ до $+3\sigma$, $p_{\Delta x}$=0,9973 ($p_{\Delta x}$≈99,7%),

от -4σ до $+4\sigma$, $p_{\Delta x}$=0,999936 ($p_{\Delta x}$≈99,99%),

от -5σ до $+5\sigma$, $p_{\Delta x}$=0,999999426 ($p_{\Delta x}$≈99,9999%).

Как видно, в интервал $\bar{x} \pm 1\sigma$ попадает примерно 68,3% членов выборки, $\bar{x} \pm 2\sigma$ примерно 95,5% членов, а в интервал $x \pm 3\sigma$ – 99,7% членов, если она подчинена нормальному закону. Это равносильно тому, что соответственно 68,3, 95,5 и 99,7% площади под кривой $f(x)$ лежит в этих пределах переменных z и x. Отсюда возникло широко известное приближенное правило «трех–пяти сигм», с которым мы познакомимся далее. В самой общей редакции оно имеет следующий смысл: «выход переменной за пределы интервала $\bar{x} \pm (3 \div 5)\sigma$ практически невозможен». Эту табличку можно видоизменить, если задать p, а затем определить соответствующий вероятный интервал Δx. Тогда получим:

$$p_{\Delta x}=0{,}95, \quad \Delta x = \bar{x} \pm 1{,}96\sigma,$$

$$p_{\Delta x}=0{,}99, \quad \Delta x = \bar{x} \pm 2{,}58\sigma,$$

$$p_{\Delta x}=0{,}999, \quad \Delta x = \bar{x} \pm 3{,}29\sigma,$$

$$p_{\Delta x}=0{,}9999, \quad \Delta x = \bar{x} \pm 3{,}89\sigma$$

Теперь можно привести приближенные критерии нормальности, которые следуют из двух последних табличек, а так же основаны на равенство нулю асимметрии и эксцесса. Приближенными критериями номальности являются два следующих отношения, которые должны выполняться одновременно:

$$\frac{|A|}{s_A} \le 3, \quad \frac{|E|}{s_E} \le 3, \tag{2.10}$$

где s_A и s_E – средние квадратические ошибки A и E.

Если неравенства (2.10) выполняются одновременно, то считается, что $A \approx 0$ и $E \approx 0$, т.е. распределение можно считать нормальным. При не выполнении (2.10) распределение следует считать не нормальным, т.е. A и E статистически значимо отличаются от нуля.

Компьютерные реализации нормального закона. В Excel нормальный закон представлен вычислением следующих функций.

1. Значений функции *плотности f(x)* по (2.1), для чего надо задать значение переменной x_i, среднее значение $\bar{x}$, стандартное отклонение (СКО) s и для логического значения *интегральная* – 0 или ложь. Синтаксис записи: НОРМ.РАСП (x_i; среднее; стандартное откл.; 0 или ложь). Например, НОРМ.РАСП (3,85; 4,65; 2,31; 0) равняется f(3,85)=0,16265.

2. Значений *интегральной* функции распределения *F(x)* по (2.2), если в качестве логического значения *интегральная* задать – 1 или истина. Например, НОРМ.РАСП (3,85; 4,65; 2,31; истина) равняется F(3,85)=0,36455.

3. Значений *обратной интегральной* функции закона *F(x)*, т.е. квантилей x_p, для чего надо задать значение $F_i(x_i)$, среднее значение $\bar{x}$, стандартное отклонение s. Синтаксис записи: НОРМ.ОБР (вероятность; среднее; стандартное откл.). Пример: НОРМ.ОБР (0,36455; 4,65; 2,31) равняется x_p=3,85.

4. Значений *интегральной* функции *F(z)* стандартного нормального закона по (2.8), для чего надо задать z_i. Синтаксис записи: НОРМ.СТ.РАСП (z_i). Например: НОРМ.СТ.РАСП (1,47) равняется F_i=0,92922.

5. Значений z_p - *обратной интегральной* функции нормального стандартного закона *F(z)*, для чего надо задать $F_i(z_i)$. Синтаксис записи: НОРМ.СТ.ОБР (вероятность). Пример: НОРМ.СТ.ОБР (0,92922) равняется z_p=1,47.

6. Значений нормированной переменной z_i по (2.5), для чего надо задать переменную x_i, среднее значение $\bar{x}$ и стандартное отклонение s. Синтаксис записи: НОРМАЛИЗАЦИЯ (x_i; среднее; стандартное откл.). Пример: НОРМАЛИЗАЦИЯ (3,47; 1,58; 1,47) равняется z_i=1,2857.

Кроме того, для существенно положительных величин x, когда исходное распределение сильно право асимметрично, часто импользуется логарифмическое преобразование

x, путем перехода к новой переменной y=lnx. В результате такого преобразования крутая левая ветвь распределения в новой переменной *y* растягивается (до $-\infty$), а длинная правая сжимается, распределение может стать близким к симметричному или симметричным. В таких случаях используется так называемое логарифмически нормальное распределение (подробнее см. [23]). Для расчетов логнормальных распределений применяются программы ЛОГНОРМ.РАСП и ЛОГНОРМ.ОБР.

2.1.3. Аппроксимация сгруппированных выборочных распределений нормальным законом с использованием Excel

Важной, *и часто конечной*, задачей статистического анализа является аппроксимация выборки (ее приближенное выравнивание или представление) тем или иным вероятностным статистическим законом. Если такая аппроксимация получена удовлетворительной с точки зрения статистических критериев, то это позволяет считать, что *выборка подчиняется* найденному теоретическому закону распределения. Тогда все необходимые для практического использования выводы делаются *исходя из найденного закона*, а выборка послужила лишь первоосновой для его получения. Покажем процедуру таких расчетов и анализа с помощью программ Excel на примере аппроксимации сгруппированной выборки средних годовых температур воздуха на метеостанции Байтык за 86-летний период наблюдений с 1915 по 2000 гг., которая уже рассматривалась в п. 1.1.4 (см. табл. 1.1 и 1.2). Все необходимые исходные данные и расчеты аппроксимации приведены в табл. 2.3.

В первых трех столбцах приведены нижние и верхние границы классов группировки температуры T°C, а также середины классов. Далее даны эмпирические частоты классов $n_{iэ}$ и их вероятности $p_{iэ}= n_{iэ}/86$ (объем выбрки). При этом на границах распределения добавлены нулевые классы температур, для которых $n_{iэ} =0$. Это надо делать всегда, так как теоретические частоты $n_{iт}$ для них будут не нулевыми. Разумеется, на исходноую выборку такие добавленные нулевые классы никакого влияния не оказывают.

Статистики выборки, рассчитанные по данным табл.1.1, оказались следующими: $T_{ср}$.=6,648°C, *СКО*=0,731°C, *A(T)*=-0,48 и *E(T)*=0,46. Следовательно, эмпирическое распределение *T* имеет умеренную отрицательную асимметрию и умеренный положительный эксцесс. Это давало повод ожидать, что аппроксимация нормальным законом может дать положительные результаты, т.е. выборка не очень существенно отичается от нормального закона, для которого *A*=0 и *E*=0.

В столбце F_{iT} – приведены рассчитанные по программе НОРМ.РАСП и значениям $T_{ср}$.=6,648°C, *СКО*=0,731°C теоретические обеспеченности верхних границ классов T_i.

Производя последовательные вычитания из последующих строк предыдущих строк получим теоретические вероятности классов $p_{im}=(F_{im}- F_{(i-1)m})$. Так, например, для класса 5-5,5°С имеем: p_{im}=0.05816-0.01208=0.04607. Исключение представляет первый класс, для которого всегда p_{iT} =0.00001-0=0.00001 (т.е. полагается, что отсутствующий предудущий класс является нулевым). Теоретические частоты классов теперь можно рассчитать по формуле $n_{im}= p_{im}$*86 (где 86 - объем выборки, равный $\sum n_{iэ}$). Можно предложить и другие варианты расчетов, но этот явяется наиболее простым и полным по представленным результатам.

Таблица 2.3

Аппроксимация сгруппированной выборки средних годовых температур на МС Байтык (табл.1.2) нормальным законом со средним значением

$T_{ср.}$= 6,448°С и $СКО$=0,731°С

Исходная выборка					Аппроксимация НЗ			Расчет
НГ,°С	ВГ,°С	СК,°С	$n_{iэ}$	$p_{iэ}$	$F_{iт}$	$p_{iт}$	$n_{iт}$	χ^2
3	3.5	3.25	0	0.0000	0.00001	0.00001	0.001	
3.5	4.0	3.75	1	0.0116	0.00015	0.00014	0.012	
	4.5	4.25	0	0.0000	0.00165	0.00150	0.129	
4.5	5.0	4.75	2	0.0233	0.01208	0.01043	0.897	
5	5.5	5.25	5	0.0581	0.05816	0.04607	3.962	1.80
5.5	6.0	5.75	18	0.2093	0.18769	0.12953	11.140	4.22
6	6.5	6.25	17	0.1977	0.41978	0.23209	19.960	0.44
6.5	7.0	6.75	26	0.3023	0.68493	0.26515	22.803	0.45
7	7.5	7.25	13	0.1512	0.87810	0.19317	16.612	3.75
7.5	8.0	7.75	3	0.0349	0.96781	0.08971	7.715	
8	8.5	8.25	1	0.0116	0.99435	0.02655	2.283	
8.5	9.0	8.75	0	0.0000	0.99935	0.00500	0.430	
9	9.5	9.25	0	0.0000	0.99995	0.00060	0.051	
9.5	10.0	9.75	0	0.0000	1.00000	0.00005	0.004	
10	10.5	10.25	0	0.0000	1.00000	0.00000	0.000	
							χ^2(эмп)	10.66

На рис. 2.1 показан график результатов апроксимации, приведенных в табл. 2.3. Из рис. 2.1 и табл. 2.3 хорошо видно, что апрокимирующая нормальная кривая дейсвительно в целом удовлетворительно описывает основные закономерности распределения вероятностей классов средних годовых температур на МС Байтык. Но при полном совпадение выборки

и нормальной кривой высоты столбцов гистограммы должны были бы точно совпадать с высотой маркеров-точек на кривой. Однако этого нет, а имеет место два весомых несоответствия между выборкой и этой кривой.

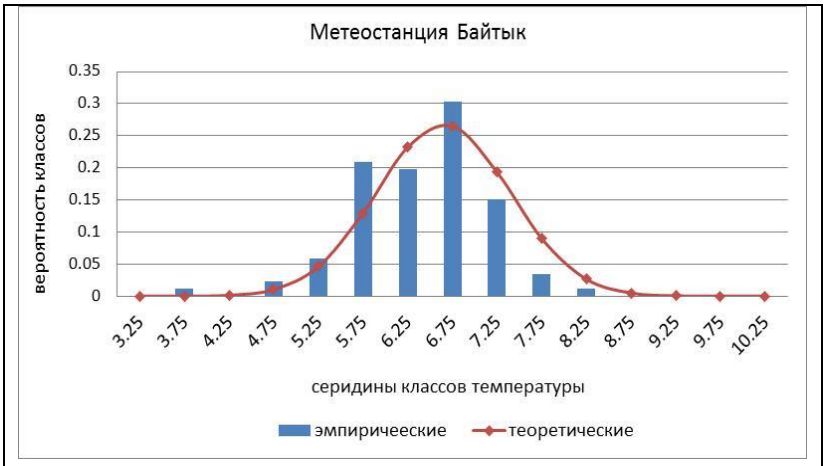


Рис. 2.1. Гистограмма эмпирического распределения средих годовых температур воздуха на метеостанции Байтык и аппроксимация его нормальным законом

Так, аппроксимация не отражает наличие второй моды, приходящийся на класс температур 5,5-6°С. Заметно также ее расхождение для правой ветви распределения: за счет левой асимметрии (A=-0,48) эмпирическая выборка здесь укорочена, теоретическая нормальная кривая лежит выше опытных данных и прослеживается вправо много дальше. Поэтому, для более точного решения вопроса о качестве аппроксимации выборки нормальным законом следует воспользоваться специальными количественнвми критерими. В качестве такого критерия обычно используется критерий χ^2-Пирсона (читается как «хи-квадрат Пирсона»), с расчетом и использованием которго познакомимся в следующем пункте.

Метеорологическими причинами таких особенностей выборки может быть то, что станция Байтык (1579 м) располагается в нижней части протяженной меридиональной склоновой долины р. Ала-Арча (северный склон Киргизского хребта), где хорошо развита горно-долинная циркуляция и ночной гравитационный сток охлажденного над склонами воздуха. Одновременно в районе станции при определенных типах синоптических процессов существенную повторяемость имеют фены. Все эти местные особенности климата, связанные с горной орографией, могли привести как к формированию второй моды, так и к зметной левой асимметрии распределения средних годовых температур.

2.1.4. Критерий χ^2-Пирсона для оценки согласования теоретического и эмпирического распределений

Рассмотрим вычисление и практическое применение критерия χ^2- Пирсона (читается как «хи-квадрат Пирсона»), который служит для оценки степени согласования теоретического и эмпирического распределений. Теория этого и других аналогичных вопросов будет рассмотрена в теме 5, посвященной проверке статистических гипотез. Основные свойства χ^2- распределения приведены в п.2.2.3.

Практическая процедура расчета и использования критерия хи-квадрат Пирсона для нашей задачи состоит в следующем.

Проверяется нулевая статистическая гипотеза *(H_0),* которая состоит в том, что эмпирический закон распределения выборки *($F_э$)* равен теоретическому закону *(F_T)*, которым она аппроксимирована, т.е. $H_0: F_э=F_T$. Критерий построен так, что суммарно оценивает насколько хорошо согласуются эмпирических частоты классов группировки $n_{iэ}$ с теоретическими частотами n_{iT}. Вычисление критерия и принятие решения по нулевой гипотезе включает 4 этапа.

1. Рассчитывается эмпирическое значение χ^2-критерия по формуле

$$\chi^2(эмп) = \sum_k \frac{\left[n_{iэ} - n_{i,T}\right]^2}{n_{i,T}}, \qquad (2.11)$$

где суммирование выполняется по всем классам группировки k.

2. Задается вероятность уровеня значимости критерия (обычно q принимается равной 0,01, 0,05 или 0,10), тем самым задается и уровень доверительной вероятности $p=(1-q)$; по определенному правилу находится число степеней свободы критерия *(CC).*
3. По погрaмме Excel «ХИ2.ОБР.ПХ» по заданному q и известному числу *CC* определяется критическое значение χ^2-критерия -$\chi^2(крит)$.
4. Если - $\chi^2(эмп)<\chi^2(крит)$, то гипотеза - $H_0: F_э=F_T$ - *принимается* на уровне значимости q (т.е. с доверительной вероятностью p); если - $\chi^2(эмп)>\chi^2(крит)$, то гипотеза - $H_0: F_э=F_T$ -*отвергается* на уровне значимости q (т.е. с доверительной вероятностью p).

Принятие нулевой гипотезу означает, что аппроксимацию выборки примененным теоретическим законом можно *принять удовлетворительной* на уровне доверительной вероятности p, т.е. с риском совершить ошибку с вероятностью q. Тогда приближенно с долей этого риска можно считать, что $F_э=F_T$. Заметим еще раз, что все выводы в матема-

тической статистике не категоричны, а имеют вероятностный характер. Не принятие гипотезы H_0 означает, что аппроксимацию следует признать не удовлетворительной.

Покажем, как эта процедура выглядит на практике на примере аппроксимации нормальным законом, приведенной в таблице 2.3.

Расчет эмпирического значения критерия -χ^2(эмп). Критерий рассчитывается по значениям эмпирических $n_{iэ}$ и теоретических n_{iT} частот, полученных для всех k-классов сгруппированного распределения. Но есть одно ограничение: частоты каждого класса $n_{iэ}$ не должны быть малы, на практике принимается, что $n_{iэ} \geq 5$. Поэтому на границах распределений, где $n_{iэ} < 5$, требуется объединение нескольких классов в один, чтобы их суммарное $n_{iэ}$ было 5 и более. Для полученного уменьшенного числа классов k_* и рассчитывается сумма (2.11), что показано в последнем столбце табл. 2.3. Это число классов k_* называется *значимыми классами*.

Так как, чаастоы первых 4-классов менее 5 (они последовательно равны 0, 1, 0, 2), то объеденим их с частотой 5-класса, равной 5. Тогда, частота первого объединеннного класса будет равна $n_{iэ} = 8$. Точно также объеденим теоретические частоты 5 первых классов и получим для них объединенную частоту $n_{iT} = 5{,}00$. Находим по этим данным выражение $(n_{iэ} - n_{iT})^2/n_{iT}$, стоящее под знаком суммы (2.11) для первого объединенного класса. Оно будет равно 1,80, запишем его в строку для 5-класса в последнем столбце табл. 2.3. Проведем точно такие же вычисления для последних 7 классов, для которых эмпирические частов равны: 13, 3, 1, 0, 0, 0, 0. В результате, искомое выражение, которое стоит под знаком суммы для этого объединенного класса, будет равно 3,75. Для остальных трех внутренних классов, эмпирические частоты которых равны 18, 17 и 26, произведем вычисление выражения $(n_{iэ} - n_{iT})^2/n_{iT}$ для каждого класса отдельно. Это будут числа 4,22, 0,44 и 0,45. Найдем теперь сумму всех 5 чисел, которые стоят в последнем столбце табл. 2.3 и которая будет равна эмпирическому значению критерия - χ^2(эмп.)=10,66, что показано в последней строке табл. 2.3.

Определение числа степеней свободы CC. Число *CC* для критерия хи-квадрат Пирсона определяется как число значимых классов k_* минус число наложенных связей m и минус 1, т.е. по формуле

$$CC = k_* - m - 1 \qquad (2.12)$$

Под *числом наложенных связей* в данном случае понимают число параметров, которое использовано при расчете аппроксимирующего распределения. Для нормального закона таких параметров два – это среднее и СКО. Таким образом, в нашем примере по (2.12) число $CC = 2$.

Задание уровня значимости критерия q. Численно *q* есть малая вероятность того, что, используя критерий, мы всегда допускаем риск совершить ошибку, т.е. забраковать гипотезу H_0, когда она на самом деле верна. Заданием *q* автоматически задается и доверительная вероятность *p=(1- q),* т.е. вероятность того, что принимаемое решение верно. Зададим в нашем примере уровень значимости *q*=0,05, тогда уровень доверительной вероятности *p*=0,95.

Оределение критического значения критеия - χ^2*(крит).* Используя программу обратного χ^2- распределения Пирсона «ХИ2.ОБР.ПХ» и, вводя в нее выбранное значение *q=0,05* и найденное число *CC=2,* получим, что χ^2(теор) = 5,99,

Заключение о качестве аппроксимации. Так как χ^2(эмп.)=10,66 > χ^2(теор)=5,99, то гипотеза H_0: $F_э=F_T$ отвергается на уровне доверительной вероятности *p*=0,95 (на уровне значимости *q=0,05*). Это означает, что с риском совершить ошибку в 5%, следует признать что выборка средних годовых температур на станции Байтык *не соответсвует* нормальному закону со средним значением 6,448°С и СКО=0,731°С, т.е. качество аппроксимации на уровне значимости *q=0,05* следует признать не удовлетворительным. Возможные метеорологические причины этого были выскажены выше.

Глава 2.2. ПОСТРОЕНИЕ ДОВЕРИТЕЬНЫХ ИНТЕРВАЛОВ ДЛЯ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ, ДИСПЕРСИИ И СКО С ПОМОЩЬЮ НОРМАЛЬНОГО ЗАКОНА, t-РАСПРЕДЕЛЕНИЯ СТЬЮДЕНТА И χ^2-ПИРСОНА

2.2.1. Построение доверительного интервала для математического ожидания при больших выборках

Выше были рассмотрены *точечные оценки* среднего и других статистик распределений, которые задавались одним числом и сопровождались значениями их средних квадратических ошибок. Если случайная величина имеет нормальное распределение, то, кроме точечных, достаточно просто можно найти надежные интервальные оценки математического ожидания, т.е. истинного среднего. Мы говорим здесь о среднем только потому, что средние значения метеорологических величин во многих случаях распределены нормально. Математическая процедура построения доверительных интервалов во всех случаях одинакова, были бы известны законы распределения статистик.

Квантильные значения случайной величины. Дадим вначале определение квантильного значения СВ, которое потребуется не только в этом разделе, но во многих других, где речь будет идти о построении доверительных интервалов для различных статистик. Квантильным значением случайной величинны *X* называется ее значение x_p *(или* x_F*)*, соответствующее заданному уровню обеспеченности, обозначаемой через *p* или *F*. Иными словами, если *F(x)* есть интегральная функция распределения, то x_F - это ее обратное значие, найденное по заданным значениям обеспеченностей *F* или *p*. Например, на рис. 2.2 показано 9 квантилей, определяемых значениями $F_1, F_2, F_3 \ldots, F_9$.

Так как доверительный интервал для среднего при больших выборках (n>30-50) строится с использованием стандартного нормального закона *F(z)*, то его квантили обозначаются через букву *z*. Для рассматриваеных ниже в п. 2.2.2. и п.2.2.3. распределений *t*-Стьюдента и χ^2-Пирсона они будут обозначаться соответственно через буквы *t* и χ (греческая буква – хи).

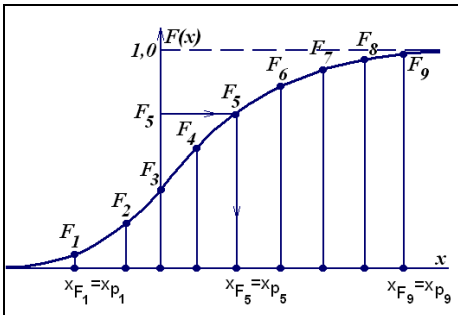


Рис. 2.2. Квантили $x_{F_i} = x_{p_i}$ случайной величины *X*.

Задавая, например, на вертикальной оси *F(x)* значение обеспеченности, равное F_5, найдем соответвующее ему квантильное значение СВ на горизонтальной оси абцис *x*.

Методика построения доверительного интервала для среднего при больших выборках. Итак, пусть выборка достаточно велика (обычно принимается n>30-50), и пусть $\bar{x}$ (оценка *мо*) есть нормально распределенная величина, для которой по выборке рассчитано значение *s* (оценка σ). Так как генеральной совокупностью мы не располагаем, то нет возможности определить точные значения параметров *мо и σ*. Однако имеется другая воз-

можность – по значениям выборочных оценок $\bar{x}$ и s определить интервал $\Delta(мо)$, в который с наперед заданной вероятностью p попадет неизвстное математическое ожидание ($мо$). Очевидно, что такой интервал будет равен:

$$\Delta_p(мо) = \bar{x} \pm z_{(1-\frac{q}{2})} \cdot \frac{s}{\sqrt{n}}, \qquad (2.13)$$

где $s/\sqrt{n}$ – есть ошибка среднего (см. (1.33)), $z_{(1-q/2)}$ –квантиль стандартного нормального распределения, соответствующий $q=(1-p)$, *который* отсекает критическую область, лежащую за пределами p и состоящую из двух, расположенных слева и справа, одинаковых по площади подобластей α *и* β, которым соответствуют квантили $z_{q/2}$ *и* $z_{(1-q/2)}$.

Отсекаемые, согласно (2.13), слева и справа области распределения $\bar{x}$, соответствующие $q/2$, называются критическими, а внутренняя область p (лежащая между ними) - областью доверительной вероятности, в которую попадает интервал $\Delta(мо)=\Delta x$. Все это показано на рис.2.3. для законоа распределения, заданного функциями $f(x)$ *и* $F(x)$.

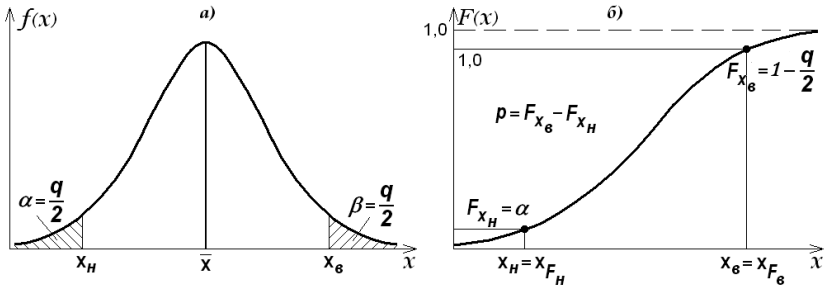


Рис. 2.3. Области доверительной вероятности p и критические области q/2 для математического ожидания (а также других статистик) относительно функций:

а) $f(x)$ и б)$F(x)$.

1) $x_н$ и $x_в$ – нижнее и верхнее значения доверительной области $\Delta x=x_в ..., x_н$, слева и справа вне ее лежит область критических значений x; 2) доверительная вероятность p – это вероятность попадания x в Δx; 3) α и β – малые вероятности, равные $q/2$, характеризующие возможный выход x за границы доверительного интервала Δx; 4) q – уровень значимости, вероятность (шанс) того, что СВ x выйдет либо слева, либо справа, либо там и там за пределы доверительного интервала $x_н ..., x_в$.

В табл. 2.4 приведены рабочие формулы определения Δx = Δ(мо) при задаваемых различных уровнях p и q. Они получены исходя из нормального закона распрееления. Пользуясь этой таблицей с легко найти интервал Δx, с требуемыми p и q. Например, для МС Фрунзе по достаточно большой выборке n=68 лет для средней годовой температуры

имеем: $\bar{T}$ =10,1^{0}C, s=0,90^{0}C. Тогда, с доверительной вероятностью p=0,99 по соответствующей формуле табл. 2.4 имеем:

$$\Delta T = 10{,}1 \pm 2{,}576 \cdot \frac{0{,}90}{\sqrt{68}} \Rightarrow \Delta T = 9{,}82..., 10{,}38^0\text{C} .$$

Таким образом, достаточно надежно можно утверждать, что истинноемноголетнее среднее годовое значение температуры на МС Бишкек (клматическая норма) с доверительной вероятностью p=0,995 лежит в интервале 9,82 ..., 10,38^{0}C. При этом, ее точечная оценка 10,1^{0}C есть середина этого интервала

Квантиль $z_{(1-\,q/2)}$ можно также найти по программе «НОРМ.СТ.ОБР», задав веорятность p. Именно по ней составлена табл. 2.4. Так, например, для p=0,975 ($q/2$=0,025) получим значения квантиля $z_{(1-\,q/2)}$ = 1,96, а для p=0,995 ($q/2$=0,0025) значение - $z_{(1-\,q/2)}$ = 2,576.

Таблица 2.4

Рабочие формулы для расчета $\Delta x = \Delta(мо)$ при различных значениях p и q для двухсторонней критической области

Доверительный интервал Δx для мо	*Доверительная вероятность p*	*Критическая вероятность q ($\alpha = q/2$; $\beta = q/2$)*
$\bar{x} \pm 1{,}645 \frac{s}{\sqrt{n}}$	0,90 (или 90%)	0,10 (или 10%)
$\bar{x} \pm 1{,}960 \frac{s}{\sqrt{n}}$	0,95 (или 95%)	0,05 (или 5%)
$\bar{x} \pm 2{,}576 \frac{s}{\sqrt{n}}$	0,995 (или 99,5%)	0,0025 (или 0,25%)
$\bar{x} \pm 3{,}2905 \frac{s}{\sqrt{n}}$	0,999 (или 99,9%)	0,001 (или 0,1%)

Ранее, для точечных оценок предлагалась запись $\bar{x} \pm s_{\bar{x}}$, т.е. для нашего примера $\bar{x}$ =10,1±0,11^{0}C. Приведение доверительного интервала для оценки генерального среднего (а так же и для других статистик) по качеству статистического решения существенно выше, чем это дает первая, минимально необходимая форма оценки точности решения.

Заметим, что часто полагается так же, что это не *мо* с заданной вероятностью p попадает в доверительный интервал Δx, а сам интервал Δx с вероятностью p накрывает *мо*. Однако практический смысл от замены одной редакции утверждения на другую, не меняется.

2.2.2. Распределение t – Стьюдета и построение доверитель-ного интервала для математического ожидания при малых объемах выборок

Выше было рассмотрено построение доверительного интервала для математического ожидания для выборок большого объема n>30–50. Решим аналогичную задачу для малой выборки, когда $n \leq 30$. Для этого применяется точно такой же подход, но только теперь надо использовать не нормальный закон, а t – распределение Стьюдета. График функция плотности этого распределения показан на рис. 2.4, где для сравнения нанесена также нормальная кривая.

Пусть имеются две СВ, одна из которых x – распределена по закону $N(0,1)$, а вторая $u=\chi^2$ – по закону χ^2 с n степенями свободы (рассматривается в следующем п. 2.2.3.). Построим новую СВ t как их отношение

$$t = \frac{x}{\sqrt{\frac{u}{n}}} = \frac{\sqrt{n}x}{\sqrt{u}},$$

где $t =]-\infty, \infty[$, т.е. изменяется на всей на бесконечном интервале.

Именно такой необычный вид случайной величины t позволяет использовать ее для построения доверительного интервала для среднего и различных критериев для проверки статистических гипотез (тема 5).

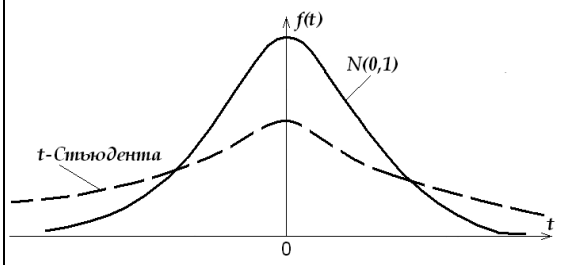


Рис. 2.4. График плотности t-распределения Стьюдента и нормального закона $N(0,1)$.

Не приводя самой формулы для плотности $f(t)$, ввиду ее некоторой сложности, отметим, что кроме переменной t, она зависит от одного праметра n, носящего название числа

степеней свободы (СС), через которые выражаются все свойтва распределения. На практике для нашей задачи - построения доверительного интервала для среднего - параметр СС для *t* – распределения определяется как объем выборки *n* минус 2*, m.e.* по формуле

$$CC = n\text{-}2 \qquad (2.14)$$

Через основные статистики основные свойства для *t* – распределения можно записать следующим образом:

$$\left.\begin{array}{l} 1.\ \bar{t} = мо = 0, \\ 2.\ \sigma_t^2 = \infty \ \text{ при } \ n = 1 \text{ и } 2, \\ \quad \sigma_t^2 = \dfrac{n}{n\text{-}2} \ \text{ при } \ n \geq 3, \\ 3.\ A(t) = 0; \quad E(t) < 0, \end{array}\right\} \qquad (2.15)$$

Из графика рис. 2.4 и формул (2.15) видно, что функция *f(t)* является симметричной (*A*=0), но при малом *n* (*n*≤30) заметно более плосковершинна (*E*<0) по сравнению с *N*(0,1); при увеличении *n* распределение достаточно быстро сходится по вероятности к *N*(0,1) так, что при больших выборках (*n*≥30-50) может заменяться нормальным распределением.

В результате, при малых выборках (*n*≤30) формулу для доверительного интервала для математического ожидания можно получить, заменив в (2.13) значение $z_{q/2}$ на квантиль $t_{n-2;1-q/2}$, т.е. в виде:

$$\Delta_p(мо) = \bar{x} \pm \left| t_{n-2;1-q/2} \right| \frac{s}{\sqrt{n}}. \qquad (2.16)$$

Значение квантилей $t_{n-2;1-q/2}$ находятся по программе Excel «СТЬЮДЕНТ.ОБР.ПХ» (или по специальным таблицам, см. [23]), где надо задать число степеней свободы *(n-2)* и значение критической вероятности *q* (подчеркнем, что именно *q,* а не *q/2,* т.к. именно это предусмотрено при составлении программы). Значения $t_{n-2;1-q/2}$ всегда несколько больше, чем $z_{q/2}$, и поэтому доверительный интервал для $\Delta_p(мо)$ получается несколько шире, чем при исползовании формулы (2.13). Однако это расширение обычно имеет практическое значение только при n<30.

Заметим, что при построении доверительного интервала для *мо* во всех случаях мы поступим правильно, если будем пользоваться *t*-распределением как для больших, так и для малых выборок. Заметим также, что исторически, пока не было ПК, решение для нормального закона по (2.13) выполнялось несколько проще, чем по (2.16) но сейчас это преимущество отсутствует.

В примере п.2.1.1 для МС Фрунзе по достаточно большой выборке n=68 лет по формуле (2.13) для средней годовой температуры $\overline{T}$ =10,1^{0}С и s=0,90^{0}С с вероятностью p=0,99 для *мо* был получен доверительный интервал 9,82…, 10,38°С.

Выполним аналогичные оценки интервала ΔT по этой станции по формуле (2.16), т.е. с использованием t-распределения, задав число CC= 68-2=66, а также другое малое число CC=10. Тогда для первого случая получим, что $t_{n-2;1-q/2}$ равно 2,652, а для второго 3,169. Оценки доверительных интервалов по (2.16) будут соответственно равны: $\Delta T_{(CC=66)}$ =9,81…,10,39°С; $\Delta T_{(CC=10)}$ = 9,23…,10.92°С. Как видно, во втором случае произошло, хотя и не очень значительное, но вполне заметное расширене доверительного интервала, тогда как в первом случае расширением можно принебречь, т.е. формулы (2.13) и (2.16) дали здесь одинаковые результаты.

2.2.3. Распределение χ^2- Пирсона и построение доверительных интервалов для дисперсии и СКО

Построение доверительных интервалов для дисперсии и среднего квадратического отклонения производится с помощью χ^2-распределения Пирсона, которое уже исползовалось в одноименном критерии, рассмотренном в п. 2.1.4. Теперь приведем оснвные свойства и этого распределения и рассмотрим его применение при построении доверительного интервала для дисперсии и СКО.

Пусть x_1, x_2, x_3 ..., x_n есть n независимых СВ, распределенных по закону N(0,1) (Использование нормировки всегда позволяет перейти к N(0,1)). Введем новую случайную величину, равную сумме квадратов:

$$u = \chi^2 = \sum_n x_i^2 ,$$

где $u \in [0, \infty[$, т.е. является существенно положительной и определена на полубесконечном интервале (дополнительное введение $u=\chi^2$ использовано исключительно для удобства записи).

Точно также, именно такой вид случайной величины χ^2 позволяет строить на основе χ^2-распределения разлияные важные статистические критерии.

Аналогично, не приводя самой формулы для плотности $f(u=\chi^2)$ *этого распределения,* отметим, что кроме переменной u она также зависит от одного праметра n, носящего название числа степеней свободы (СС), через которые выражаются все свойства распределения. На практике для нашей задачи число CC для χ^2– распределения определяется как объем выборки n минус 1, *т.е.* по формуле

$$CC = n-1 \qquad (2.17)$$

График функции плотности этого распределения показан на рис. 2.5, а основные свойства, выражаемые через статисики, описываются формулами (2.18):

$$\begin{aligned} &мо(\chi^2) = \bar{u} = n, \\ &\sigma^2(u) = 2n, \\ &мода = (n-2) \text{ при } n \geq 3. \end{aligned} \qquad (2.18)$$

Видно, что чем больше *n*, тем положе и симметричнее становится *f(u, n)* и при $n \to \infty$ функция *f(u, n)* сходится по вероятности к нормальному закону. Однако эта сходимость достаточно медленнаая. Во многих статистических руководствах приводятся таблицы хи-квадрат распределения для числа СС=1, 2, 3 ..., 30. Но для нашей задачи удобнее пользоваться для этого прогаммой Excel «ХИ2.ОБР.ПХ».

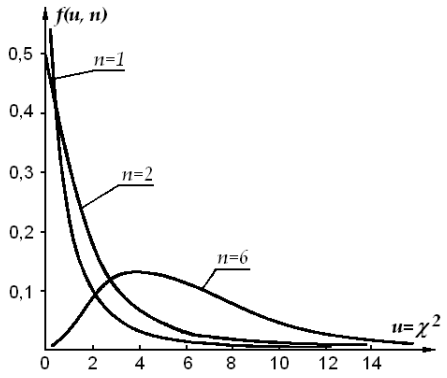


Рис. 2.5. График функции плотности χ^2-распределения при числе степеней свободы *n*=1, *n*=2 и *n*=6 (при *n*≥3 кривая имеет моду в точке *u=n-2*).

Формулу для доверительного интервала неизвесной гинеральной дисперсии σ^2, получим задавая уровень доверительной вероятности *p* и, следовательно, уровень ее критических значений *q=1-p*. При этом левую α и правую β подобласти критических значений возьмем, как обычно, одинаковымии по вероятности и равными q/2, т.е. примем, что $\alpha = \frac{q}{2}$ и $\beta = 1 - \frac{q}{2}$). Тогда, можно записать, что доверительный интервал для неизвестной σ^2 будет определяться неравенством:

$$\frac{s^2(n-1)}{\chi^2_{n-1;\alpha}} < \sigma^2 < \frac{s^2(n-1)}{\chi^2_{n-1;\beta}}, \qquad (2.19)$$

где в числителях неравенства стоят выборочные оценки дисперсии s^2, а в знаменателях критические значения или квантили χ^2, определяемые по числу степеней свободы *CC=n-1* и вероятностям α =q/2 и β=*1*-q/2.

Для того, чтобы получить доверительный интервал для среднего квадратического отклонеия σ надо извлечь корень из левой и правой части неравеста (2.19) и получить требуемую формулу (2.20):

$$\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{n-1;\alpha}}} < \sigma < \frac{s\sqrt{(n-1)}}{\sqrt{\chi^2_{n-1;\beta}}} \qquad (2.20)$$

На рис. 2.6 показана геометрическая интерпретация использования χ^2-распределения для построения области доверительных значений и критических областей для дисперсии.

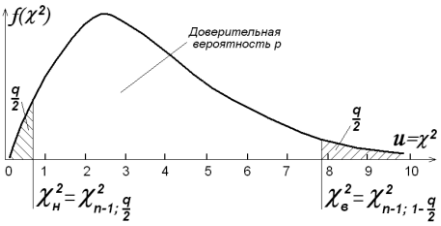


Рис. 2.6. Использование функции плотности $f(\chi^2)$ для построения доверительного интервала для генеральной дисперсии σ^2 по ее выборочной оценке s^2. Здесь $\chi^2_н$ и $\chi^2_в$ нижние и верхние критические значения χ^2, подставляемые в (2.19), которые определяются по таблицам $F(\chi^2)$, заштрихованные площади – значение $\frac{q}{2}$, не заштрихованная площадь под кривой $f(\chi^2)$ – есть доверительная вероятность *p*.

Так как χ^2-распределение несимметрично, то точечные оценки параметров σ^2 и σ лежат не точно в середине доверительного интервала (в отличие от оценки интервала для мо).

Покажем на конкретном примере как построить критическую область для дисперсии и СКО. Из климатического справочника для станции Фрунзе для средней годовой температуры воздуха T^0C имеем, что $s^2(T)$=0,81 и $s(T)$=0,90; число лет наблюдений *n*=68, т.е. *CC=n-1*=67. Зададим доверительную вероятность *p*=0,95, а критический уровень *q=(1-p)*=0,05. Тогда, по по программе «ХИ2.ОБР.ПХ» при *CC=67* получим квантили: $\chi^2_{67;0,025}$=91,52, $\chi^2_{67;0,975}$=46,26.

Подставляя найденные значения в (2.19) получим:

$$\frac{0{,}81 \cdot 67}{91{,}52} \le \sigma^2 \le \frac{0{,}81 \cdot 67}{46{,}26} \Rightarrow 0{,}59 \le \sigma^2 \le 1{,}17 .$$

Извлекая квадратный корень из обеих частей неравенства получим для СКО - $0{,}77 \le \sigma \le 1{,}08$.

Таким образом, для станции Фрунзе неизвестная нам генеральная дисперсия σ^2 лежит в интервале 0,59 .., 1,17 $(°C)^2$, а неизвестное СКО в интервале 0,77 ..., 1,08 (°C) при доверительной вероятности p=0,95 и критическом значении q=0,05.

ТЕМА 3. НЕКОТОРЫЕ ТЕОРЕТИЧЕСКИЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЙ, ПРИМЕНЯЕМЫЕ В МЕТЕОРОЛОГИИ ДЛЯ АППРОКСИМАЦИИ ОДНОМЕРНЫ СЛУЧАЙНЫХ ВЕЛИЧИН

Важным методом математической статистики является приближенное описание (аппроксимация) тем или иным теоретическим законом распределения результатов наблюдений метеостанций, представленных в виде эмпирических рядов или выборок. В случае положительного качества такой аппроксимации есть основание считать, что в основе формирования выборки (т.е. климатического режима рассматриваемой величины) и использованного теоретического закона лежат схожие или даже одни и теже факторы. Это значительно повышает практическую обоснованность климатических выводов, которые можно получить из статистического анализа метеорологического ряда. Более того, теперь анализ выборки, по существу, можно заменить анализом подобранного для ее описания теоретического закона, что дает гораздо более широкие возможности, чем ее прямой анализ. Например, по выборке, в силу ее ограниченности, обычно невозможно рассчитать вероятности появления редких экстремалей, вероятных 1 раз в 10, 50, 100 лет и реже. Но по найденной аппроксимации это делается легко. Тема 3 посвящена рассмотрению основных теоретических законов, которые используются в метеорологии в аппроксимационных целях и возможностях климатического анализа при их применении.

Глава 3.1. ЗАКОНЫ РАСПРЕДЕЛЕНИЙ ДИСКРЕТНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

3.1.1. Закон редких событий Пуассона

Пусть СВ дискретна и принимает значения 0, 1, 2,..., n. Здесь n – объем выборки независимых событий x=0, x=1, x=2 ..., т.е. событие состоит в том, что x принимает в точности одно из своих возможных дискретных значений. Средняя вероятность или *вероят-*

ность события в выборке $p = \bar{x}/n$, где $\bar{x}$ – средняя частота события, т.е. первый начальный момент или выборочное среднее.

Пусть $n \to \infty$, $p \to 0$, но их произведение $np \to \lambda = const$. Тогда, вероятность $p_n(x)$ того, что СВ в результате испытания в выборке объема n примет значение x равна:

$$p_n(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad (3.1)$$

где $\lambda = \bar{x}$ – единственный параметр распределения $p_n(x)$.

Формула (3.1) показывает, как распределены отдельные вероятности $p_n(x)$, если x будет принимать значения, равные 0, 1, 2 ... n. Эти вероятности быстро убывают по обратному экспоненциальному закону благодаря тому, что в знаменателе стоит очень быстро растущая величина $x! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot x$ *(x! - читатся как «икс-факториал»)*

Интегральным законом Пуассона будет функция $F_n(x)$:

$$F_n(x) = \sum_n p_n(x). \qquad (3.2)$$

Функции плотности у дискретного закона (3.2) нет.

Функция $p_n(x)$ по (3.1) в зависимости от $\bar{x}$ может иметь вид одномодальной правоасимметричной кривой или обратной экспоненты, как это показано на рис. 3.1. С ростом среднего значения $\bar{x}$ асимметрия стремится к нулю, а (3.2) к нормальному распределению.

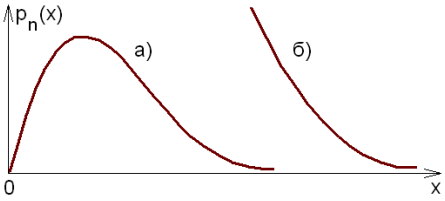


Рис. 3.1. Два характерных вида графика $p_n(x)$: а) одномодальный, $A>0$; б) обратно экспоненциальный A>>0.

В табл. 3.1 приведены распределения Пуассона для некоторых λ=0,2, 0,5 ..., 8, из которых наглядно видны эти его особенности.

Все свойства закона Пуассона выражаются через единственную статистику – среднее значение $\bar{x}$=λ, что обеспечивает легкость его использования при ручном счете:

$$\bar{x} = np = \lambda, \qquad (3.3)$$

$$\mu_2 = \mu_3 = \bar{x} = \lambda, \qquad (3.4)$$

$$A = \frac{\mu^3}{\sigma^3} = \frac{1}{\sqrt{\bar{x}}}, \qquad (3.5)$$

$$E = 1/\bar{x}. \qquad (3.6)$$

Таблица 3.1

Распределение пуассоновских вероятностей $p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$

для x=0 …, 22 и некоторых значений $\lambda \leq 8$.

λ / x	*0,2*	*0,5*	*0,8*	*1*	*3*	*5*	*8*
0	0,8187	0,6065	0,4493	0,3679	0,0498	0,0067	0,0003
1	0,1637	0,3033	0,3595	0,3679	0,1494	0,0337	0,0027
2	0,0164	0,0758	0,1438	0,1839	0,2240	0,0842	0.0107
3	0,0011	0,0126	0,0383	0,0613	0,2240	0,1404	0,0286
4	0,0001	0,0016	0,0077	0,0153	0,1680	0,1755	0,0573
5	0,0000	0,0002	0,0012	0,0031	0,1008	0,1755	0,0916
6		0,0000	0,0002	0,0005	0,0504	0,1462	0,1221
7			0,0000	0,0001	0,0216	0,1044	0,1396
8				0,0000	0,0081	0,0653	0,1396
9					0,0027	0,0363	0,1241
10					0,0008	0,0181	0,0993
11					0,0002	0,0082	0,0722
12					0,0001	0,0034	0,0481
13					0,0000	0,0013	0,0296
14						0,0005	0,0169
15						0,0002	0,0090
16						0,0000	0,0045
17							0,0021
18							0,0009
19							0,0004
20							0,0002
21							0,0001
22							0,0000

Компьютерная реализация распределения Пуассона в Excel представлена вычислением двух функций в программе «ПУАССОН.РАСП»:

1. Значений интегральной функции распределения (3.2) $F_n(x)$, для чего надо задать ожидаемое число появления события x, среднее значение $\bar{x} = \lambda$ по выборке, а также для логического значения *интегральная* указать – 1 или *истина*. Синтаксис записи: ПУАССОН.РАСП (x; $\bar{x}$; истина). Например, ПУАССОН (10; 4; истина) равняется $F(10)$=0,9972.

2. Значений вероятностей $p_n(x)$ по (3.1), если для логического *интегральная* указать – 0 или *ложь*. Синтаксис записи: ПУАССОН (x; $\bar{x}$; ложь). Например, ПУАССОН (10; 4; ложь) равняется $p(10)$=0,0053.

Исходя из вероятностных свойств закона, его целесообразно применять для описания распределения числа появления относительно редких метеорологических явлений в году, за месяц или сезон: числа гроз, градобитий, туманов и др. Вторым типом задач применения закона Пуассона является его использование для аппроксимации *усеченных выборок* экстремальных величин: максимальных значений скоростей ветра, гололедных отложений [23] и некоторых других метеорологических величин, распределения которых ограничены слева нулевыми значениями. Приближенным критерием применимости закона является примерное равенство средней и дисперсии по (3.4), т.е. $\bar{x} \approx s^2$.

3.1.2. Аппроксимация законом Пуассона усеченной сгруппированной выборки. Вычисление экстремальных вероятностных значений СВ с заданным периодом повторения

Аппроксимация законом Пуассона усеченной сгруппированной выборки. Рассмотрим как решаются с помощью закона Пуассона задачи второго типа – выравнивание усеченной выборки экстремальных значений метеорологической величины. Под усеченной в данном случае понимается выборка, начинающаяся не с нуля, как это предусматривается для закона Пуассона, а с некоторого требуемого значения. Например, если рассматривать распределение максимальных скоростей в бурях, то вкачестве критерия бурь надо взять буревые скорости, начиная с 10-15 м/с. Одновременно покажем как можно аппроксимировать сгруппированную выборку в том числе и *непрерывной СВ* (какой является, например, скорость ветра) дискретным законом Пуассона. Тем самым значительно расширяется практическое применение этого закона.

Рассмотрим в качестве примера распределение максимальных скоростей ветра в бурях на МС Фрунзе, которое уже приводилось в п. 1.1.2. В первых 4 столбцах табл. 3.1 приведена исходная сгруппированная выборка из 211 бурь (скорость 15 м/с и более), наблюдавшихся за 20 лет: границы классов скоростей V_i (первый столбец), середины классов $\overline{V}_i$ (второй столбец), частоты попадения скоростей n_i в каждый класс (третий столбец)

и вероятности классов $p_i = n_i/211$ (4 столбец). Середина классов представляют собой исходную натуральную переменную, которая начинается с 16 м/с, что соответсвует первому классу. Непосредственно к этой переменной закон Пуассона применить нельзя. Поэтому линейным преобразованием, которое не изменяет асимметрию и эксцесс исходного распределения (т.е. его форму), перейдем к новой цельночисленной переменной x_i=0, 1, 2 ..., 7 по формуле:

$$x_i = \frac{\overline{V}_i - 16}{\Delta V},$$

где ΔV – ширина классов группировки (в нашем случае ΔV=2 м/с), 16- серидина первого класса.

В новой переменной x_i рассчитаем два начальных момента $m_1 = \frac{1}{n}\sum n_i x_i$ и $m_2 = \frac{1}{n}\sum n_i x_i^2$. Они оказались равными: m_1=0,483; m_2=0,853. Так как $\mu_2 = s^2 = m_2 - m_1^2$, то s^2=0,620. Имеет место относительная близость $\bar{x}$=0,483 и s^2=0,620, что ориентировочно указывает на возможную успешную аппроксимацию выборки законом Пуассона.

Рассчитаем по (3.1) вероятности $p_n(x)$, обеспеченности $F_n(x)$ по (3.2) по прогамме ПУАССОН.РАСП, а теоретические частоты классов найдем как произведения $n_{i,т}=n \cdot p_n(x)$. Результаты этих расчетов приведены в 6-8 столбцах табл. 3.3. На рис. 2.2 показана гистограмма повторяемостей классов исходной выборки и рассчитанная по закону Пуассона выравнивающая кривая для вероятностей классов.

Таблица 3.3

Выравнивание распределения максимальных скоростей ветра в бурях на МС Фрунзе распределением Пуассона

Исходная выборка				*Аппроксимация-Пуассон*				*Хи-кв.*	T_i, *лет*
V_i, *м/с*	$\overline{V}_i$, *м/с*	n_i	p_i	x_i	$F_n(x)$	$p_n(x)$	$n_{i,т}$		
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
15-17	16	134	0,6351	0	0,6169	0,6169	130,2	0,11	0,25
17-19	18	59	0,2796	1	0,9149	0,2980	62,9	0,24	1,11
19-21	20	15	0,0711	2	0,9869	0,0720	15,2		7,24
21-23	22	1	0,0047	3	0,9985	0,0116	2,5		63,2
23-25	24	1	0,0047	4	0,9999	0,0014	0,3	0	
25-27	26	0	0	5	1,000	0,0001	0,02		
27-29	28	1	0,0047	6		0			

29-31	30	0	0	7		0			

$\sum$=211 $\chi^2_{эмп}$=0,35

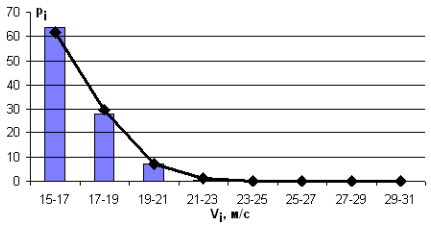


Рис. 3.2. Эмпирическая гистограмма p_i и выравнивающая кривая

Как следует из табл. 3. 3, согласование частот n_i и $n_{i,т}$ выглядит очень хорошим. Особенно наглядно это видно на рис. 3.2, где приведена эмпирическая гистограмма p_i и выравнивающая кривая $p_n(x)$. Расчет критерия согласия χ^2-Пирсона при числе степеней свободы k=3-1-1=1 дает следующее (см. п. 2.1.4): $\chi^2_{эмпир}$=0,35; $\chi^2_{1,q=0,05}$=3,84, т.е. $\chi^2_{эмпир} < \chi^2_{1,q=0,05}$=3,84 и гипотеза о соответствие эмпирического распределения скоростей на МС Фрунзе закону Пуассона может быть принята на уровне доверительной вероятности p=0,95, т.е. можно считать, что распределение скоростей ветра в бурях здесь близко соответствует найденному закону.

Вычисление экстремальных вероятностных значений СВ с заданным периодом повторения. Для решения многих практических задач важно знать период повторения экстремальных вероятностных значений СВ. Например, для расчета ветровых нагрузок на сооружения [12] требуются иметь данные о периодах скоростей ветра, превышающих определенную нижнюю границу 1 раз в 5, 10, 15, 25 лет. Эта нижняя граница наилучшим образом может быть вычислена по найденному закону *F(x), если исходная выборка представляет собой временной ряд*. Значению СВ, соответствующей границе превышения, приписывают уровень, который она может принимать 1 раз в заданное число лет. Разъясним этот подход на конкретном примере.

Пусть $F_i(V_i)$ функция обеспеченности скорости ветра, т.е. $F_i(V_i)=p(V \le V_i)$. Тогда, $F_{*i} = 1 - F_i$ функция превышений ($F_{*i} = p(V > V_i)$, т.е. вероятностей того, что скорость превысит значение V_i. Пусть среднее годовое число наблюдений равно $\bar{n}$ (это средний годовой объем выборки). Тогда, из n случаев наблюдений в году в $m = \bar{n}(1 - F(V_i))$ случаях, скорость ветра превысит уровень V_i, т.е. m есть частота превышений V_i в году.

В результате, период превышений V_i, равный T_i (в годах), будет:

$$T_i = \frac{1}{\bar{n}(1 - F_i(V_i))}. \qquad (3.7)$$

Покажем вычисление T_i на примере табл. 3.3, которого имеем: n=211, число лет N=20, $\bar{n}$ =211/20=10,55. Тогда для *верхних границ* классов скоростей в табл. 3.3 получим по (3.7) периоды их повторения T_i, приведенные в последней ее колонке. Подчеркнем особо, что найденные T_i должны быть отнесены к верхним границам своих классов. Так, ско-

рость 17 м/с имеет период повторения 0,25 года, 19 м/с – 1,11, 21 м/с – 7,24 и 23 м/с – 63,2 года. Это означает, например, что 1 раз в 63,2 года скорость ветра достигнет или превысит значение 23 м/с. Вычисления периодов можно было бы продолжить и далее, но они не потребуются, так как последний нужный нам период равен 25 лет.

Используя эти дробные значения периодов повторения T_i линейной интерполяцией можно рассчитать скорости ветра, вероятные 1 раз в год, 5, 10 и 20 лет (или другие периоды, какие требуются). Для указанных периодов максимальные скорости ветра в бурях для двух-минутного осреднения оказались равными: год – 18,8 м/с; 5 лет – 20,3; 10 лет – 21,1; 20 лет – 21,5 м/с. Эти значения и следует использовать как верятностные оценки максимальных скоростей ветра в различного рода практических приложениях. Они показывают, то в Чуйской долине Тянь-Шаня, где расположена МС Фрунзе, уровень максимальных скоростей низок и по СНИП соответсвует первому-второму ветровому району. Например, в гребневых зонах гор для периода повторения в 10 лет скорости могут достигать 40 и даже 50 м/с [23].

Отметим особо, что, не проведя выравнивание выборки законом Пуассона, мы, по-существу, не смогли бы рассчитать вероятные значения максимальных скоростей такой редкой повторяемости как в 10 и 20 лет, т.к. эмпирические частоты классов 21–23 м/с и выше неустойчивы и. непосредственно по выборке этого сделать нельзя.

Аналогичным образом рассчитывается период повторения для вероятностных экстремальных значений любых метеорологических величин с использованием для выравнивания эмпирических распределений любого теоретического закона.

В заключение подчеркнем, что формула (3.7) имеет общий характер, и описанный способ расчета может быть использован для любой метеорологической величины, оппраксимированной любым теоретическим законом распределения.

3.1.3. Биномиальный закон Бернулли

Пусть СВ дискретна и может принимать значения x=0, 1, 2 ..., n, где n – объем выборки. Пусть, как и в законе Пуассона, вероятность события в выборке $p = \bar{x}/n$, где $\bar{x}$ – средняя частота события по выборке. Вероятность отсутствия (не появления) события равна $q=1-p$. Других условий, как, например, в законе Пуассона, на вероятностную схему не накладывается. Тогда, вероятность $p_n(x)$ того, что событие в выборке объема n произойдет ровно x, раз равна:

$$p_n(x) = c_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, \qquad (3.8)$$

где n и p – параметры закона ($q=1-p$ не является дополнительным параметром, так как определяется заданием p).

Закон распределения (3.8), установленный Бернулли, получил название биномиального от входящего в него биномиального коэффициента – числа сочетаний из n по x:

$$c_n^x = \frac{n!}{x!(n-x)!}.$$

Функция распределения закона имеет вид:

$$F_n(x) = \sum_n c_n^x p^x q^{n-x}. \qquad (3.9)$$

Закон Бернулли характеризуется следующими свойствами:

$$мо = \bar{x} = np, \qquad (3.10)$$

$$\sigma^2 = \mu^2 = npq, \quad c = \left(\frac{q}{np}\right)^{\frac{1}{2}}, \qquad (3.11)$$

$$A = \frac{q-p}{(npq)^{1/2}}, \qquad (3.12)$$

$$E = \left(1 - 6p + 6p^2\right)/npq. \qquad (3.13)$$

Из (3.12) следует, что при $p=q=0{,}5$ распределение (3.8) будет симметрично ($A=0$), при $p>0{,}5$ значение $A<0$ – наблюдается левая асимметрия, при $p<0{,}5$ значение $A>0$ – правая асимметрия.

Биномиальное распределение имеет *два предельных перехода* к нормальному закону и закону Пуассона.

При

$$n \to \infty, \quad A \to 0 \quad и \quad E \to 0, \qquad (3.14)$$

практически при

$$npq \geq 9, \qquad (3.15)$$

распределение (3.8) переходит в нормальный закон.

При

$$n \to \infty, \quad p \to 0 \quad и \quad np \to \lambda = const, \qquad (3.16)$$

практически при

$$n > 60 \quad \text{и} \quad \bar{x} \approx \sigma^2, \qquad (3.17)$$

распределение (3.8) переходит в закон редких событий Пуассона.

Из этих свойств легко представить себе все возможное многообразие вида функции распределения (3.8), что предопределяет возможности широкого использования биномиального закона на практике.

Уже при не очень больших $n \geq 20$ возникают трудности вычисления биномиального коэффициента c_n^x за счет факториалов, приводящих к большим числам *x!*, *(n-x)!* и *n!*. Эти трудности снимаются, прежде всего, двумя предельными переходами (3.14), (3.15) и (3.16), (3.17), когда распределение заменяется нормальным или пуассоновским.

Компьютерная реализация биномиального распределения в Excel представлена вычислением двух функций.

1. Значений интегральной функции распределения (3.9) $F_n(x)$, для чего надо задать x (ожидаемое число появления события в выборке), число испытаний n (объем выборки), *среднюю* вероятность p события в выборке ($p = \bar{x}/n$), а также для логического значения *интегральная* указать – 1 или *истина*. Синтаксис записи: БИНОМ.РАСП (x; n; p; истина). Например, БИНОМ.РАСП (12; 20; 0,35; истина) равняется $F_{20}(12)$=0,99398473.

2. Значений вероятностей $p_n(x)$ по (3.8), если для логического *интегральная* указать – 0 или *ложь*. Синтаксис записи: БИНОМ.РАСП (x; n; p; ложь). Например, БИНОМ.РАСП (7; 20; 0,35; ложь) равняется $p_{20}(7)$=0,184401186.

Использование этой программы Excel снимает все трудности вычислительного характера при расчетах по биномиальному закону.

Закон Бернулли, как и закон Пуассона, в метеорологических задачах в основном применяется для описания распределения числа дней в году с какими-либо явлениями, например, сухих, влажных и т.д. Возможна также оценка каких-либо редких событий, которые могут произойти за несколько лет, или единиц десятков лет.

Таблица 3.4

Выравнивание распределения максимальных скоростей ветра в бурях на МС Фрунзе биномиальным распределением

Исходная выборка				*Аппроксимация-Бернулли*				T_i, *лет*
V_i, *м/с*	$\bar{V}_i$, *м/с*	n_i	p_i	x_i	$F_n(x)$	$p_n(x)$	$n_{i,т}$	
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*10*
15-17	16	134	0,6351	0	0.6166	0.6166	130.1	0.25
17-19	18	59	0,2796	1	0.9151	0.2985	62.98	1.12
19-21	20	15	0,0711	2	0.9870	0.0719	15.17	7.29
21-	22	1	0,004	3	0.9985	0.0115	2.43	62.6

23			7					7
23-25	24	1	0,0047	4	0.9999	0.0014	0.29	
25-27	26	0	0	5	1.0000	0.0001	0.03	
27-29	28	1	0,0047	6	1.0000	0.0000	0.00	
29-31	30	0	0	7	1.0000	0.0000	0.00	

Техника расчетов по закону Бернулли полностью аналогична закону Пуасона. В качестве примера в табл. 3.4 приведена аппроксимация выборки максимальных скоростей в бурях по табл. 3.3, которая была ранее выравнена в ней законом Пуассона. Для этого используем статистики и параметры: $\bar{x}$=0,483, n=211, p=0,483/211=0,002289. Вводя эти данные в программу БИНОМ.РАСП получим значения обеспеченностей и вероятностей классов по биномиальному закону, которые приведены в столбцах 6 и 7 табл. 3.4. Далее обычным путем рассчитаем теоретические частоты классов $n_{i,т}$, и по (3.7) периоды повторения скоростей ветра $T_{i,}$, соответствующие верхним границам классов.

Как видно, результаты аппроксимации по закону Бернулли практически совпали с аналогичными по закону Пуассона. Это не удивительно, так как аппроксимировалась одна и таже выборка, и использованные законы описали ее идентичным образом.

Глава 3.2. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

3.2.1. Экспоненциальное распределение

Экспоненциальное распределение в метеорологии находит применение для аппроксимации длительностей метеорологических явлений: грозы, тумана, сильных осадков, бурь и др. Однако автор с успехом применял его для описания распределений величин экстремалей: годовых максимумов скоростей ветра и гололедных отложений на проводах, годовых максимумов промерзаний почвы, годовых максимумов водозапасов снежного покрова [23]. Это распределение является частным случаем гамма-распределения и еще более общего – распределения Вейбулла, которые рассматриваются ниже.

Случайная переменная, подчиняющаяся экспоненциальному распределению, *непрерывна* и принимает только *неотрицательные значения*, т.е. $x \geq 0$.

Плотность распределения имеет вид:

$$f(x) = \lambda e^{-\lambda x}, \tag{3.18}$$

где λ – единственный параметр распределения.

Оно обладает следующими свойствами:

среднее значение $\bar{x} = \frac{1}{\lambda}$, (3.19)

мода $x_м = 0$, (3.20)

медиана $x_{ме} = \frac{1}{\lambda} \ln 2$, (3.21)

дисперсия $\sigma^2 = \frac{1}{\lambda^2}$, (3.22)

асимметрия $A = 2$, (3.23)

эксцесс $E = 6$. (3.24)

Таким образом, это распределение, как и закон Пуассона, описывается единственным параметром $\lambda = \frac{1}{\bar{x}}$, что делает его очень простым в практическом применении. Общий вид функции плотности (обратная экспонента) показан на рис. 3.3. Мода кривой плотности всегда находится в точке *x=0, а* асимметрия и эксцесс *постоянны* и равнысоответственно 2 и 6.

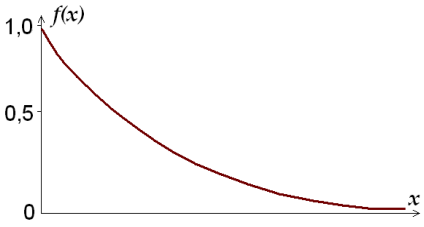


Рис. 3.3. График функции плотности для экспоненциального распределения.

Заменяя λ на $\frac{1}{\bar{x}}$ в (3.18) получим рабочие формулы для расчета плотности *f(x)* и функции распределения *F(x)*

$$f(x) = \frac{1}{\bar{x}} e^{-x/\bar{x}} = \frac{1}{s} e^{-x/s}, \tag{3.25}$$

$$F(x) = -1 - e^{\tilde{o}/\bar{x}} = 1 - e^{-x/s}. \tag{3.26}$$

Приближенным критерием применимости экспоненциального закона является

$$\bar{x} \approx s. \tag{3.27}$$

Расчет по (3.25), (3.26) не встречает каких-либо технических трудностей. Обратим только внимание на то, что, так как переменная *x* теперь является непрерывной, то это значит, что аппроксимацию можно проводить как в натуральной, так и условной переменной, причем как для не сгруппированной выборки (когда значения *x>0* могут быть дробными числами), так и для сгруппированной выборки по правилам п. *3.1.2.* При этом для *x=0 всегда* значение *f(x)* максимально*, а* значение *F(x)=0.*

Компьютерная реализация экспоненциального распределения в программах *Excel* представлена вычислением двух функций.

1. Значений интегральной функции распределения *F(x)* по (3.26), для чего надо задать переменную *x*, параметр $\lambda = 1/\bar{x}$ ($\bar{x}$ определяется по выборке), а также для логического значения *интегральная* указать–1 или *истина*. Синтаксис записи: ЭКСП.РАСП (*x*, $\lambda = \frac{1}{\bar{x}}$; интегральная 1). Например, ЭКСП.РАСП (86; 0,0552; истина) равняется F(86)=0,9913.

2. Значений функции плотности *f(x)* по (3.25), если для логического *интегральная* указать – 0 или *ложь*. Например, ЭКСПРАСП (86; 0,0552; ложь) равняется f(86)=0,0004789.

3.2.2. Гамма-распределение

Плотность рассмотренного выше экспоненциального распределения имеет моду в точке 0. Однако во многих задачах достаточно очевидно, что мода находится не в точке 0, а сдвинута вправо и часто ярко выражена. Такие выборки могут успешно описываться гамма-распределением. В метеорологии гамма-распределение может применяться для аппроксимации правоасимметричных выборок, ограниченных слева нулем: численных значений осадков, скорости ветра, давления водяного пара, длительности метеорологических явлений и др.

Функция плотности *двух параметрического* гамма-распределения имеет вид:

$$f(x,\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\beta x}, \qquad (3.28)$$

где СВ величина непрерывна и ограничена нулем слева, $0 \le x < \infty$; $\alpha > 0$ – безразмерный параметр «формы» распределения, от которого зависит его вид; $\beta > 0$ – параметр «масштаба» распределения, т.е. его размаха, (имеет размерность x^{-1}); $\Gamma(\alpha)$ – Г-функция Эйлера (откуда распределение и получило свое название).

В частном случае, когда α=1, гамма-распределение переходит в рассмотренное выше экспоненциальное с $\beta = 1/\lambda$. Если α – целое положительное, то частным случаем является распределение Эрланга.

Статистики, параметры и свойства гамма-распределения следующие:

среднее значение $\bar{x} = \frac{\alpha}{\beta}$, (3.29)

мода $x_{м} = \frac{\alpha - 1}{\beta}$ *(при* $\alpha \ge 1$*)*, (3.30)

дисперсия $\sigma^2 = \frac{\alpha}{\beta^2}$, (3.31)

асимметрия $A = 2/\sqrt{\alpha}$, (3.32)

эксцесс $E = 6/\alpha$. (3.33)

На рис. 3.5 показан типичный вид графика плотности гамма-распределения, из которого видно, что это модальная правоасимметричная кривая, но, в отличие от экспоненциального, в общем случае не имеющая начало в точке (0.0).

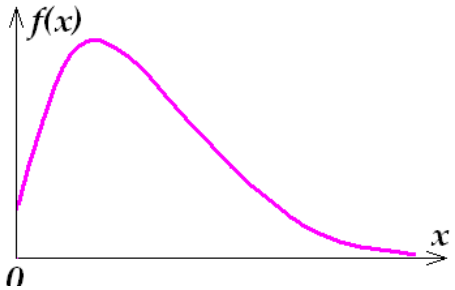


Рис. 3.5. График функции плотности гамма-распределения с параметром формы α=3 и параметром масштаба β=1.

Формулы (3.29)-(3.31) позволяют, оценив по выборке первые два момента – среднее $\bar{x}$ и дисперсию s^2, без труда рассчитать по ним все параметры и статистики гамма-распределения:

$$\alpha = (\bar{x}/s)^2, \quad \beta = \bar{x}/s^2. \quad (3.34)$$

Ручной счет гамма-рас-пределения достаточно сложен, т.к. требует использования таблиц гамма-функции, и мы его рассматривать не будем. В этом обычно нет надобности, учитывая наличие компьютерных программ.

Если от x перейти к безразмерной переменной $u = \beta x = \frac{x\bar{x}}{s^2}$, то β=1 и гамма-распределение принимает *стандартный вид*:

$$f(u,\alpha) = \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}, \quad (3.35)$$

$$F(u,\alpha) = \int_0^u \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}du, \quad (3.36)$$

где $f(u, \alpha)$ и $F(u, \alpha)$ – зависят только от одного параметра формы α.

Именно переход к этим формулам используется для ручного счета с применением таблиц стандартной гамма-функции $\Gamma(\alpha)$.

Компьютерная реализация гамма-распределения в программах Excel представлена вычислением следующих функций **(*Внимание:* в различных версиях Excel в качестве параметра возможно как использование β по (3.34), так и $\beta^*=1/\beta$)**.

1. Значений интегральной функции распределения $F(x)$, для чего надо задать переменную x, параметры α и β, а также для логического значения *интегральная* указать – 1 или *истина*. Синтаксис записи: ГАММА.РАСП (x, α, β, интегральная). Например, ГАММА.РАСП (6; 3,38; 0,788; истина) равняется F=0,97092.

2. Значений функции плотности *f(x)* по (3.28), если для логического *интегральная* указать – 0 или *ложь*. Например, ГАММА.РАСП (6; 3,38; 0,788; ложь) равняется *f(x)*=0,02690.

3. Если α=1, то вычисляется экспоненциальное распределение с $\lambda=1/\beta$.

4. Если α целое, то вычисляется распределение Эрланга.

5. Если β=1, то вычисляются функции *f(u)* и *F(u)* стандартного гамма-распределения по (3.35) и (3.36).

6. Значений x_p (квантилей), т.е. обратной функции для гамма-распределения. Синтаксис записи: ГАММ.АОБР (Вероятность *F(u)*; α; β). Например, ГАММА.ОБР (0,96; 3,38; 0,788) равняется x_p=5,652.

В качестве примера в табл. 3.5 приведена процедура выравнивания гаммма-распределением выборки спектра размеров облачных капель, а на рис. 3.6 показаны полученные эмпирическая и теоретические гистограммы распределений. В качестве натуральной переменной в табл. 3.6 взяты середины классов группирования (столбец 3), которые обозначены через x_i. В столбце 4 и 5 приведены эмпирические частоты классов $n_{iэ}$ и эмпирические вероятности $p_{iэ}= n_{iэ}/399$ (объем выборки).

Таблица 3.5

Аппроксимация сгруппированной выборки спектра радиусов облачных капель (НГ, ВГ - границы классов, мкм) гамма-распределением.

Выборка спектра капель					Расчет m_1 и m_2		Аппроксимация гамма-распределением			χ^2
НГ	ВГ	x_i	$n_{iэ}$	$p_{iэ}$	$x_i*n_{iэ}$	$x^2_i*n_{iэ}$	$F_{iт}$	$p_{iт}$	$n_{iт}$	
1	2	3	4	5	6	7	8	9	10	11
0	4	2	55	0.138	110	220	0.17351	0.17351	69.23	2.93
4	8	6	140	0.351	840	5040	0.46723	0.29372	117.2	4.44
8	12	10	90	0.226	900	9000	0.69622	0.22899	91.37	0.02
12	16	14	50	0.125	700	9800	0.83870	0.14248	56.85	0.83
16	20	18	30	0.075	540	9720	0.91824	0.07953	31.73	0.09
20	24	22	18	0.045	396	8712	0.95988	0.04164	16.62	0.12
24	28	26	8	0.020	208	5408	0.98078	0.02090	8.34	0.01
28	32	30	4	0.010	120	3600	0.99096	0.01018	4.06	0.02
32	36	34	2	0.005	68	2312	0.99581	0.00485	1.93	
36	40	38	1	0.003	38	1444	0.99808	0.00227	0.91	

40	44	42	1	0.003	42	1764	0.99913	0.00105	0.42	
44	48	46	0	0.000	0	0	0.99961	0.00048	0.19	
48	52	50	0	0.000	0	0	0.99983	0.00022	0.09	
52	56	44	0	0.000	0	0	0.99992	0.00010	0.04	
56	60	58	0	0.000	0	0	0.99997	0.00004	0.02	
			399		9.930	142.9			$\chi 2$(э)	8.45
					m_1	m_2			$\chi 2$(т)	11.1

Параметры гамма-распределения: α= 2.22547; β* = 4.46190

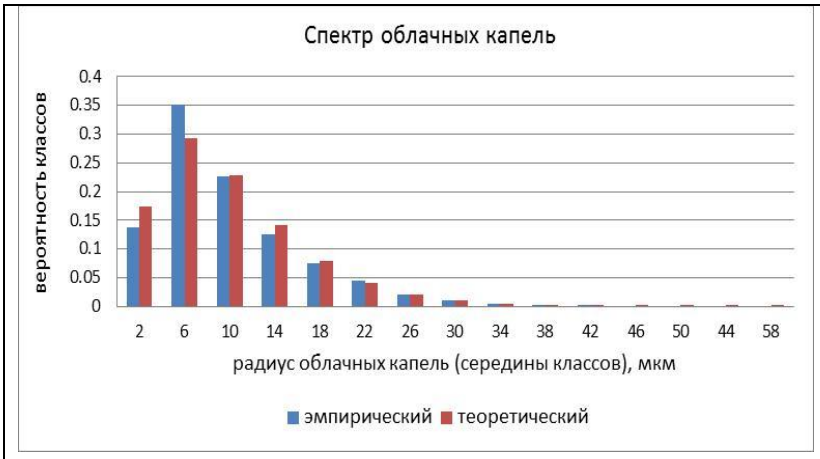


Рис.3.6 Эмпирическая и теоретическая (гамма-распределение) гистограммы спектра размеров облачных капель по данным табл. 3.5.

В столбцах 6 и 7 произведен расчет первого (m_1=9.930) и второго (m_2=142.91) начального моментов, по которым по (3.34) рассчитаны параметры гамма-распределения, оказавшиеся равными: α= 2.22547; β* = 4.46190. По значению этих параметров по программе ГАММА.РАСП определены теоретические обеспеченности $F_{iт}$ верхних границ классов группирования (столбец 8). По значениям $F_{iт}$ последовательным вычитанием (см.п. 2.1.3) найдены теоретические вероятности классов $p_{iт}$, по которым рассчитаны их теоретические частоты $n_{iт}= p_{iт}*399$. В последнем столбце 11 по правилам п. 2.1.4 рассчитано эмпирическое значение критерия согласия χ^2-Писона (χ^2(эмп) = 8.45). По прогпрамме ХИ2ОБР.ПХ по заданному значению q=0,05 и числу степеней свободы СС=5 найдено, что χ^2(теор) = 11.07. Так как χ^2(эмп)< χ^2(теор), то качество аппроксимации выборки следует признать удовлетворительным на уровне доверительной вероятности p=0,95. Визуально это видно из близкого совпадения высот столбиков эмпирической и теоретической гистограмм на рис. 3.6 для всех областей спектра капель. При этом теоретическая гисто-

грамма прослеживается вправо существенно дальше, чем эмпирическая, что говорит о том, что в этой области спектра могут существовать более крупные капли (хотя и с малой вероятностью), чем это имеет место в выборке.

Таким образом, при использовании гамма-распределения, как и экспоненциального, надо помнить, что переменная $x>0$ *и* является непрерывной. Поэтому она *не обязательно* должна принимать только целочисленные значения: 0, 1, 2, 3 и т.д. Это значит, что аппроксимацию можно проводить как в такой условной целочисленной переменной, так и в натуральной, возможно дробной, переменной. Причем это можно делать как для не сгруппированной выборки, так и для сгруппированной выборки, как это показано в табл. 3.6. За счет того, что гамма-распределение двух параметрическое оно работает более гибко, чем одно параметрические распределения экспоненциальное и Пуассона.

3.2.3. Распределение Вейбулла

Распределение Вейбулла широко используется в технике при рассмотрении вопросов надежности изделий. В метеорологии оно нашло широкое применение для расчета различного рода экстремалей метеорологических величин. В некоторых работах распределение Вейбулла называют распределением Гудрича Р.Д., который пришел к нему эмпирическим путем.

Вероятностную схему образования распределения Вейбулла можно представить следующим образом. Пусть, например, производятся наблюдения за скоростью ветра в течение каждых суток года. Тогда, наблюденные значения суточных максимумов скоростей будут V_1, V_2, V_3 ..., V_{365}. Выберем из них наибольшее за этот год значение, например, $V_{макс,127}$. Проведя, таким образом, наблюдения за n лет мы получим n годовых значений $V_{макс,i}$. Сформируем из них выборку годовых максимумов скорости. При $n\to\infty$ получим предельное распределение, соответствующее распределению Вейбулла. Подобную схему можно распространить на любые экстремали, важно только, чтобы n было большим и чтобы $F_*(x)=1-F(x)$ при $n\to\infty$ достаточно быстро стремилось к нулю.

Функция плотности распределения Вейбулла для непрерывной, ограниченной слева нулем, переменной СВ $x\geq 0$ имеет вид:

$$f(x)=\alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}, \qquad (3.37)$$

или, полагая $\lambda=\frac{1}{\beta^{\alpha}}$,

$$f(x)=\frac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \qquad (3.38)$$

где $\alpha>0$ – безразмерный параметр, определяющий форму кривой; β – параметр масштаба, имеющий размерность x.

Функция плотности $f(x)$ гибко меняет свою форму в зависимости от параметра α. Она может быть правоасимметричной кривой при $\alpha>1$, переходить в экспоненциальное распределение при $\alpha=1$ и иметь вид резко выраженной обратной экспоненты при $0<\alpha<1$ (рис. 3.7).

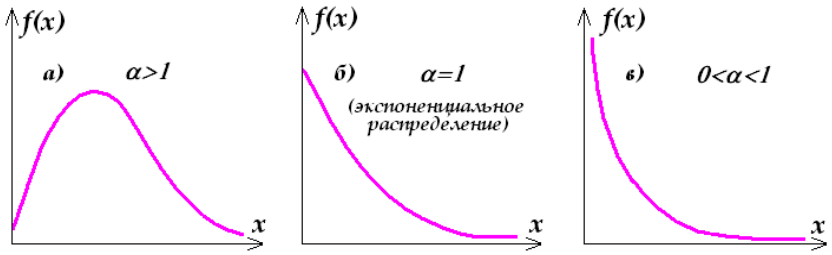


Рис. 3.7. Графики плотности $f(x)$ распределения Вейбулла при трех градациях α.

Интегральная функция распределения Вейсбулла $F(x)$ выражается формулой:

$$F(x)=1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}}. \qquad (3.39)$$

Параметры распределения Вейсбулла α и β выражаются через среднее, СКО, коэффициент вариации c и гамма-функцию Эйлера $\Gamma(\alpha)$ следующими соотношениями:

$$\bar{x}=\beta\Gamma\left(1+\frac{1}{\alpha}\right)=\beta k_{\alpha}, \qquad (3.40)$$

$$s=\beta\left[\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^{2}\left(1+\frac{1}{\alpha}\right)\right]^{0,5}=\beta c_{\alpha}, \qquad (3.41)$$

$$c=\frac{s}{\bar{x}}=\frac{c_{\alpha}}{k_{\alpha}}. \qquad (3.42)$$

Эти параметры вычисляются по специальным программам, которых, к сожалению, в Excel нет. Значения α, k_{α} и c_{α}, в зависимости от выборочного коэффициента вариации $c=s/\bar{x}$, приведены в табл.3.6. Они могут быть использованы для приближенной оценки параметров α и β по выборочным оценкам $\bar{x}$, s и c.

Например, по данным многолетних наблюдений за осадками на МС Пржевальск в январе получено: $\bar{x}$=56 мм; s=26,9 мм и, следовательно, $c=s/\bar{x}$=0,48. По табл. 3.6 имеем для c=0,48: α=2,2; k_{α}=0,886; c_{α}=0,425. По правой чсти формулы (3.40) получим: β=63,21

мм. Таким образом, параметры α и β для распределения Вейбулла будут равны: α=2,2; β=63,21 мм. Определив таким путем параметры распределения Вейбулла можно далее производить аппроксимацию выборок этим законом, используя программы Excel.

Таблица 3.6

Параметр α и коэффициенты k_α и c_α распределения Вейбулла в зависимости от c

α	k_α	c_α	$c=s/\bar{x}$	α	k_α	c_α	$c=s/\bar{x}$
0,2	120	1900	15,83	1,6	0,897	0,574	0,640
0,3	8,86	46,9	5,29	1,7	0,892	0,540	0,605
0,4	3,32	10,4	3,14	1,8	0,889	0,512	0,575
0,5	2,00	4,47	2,24	1,9	0,887	0,485	0,547
0,6	1,50	2,61	1,74	2,0	0,886	0,463	0,523
0,7	1,27	1,86	1,46	2,1	0,886	0,441	0,498
0,8	1,13	1,43	1,26	2,2	0,886	0,425	0,480
0,9	1,05	1,17	1,11	2,3	0,886	0,409	0,461
1,0	1,00	1,00	1,00	2,4	0,887	0,394	0,444
1,1	0,965	0,878	0,910	2,5	0,887	0,380	0,428
1,2	0,941	0,787	0,837	3,0	0,893	0,326	0,365
1,3	0,924	0,716	0,775	3,5	0,900	0,285	0,316
1,4	0,911	0,659	0,723	4,0	0,906	0,255	0,281
1,5	0,903	0,612	0,678				

$\bar{x} = \beta k_\alpha$, $s = \beta c_\alpha$

Компьютерная реализация распределения Вейбулла в программах Excel представлена вычислением двух функций.

1. Значений интегральной функции распределения *F(x)* по (3.39), для чего надо задать переменную *x*, параметры α и β, а также для логического значения *интегральная* указать 1 или *истина*. Синтаксис записи: ВЕЙБУЛЛ (*x*, α, β, интегральная). Например, ВЕЙБУЛЛ (105; 20; 100; истина) равняется *F*=0,92958.

2. Значений функции плотности *f(x)* по (3.37), если для логического *интегральная* указать – 0 или *ложь*. Например, ВЕЙБУЛЛ (105; 20; 100; ложь) равняется *f(x)*=0,035589.

3. При α=1 вычисляются *F(x)* и *f(x)* для экспоненциального распределения.

Расчеты по распределению Вейбулла с использованием Excel, если известы α и β, элементарны и не содержат каких-либо трудностей. Точно так же, как и для экспоненци-

ального и гамма распределений, обратим внимание на то, что, так как переменная *x* является непрерывной, то это значит, что аппроксимацию можно проводить как в натуральной, так и условной переменной, причем как для не сгруппированной выборки (когда значения *x>0* могут быть дробными числами), так и для сгруппированной выборки по правилам п. *3.1.2.* За счет того, что распределение Вейбулла двух параметрическое оно работает более гибко, чем одно параметрические распределения экспоненциальное и Пуассона.

В заключение заметим, что все рассмотренные теоретические распределения, кроме биномиального, являются правоасимметричными. Это накладывает ограничения на их использование, так как ими заведомо нельзя аппроксимировать выборки с левой асимметрией. Только биномиальный закон может выполнять эту роль. В этом смысле он более универсален по применению, чем все другие рассмотренные законы распределений.

ТЕМА 4. ЭМПИРИЧЕСКИЕ СВЯЗИ И ЗАВИСИМОСТИ СЛУЧАЙНЫХ ВЕЛИЧИН

Во многих случаях между метеорологическими случайными величинами (в простейшем случае – двумя переменными y и x) существует статистическая зависимость, которая принципиально отличается от функциональной зависимости. Если рассматривать x как независимую переменную, а y – зависимую, то при функциональной зависимости любому значению x «соответствует одно и только одно значение y» (y – есть функция от x). При статистических связях, напротив, любому значению x соответствуют несколько (множество) значений y. Например, температура воздуха T в горах понижается с высотой места z. Но на эту основную причинную зависимость $T(z)$ накладывается множество других случайных факторов – географическое положение горной страны, конкретная орография местности, конкретная синоптическая ситуация и т.д. В результате, обычно имеет место только более или менее четко выраженная тенденция зависимости y от x, которая в целом, наряду с главными причинами, определяется множеством случайных факторов, большинство из которых количественно описать обычно просто невозможно. Такие нефункциональные зависимости носят название *статистических*. Они выявляются тем надежнее, чем больше объем совместной выборки (x, y) и чем сильнее проявляется главная причинная связь, на которую накладывается искажающий ее "случайный шум" от влияния случайных факторов.

В главах темы 4 рассмотрены вопросы *оценки зависимостей между СВ* (уравнений связывающих СВ) и *оценки силы связи*, т.е. различные индексы, которые численно показывают как сильна эта связь. Вначале рассматриваются зависимости и связи между двумя переменными (*парные зависимости и связи*), затем они будут распространены на несколько СВ (*множественные связи и зависимости*).

Глава 4.1. МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ: МАТЕМАТИЧЕСКАЯ ФОРМУЛИРОВКА ЗАДАЧИ, ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ, ОЦЕНКА ДОСТОВЕРНОСТИ, ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ

4.1.1. Математическая модель парной линейной регрессии и точечная оценка ее параметров

Пусть имеется совместная выборка двух метеорологических величин x и y, которую можно назвать системой двух СВ и записать этот факт как (x, y), что означает, что система (x, y) получена путем совместных измерений x и y и будет статистически анализироваться далее также совместно. Здесь могут встретиться два случая: 1) x и y – обе являются СВ (например, x – температура воздуха, y – осадки); 2) y – есть СВ, а x – неслучайная (например, y – температура воздуха, x – высота места). Применяемый математический аппарат в обоих случаях одинаков. Поставим вопрос как отыскать зависимость y от x, которую обозначим как $y=y(x)$, помня, что это не функциональная, а статистическая зависимость.

В простейшем случае, когда причинная связь y и x линейна, модель этой зависимости будет иметь вид (волнистая черта над y специально подчеркивает, что это не наблюденное, а определенное по модели значение $\tilde{y}$):

$$\tilde{y}_i = b_0 + b_1 x_i \pm \sigma_м , \qquad (4.1)$$

где $\sigma_м$ – случайная ошибка модели, которая с различной вероятностью может принимать определенные значения.

Таким образом, коренным отличием (4.1) от функциональной зависимости является наличие случайной ошибки модели $\sigma_м$ и способ определения параметров b_0 и b_1. Параметры модели (4.1) определяются по выборке (x, y) объема n, т.е. используя наблюденные пары (x_i, y_i) методом наименьших квадратов (введем для него специальное обозначение МНК). Его суть состоит в том, что минимизируется сумма квадратов уклонений $(y_i - \tilde{y}_i)^2$, т.е. используется в качестве исходного выражение:

$$u_{мин} = \sum_n (y_i - \tilde{y}_i)^2 . \qquad (4.2)$$

Для вычисления параметров b_0 и b_1 с использованием МНК, надо в (4.2) подставить (4.1) (без учета $\pm\sigma_м$), продифференцировать полученное выражение по b_0 и b_1, приравнять частные производные нулю, а затем выразить b_0 и b_1. Согласно этой процедуре получим

следующую систему нормальных уравнений МНК, служащих для определения параметров b_0 и b_1:

$$b_0 = \frac{\sum y_i \cdot \sum x_i^2 - \sum (x_i y_i) \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, \tag{4.3}$$

$$b_1 = \frac{n \sum (x_i y_i) - \sum x_i \cdot \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, \tag{4.4}$$

где x_i и y_i – наблюденные значения x и y в не сгруппированной выборке.

Если определить b_1 по (4.4), то из (4.1) можно также получить:

$$b_0 = \bar{y} - b_1 \bar{x}, \tag{4.5}$$

где $\bar{x}$ и $\bar{y}$ - средние значения x и y, определнные по выбрке.

Эти формулы получены для несгруппированной выборки, однако они без труда обобщаются на случай сгруппированной выборки. На практике все парамеры регрессии (4.1) рассчитываются по программам Excel, что снимает все трудности вычислительного характера.

После того, как по (4.3)–(4.5) получены b_0 и b_1 модель (4.1), найденная по МНК, принимает конкретный вид. Задавая наблюденные значения x_i, по (4.1) можно вычислить все n значений $\tilde{y}_i$. Это дает возможность найти несмещенную среднюю квадратическую ошибку модели $\sigma_м$, равную:

$$\sigma_м \approx s_м = \left[\frac{1}{n-2} \sum_n (y_i - \tilde{y}_i)^2 \right]^{0,5}, \tag{4.6}$$

где число степеней свободы равно $(n-2)$, т.к. для вычисления модели использованы два параметра b_0 и b_1, рассчитанные по выборке, а σ по этой же причине заменено на s (выборочную оценку СКО).

Интересно представить, как может выглядеть опытное поле точек (x_i, y_i) при различных статистических зависимостях на диаграмме рассеивания, если по одной оси отложить x_i, а по другой y_i. Этот вид показан на графиках рис. 4.1 и 4.2.

На рис. 4.1 приведен конкретный пример вида опытного поля точек с нанесенной на него линиией регрессии и ее уравнением, показывающий высотную зависимость средних годовых температур воздуха по данным 33 метеостанций Кыргызстана. Эта высотная регресия имет вид:

$$y=-5.143*x+14.88\pm2.6,$$

где b_1=-5.143°С/км – есть угловой коэффициент, равный вертикальному градиенту температуры, который показывает, что средняя годовая температура в Кыргызстане понижатся в среднем на -5,143°С на каждый км высоты; b_0=14.88°С – остаточный член, который

представляет собой отрезок, отскаемый линией регрессии (если ее продлить влево до нулевой высоты) на вертикальной оси температур; $s_м$=±2,6°С - средняя квадратическая ошибка урвнения регрессии, характеризующая разброс данных отдельных станций около линии регресии.

Из рис. 4.1 также видно, что опытное поле точек имеет вид линейной полосы, проходящей примерно по диагонали из левого верхнего угла графика в правый нижний. При этом прямая регрессии адекватно описывает усредненную зависимость понижения температуры с высотой, проходя интерполяционным образом через поле точек.

Большое значение имеет величина ошибки регрессии, которая численно показывает, как сильно на изменение средних годовых температур, кроме высоты места, влияют другие случайные факторы, которые не учитываются регрессией. К таким факторам, прежде всего, относятся индивидуальные орографические условия расположения станций, сильно влияющие на режим температуры. Можно отметить поэтому, что, с учетом этого обстоятельства, точность полученного уравнения регрессии является вполне удовлетворительной, и ее можно с успехом использовать на практике для приближенных климатических расчетов.

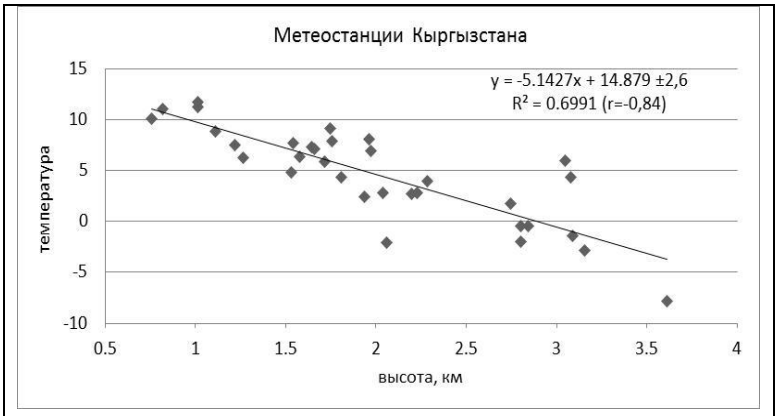


Рис.4.1 Линейная регрессия зависимости средних годовых температур от высоты места, полученная по данным 33 метеостанций Кыргызстана

На рис. 4.2 показана возможная геометрия опытного поля точек при различной силе и направленности корреляционной связи. Как видно, для линейной (положительной и отрицательной) статистической зависимости *y* от *x* опытное поле точек должно быть вытянуто примерно вдоль диагонали плоскости *yox в виде линейной полосы*. Чем уже эта по-

лоса, тем сильнее зависимость y от x. Математическая модель (4.1), если построить ее график, пройдет через опытное поле точек интерполяционным образом и обязательно через точку $(\bar{x}, \bar{y})$. Значение b_0 будет численно равно отрезку, отсекаемому на оси y, а b_1 – тангенсу угла наклона $\tilde{y}(x_i)$ к оси x. Значение $s_м$ (ошибка модели) характеризует разброс опытных точек по оси y относительно прямой $\tilde{y}_i(x_i)$.

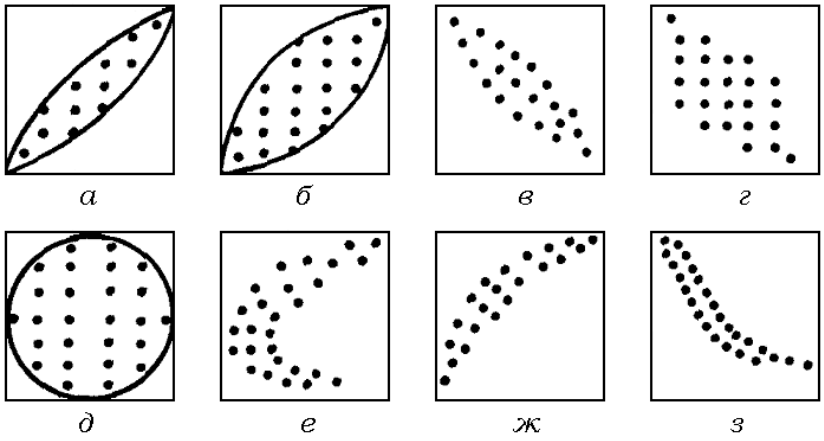


Рис. 2. Вид опытного поля точек на диаграмме рассеивания:
а – сильная линейная положительная зависимость y_i от x_i; *б* – умеренная линейная положительная зависимость; *в* – сильная линейная отрицательная зависимость; *г* – умеренная линейная отрицательная зависимость; *д* – отсутствие какой-либо статистической зависимости; *е* – неопределенная нелинейная статистическая зависимость (может быть описана двумя уравнениями); *ж* – сильная положительная нелинейная зависимость; *з* – сильная отрицательная нелинейная зависимость.

Модель (4.1), полученная таким образом, носит название *парной линейной регрессии y на x*. Надо ясно понимать, что поскольку все параметры модели b_0, b_1 и $\sigma_м$=$s_м$, получены по выборке, то они сами являются СВ. Очевидно, для того чтобы уверенно применять регрессию на практике надо, прежде всего, убедиться, что она статистически значима, т.е. полученная зависимость не может быть объяснена чисто случайными причинами (например, малым числом опытных точек), а отражает реальные связи x и y. Кроме того полученное уравнение регрессии следует использовать на практике только в пределах значений независимой переменной x, которые получены в выборке. Возможен лишь небольшой выход за этот диапазон, т.к. неизвесто как бы располагались там недостающие точки наблюдений. Например, уравнение высотной регрессии на рис. 4.1 можно применять округленно в диапазоне высот 0,5-4 км. Подробнее все это будет рассмотрено в следующих пунктах темы 4.

4.1.2. Три источника дисперсий в регрессионном анализе, связь трех дисперсий между собой. Коэффициент детерминации. Доверительный интервал для линии регрессии

Пусть, как и прежде, x_i, y_i – наблюденные в выборке значения переменных, $\bar{x}$ и $\bar{y}$ – их средние значения, рассчитанные по выборке, $\tilde{y}_i$ – рассчитанные по найденной модели парной линейной регрессии значения y по заданным x_i. Тогда общая дисперсия y в выборке (обозначим ее точное значение σ_y^2, а оценку как s_y^2) будет определяться суммой Σ_y:

$$\Sigma_y = \Sigma_y (y_i - \bar{y})^2, \tag{4.7}$$

$$\sigma_y^2 \approx s_y^2 = \frac{1}{n-1}\Sigma_1 (y_i - \bar{y})^2. \tag{4.8}$$

Очевидно, что часть этой дисперсии определяется зависимостью y от x по (4.1) и может быть *объяснена этой зависимостью*. Обозначим эту закономерную составляющую общей дисперсии через $\sigma_1^2 \approx s_1^2$. Значение s_1^2 будет связано с суммой Σ_1:

$$\Sigma_1 = \Sigma_1 (\tilde{y}_i - \bar{y})^2, \tag{4.9}$$

$$\sigma_1^2 \approx s_1^2 = \frac{1}{n-1}\Sigma_1 (\tilde{y}_i - \bar{y})^2. \tag{4.10}$$

Но регрессия (4.1) полностью не объясняет всю общую дисперсию y, иначе бы ошибка модели $\sigma_м$=0, т.е. всегда $\Sigma_y > \Sigma_1$ и $\sigma_y^2 > \sigma_1^2$. Не объясненная регрессией часть общей дисперсии носит название *остаточной дисперсии*, которая собственно определяет ошибку модели (4.1). Она определяется через Σ_2:

$$\Sigma_2 = \Sigma_2 (y_i - \tilde{y}_i)^2, \tag{4.11}$$

$$\sigma_2^2 = s_м^2 \approx \frac{1}{n-2}\Sigma_2 (y_i - \tilde{y}_i)^2. \tag{4.12}$$

В результате можно записать очень важные соотношения между введенными тремя суммами и дисперсиями:

$$\Sigma_y = \Sigma_1 + \Sigma_2; \quad s_y^2 = s_1^2 + s_2^2. \tag{4.13}$$

Именно на соотношениях (4.13) базируется оценка статистияеской значимости регрессии и корреляции с использованием F–критерия Фишера, что будет рассмотрено ниже.

Графической иллюстрацией взаимосвязи трех сумм и дисперсий является рис. 4.3, на котором показаны три источника дисперсий для 6 опытных точек (x_i, y_i), по которым получена линейная регрессия $\tilde{y}_i(x_i)$.

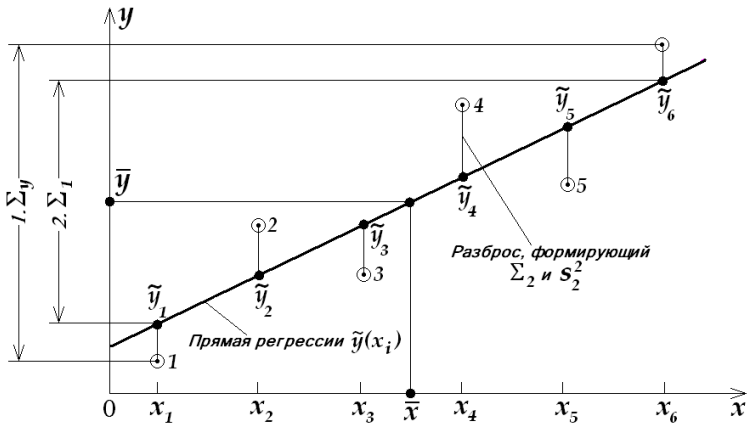


Рис. 4.3. Иллюстрация возникновения трех видов дисперсий:

1) общий разброс y_i в выборке, характеризуемый Σ_y; 2) разброс $\tilde{y}_i$, полученных по регрессии $\tilde{y}_i(x_i)$, характеризуемый Σ_1; 3) вертикальными отрезками показаны уклонения $(y_i - \tilde{y}_i)$, сумма квадратов которых составляет Σ_2 (эта сумма определяет среднеквадратическую ошибку регрессии); 1–6 – опытные точки (x_i, y_i) по которым рассчитана регрессия $\tilde{y}_i(x_i)$.

Таким образом, все сказанное можно представить табличкой-схемой:

Источник дисперсии	*Описание источника*	*Формула дисперсии*
1. Общая	Общий разброс зависимой переменной y в выборке	$\sigma_y^2 \approx s_y^2 = \frac{1}{n-1}\Sigma_1(y_i - \bar{y})^2$
2. Объясняемая или детерминированная	Разброс зависимой переменной y, который количественно описывается регресссией	$\sigma_1^2 \approx s_1^2 = \frac{1}{n-1}\Sigma_1(\tilde{y}_i - \bar{y})^2$
3. Остаточная или чисто случайная	Разброс зависимой переменной, вызванный не учитываемыми регрессией случайными факторами	$\sigma_2^2 = s_м^2 \approx \frac{1}{n-2}\Sigma_2(y_i - \tilde{y}_i)^2$

Коэффициент детерминации. Введем теперь на основании (4.13) весьма важное понятие – коэффициент детерминации B как отношение:

$$B = \frac{\Sigma_1}{\Sigma_y} = \frac{s_1^2}{s_y^2}. \qquad (4.14)$$

Величина *B* характеризует качество найденного уравнения регрессии. Она показывает, какая доля общей дисперсии s_y^2 в выборке объясняется (т.е. определяется или детерминируется) регрессией. Например, если *B*=0,4, то это значит, что 40% общей дисперсии s_y^2 описывается (т.е. объясняется) регрессией, а 60% соответствуют не учитываемым регрессией случайным факторам. Если же *B*=0,80, то только 20% дисперсии можно отнести к влиянию случайных факторов, тогда как 80% – объясняются регрессией и теперь эта регрессия отражает достаточно хорошо имеющуюся в выборке закономерность $\tilde{y} = \varphi(x_i)$.

Величина B характеризует в этом плане качество той конкретной регрессии, которая найдена. Поэтому она используется в качестве показателя для любого вида регрессий – линейных, нелинейных и множественных. В Excel параметр *B* называется показателем *достоверности* регрессии.

Доверительный интервал для линии регрессии. Вычисляя $\tilde{y}_i$ по регрессии (4.1), мы всегда по заданным x_i получаем оценку средних значений $\bar{\tilde{y}}_i$, лежащих налинии регрессии. Чтобы получить

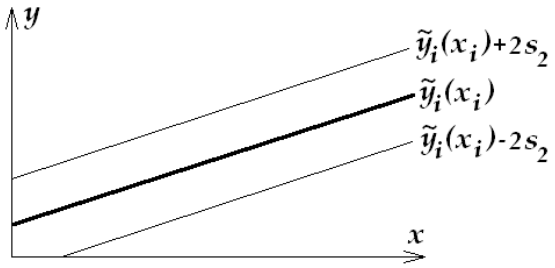


Рис. 4.4. Построение доверительного интервала $\tilde{y}_i(x_i) \pm 2s_2$

для линии регрессии при *p*=0,95 и *q*=0,05.

доверительный интервал, в который могут попасть действительные (индивидуальные) значения y_i надо построить ниже и выше прямой $\tilde{y}_i = b_0 + b_1 x_i$ две линии, отстоящие от нее при $n>30$ на $\pm z_{q/2} s_2$, где q – уровень значимости, связанный с доверительной вероятностью $p=1-q$. Если объем выборки мал ($n<30$), то вместо $z_{q/2}$ надо взять квантиль t-распределения с (n-2) степенями свободы, т.е. $t_{(n-2),\, q/2}$. На рис. 4.4 показан пример такого построения доверительного интервала для регрессии $\tilde{y}_i(x_i)$.

В этом примере $z_{q/2}$=1,96≈2, что соответствует *p*=0,95 и *q*=0,05. Это одновременно означает, что 95% опытных точек (x_i, y_i) выборки должны лежать внутри полосы, ограниченной тонкими прямыми графика рис 4.4.

4.1.3. F–распределение Фишера и оценка статисической значимости парной линейной регрессии

F-распределение Фишера играет исключительно большую роль в дисперсионном и регрессионном анализе для построения различных статистических критериев. Заметим, что здесь *F* – это случайная величина, которую мы раньше обозначали буквами *x*, *t*, *z* и т.д. Буквой *F* обозначалась *интегральная функция распределения*. Но исторически сложилось так, что для распределения Фишера во всех статистических справочниках, учебниках и таблицах за буквой *F закреплено обозначение СВ* и, конечно, мы не вправе вносить здесь изменения.

Пусть имеются две СВ u_1 и u_2 распределенные по χ^2 соответственно с n_1 и n_2 степенями свободы. Тогда, их отношение *F*, равное:

$$F = \frac{u_1}{n_1} \Big/ \frac{u_2}{n_2} = \frac{n_2 u_1}{n_1 u_2}, \qquad (4.15)$$

подчинено *F*-распределению Фишера, которое зависит только от числа степеней свободы числителя – n_1 и знаменателя – n_2.

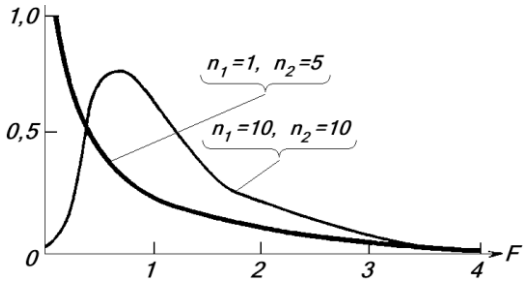


Рис. 4.5. Плотность вероятности *F*-распределения: F (n_1=1; n_2=5) и F (n_1=10; n_2=10).

Кривая функции плотности $f(F, n_1, n_2)$ имеет характерный вид, показанный на рис. 4.5. Это модальная, сильно правоасимметричная кривая, правая ветвь которой асимптотически медленно стремится к оси абсцисс. При n_1=1 она имеет вид обратной экспоненты.

F - распределение обладает следующими свойствами.

1. СВ $F \geq 0$ – существенно положительная величина.

2. $мо = \overline{F} = \frac{n_2}{n_2 - 2}$ (при n_2>2).

3. $\sigma^2(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$ (при n_2>4).

4. Мода $F_м = \frac{n_2(n_1 - 2)}{n_1(n_2 + 2)}$ (при n_1>2) и находится вблизи *F*=1.

5. Коэффициент асимметрии *A*:

$$A = \frac{(2n_1 + n_2 - 2)\sqrt{8(n_2 - 4)}}{(n_2 - 6)\sqrt{n_1 + n_2 - 2}}$$ (при n_2>6).

В Excel *F*-распределение представлено вычислением нескольких функций. Из них для оценки статистической значимости парной линейной регрессии следует использовать программу F.ОБР.ПХ. С остальными программами будем знакомиться по мере необходимости при решении с их помощью возникающих других задач.

Итак, требуется проверить нулевую статистическую гипотезу *(H_0),* которая состоит в том, что на заданном уровне значимости *q* (на уровне доверительной вероятности *p=1-q*), найденная по выборке эмпирическая регрессиия является статистически *незначимой*. Это равносильно тому, что закономерная по (4.9) сумма $\Sigma_1 = \Sigma_1(\tilde{y}_i - \bar{y})^2$, *незначимо* отличается от остаточной случайной по (4.11) суммы $\Sigma_2 = \Sigma_2(y_i - \tilde{y}_i)^2$. Поэтому эмпирическое значение *F*-критерия (с учетом свойств *F*-распределения) будет равно:

$$F(эмп) = \frac{\Sigma_1 : k_1}{\Sigma_2 : k_2} \qquad (4.16)$$

где k_1 – есть число степеней свободы числителя (CC_1), равное числу независимых переменных в уравнении регрессии; $k_2=(u-k_1-1)$ – есть число степеней свободы знаменателя (CC_2), *n* – объем выборки.

В рассматриваемой задаче для парной регрессии число независимых переменных равно 1, поэтому имеем:

$$k_1 \text{ (или } CC_1) = 1, \qquad k_2 \text{ (или } CC_2) = n-2 \qquad (4.17)$$

Вычисление *F*-критерия и принятие решения по нулевой гипотезе включает следующие этапы.

1. По (4.16) рассчитывается эмпирическое значение *F*-критерия (на практике для вычисления параметров регрессии используется программа ЛИНЕЙН, по которой находится и *F(эмп)*, с чем познакомимся в п.4.2.2).
2. *З*адается вероятность уровеня значимости критерия (обычно вероятность *q* принимается равной 0,01, 0,05 или 0,10), тем самым задается и уровень доверительной вероятности *p=(1-q);* по правилам (4.17) число степеней свободы числителя CC_1= 1, а число степеней свободы знаменателя CC_2= *n-2*.
3. По пограмме Excel «F.ОБР.ПХ» по заданному *q (*примем *q* =0,05) и известным CC_1=1 и CC_2= *n-2*. определяется критическое значение критерия, т.е. квантиль - *F(крит)*.
4. Если - *F(эмп)< F(крит)*, то гипотеза - *H_0: «регрессия не значима» - принимается* на уровне значимости *q* (т.е. с доверительной вероятностью *p*); если - *F(эмп)>F(крит)*, то гипотеза - *H_0: отвергается* на уровне значимости *q* (т.е. с до-

верительной вероятностью *p*) и принимается альтернативная гипотеза - *«регрессия значима»*.

Принятие альтернативной гипотезы - *«регрессия значима»* - означает, что на уровне доверительной вероятности *p (*т.е. с риском совершить ошибку с вероятностью *q=5%),* регрессия может быть тспользована на практике для прикладных климатических расчетов. Так, для примера на рис. 4.1, где рассматривалась высотная зависимость средних годовых температур на территории Кыргызстана, для $q=0,05$, $CC_1=1$ *и* $CC_2=31$ имеем: *F(эмп)= 72,03 и F(крит)=4.16; так как F(эмп)> F(крит),* то высотная регрессия является статистически значимой на уровне доверительной вероятности p=0,95 и может быть использована на практике.

Отметим также, что статистическая значимость регрессии означает, что ее угловой коэффициент b_1 значимо отличается от нуля и, следовательно, регрессия не параллельна горизонтальной оси температур. Поэтому нулевую гипотезу можно было бы также сформулировать как H_0: $b_1=0$ и проверить с помощью другого критерия, основанного на *t*-распределении Стьюдента (это будет рассмотрено в теме 5).

Кроме того, можно построить еще одно полезное *F*-отношение Фишера, равносильное по использованию (4.16). Для этого надо сравнить общую дисперсию (4.8) по выборке - s_y^2 и остаточную дисперсию по (4.12) - s_2^2, т.е. получить *F (эмп)* как

$$F\ (эмп) = s_y^2 / s_2^2 \qquad (4.18)$$

Значение *F (крит)* в этом случае также определяется по программе «F.ОБР.ПХ», но по другим степеням свободы, равным:

$$CC_1\ (\text{числитель}) = (n-1),\ CC_2\ (\text{знаменатель}) = (n-2). \qquad (4.19)$$

Так, для регрессии на рис. 4.1 имеем: s_y^2 =21,334, s_2^2=6,626, а их отношение *F(эмп)= 3,22.* Для q=0.05 и CC_1=32 и CC_2=31 по программе F.ОБР.ПХ .получим *F(крит)=1,82. Таким образом, как и следовало ожидать, п*о варианту *F*-критерия (4.18) точно также получено, что на уровне *q=0.05* высотную регрессию следует признать статистически значимой.

Заметим также, что отношение (4.18) численно показывает, во сколько раз значения $\tilde{y}_i$, получаемые по уравнению регрессии, по качеству предсказания лучше, чем при использовании прогноза по среднему по выборке значению $\bar{y}$. Так, среднее значение годовых температур по 33 станциям Кыргызстана по выборке на рис. 4.1 равно 4,58°С. Эту величину можно взять в качестве грубой оценки средней годовой температуры в Кыргызстане независимо от высоты места. Если, однако, для таких оценок использовать расчет-

ные значения по найденной высотной регрессии, то точность их будет в 4,6 раза выше, чем эта средняя грубая оценка, полученная без учета высоты места.

Глава 4.2. ЛИНЕЙНАЯ КОРРЕЛЯЦИОННАЯ СВЯЗЬ МЕЖДУ ДВУМЯ СЛУЧАЙНЫИ ВЕЛИЧИНАМИ

В главе 4.1 было показано, как по МНК можно установить линейную зависимость *y* от *x*, не ставя вопрос о причине этой зависимости и ограниченно решая его о силе связи. Косвенно о силе связи можно было судить по средней квадратической ошибке регрессии s_2.

Вопрос о причинности связей *y* и *x* лежит за рамками статистической теории (он должен решаться предметными науками), но все же именно методы статистики помогают во многих случаях установить эту причинность и выявить силу связи. Последнее достигается использованием второй формы анализа зависимости *y* и *x* – *корреляционного анализа силы связи между ними*, для чего используются безразмерные показатели силы связи:

- коэффициент линейной корреляции (или просто – коэффициент корреляции),
- индекс корреляции или специализированный коэффициент корреляции для заданного типа уравнения регрессии.

Методы регрессионного и корреляционного анализов тесно связаны между собой и обычно их выполняют совместно. При этом для нормально распределенной системы СВ (*x, y*) имеет место только линейная корреляционная связь.

4.2.1. Коэффициент линейной корреляции *r*, его свойства и связь с регрессией, оценка значимости r

В п.1.2.1 наряду с начальными и центральными моментами было введено понятие *смешанных начальных* m_k^* и *центральных* μ_k^* моментов СВ (формулы (1.17) и (1.18)).

Нас будет интересовать только второй смешанный центральный момент, который согласно (1.18) при k_1=1 и k_2=1 имеет вид:

$$\mu_2^* = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}), \qquad (4.20)$$

Он носит название ковариационного момента (или просто – ковариации) и характеризует *силу линейной корреляционной связи x и y в выборке (x, y)*. Неудобство непосред-

ственного использования (4.20) состоит в том, что значение μ_2^* зависит не только от силы связи *x* и *y*, но и от единиц измерения обеих СВ. Например, x_i – может измеряться в мм, см, м и т.д., а y_i , - г, кг, т и т.д. Чтобы исключить влияние единиц измерения переходят к безразмерному параметру *r, разделив* (4.20) на $\sigma_x * \sigma_y$

$$r = \frac{1}{n-1} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad (4.21)$$

Безразмерное выражение (4.21) носит название парного линейного коэффициента корреляции *r(x,y)* или просто *r*. Для точечной оценки *r* могут быть использованы и другие равносильные формулы, которые мы приводить не будем, их можно найти, наример, в [23]. Коэффициент корреляции *r* обладает следующими свойствами.

1. Величина *r* есть показатель силы *линейной* корреляционной связи, т.е. он отражает только линейную составляющую связи *x* и *y* в выборке. Его значение меняется в пределах:

$$-1 \le r \le 1: \qquad (4.22)$$

- $r=\pm 1$ – линейная корреляционная связь переходит в аналогичную чисто функциональную;
- $r=0$ – линейная корреляционная связь отсутствует (но может быть иная, нелинейная связь, факт того, что $r=0$ этого не отрицает, кроме случая нормальной системы *(x, y)*, когда $r=0$ означает отсутствие любого вида связи);
- знак *r* указывает на направление связи: если знак +, то связь прямая (регрессия – прямо пропорциональная зависимость), если знак –, то связь обратная (регрессия – обратно пропорциональная зависимость)

2. Величина $r^2=B$ есть коэффициент детерминации (4.14).

3. Средняя квадратическая ошибка r выражается формулой

$$\sigma_r = \frac{1-r^2}{\sqrt{n-2}} \qquad (4.23)$$

4. В случае, когда *x* и *y* обе являются СВ, имеет место не одно, а два уравнения регрессии: первое – уже рассмотренная регрессия «*y* по *x*» (коротко – *y/x*), когда роль независимой переменной выполняет *x*, и второе – регрессия «*x* по *y*» (коротко – *x/y*), когда независимой переменной является *y*, а *x* предсказывается по *y*. Через *r* оба эти уравнения записываются так:

$$\tilde{y}_i - \bar{y}_i = r \frac{\sigma_y}{\sigma_x} (x_i - \bar{x}) \pm \sigma_{2,y/x}, \qquad (4.24)$$

$$\tilde{x}_i - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y_i - \bar{y}) \pm \sigma_{2,x/y}. \qquad (4.25)$$

После подстановки в них численных значений $\bar{x}$, $\bar{y}$, r, σ_y и σ_x они приводятся к обычному виду:

$$\left.\begin{aligned} \tilde{y}_i &= b_{1,y/x} x_i + b_{0,y/x} \pm \sigma_{2,y/x}, \\ \tilde{x}_i &= b_{1,x/y} y_i + b_{0,x/y} \pm \sigma_{2,x/y}. \end{aligned}\right\}, \qquad (4.26)$$

и, следовательно,

$$r = \sqrt{b_{1,y/x} \cdot b_{1,x/y}} \; . \qquad (4.27)$$

Наличие двух уравнений означает, что они не могут выражаться одно через другое (как при функциональной зависимости x и y), а, согласно МНК, должны строиться отдельно. При этом, разумеется, что брать за x, а что за y – дело исследователя и специфики решаемой задачи.

Наглядно положение двух прямых регрессий (4.26) показано на рис.4.6. Обе регрессии образуют «ножницы» с углом γ. Чем ближе $|r|$ к 1, тем меньше γ ($\gamma \to 0$ при $|r| \to 1$), при $r=\pm 1$ угол $\gamma=0$ и обе регрессии, сливаясь в одну, переходят в чисто функциональную линейную зависимость. При $r=0$ угол $\gamma=\pi/2$ и регрессии пересекаются под прямым углом.

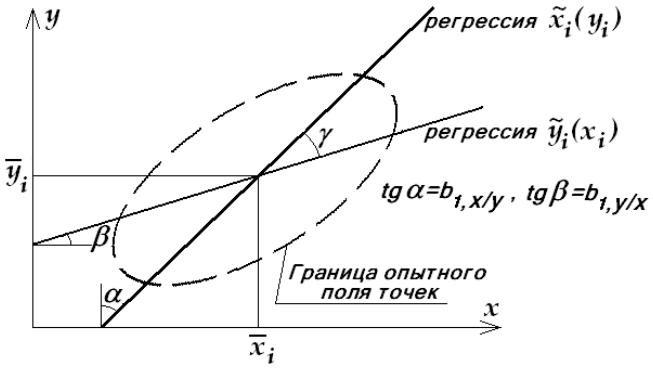


Рис. 4.6. Графики прямых регрессий $\tilde{y}_i(x_i)$ и $\tilde{x}_i(y_i)$.

Таким образом, согласно МНК, для наилучшего предсказания y по x надо брать регрессию (4.24), а для наилучшего предсказания x по y - регрессию (4.25).

5. Ошибки регрессий σ_2 через r могут быть вычислены по формулам:

$$\sigma_{2,y/x} = \sigma_y \sqrt{1-r^2} \; , \qquad (4.28)$$

$$\sigma_{2,x/y} = \sigma_x \sqrt{1-r^2} \; , \qquad (4.29)$$

где σ_y^2 и σ_x^2 – полные дисперсии y и x.

6. Доверительные границы для каждой из регрессий строятся обычным способом (см. п. 4.1.2) с использованием их стандартных ошибок по (4.28) и (4.29).

7. На практике считают обычно, что x и y достаточно тесно связаны между собой, если $|r| \geq 0{,}7$ (в этом случае $B \geq 0{,}49$). Однако *всегда* можно говорить о наличие связи и использовать ее практически и при $|r| < 0{,}7$, если ее удается объяснить физическими причинами.

8. Оценка статистической значимости r выполняется автоматически при оценке значимости регрессии по F-критерию (4.16): если регрессия на уровне заданной критической вероятности q является значимой/(не значимой), то это одновременно означает и значимось/(не значимость) коэффициента r. Это следует из того, что угловые коэффициенты двух регрессий b_1, согласно (4.24) и (4.25), равны:

$$b_{1(y/x)} = r\frac{\sigma_y}{\sigma_x}\ , \qquad b_{1(x/y)} = r\frac{\sigma_x}{\sigma_y}$$

Следовательно, равенство/(не равенство) b_1 нулю возможно только при равенстве/(не равенстве) нулю значения r. Разумеется, есть и прямой способ проверки значимости коэффициента корреляции, основанный на использовании t- распределения Стьдента, с которым мы познакомимся в теме 5.

4.2.2. Причинность корреляционно-регрессионных связей, ложная корреляция. Компьютерная реализация парной линейной корреляции и регрессии в Excel

Существо и причины корреляционно-регрессионных связей лежат вне статистических методов и должны объясняться исходя из предметного анализа конкретных наук. Установленная статистическая зависимость между y и x ничего не говорит об их причинной зависимости. Например, в любом крупном городе или стране количество потребляемых лекарств хорошо коррелирует с количеством смертей, а количество школ с числом душевнобольных. Точно так же у ребенка размер ладони коррелируют с умственным развитием. В этих примерах «причина парадокса» лежит на поверхности: и x и y коррелируют с третьей СВ – числом жителей для первых двух примеров и с возрастом ребенка, обусловливая их ложную корреляцию между собой.

Коррелирование x и y с третьей СВ z – один из самых распространенных случаев ложной корреляции. Он особенно «опасен» тем, что уровень ложной корреляции между x и y может быть высок, если высока внутренняя корреляция для систем *(x, z)* и *(y, z)*. Можно привести такой чисто метеорологический пример: давление воздуха p в горах понижается с высотой z, также понижается с высотой и температура T. Следовательно, хорошо будут коррелировать между собой для горных станций p и T. Если установленную для них

зависимость перенести на прогноз $T=\varphi(p)$ для равнинных территрий, т.е мало меняющихся высот, то результат будет отрицательным, т.к. изменение температуры здесь преимущественно зависит от других факторов. Сильная ложная корреляция может возникнуть за счет нелинейного преобразования переменных x и y. Например, пусть имеются две случайные величины (x_1, x_2) и $r(x_1, x_2)=0$. Перейдем к новой СВ $y=x_1/x_2$ (заметим, что x_1 и x_2 обе переменные и результат их деления также переменная). Теперь корреляция между y и x_2 или y и x_1 может достигнуть $|r|$=0,7-0,9. Это очень распространенный случай ошибки, допускаемый при обработке опытных данных. В этом случае экспериментатор по одной оси откладывает y_i, а по другой – отношение y_i/x_i или x_i/y_i. Ясно, что здесь ложная корреляция возникает за счет корреляции y_i с y_i или y_i с $1/y_i$. В этом плане недопустимы любые другие виды нелинейных преобразований переменных, когда надо исключить возникновения ложной корреляции.

Ложная корреляция может возникнуть за счет неравномерности распределения признака в координатной плоскости. Это также весьма типичный случай. Например, расчет коэффициента корреляции для 58 МС Киргизии между их широтой φ и долготой λ дал $r(\varphi, \lambda)$=0,54, а φ и z – $r(\varphi, z)$=-0,38. В этих случаях она возникла чисто случайно из-за неравномерного распределения их координат в пространстве.

Поэтому, рассчитав формальным путем корреляцию и регрессию надо убедиться, что корреляция между x и y имеет действительно причинный характер, для чего надо привлечь все возможности профессионального предметного анализа. Во многих случаях ложную корреляцию, когда она является следствием влияния третьих величин на x и y, можно исключить, вычислив частные коэффициенты корреляции. К сожалению, тема частной корреляции выходит за рамки настоящего учебника.. В любом случае, как уже отмечалось, объяснение корреляции и регрессии есть дело предметных наук, а не математической статистики.

Вычисление параметров парной линейной корреляции и регрессии в Excel. В Excel имеются более 10 программ для вычисления различных статистик парной линейной корреляции и регрессии. Из них наиболее полной является программа «ЛИНЕЙН», которая вычисляет значения следующих статистик: b_1 по (4.4) , s_{b_1} (средняя квадратическая ошибка b_1), b_0 по (4.3), s_{b_0} (средняя квадратическая ошибка b_0), $B=r^2$ по (4.14), s_2 по (4.6), $F(эмп)$ по (4.16), число степеней свободы $CC_2 = k_2 = (n-2)$, а также суммы Σ_1 и Σ_2 по (4.9) и (4.11).

Таблица 4.1

Значения средниех годовых температур по данным 33 разновысотных метеостанций Кыргызстана.

Станция	Z, км	T°C	Станция	Z, км	T°C
1. Фрунзе	0,756	10,1	18. Кара-Куджур	2,800	-0,5
2. Новороссийка	1,532	4,8	19. Ат-Ойнок	3,050	6.0
3. Чон-Арык	1,110	8,8	20. Нарын	2,040	2,8
4. Байтык	1,579	6,3	21. Казарман	1,266	6,2
5. Тюя-Ашу сев/	3,090	-1,4	22. Чаткал	1,937	2,4
6. Талас	1,217	7,5	23. Устье р. Тос	1,759	7,9
7. Токтогул	0,821	11,0	24. Ангрен	2,286	3,9
8. Чолпон-Ата	1,645	7,3	25. Ак-Терек-Гава	1,748	9,1
9. Пржевальск	1,716	5,9	26. Чаар-Таш	2,748	1,7
10. Рыбачье	1,660	7,1	27. Узген	1,012	11,2
11. Койлю	2,800	-2,0	28. Ош	1,016	11,7
12. Тамга	1,960	8,1	29. Кызыл-Джар	2,230	2,8
13. Тянь-Шань	3,614	-7,8	30. Гульча	1,542	7,7
14. Ак-Шийряк	2,844	-0.5	31. Хайдаркан	1,970	6,9
15. Каракольская	3,08	.4.3	32. Сары-Таш	3,155	-2,9
16. Кочкорка	1,810	4,3	33. Дараут-Коргон	2,220	2,7
17. Сусамыр	2,061	-2,1			

Покажем порядок работы с этой программой на примере расчета высотной регрессии для средних годовых температур по данным 33 метеостанций Кыргызстана, расположенных на высотах от 0,6 до 3,61 км, график которой показан на рис. 4.1. Исходные данные взяты из климатического справочника 1989 г. и приведены в табл. 4.1.

Результаты расчетов статистик корреляции и регрессии по программе ЛИНЕЙН выдаются в форме специальной таблицы, имеющий размер «2 столбца* на 5 строк». Поэтому для записи статистик надо заранее *выделить необходимый диапазон из 10 ячеек*. Например, пусть это будет диапазон ячеек, показанный в таблице 4.2. Рассчитанные статистики будут записаны в ячейки табл. 4.2, согласно приведенных буквенных обозначений.

Таблица 4.2

Схема записи параметров, рассчитываемых программой ЛИНЕЙН (для записи выделен залитый синим диапазон из 10 ячеек)

F	G
b_1	b_0
s_{b_1}	s_{b_0}
R^2	s_2
F	k_2
Σ_1	Σ_2

Дальнейшая процедура работы с программой состоит в следующем.

1. Через кнопку *fx* (мастер функций) запустить программу ЛИНЕЙН, для чего выделить в окне перечня функций «*статистические*», а затем программу ЛИНЕЙН.

2. Через кнопку ОК открыть окно *Аргументы функции* и последовательно ввести (для нашего примера):

- Известные значения *y* – это *массив* (весь столбец) температур *T* по табл. 4.1.
- Известные значения *x* – это *массив* (весь столбец) высот z по табл. 4.1.
- Константа – задается как *1 (истина)* или как *0 (ложь);* если будет введено 0 (ложь), то программа *задаст* статистику $b_0=0$, если будет введено 1 (истина), то программа будет *вычислять* b_0 обычным способом.
- Статистика – задается как *1 (истина)* или как *0 (ложь);* если будет введено 0 (ложь), то программа выдаст только одну станистику - b_1, если будет введено 1 (истина), то программа выдаст весь набор статистик по табл. 4.2.

6. Щелкнуть кнопку ОК, затем ввести курсор мыши *в верхнюю строку формул* и щелкнуть ЛКМ.

7. Одновременно нажать три клавиши: Ctrl+Shift+Enter (рекомендуется для отсутсвия сбоя сначала нажать первые две кклавиши Ctrl+Shift, а затем дожать третью - Enter).

8. В выделенном диапазоне ячеек появится массив результатов расчетов статистик по схеме табл. 4.2.

Так, для нашего расчетного примера получим следующий массив статистик корреляции и регрессии

-5.14265	14.87857
0.605962	1.293045
0.699102	2.574176
72.02495	31
477.2647	205.4178

Заметим, что в расчетном массиве статистик нет коффициента корреляции r. Однако его модуль легко получить извлеченем квадратного корня из значения B, а знак r совпадает со знаком b_1. Так, в нашем приере получим, что r= -0,84.

Подчеркнем еще раз, что значения *y* и *x* вводятся именно как *массивы* выделением одновременно всего столбца.

Кроме того отдельные параметры корреляции и регрессии в Excel вычисляются в следующих программах статистических функций:

- *ТЕНДЕНЦИЯ*: вычисляется массив точек *y* по заданному *x* массива *(x, y)*, по которому рассчитывается линейная регрессия,
- *СТОШУХ*: вычисляется ошибка линейной регрессии *y(x)* по заданному массиву точек *(x, y)*,
- *ПРЕДСКАЗ*: вычисляются предсказываемое значение *y* в заданной точке *x*, рассчитанное по линейной регрессии для заданного массива точек *(x, y)*,
- *ПИРСОН*: вычисляется коэффициент корреляции по заданному массиву точек *(x, y)*,
- *ОТРЕЗОК*: вычисляется отрезок b_0 регрессии по заданному массиву точек *(x, y)*,
- *НАКЛОН*: вычисляется коэффициент b_1 регрессии (равный ее тангенсу угла наклона) по заданному массиву точек *(x, y)*,
- *КОРРЕЛ*: вычисляется коэффициент корреляции по заданным массивам интервалов ячеек,
- *КОВАР*: вычисляется ковариация $\mathrm{cov}(x,y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$ по заданному массиву точек *(x, y)*,
- *КВ ПИРСОН*: вычисляется квадрат коэффициента корреляции, т.е. коэффициент детерминации $B=r^2$ по заданному массиву точек *(x, y)*.
- *ФИШЕР И ФИШЕРОБР* – прямое и обратное преобразования Фишера

Глава 4.3. НЕЛИНЕЙНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ: ПАРАБОЛИЧЕСКАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ, ДРУГИЕ ВИДЫ НЕЛИНЕЙНОЙ КОРРЕЛЯЦИИ И РЕГРЕССИИ В EXCEL, ИХ ВЫЧИСЛЕНИЕ И ИСПОЛЬЗОВАНИЕ

Равенство коэффициента корреляции r нулю означает только, что линейная корреляция и регрессия отсутствуют. Но связь может быть нелинейной и даже сильной, тогда как r, в качестве показателя только линейной связи, ничего не говорит об этом. Возможно также, что линейная составляющая связи может быть слабой, а нелинейная – сильной. Ясно, что при анализе выборки надо использовать все ее возможности относительно связи x с y, независимо от вида связи. Это можно выполнить, применив специальные методы анализа, которые рассматриваются ниже.

Однако здесь надо сделать важное замечание *принципиального* характера: МНК, который мы хотели бы использовать для оценки нелинейной корреляции и регрессии, применим только для уравнений регрессий, в которые искомые статистики (параметры уравнения) *входят линейно* (при этом сама переменная x может входить нелинейно).

Если же параметры регрессии входят в ее уравнение *нелинейно*, то метод МНК не применим для их отыскания. Наиболее распространенным путем решения задачи в этом случае является такое преобразование переменных, при котором искомые параметры будут входить в регрессию линейно. Однако надо помнить, что в этом случае по (4.2) минимизируется сумма квадратов уклонений не исходных переменных, а их преобразованных функций. Поэтому полученные коэффициенты регрессий не будут в строгом смысле соответствовать МНК, но вполне пригодны как приближение к их отыскиваемым наилучшим оценкам.

Чтобы отличить нелинейную регрессию от линейной и множественной, которая будет рассмотрена в главе 4.4, изменим систему обозначений для коэффициентов при x, обозначая их буквами a, b, c, d ... (а не b_0, b_1, b_2 ..., как в линейной и множественной регрессиях).

4.3.1. Полиномиальная (параболическая) корреляция и регрессия

Полиномиальная или параболическая корреляция, особенно второй степени, широко применяется на практике, вследствие своей математической универсальности: графики

параболы второго порядка могут иметь выпуклости вверх или вниз различной кривизны, что хорошо согласуется во многих случаях с опытными данными и находит простое физическое объяснение.

Формулы для парной параболической регрессии второго, третьего и четвертого порядков можно записать в виде:

$$\tilde{y} = a + bx + cx^2 \pm \sigma_2, \tag{4.30}$$

$$\tilde{y} = a + bx + cx^2 + dx^3 \pm \sigma_2, \tag{4.31}$$

$$\tilde{y} = a + bx + cx^2 + dx^3 + ex^4 \pm \sigma_2, \tag{4.32}$$

где *a, b, c, d* и *e* – параметры уравнений (статистики), которые подлежат определению по МНК по выборке; $\sigma_2 \approx s_2$ – средняя квадратичная ошибка модели регрессии, определяемая в общем случае по (4.6)

Теоретически можно повышать степень параболы, однако, как показал опыт, на практике обычно не приходится использовать параболу выше третьего порядка. Здесь дело не только в быстро возрастающем объеме вычислений, но, прежде всего, в физической (т.е. причинной) не объяснимости получаемых уравнений. В результате они обычно превращаются в простые интерполяционные формулы. В программах Excel расчеты ограничены вычислением параболы 6-порядка. Что же касается ручного счета, то он исчерпывает свои технические возможности на параболе второго–третьего порядка. Таким образом, реально, чаще всего, можно говорить о вычислении и применении парабол второго (4.30) и реже третьего (4.31) порядков.

Коэффициенты *a, b, c* и *d* парабол (4.30)-(4.32) и более высоких порядков входят в модели регрессий линейно. Поэтому МНК здесь применим для их отыскания в обычном классическом виде. Например, для параболы второго порядка требуемая система уравнений имеет вид:

$$\left.\begin{aligned} an + b\Sigma x + c\Sigma x^2 &= \Sigma y, \\ a\Sigma x + b\Sigma x^2 + c\Sigma x^3 &= \Sigma xy, \\ a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 &= \Sigma x^2 y. \end{aligned}\right\} \tag{4.33}$$

Аналогичным образом можно записать более сложные системы нормальных уравнений для парабол любого порядка. Решая эти системы относительно параметров уравнений *a, b, c* и т. д. получим их выборочные оценки. При этом все значения сумм, входящих в уравнения вида (4.33), есть числа, которые находятся по выборке. Не будем приводить здесь формулы для таких расчетов, т.к. необходимые вычисления выполняются по программам Excel и ручной счет не потребуется.

Силу корреляционной параболической связи для найденного *конкретного уравнения регрессии* можно оценить, если использовать для этого приведенные ранее в п. 4.1.2 основные соотношения регрессионного анализа:

$$\Sigma_y = \Sigma_1 + \Sigma_2; \quad s_y^2 = s_1^2 + s_2^2, \tag{4.34}$$

$$\Sigma_1 = \Sigma_1(\tilde{y}_i - \bar{y})^2; \quad \Sigma_2 = \Sigma_2(y_i - \tilde{y}_i)^2; \quad \Sigma_y = \Sigma_y(y_i - \bar{y})^2, \tag{4.35}$$

где $\bar{y}$ – рассчитанное по выборке среднее значение y; y_i – фактические значения y в выборке; $\tilde{y}_i$ – значения y, рассчитанные по найденной регрессии по заданным x_i.

Назовем *индексом корреляции* величину r_c, введенную соотношениями:

$$r_c^2 = \Sigma_1/\Sigma_y = 1 - \Sigma_2/\Sigma_y, \tag{4.36}$$

$$r_c^2 = s_1^2/s_y^2 = 1 - s_2^2/s_y^2, \tag{4.37}$$

$$r_c = \sqrt{r_c^2}. \tag{4.38}$$

Как видно, $r_c^2 = B$ есть по прежнему коэффициент детерминации, показывающий долю закономерной дисперсии s_1^2 в общей s_y^2. Этим объясняется введение индекса корреляции $r_c = \sqrt{B}$, как *специализированного* показателя силы корреляционной связи, соответствующей найденному по выборке уравнению регрессии. Величину r_c называют: индексом корреляции или специализированным коэффициентом корреляции (например, параболическим коэффициентом корреляции второго, третьего порядка и т.д.).

Индекс корреляции r_c, исходя из (4.36)–(4.38), обладает следующими очевидными свойствами:

1) $0 \le r_c \le 1$ и показывает силу нелинейной корреляционной связи x и y в выборке, соответствующей конкретному найденному уравнению регрессии;

2) если $r_c=0$, то соответствующая корреляционная связь отсутствует, при $r_c=1$ она переходит в чисто функциональную связь заданного вида;

3) всегда имеет место $r_c \ge r$, при этом при $r_c \approx r$ (лучше, $r_c^2 \approx r^2$) – связь близка к линейной и в точности линейна, когда $r_c = r$ ($r_c^2 = r^2$);

Средняя квадратическая ошибка значения r_c равна

$$\sigma_{r(c)} = (1 - r_c^2)/\sqrt{(n-2)} \tag{4.39}$$

4) $r_c^2 = B$ – соответствует коэффициенту детерминации;

5) значение r_c (как и r) ничего не говорит о том, существует ли в выборке какая-либо другая, более сильная связь, чем соответствующая найденной по конкретной заданной зависимости;

6) относительно r_c может быть проверена гипотеза о его значимости с использованием F-критерия Фишера (4.16) точно таким же образом, как это показано для линейной регрессии и корреляции в п.4.1.3. Однако теперь надо вручную посчитать $\sum_1$ и $\sum_2$, т.к. в программе для параболической корреляции Excel они не вычисляются

Формулу для средней квадратичесой ошибки уравнения параболической регрессии в (4.30)-(4.32) теперь можно записать как

$$\sigma_2 = s_2 = \sigma_y \sqrt{1 - r_c^2}\,, \qquad (4.40)$$

$\sigma_y = s_y$ – есть среднее квадратическое отклонение по выборке.

Отметим особо, что все сделанные выкладки и заключения (4.34)-(4.40) будут также полностью справедливы и для любых других типов нелинейных уравнений, а не относятся только к их параболическому виду.

Расчет параболической корреляции 2-6 порядков предусматривается в прогаммах Excel при построении корреляционных графиков и показан в слкдующем п. 4.3.2 (рис. 4.8).

4.3.2. Построение корреляционных графиков и расчет нелинейных корреляционных зависимостей в Excel

В статистических программах Excel по МНК рассчитываются следующие нелинейные регрессии.

1. Экспоненциальная зависимость - $y=ae^{bx}$
   где a и b –параметры, определяемые по выборке, а e= 2,718…
2. Логарифмическая зависимость – $y=a\ln x+b$
   где a и b –параметры, определяемые по выборке.
3. Полиномиальная (рассмотренная в п.4.3.1 параболическая) –
   $y=a+b_1x+b_2x^2+b_3x^3+b_4x^4+b_5x^5+b_6x^6$,
   где параметры а и b_i определяются по выборке, а показатель степени при x задается, равным от 2 до 6.
4. Степенная – $y=ab^x$,
   где a и b –параметры, определяемые по выборке.
5. Логарифмическое приближение - $y=ab^x$ и $y=b^x$,
   где a и b –параметры, определяемые по выборке.

Расчет зависимостей 1-4 предусмотрен в программах построения графиков, а расчет двух зависимостей 5 –задается в перечне функций «статистические» программой

ЛГРФПРИБЛ. При этом работа с программой ЛГРФПРИБЛ полностью аналогично с программой ЛИНЕЙН, рассмотренной в п. 4.2.2.

Подчеркнем еще раз, что все свойства параболической корреляции и регресии, рассмотренные в предыдущем п. 4.3.1, полностью распространяются и на все другие виды нелинейных зависимостей, в том числе и перечисленные выше типы 1-5.

Рассмотрим порядок построения корреляционных графиков в Excel с одновременным получением формул зависимостей 1-4 на примере выборки табл. 4.1, где приведены средние годовые температуры по 33 разновысотным станциям Кыргызстана.

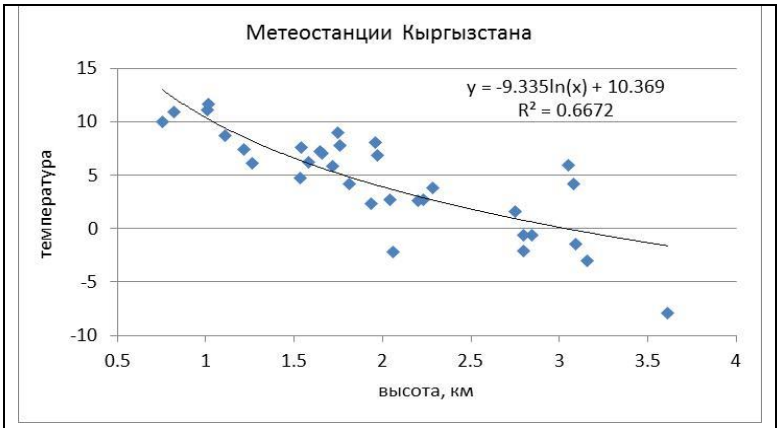


Рис. 4.7 Логарифмическая регрессия зависимости средних годовых температур от высоты места, полученная по данным 33 метеостанций Кыргызстана

На рис. 4.7 показано опытное поле точек, характеризующее высотное распределение средних годовых температур для 33 метеостанций Кыргызстана и рассчитанный по МНК график логарифмической регрессионной зависимоти от высоты места. Уравнение этой зависимости имеет вид:

$$y = -9.335\ln(x)+10.369,$$

где y,°С – средняя годовая температура; x, км – высота

Это уравнение взято с графика рис. 4.7, где оно получено следующим образом. Вначале, используя «Вставку» было постоено опытное поле точек по данным табл. 4.1 (для этого лучше использовать вставку *точечной* диаграммы). После этого, для получения уравнения регрессии, которое будет записано в поле графика, надо щелкнуть ПКМ (правой кнопкой мыши) на любую из точек графика. В появившемся диалоговом окне надо ЛКМ (левой кнопкой мыши) щекнуть на опции «добавить линию тренда». В появившемся *окне* вверху дается следующий перечень регрессионных зависимостей, которыми можно аппроксимировать выборку: экспоненциальная, линейная, логарифмическая, полино-

миальная и степенная. Надо выбрать из них требуюмую и щелкнуть на ней ЛКМ (если выбрана «полиномиальная», то дополнительно надо щелкнуть ЛКМ на задаваемом справа значении показателя степени от 2 до 6). После этого внизу *этого же окна* надо щелкнуть ЛКМ на кнопках: «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (B^2)». Закрыть окно. В поле графика будут записаны все эти результаты, как это показано на рис. 4.7.

Как видно из полученных данных (рис. 4.7), коэффициент детерминации для логарифмической регрессии B=0,6672, что дает специализированный коэффициени корреляции r_c= 0,82. Рассчитанная по общей формуле (4.6)

$$\sigma_м \approx s_м = \left[\frac{1}{n-2} \sum_n (y_i - \tilde{y}_i)^2 \right]^{0,5}$$

средняя квадратическая ошибка логарифмической регрессии s_2=2,71°C.

При этом чисто глазомерно видно, что логарифмичесая кривая вполне хорошо описывает распределение опытного поля точек.

Однако на рис. 4.1 это же поле точек не менее хорошо описывалось более простой линейной регрессией. Для того чтобы сделать выбор между этими двумя регрессиями надо сравнить численные значения полученных статистик. Для линейной регрессии они были равны: B=0,6991, r=-0,84, s_2=2,57°C. Хорошо видно, что переход от линейной к более сложной логарифмической зависимости не привел к более высокому коэффициенту корреляции и не сизил ошибку регресии. Напротив, эти показатели даже незначительно ухудшились. Одновременно в геометрии опытного поля точек на рис.4.1 и 4.7 заметно не прослеживается какой-либо нелинейности в их расположении. Особенно хорошо это видно на рис. 4.1. Все это позволяет уверенно заключить, что линейная регрессия, полученная в п.4.1.1, является более предпочтительной для практического использования по сравнению с логарифмической. Такое заключение основывается на общем статистическом правиле: *если переход к более сложной статистической модели не дает существенного (статисически значимого) улучшения результатов, то выбор надо сделать в пользу блее простой модели.* Причем во многих случаях статистическую значимость улучшения результатов удается проверить критериально. В данном случае в такой проверке нет необходимости, т.к. более сложная логарифмичесская корреляция привела не к улучшению, а к небльшому ухудшению результатов.

Приведем также на рис. 4.8 в чисто учебных целях результаты аппроксимации этой же выборки средних годовых температур параболическими зависимостями 3 и 6 порякков. Сравнение коэффициентов детерминации сразу же показывает, что статистическик качества регрессий по сравнению с линейной существенно не возросли. Так, для параболы 3-

порядка B=0,7116 и r_c=0,84, а для параболы 6-порядка B= 0,7563 и r_c=0,87. Однако главной трудносью принятия этих форм корреляции и регрессии для практике является другое: невозможность сколько-нибудь правдоподобным образом объяснить такой сложный вид высотной зависимости в распределении температуры. Поэтому еще раз подчеркнем, что приведение этотого графика преследует важную учебную цель – показать ненужность и ошибочность использования более сложных статистических моделий, там, где это не находит смысла.

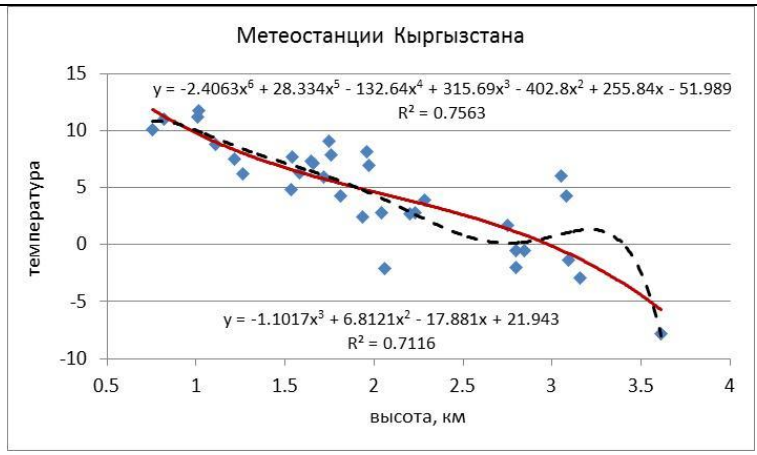


Рис. 4.8 Аппроксимация высотного распределения средних годовых температур по данным 33 метеостанций Кыргызстана экспоненциальной регрессионной зависимостью 3 (сплошная линия) и 6 (пунктир) степени.

В заключение заметим, что применение параболы второго порядка в данном случае привело к практическому совпадению линейного графика на рис. 4.1 и параболического второго порядка, ввиду почти не заметной кривизны последнего. Таким образом, в линейном характере понижения в Кыргызстане средних годовых температур с высотой нет никаких сомнений, по крайней мере, если судить по данным использованных метеостанций.

Глава 4.4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Весьма важной и распространенной является статистическая задача нахождения корреляционной связи и регрессионной зависимости не между двумя, а между тремя и более переменными $x_0, x_1, x_2 \ldots, x_k$. При этом одна из них, обычно это $x_0=y$, выступает в роли зависимой переменной или предиктанта (предсказываемой или прогнозируемой СВ вели-

чины), а остальные в роли независимых СВ-предикторов (предсказателей). Таким образом, задача сводится к нахождению уравнения функции в виде $x_0=y=\varphi(x_1, x_2, ..., x_k)$ и совокупной корреляционной связи $x_0=y$ с остальной системой СВ ($x_1, x_2, x_3 ..., x_k$). В наиболее простом виде она решается для линейной регрессии и линейной корреляции, где с успехом может быть использован МНК, как он применялся для случая двух переменных. Решение этой задачи и рассматривается в настоящей главе.

4.4.1. Множественная линейная регрессия, оценка ее параметров по МНК и основные свойства

Пусть имеется система СВ ($x_0, x_1, x_2 ..., x_k$), и пусть в ней $\tilde{x}_0 = \tilde{y}$ есть предиктант (предсказываемая или зависимая СВ), а $x_1, x_2, x_3 ..., x_k$ – независимые переменные или предикторы (предсказатели), по значениям которых в совокупности предсказывается СВ $\tilde{y} = \tilde{x}_0$. Тогда, для случая линейной зависимости можно записать уравнения множественной линейной регрессии в виде:

$$\tilde{y}(x_1, x_2, x_3 ..., x_k) = \tilde{x}_0 = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \pm s_2, \qquad (4.41)$$

где $b_0, b_1, b_2 ..., b_k$ – коэффициенты регрессии, которые надо определить по выборке с помощью МНК; $s_2=s_{рег}$ – среднее квадратическое значение ошибки регрессии, которая характеризует точность полученного уравнения.

Поскольку для нахождения коэффициентов применяется МНК, то принципиальный подход здесь ничем не отличается от рассмотренного для парной регрессии в главе 4.1. Он становится только более громоздким технически и это его свойство быстро растет с увеличением числа независимых переменных $x_1, x_2, x_3 ...$.

Точно также минимизируется сумма квадратов уклонений

$$\Sigma(y_i - \tilde{y}_i)^2, \qquad (4.42)$$

в результате чего находится система нормальных линейных уравнений, по которой вычисляются параметры уравнения (4.41).

Так, для регрессии из двух и трех независимых переменных:

$$y=b_0+b_1x_1+b_2x_2, \qquad (4.43)$$

$$y=b_0+b_1x_1+b_2x_2+b_3x_3, \qquad (4.44)$$

имеем соответственно следующие системы нормальных уравнений:

$$\left.\begin{aligned} b_0 n + b_1 \Sigma x_1 + b_2 \Sigma x_2 &= \Sigma y, \\ b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 &= \Sigma y x_1, \\ b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 &= \Sigma y x_2. \end{aligned}\right\} \qquad (4.45)$$

$$\left.\begin{aligned} &b_0 n + b_1 \sum x_1 + b_2 \sum x_2 + b_3 \sum x_3 = \sum y, \\ &b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + b_3 \sum x_1 x_3 = \sum x_1 y, \\ &b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 + b_3 \sum x_2 x_3 = \sum x_2 y, \\ &b_0 \sum x_3 + b_1 \sum x_1 x_3 + b_2 \sum x_2 x_3 + b_3 \sum x_3^2 = \sum x_3 y. \end{aligned}\right\} \quad (4.46)$$

При этом в этих уравнениях x_1, x_2, x_3 ..., x_k – это различные СВ, а их отдельные численные значения в выборке будут иметь обозначения как x_{1i}, x_{2i}, x_{3i} ..., x_{ki}, т.е. индекс 1, 2, 3 ..., k соответствует номеру СВ или фактора, а индекс i-численному i-тому значению СВ. Здесь этот индекс опущен, чтобы не загромождать выражения для формул.

Если, например, рассмотреть зависимость средних температур воздуха T в горных местностях от наиболее важных факторов, то в роли таковых, прежде всего, выступают: высота места z, широта – φ и долгота – λ. Тогда, уравнение (4.41) можно записать в виде:

$$T=b_0+b_1z+b_2\varphi+b_3\lambda \pm s_2, \quad (4.47)$$

где $z=x_1$, $\varphi=x_2$ и $\lambda=x_3$.

Следует отметить, что факторы z, φ и λ не вполне случайны для горных МС, т.к. размещение сети гидрометслужбы происходит по каким-то определенным правилам. Но вот предиктант T – есть чисто СВ, на формирование которой оказывают влияние и многие другие, не учитываемые регрессией (4.47), факторы. Однако, как и в случае линейной регрессии, применение математического аппарата МНК здесь остается в силе независимо от того, случайны или не случайны предикторы.

Для множественной линейной регрессии (4.41) справедливы основные соотношения регрессионного анализа п. 4.1.2:

$$\Sigma_y=\Sigma_1+\Sigma_2, \quad (4.48)$$

где

$$\Sigma_y = \Sigma(y_i - \bar{y})^2, \quad \Sigma_1 = \Sigma_1(\tilde{y}_i - \bar{y})^2, \quad \Sigma_2 = \Sigma_2(y_i - \tilde{y}_i)^2, \quad (4.49)$$

и соответственно:

$$s_y^2 = s_1^2 + s_2^2, \quad (4.50)$$

$$s_y = \left[\frac{1}{n-1}\Sigma(y_i - \bar{y})^2\right]^{0,5}, \; s_1 = \left[\frac{1}{n-k}\Sigma(\tilde{y}_i - \bar{y})^2\right]^{0,5},$$

$$s_2 = \left[\frac{1}{n-k-1}\Sigma(y_i - \tilde{y}_i)^2\right]^{0,5}, \quad (4.51)$$

где k – число использованных независимых переменных.

Тогда, коэффициент детерминации B будет равен:

$$\left.\begin{aligned} &B = \Sigma_1/\Sigma_y = 1 - \Sigma_2/\Sigma_y, \\ &B = s_1^2/s_y^2 = 1 - s_2^2/s_y^2. \end{aligned}\right\} \quad (4.52)$$

Он также показывает, какая доля общей дисперсии s_y^2 объясняется регрессией, т.е. найденной зависимостью y от (x_1, x_2, x_3 ..., x_k). Аналогично, отношение $F=\Sigma_y/\Sigma_2$ показывает, во сколько раз регрессия предсказывает результат лучше по сравнению с предсказанием $y = \bar{y}$.

Проверка статистической значимости регрессии может быть сделана по F-критерию Фишера:

$$F(эмп) = \frac{\sum_1 : k_1}{\sum_2 : k_2}, \qquad (4.53)$$

порядок использования которого описан в п. 4.1.3 полностью применим для множественной регрессии. При этом одновременно оценивается и значимость коэффициента множественной корреляции R, свойства которого рассматриваются в следующем п. 4.4.2.

Надо тоьлько помнить, что k_1 – есть число степеней свободы числителя ($СС_1$), равное числу *независимых* переменных в уравнении регрессии; $k_2=(u-k_1-1)$ – есть число степеней свободы знаменателя ($СС_2$), n – объем выборки. Теперь, в отличие от парной регрессии (где $k_1=1$), значение $k_1 \geq 2$, т.к. в уравнении множественной регрессии число независимыж переменных два и более.

Если $F_{эмп}$ по (4.53) больше $F_{крит}$, найденного по программе F.ОБР.ПХ (по заданным значениям вероятности q, $СС_1$ и $СС_2$) то регрессию на уровне доверительной вероятности $p = 1- q$ следует признать статистически значимой и наоборот. При этом принятие гипотезы о статистической незначимости множественной регрессии означает, что все ее угловые коэффициенты равны нулю, т.е.

$$b_1=b_2=b_3= ..., b_k=0.$$

В целом вопрос о выборе наилучшего подмножества предсказателей-факторов для множественной регрессии и возможностей ее применимости на практике значительно сложнее, чем для парной регрессии, что рассматривается ниже.

Все вычисления параметров множественной регрессии будут выполняться нами по прогамме ЛИНЕЙН, что позволит полностью избежать трудностей технческих характера, неизбежных при ручном счете.

4.4.2. Коэффициент множественной линейной корреляция R и его свойства

Наряду с рассмотренной множественной регрессией, характеризующей зависимость $y=x_0$ от (x_1, x_2, x_3 ..., x_k), можно охарактеризовать и силу линейной корреляционной связи

между ними, введя специальный показатель – коэффициент множественной линейной корреляции *R*:

$$R = \sqrt{B} = \left[\frac{\sum_1 (\tilde{y}_i - \bar{y})^2}{\sum_y (y_i - \bar{y})^2} \right]^{0,5}, \qquad (4.54)$$

где все обозначения те же, что и в предыдущих разделах темы.

Определенный таким образом коэффициент множественной линейной корреляции *R* обладает следующими свойствами.

1. *R* – есть мера линейной зависимости $y=x_0$ от совокупности остальных переменных ($x_1, x_2, x_3 ..., x_k$).

2. Значение *R* меняется в пределах $0 \le R \le 1$,

- если $R=0$, то между $y=x_0$ и остальными переменными в совокупности нет линейной корреляционной связи (но может быть нелинейная множественная корреляция),
- если $R=1$, то связь между $y=x_0$ и остальными переменными переходит в чисто функциональную линейную.

3. Для случая одной независимой переменной $R_{x,y}= |r_{x,y}|$, где r_{xy} – обычный парный коэффициент корреляции.

4. *R*, как и *r*, не зависит от линейного преобразования переменных.

5. Стандартная ошибка *R* выражается формулой:

$$s_R = \frac{1 - R^2}{\sqrt{n - k - 1}}, \qquad (4.55)$$

где *k* – числонезависимых переменных в выборке

6. Стандартная ошибка уравнения регрессии выражается через *R* как:

$$s_2 = s_{рег} = s_y \sqrt{1 - R^2}\ . \qquad (4.56)$$

7. Статистическая значимость *R* оценивается одновременно состатистической значимостью множественной линейной регрессии по *F*-критерию Фишера (см. п. 4.4.1).

Другие методы оценки значимости *R* рассматриваются в теме 5.

Коэффициент множественной корреляции *R и все статистики* множественной корреляции и регрессии, могут быть выражены также через обычные парные коэффициенты корреляции *y* с каждой из переменных x_k и различных пар переменных между собой, что является следствием связи множественной корреляции и регрессии. Однако рассмотрение этого вопроса выходит за рамки нашего учебника. Точно также здесь не рассматриваются вопросы частной корреляции между лвумя различными переменными-

предикторами, когда исключается одновременное влияние на них других переменных-предикторов.

4.4.3. Вычисление параметров множественной линейной регрессии в Excel

В Excel с помощью прогрпммы «ЛИНЕЙН» выполняется вычисление значений параметров уравнения множественной регрессии (4.41)

$$\tilde{y}(x_1, x_2, x_3..., x_k) = \tilde{x}_0 = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \pm s_2 ,$$

а также другие необходимые статистики, а имнено:

- угловые коэффициенты регрессии: b_1, b_2 ..., b_k и остаточный член b_0 (первая строка);
- стандартные ошибки угловых коэффициентов b: s_{2b0}, s_{2b1}, s_{2b2}.., s_{2bk} (вторая строка);
- коэффициент детерминации B=R^2;
- стандартная ошибка уравнения регрессии – s_2;
- значение $F_{эмп}$ по формуле (4.53);
- число степеней свободы $CC_2=k_2=(n-k_1-1)$, где n – объем выборки, k_1 – число независимых переменных;
- Σ_1 - регрессионная сумма квадратов по формулам (4.49) п. 4.4.1;
- Σ_2 - остаточная сумма квадратов по формулам (4.49) п. 4.4.1.
- В пустые ячейки будет записано: нет данных - #Н/Д

Результаты расчетов выдаются на печать в форме табл. 4.3, которая представляет собой расширенный вариант табл. 4.2, соответству аналогичным расчетам для парной линейной регрессии.

Таблица 4.3

Форма выдачи расчетов статистик по программе «ЛИНЕЙН» Excel для множественной линейной регрессии

b_k	b_{k-1}	……	b_2	b_1	b_0
s_{2bk}	$s_{2b(k-1)}$		s_{2b2}	s_{2b1},	s_{2b0}
$B=R^2$	s_2		#Н/Д	#Н/Д	#Н/Д
$F_{эмп}$	CC_2		#Н/Д	#Н/Д	#Н/Д
Σ_1	Σ_2		#Н/Д	#Н/Д	#Н/Д

Работа с программой ЛИНЕЙН в случае множественной регрессии полностью аналогична, описанной в п. 4.2.2 для парной регрессии.

Только теперь мы имеем дело с двумя и более независимыми переменными и поэтому под запись результатов надо выделить массив ячеек, содержащий так же 5 строк, но число столбцов должно быть на единицу больше числа независимых переменных.

Таблица 4.4

Значения средниех годовых температур и координат места (высота, широта и долгота) для 33 метеостанций Кыргызстана.

Метеостанция	Высота, км	Широта,°	Долгота,°
1. Фрунзе	0,756	42,85	74,53
2. Новороссийка	1,532	42,73	76,07
3. Чон-Арык	1,110	42,7	74,03
4. Байтык	1,579	42,68	74,5
5. Тюя-Ашу сев/	3,090	42,7	73,82
6. Талас	1,217	42,52	72,22
7. Токтогул	0,821	41,82	72,83
8. Чолпон-Ата	1,645	42,6	76,93
9. Пржевальск	1,716	42,5	78,43
10. Рыбачье	1,660	42,45	76,18
11. Койлю	2,800	42,2	79,02
12. Тамга	1,960	42,18	77,55
13. Тянь-Шань	3,614	41,92	78,23
14. Ак-Шийряк	2,844	41,82	78,75
15. Каракольская	3,080	41,52	77,45
16. Кочкорка	1,810	42,22	75,73
17. Сусамыр	2,061	42,15	74,02
18. Кара-Куджур	2,800	41,93	76,3
19. Ат-Ойнок	3,050	41,65	74,53
20. Нарын	2,040	41,43	75,98
21. Казарман	1,266	41,07	74,03
22. Чаткал	1,937	41,9	71,32
23. Устье р. Тос	1,759	41,58	71,67
24. Ангрен	2,286	41,53	70,75
25. Ак-Терек-Гава	1,748	41,27	72,82
26. Чаар-Таш	2,748	41	73,8

27. Узген	1,012	40,77	73,3
28. Ош	1,016	40,63	72,8
29. Кызыл-Джар	2,230	40,32	74,25
30. Гульча	1,542	40,32	73,45
31. Хайдаркан	1,970	39,95	71,35
32. Сары-Таш	3,155	39,72	73,25
33. Дараут-Коргон	2,220	39,55	72,18

Покажем все это на примере расчета множественной регрессии, которая описывает зависимость средних годовых температур (y) для 33 метеостанций Кыргызстана от высоты (x_1=z, км), широты (x_2= φ°) и долготы места (x_3= λ°). Исходные данные приведены в табл.. 4.4.

Выделяем массив из 10 ячеек для заприси результатов, состоящий из 5 строк и 4 столбцов, т.к. число независимых перменных равно трем. Дальнейшая процедура работы с программой состоит в следующем.

1. Через кнопку *fx* (мастер функций) запустить программу ЛИНЕЙН, для чего выделить в окне перечня функций «*статистические*», а затем программу ЛИНЕЙН.

2. Через кнопку ОК открыть окно *Аргументы функции* и последовательно ввести (для нашего примера):

- Известные значения *y* – это *массив* (весь столбец) температур *T* по табл. 4.4.
- Известные значения *x* – это *массив* (три столбца одновременно), включающий столбцы высот (x_1) широт (x_2), и долгот (x_3) по данным табл. 4.4.
- Константа – задается как *1 (истина)* или как *0 (ложь);* если будет введено 0 (ложь), то программа *задаст* статистику b_0=*0*, если будет введено 1 (истина), то программа будет *вычислять* b_0 обычным способом.
- Статистика – задается как *1 (истина)* или как *0 (ложь);* если будет введено 0 (ложь), то программа выдаст только станистики - *b,* если будет введено 1 (истина), то программа выдаст весь набор статистик по схеме табл. 4.3.

6. Щелкнуть кнопку ОК, затем ввести курсор мыши *в верхнюю строку формул* и щелкнуть ЛКМ.

7. Одновременно нажать три клавиши: Ctrl+Shift+Enter (рекомендуется для отсутсвия сбоя сначала нажать первые две кклавиши Ctrl+Shift, а затем дожать третью - Enter).

8. В выделенном диапазоне ячеек появится массив результатов расчетов статистик по схеме табл. 4.3.

Так, для нашего расчетного примера получим следующий массив статистик множественной регрессии

-0.0235854	-0.473543	-5.206232	36.484690
0.25046312	0.59013988	0.7022663	21.801962
0.70920193	2.61641195	#Н/Д	#Н/Д
23.5751859	29	#Н/Д	#Н/Д
484.159691	198.522734	#Н/Д	#Н/Д

Следует заметить, что в исходной табл. 4.4 независимые переменные занимали столбцы в порядке «слева на право», т.е. сначала шел столбец высот, затем правее него - широт и еще правее - долгот. В полученной таблице статистик они располагаются наоборот – «справа налево». Так, для первой строки, где приведены угловые коэффициенты, имеем: в ячейке первого столбца *слева* записано значение b_0, во второй ячейке левее – знаяение b_1, затем еще левее идут b_2 и b_3. Разумеется, во второй строке в такой же последовательности располагаются и средние квадратичесские ошибки этих параметров. Остальные параметры располагаются так же, как и для парной регрессии.

Заметим, что, как и для парной регрессии, в расчетном массиве статистик нет коффициента корреляции *R*. Однако его легко получить извлеченем квадратного корня из значения *B*, помня, что *R* принимает только положительные значения от нуля до 1. Так, в нашем приере получим, что $R = 0{,}84$.

Полученное уравнение множественной регрессии, отражающее зависимость средних годовых температур на территории Кыргызстана от высоты (z. км), широты (φ,°) и долготы (λ,°) места имеет вид:

$$y=36.48-5.206z-0{,}474\varphi-0{,}0236\lambda\pm2{,}62$$

Средняя квадратическая ошибка уравнения регрессии равна ±2,62°С. Уравнение статистически значимо на уровне доверительной вероятности p=0,95 (риск ошибки 5%), так как *F(эмп)=23,56*, а рассчитанное по программе «F.ОБР.ПХ» значение *F(крит)=2,93* при $q=0{,}05$, $CC_1=3$ и $CC_2=29$.

Теперь осталось критически рассмотреть полученное уравнение множественной регрессиии с тем, чтобы установить: все-ли три переменные являются статистически значимыми и нельзя-ли его упростить, исключив незначимые переменные, если такие окажутся. Эти имеет большое значение для практике, так как некоторые переменные могут дублировать друг-друга, являясь не эффективными и не нужными. Рассмотрим это подробнее в следующем п. 4.4.4.

4.4.4. Правила формирования и использования множественной линейной регрессии и корреляции

В отличие от парной линейной регрессии и корреляции при использовании множественной модели, возникает ряд сложностей, преодоление которых может быть достигнуто теми или иными методами. Ниже мы кратко коснемся наиболее важных из них, что позволит при необходимости познакомиться с ними глубже по первоисточникам.

1. *Нормирование переменных*. В уравнении регрессии (4.41) п. 4.5.1 значения угловых коэффициентов b_k будут зависеть не только от влияния фактора x_k на у, но и от единиц измерения x_k. Поэтому по численным значениям b_k трудно судить о вкладе фактора x_k в регрессию. Этого можно избежать, если нормировать переменные, перейдя к СВ u_k:

$$u_k = \frac{x_k - \bar{x}_k}{\sigma_k}. \qquad (4.57)$$

Тогда, формула (4.41) п. 4.5.1 запишется в следующем нормированном виде [1]:

$$\tilde{u}_0 = \alpha_1 u_1(x_1) + \alpha_2 u_2(x_2) + \ldots + \alpha_k u_k(x_k). \qquad (4.58)$$

Таким преобразованием единицы измерения разных факторов приводятся к одному безразмерному масштабу. Это позволяет напрямую сравнивать численное значение a_k в (4.58) и отбрасывать те факторы, значения a_k для которых принебрежимо малы.

В качестве примера приведем соотношения обычных и нормированных угловых коэффициентов для множественной регрессии, построенной по трем факторам для прогноза урожайности зерновых культур в Западном Казахстане [29]. В табличке приведены три основных фактора, влияющих на урожайность и соответствующие им значения обычных угловых коэффициентов b_k, полученных для уравнения вида (4.41) и нормированных α_k, соответствующих уравнению (4.58):

Фактор	*Обычный* b_k	*Нормированый* α_k
1. Сумма осадков за апрель	0,065	0,267
2. Влагообеспеченность июня	1,066	0,366
3. Осадки холодного периода	0,014	0,474
	Свободный член=–2,390	

Если оценивать по исходным значениям b_k вклад этих трех факторов в регрессию (4.41), то явно преобладающим является вклад фактора 2 (влагообеспеченность июня), который в 16 раз больше вклада фактора 1 (сумма осадков за апрель) и в 76 раз больше вклада фактора 3 (осадки холодного периода).

Однако анализ нормированных значений α_k говорит совсеем другое: 1) прежде всего, все три фактора являются значимыми, т.к. их отношения составляют - 0,56, 0,77 и 1,00, 2) казалось бы, первоначально совершенно незначащий фактор 3 (осадки холодного периода) с b_3=0,014 на самом деле оказался наиболее весомым, и в нормированном уравнении имеет самое высокое значение α_3=0,474.

Этот пример наглядно показывает большую полезность перехода к нормированным переменным при сильно разномасштабных по размаху факторах, что позволяет оставлять в уравнении действительно важные факторы и исключать малозначащие.

2. *Исключение дублирующих и не эффективных аргументов*. Если в аргументах x_1, x_2 ..., x_k некоторые факторы сильно коррелируют между собой, то формально это означает, что они дублируют друг друга и поэтому достаточно оставить один из них, либо имеющий более высокую корреляцию с y, либо физически более обоснованный. Без такого исключения уравнение (4.41), будет неустойчивым. Это означает, что значения его коэффициентов b_k могут сильно меняться, даже при небольшом изменении объема выборки. Исключение дублирующих коэффициентов b_k незначительно уменьшит объем информации, учитываемой регрессией при прогнозировании y, но зато повысит устойчивость уравнения регрессии относительно значений угловых коэффициентов

При поиске наилучшего набора предикторов, когда их общее число достаточно велико (k≥10) используются различные пошаговые процедуры включения и исключения различных предикторов [2]. В качестве простых *пошаговых правил*, позволяющих решить задачу ограничения переменных в первом приближении, следует рекомендовать следующие:

- на шаге 1 выбирается одна независимая переменная x_1, которая наиболее физически обоснована и имеет наиболее высокий коэффициент корреляции с y; находится парная линейная регрессия $\tilde{y} = b_0 + b_1 x_1 \pm s_2$;
- на шаге 2 вводится вторая, наиболее обоснованная таким же образом переменная x_2, и находится уравнение регрессии для двух переменных; если вновь вводимая переменная существенно увеличивает R и одновременно снижает ошибку уравнения s_2, то она включается в уравнение как значимый фактор; если значене R не повышается, а значение s_2 не снижается, то переменная x_2 исключается из уравнения;
- на последующи шагах 3, 4 и т.д. эта процедура включения и исключения переменных продолжается, пока не будет окончательно сформировано рабочее уравнение множественной регрессии.

Приведем пример, иллюстрирующий применение этих правил на основании анализа полученной выше множественной регрессии, построенной по данным табл. 4.4 и

описывающей зависимость средних годовых температур от высоты, широты и долготы места для Кыргызстана. Напомним, что 33 использованные метеостанции располагаются в значительном диапазоне высот, от 0,6 до 3,61 км, а территориия Кыргызстана простирается с запада на восток всего на 925 км и с севера на юг на 400 км, имея сложную горную орографию. Исходя из климатических закономерностей и этих данных, разумно предполагать, что главным фактором изменния температуры в Кыргызстане является высота места. Влияние широты и долготы места не может быть большим, учитывая малые размеры территории. Это подтверждают коэффициенты корреляции между средними годовыми температурами на 33 станциях и высотой, широтой и долготой места. Они оказались соответственно равными: r(T,z)=-0,84, r(T,φ)=-0,15 и r(T,λ)=-0,35. Однако следует учесть, что за счет неравномернсти расположения станций по территории возможно влияние ложной корреляции на степень коррелированноси температур с широтой и высотой места, т.к. не исключено, что высота z коррелирует с φ и λ.

Так как у нас всего три переменных, то приведем следующий набор из 4 возможных шагов.

Шаг 1. Вводим одну переменную: $x_1=z$, получаем парную регрессию

$T^0C=14,88-5,143z \pm 2,57$,

$R=0.836;\ s_2=2.57°C.$

Шаг 2. Вводим две переменные: $x_1=z$ и $x_2=\varphi$, получаем множественную регрессию

$T^0C=35,00-5,237z\ -0,5026\varphi \pm 2,57$

$R=0.842;\ s_2=2.57°C.$

Шаг 3. Вводим две переменные: $x_1=z$ и $x_3=\lambda$, получаем множественную регрессию

$T^0C=24,21-5,006z\ -0,1287\lambda \pm 2,60$

$R=0.838;\ s_2=2.60°C.$

Шаг 4. Вводим все три переменные: $x_1=z$, $x_2=\varphi$ и $x_3=\lambda$, получаем множественную регрессию

$T^0C=36,49-5,206z\ -0,4735\varphi\ -0,0236\lambda \pm 2,57$

$R=0.842;\ s_2=2.62°C.$

На шаге 1 была получена статистически значимая, на уровне доверительной вероятности p=0,95, парная линейная регрессия с коэффициентом корреляции r=-0,836 и стандартной ошибкой $s_2=2.57$°C. Введение широты места на шаге 2 повысило модуль коэффициента корреляции всего на 0,06, что следует считать незначимым увеличением. При этом ошибка регрессии осталась без изменений. Это говорит о том, что в данном случае влияние широты места является статистически несущественным фактором и эту переменную следует исключить из уравнения регресии. Дльнейшие шаги с использованием двух пере-

менных z *и* λ, а также всех трех переменных - z, φ *и* λ - привели даже к тенденции увеличения s_2 при несущественных изменениях R. Все это говорит о том, что, если исходить из выборки табл. 4.4, то, для практического использоваия, следует рекомендовать для Кыргызстана наиболее простую парную регрессию, которая была получена на шаге 1

$$T^0\text{C}=14{,}88-5{,}143z \pm 2{,}57.$$

Анализ регрессионных остатков. Построение уравнения регрессии путем подбора различных моделей есть этапы моделирования статистического описания метеорологического явления, представленного выборкой исходных данных. Качество модели для оценки возможностей ее практической пригодности должно быть проверено, исходя из различных подходов. Одной из таких процедурных проверок модели является анализ регрессионных остатков d_i:

$$d_i = y_i - \tilde{y}_i, \qquad (4.59)$$

где y_i – наблюденное, а $\tilde{y}_i(x_i)$ – рассчитанное по регрессии значение предиктанта.

Эти остатки были использованы выше для определения остаточной дисперсии, т.е. стандартной ошибки модели. Но они содержат в себе и другую информацию. Остатки – это то, что нельзя объяснить уравнением регрессии, их можно классифицировать как метеорологический (климатический) шум, помехи или погрешности.

При проведении регрессионного анализа считают, что эти погрешности независимы, имеют нулевые средние ($\Sigma d_i=0$), одинаковую (постоянную) дисперсию, не зависящую от значений, принимаемых предикторами (x_{1i}, x_{2i}, x_{3i} ..., x_{ki}), и подчиняются нормальному закону распределений. Только при соблюдении (подтверждении) всех этих условий можно говорить, что модель построена статистически правильно, и с этой точки зрения, безупречна для использования.

Нормальность остатков d_i можно проверить любым из изложенных ранее методов:

- оценкой эмпирических значений асимметрии и эксцесса остатков;
- аппроксимации их распределения нормальной кривой с последующей оценкой по критерию согласия χ^2-Пирсона.

Кроме того, можно построить графики остатков d_i следующих типов [2, 20, 29]:

1) в зависимости от x_i, если x – время или известная последовательность наблюдений;

2) в зависимости от предсказываемых значений $\tilde{y}_i$;

3) в зависимости от каждого фактора x_i $(i=\overline{1,k})$;

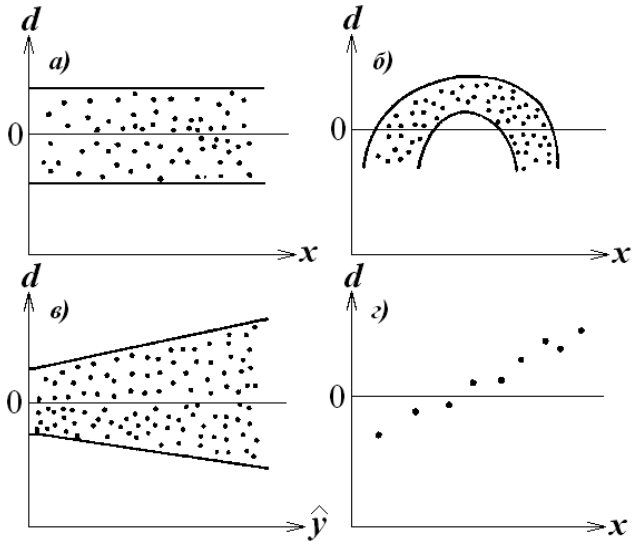


Рис.4.9. Виды графиков регрессионных остатков модели:

а) при постоянной дисперсии s_2^2; *б*) без учета нелинейности модели;

в) при увеличении дисперсии s_2^2; *г)* без учета линейного тренда *y* от *x*.

Такие графики позволяют выявить адекватность (неадекватность) модели выборке, наличие не включенного в регрессию линейного или иного тренда, непостоянство дисперсии ошибок регрессии (т.е. s_2^2), наличие выбросов, необходимость включения в уравнение регрессии дополнительных факторов и другое. На рис. 4.9 показаны основные виды графиков регрессионных остатков и указано на те задачи, которые они позволяют решить. При этом надо учитывать, что если зависимость *y* от *x* в выборке соответствует определенному *нелинейному* типу регрессии, а использована *линейная* модель, то полученный график остатков $dy_i(x)$ будет иметь форму, близкую к типу фактической зависимости *y(x)*, т.е. в этом случае формы зависимости *y(x)*, в выборке и формы вида остатков *совпадают.*

1. Адекватность регрессионной модели и выборки (рис. 4.9а). Если остатки располагаются вдоль оси *x* в форме полосы примерно постоянной ширины, которая не зависит от *x*, то это говорит о том, что:

- выбранная регрессия как модель адекватна зависимости *y(x)* в выборке;
- остаточная дисперсия s_2^2 не зависит от *x*, т.е. постоянна;
- если отдельные точки d_i далеко выходят за пределы этой равномерной полосы, то надо обратить внимание на эти члены выборки и установить причину такого факта.

2. Необходимость включения в уравнение регрессии нелинейного члена (рис.4.9б). Если остатки располагаются вдоль оси *x* в форме нелинейной полосы определенного вида, то это говорит о том, что в уравнение регрессии надо включить дополничельный нелинейый фактор (переменную) такого же вида.

3. Непостоянство остаточной дисперсии (рис. 4.9в). В случае увеличения остаточной дисперсии при увеличении *x* или, когда график остатков будет иметь вид раструба, необходимо: 1) либо преобразовать исходные переменные *x* или *y*, или *x* и y, так, чтобы «раструб остатков» исчез; 2) либо ввести в модель дополнительные факторы для этой же цели; имеющуюся регрессионную модель до такого преобразования и исключения «раструба остатков» следует считать неадекватной.

4. Необходимость введения в модель дополнительного линейного предиктора (рис. 4.9*г*). Если графики остатков обнаруживают линейный тренд, то в уравнение регрессии надо ввести дополнительный линейный член, который учтет линейный тренд, который имется в выборке.

Например, график зависимости от высоты регрессионных остатков для линейной регрессии, полученной в примере выше на шаге 1, показан на рис. 4. 10. Видно, что остатки имееют вид линейной полосы, вытянутой вдоль оси высот. Это говорит о том, что линейная высотная регрессия является адекватной выборке, и добавление других факторов (широта, долгота) не требуется, что было получено выше другим путем.

Одновременно три точки на графике дают значительные выбросы регрессионных остатков. Это котловинная станция Сусамыр (2,061 км) и станции Ат-Айнок (верхняя часть склоновой долины, 3,05 км) и Каракольская (верхняя часть межгорной долины р. Нарын, 3,08 км). Аномально большие выбросы на них объясняются этими орографическими условиями расположеня станций, которые не учитываются чисто высотной регрессией. К сожалению, включить орографию как количественный фактор в уравнение регрессии не представляется возможным.

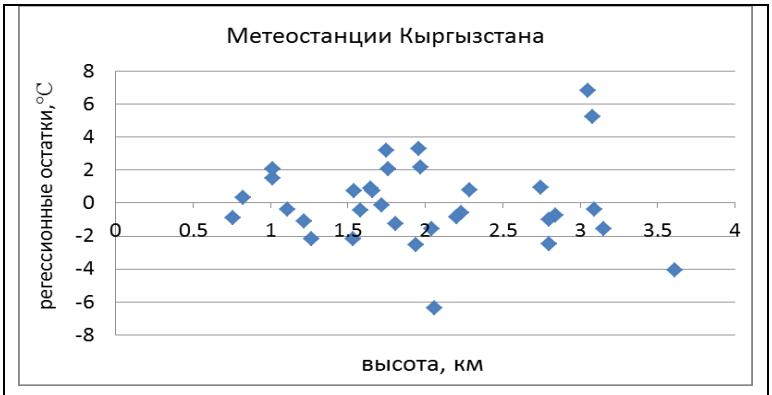


Рис. 4.10. Зависимость регрессионных остатков от высоты места для высотной линейной регрессии средних годовых тмператур для территории Кыргызстана

Анализ графиков остатков является простым и наглядным методом, позволяющим судить об адекватности выбранной модели зависимости y от предиктора (предикторов) в выборке. Часто достаточно ограничиться анализом таких графиков без использования строгих статистических критериев.

Тема 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Важным разделом математической статистики является раздел, где рассматриваются статистические критерии, методология и техника проверки статистических гипотез. В предыдущих главах мы уже коснулись, правда фрагментарно, изложения ряда вопросов этой темы: критерия согласия эмпирических и теоретических распределений χ^2-Пирсона и F- критерий Фишера для оценки значимости регрессии и корреляции. Рассмотрим теперь эти и другие вопросы более подробно на основе накопленного статистического материала и с привлечением соответствующих положений теории. Как и прежде, основной задачей изложения является последовательное обучение, привитие статистической культуры, демонстрация всех предлагаемых решений на конкретных задачах. Сложные теоретические выводы опущены и заменены изложением принципиальных основ получаемых решений так, чтобы научить квалифицированно использовать предлагаемые методы.

Глава 5.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ И ПОНЯТИЯ МЕТОДОВ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

5.1.1. Нулевая и альтернативные гипотезы, задачи, решаемые с помощью проверки нулевых гипотез, последовательность анализа

Статистической гипотезой называется *любое предположение* относительно статистических свойств выборки. Особое место среди статистических гипотез занимают *нулевые гипотезы* (обозначим их через H_0) и противоположные им альтернативные гипотезы (обозначим их через $\overline{H}$).

Нулевая гипотеза H_0, которая *проверяется*, утверждает:

- *отсутствие* проверяемых свойств *в выборке*, т.е. равенство их нулю (например, линейная корреляция в выборке отсутствует и r=0);
- *равенство* проверяемых свойств *в различных выборках*, т.е. равносильность (равенство) выборок по проверяемому свойству например, средние значения двух выборок x_1*(ср. знач.)=* x_2*(ср. знач.).*

Таким образом, нулевая гипотеза утверждает, что либо в изучаемой одной выборке данное свойство *отсутствует* (равно нулю), либо в изучаемых нескольких выборках это свойство *одинаково*, а наблюдаемые количественные отличия выборок по этим свойствам чисто случайны, т.е. статистически незначимы (не существенны).

Например, рассматривается корреляция зимних температур воздуха с высотой места на горных станциях. Получен эмпирический коэффициент корреляции r=-0,21, т.е. связь (T, z) слабая. Основная причина ясна – зимние инверсии температуры, которые нарушают закономерное понижение T с z. Встает вопрос: можно ли считать связь с r=-0,21 статистически значимой или ею можно пренебречь. Для этого надо *проверить* нулевую гипотезу об отсутствии корреляционной связи, т.е., что r=0. Тогда, H_0: r=0. Именно так она записывается в краткой форме, которой мы будем придерживаться далее.

Или другой пример. Средняя годовая температура на МС Фрунзе (756 м), расположенной на подгорной равнине, равна $T_Ф$=10,1^{0}С, а на МС Чон-Арык (1110 м), расположенной в зоне подножий – $T_{ЧА}$=8,8^{0}С. Вопрос: действительно ли средние температуры на этих станциях различны из-за разницы высот или наблюдаемое различие объясняется чисто случайными факторами, например, ограниченностью рядов наблюдений на станциях? Нулевая гипотеза, соответствующая этому вопросу, записывается так, H_0: $T_Ф=T_{ЧА}$ (или H_0: $|T_Ф-T_{ЧА}|=0$).

Альтернативной нулевой гипотезе (или альтернативной гипотезой) называется любая гипотеза, противоположная нулевой. Каждой нулевой гипотезе H_0, может быть противопоставлено три альтернативных: $\overline{H}_1$, $\overline{H}_2$ и $\overline{H}_3$. Так, для второго примера имеем:

- нулевая гипотеза H_0: $T_Ф=T_{ЧА}$,
- альтернативные $\overline{H}_1$: $T_Ф \neq T_{ЧА}$; $\overline{H}_2$: $T_Ф > T_{ЧА}$; $\overline{H}_3$: $T_Ф < T_{ЧА}$.

Заметим, что по выборке $T_Ф$=10,1>$T_{ЧА}$=8,8, но мы вправе рассмотривать возможность противоположного неравенства вследствие исходного предположения по H_0 о чисто случайном различие температур.

При проверке нулевых гипотез можно противопоставить H_0 любую из альтернативных $\overline{H}$. При этом преследуется цель:

- *либо отвергнуть* H_0 на определенном уровне значимости q, когда численное значение критерия попало в область его критических значений (т.е. забраковать ее и принять $\overline{H}$ (например, $\overline{H}_3$: $T_Ф > T_{ЧА}$ в соответствии с высотами расположения станций),
- *либо принять* H_0 на определенном уровне значимости q, (принятие H_0 менее категорично и только соответствует утверждению, что численное значение критерия не противоречит H_0)

Надежность, т.е. правильность решения во всех случаях будет зависеть от того, насколько сделано все возможное, чтобы забраковать H_0, т.к. именно квалифицированное отклонение нулевой гипотезы есть одновременно путь к правильному принятию альтернативной $\overline{H}$.

С помощью проверки нулевых гипотез можно решить следующие задачи.

1. Равенства выборочных *средних* значений разных k выборок, т.е. равенство выборок по математическому ожиданию

$$\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = ... = \bar{x}_k . \quad (5.1)$$

2. Равенства выборочных *дисперсий* s^2 различных k выборок, т.е. равенство выборок относительно их дисперсий

$$s_1^2 = s_2^2 = s_3^2 = ... = s_k^2 . \quad (5.2)$$

3. Равенства выборочных законов распределений различных k выборок

$$F_1(x_1)=F_2(x_2)=F_3(x_3)= ..., =F_k(x_k), \quad (5.3)$$

т.е. равенство выборок в целом по всем статистическим свойствам

4. Отсутствия в выборке парной и множеественной корреляции и регрессии

$$r=0,\ b_1; \quad R=0,\ b_1=b_2=b_3= ...=b_k=0. \quad (5.4)$$

5. Соответствия выборочного закона распределения $F_{эм}$ определенному теоретическому закону

$$F_{эм.}(x)=F_{теор.}(x). \quad (5.5)$$

6. Оценки случайности выборки.

Первые три задачи называются задачами оценки *однородности выборок*.

Чтобы проверять H_0 необходимо использовать тот или иной статистический критерий или тест, эмпирические значения которых вычисляются *по определенным формулам*. При этом рассчитанное эмпирическое значение критерия $y_{эм}$ *само является случайной величиной.* Процедура проверки H_0 заключается в следующем:

- формулируется H_0 и конкретная альтернатива $\overline{H}_i$, выбирается наиболее подходящий критерий и задается уровень значимости q (тем самым и уровень доверительной вероятности $p=1- q$);

- по формуле критерия рассчитывается его эмпирическое значение $y_{эм}$, а по его теоретическому распределению находится критическое значение – $y_{кр}$; это делается с помощью компьютерных программ или по специальным таблицам;
- сравниваются $y_{эм}$ и $y_{кр}$: 1) если $y_{эм}$ попадает в область критических значений, то H_0 отвергается (наблюдаемые различия не случайны); 2) если $y_{эм}$ попадает в область доверительных значений критерия, то H_0 принимается (наблюдаемые различия случайны).

Подчеркнем еще раз, что все выводы из процедуры проверки H_0 носят вероятностный, а не полностью категоричный характер. На самом деле H_0 категорически никогда не отвергается и, тем более, не принимается, а делается только вероятностный вывод о ее соответствии или несоответствии конкретному эмпирическому материалу на заданном уровне доверительной вероятности. Увеличивая объем выборок мы, видимо, всегда на каком-то этапе сможем забраковать H_0.

5.1.2. Уровень значимости критерия, критические области и области доверительных значений критерия

Уровнем значимости q для любого критерия *y* называется такое малое значение вероятности его появления, при котором *практически* можно считать это событие *невозможным*. Как уже отмечалось (см. вводную главу), статистическая теория *в принципе* не может ответить на вопрос какое событие считать невозможным, это дело практики и предметных наук. Обычно за уровень значимости q берут одну из вероятностей 0,1, 0,05 или 0,01, в климатологии чще всего принимается q=0,05.

Как известно, с уровнем значимости q связан уровень доверительной вероятности p (который дополняет его до 1):

$$p=1-q. \qquad (5.6)$$

Простой смысл задания p и q при проверкие нулевых гипотез состоит в следующем: задав q мы тем самым задаем, что принятие H_0 будет происходить не со 100% гарантией, а с риском совершить ошибку в q% случаев (ошибка 1-рода, см. ниже). Таким образом, в среднем мы будем принимать правильные решения в p% случаев, в q% - ошибочно отвергать H_0. Это не недостаток статистических методов, а следствие случайной природы самих изучаемых явлений.

Задавая уровни p и q, можно построить соответствующие им четыре различных области доверительных значений критерия и области его критических (практически невозможных) значений. Если рассчитанное по выборке эмпирическое значение критерия $y_{эм}$ попадает в область допустимых значений y, то говорят, что нулевая гипотеза *не противо-*

речит опытным данным и, в этом смысле, может быть *принята* на уровне доверительной вероятности *p*. Если $y_{эм}$ попадает в область критических значений *y*, то H_0 *отвергается* на уровне значимости *q*.

Вид четырех областей доверительных и критических значений критерия *y* показан на рис. 5.1, где по оси абсцисс отложены значения критерия *y*, а по оси ординат функция плотности распределения критерия *f(y)*. Это: 1) односторонняя область больших отрицательных значений критерия; 2) односторонняя область его больших положительных значений; 3) двухсторонняя область больших по абсолютной величине значений критерия; 4) двухсторонняя область малых по абсолютной величине значений критерия.

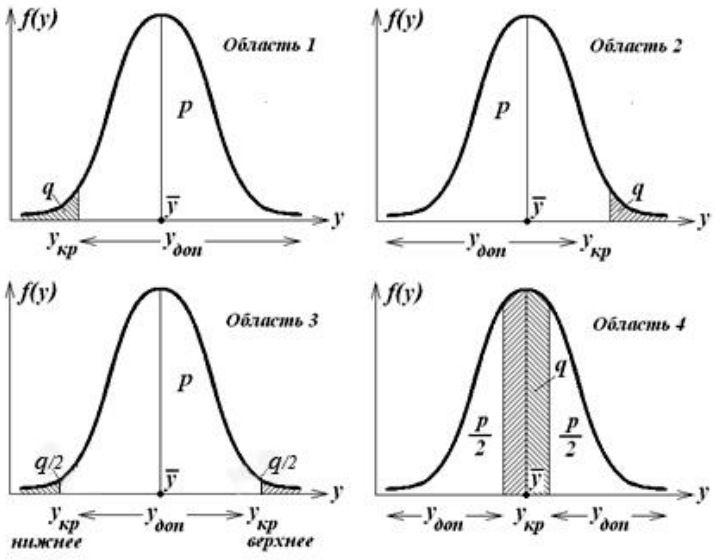


Рис. 5.1.Четыре типа области допустимых значений критерия ($y_{доп}$) и его критических значений ($y_{кр}$).

1. *Односторонняя критическая область больших отрицательных значений критерия*. Она характеризуется *отрицательными относительно* $\bar{y}$ и большими по абсолютной величине значениям $y_{кр}$, которым соответствует вероятность *q* (заштрихованная область). Область доверительной вероятности *p* – это не заштрихованная площадь под кривой *f(y)* (аналогично для областей 2, 3 и 4). Если рассчитанное эмпирическое значение $y_{эм}$ попадет левее $y_{кр}$, то оно окажется в критической области недопустимых значений, если правее $y_{кр}$, то оно окажется в области допустимых значений. Таким образом, области критических значений *y* соответствует вероятность *q*, а области допустимых значений *y* – вероятность *p=1- q*.

Для области 1 нулевая и альтернативная гипотезы записываются так:

$$H_0: y_1=y_2 \text{ (или } |y_1-y_2|=0);\ \bar{H}_1: y_1<y_2, \qquad (5.7)$$

где y_1 и y_2 соответствуют выборкам 1 и 2 (что обозначать за выборку 1, а что за 2 – безразлично).

2. *Односторонняя критическая область больших положительных значений критерия*. Она характеризуется положительными относительно $\bar{y}$ и большими по величине значениями $y_{кр}$. В остальном ее устройство аналогично области 1.

Для области 2 нулевая и альтернативные гипотезы имеют запись:

$$H_0: y_1=y_2 \text{ (или } |y_1-y_2|=0);\ \bar{H}_2: y_1>y_2. \qquad (5.8)$$

3. *Двухсторонняя критическая область больших по абсолютной величине значений критерия*. Ее отличительной особенностью является то, что область критических значений разбивается на две подобласти, лежащих левее и правее $\bar{y}$, а область допустимых значений симметрична относительно $\bar{y}$. Теперь левой подобласти критических значений соответствует $q/2$, а правой – также $q/2$, так, что в целом критическая вероятность равна q. (Левые и правые критические подобласти не обязательно брать одинаковыми по критической вероятности $q/2$, но на практике используются именно такие симметричные подобласти).

Для области 3 нулевая и альтернативная гипотезы записываются в виде:

$$H_0: y_1=y_2 \text{ (или } |y_1-\bar{y}|=0;\ \bar{H}_3: y_1\neq y_2. \qquad (5.9)$$

При практическом использовании трех рассмотренных критических областей надо помнить, что хотя нулевые гипотезы для них формулируются одинаково, альтернативные строятся так, чтобы была *максимальная возможность отклонить* H_0. Например, когда в п. 5.1.1 говорилось о формулировке H_0 и $\bar{H}_i$ для сравнения средних годовых температур на МС Фрунзе (z=756 м, T=10,1^{0}С) и МС Чон-Арык (z=1110 м, T=8,8^{0}С), то H_0: $T_{Фр}=T_{ЧА}$ обязательно должна быть противопоставлена альтернатива $\bar{H}_2: T_{Фр}>T_{ЧА}$, а не альтернатива $\bar{H}_3: T_{Фр}\neq T_{ЧА}$. Только в этом случае достигается условие максимальной возможности забраковать H_0 (и в результате сделать правильный вывод), т.к. исходя из высот станций ясно, что, скорее всего, $T_{Фр}>T_{ЧА}$.

Другой, чисто технически важный вопрос при использовании области 3 – это правильное определение по программам Excel (или статистическим таблицам) критических значений $y_{кр}$ для левой α и правой β подобластей, равных $q/2$. Для левой подобласти надо находить квантиль $y_{кр}(\alpha)=y_{(q/2)}$, а для правой – квантиль $y_{кр}(\beta)=y_{(1-q/2)}$.

Критическая область 4 нами использоваться не будет, поэтому и не будем ее подробно рассматривать.

5.1.3. Ошибки первого и второго рода. Типы статистических критериев

С критическими областями связаны возникающие при использовании любых критериев ошибки 1 и 2 рода.

Ошибка первого рода состоит в том, что на самом деле проверяемая нулевая гипотеза H_0 *верна, но она ошибочно отклоняется* (т.е. принимается неверная альтернативная гипотеза $\overline{H}_i$). Это может произойти вследствие того, что критерий *случайно попал* в критическую область. Хотя вероятность попадания в нее критерия и мала (она равна q), но все же это возможно, т.к. обычно q=0,01 ..., 0,05. Не имея более тонкого инструмента судить о случайности (не случайности) такого значения критерия, мы поступаем формально и бракуем H_0. Таким образом, шанс или вероятность совершить ошибку первого рода всегда в точности равна *уровню значимости* q.

Казалось бы, если мы расширим область доверительных значений, т.е. увеличим p и тем самым снизим q, то уменьшим шанс совершить ошибку первого рода. И это действительно так: чем меньше q, тем меньше риск совершить ошибку первого рода. Но здесь подстерегает другая опасность – увеличивается вероятность совершения ошибки второго рода.

Ошибка второго рода состоит в том, что проверяемая нулевая *гипотеза* H_0 *не верна, но она ошибочно принимается* (т.е. на самом деле верна альтернативная $\overline{H}_i$, которая отклоняется). Вероятность совершить ошибку второго рода зависит от двух причин: 1) от уровня значимости q - она тем больше, чем меньше q; 2) от мощности используемого критерия (т.е. его качества, характеризующегося шансом не допустить ошибку второго рода, лучше «разделить» нулевую и альтернативные гипотезы). Если через g обозначить вероятность совершить ошибку второго рода, то мощность критерия будет равна вероятности $u=(1-g)$. Мощностью критерия, следовательно, есть вероятность отклонения нулевой гипотезы, когда верна альтернативная. Мощность критерия увеличивается с увеличением объема выборки n и поэтому чем больше выборка, тем надежнее получаемое решение.

Полностью избежать ошибок 1 и 2 рода не удается. Уменьшая одну из них, мы увеличиваем вторую. На практике выбирается некоторый баланс в вероятностях совершения этих ошибок. С одной стороны уровень значимости обычно принимается оптимальным, q=0,01 ..., 0,05, т.е. относительно не очень малым и не очень большим. С другой стороны,

чтобы снизить вероятность совершения ошибки второго рода, используются наиболее совершенные критерии, характеризующиеся более *высокой мощностью*.

Типы статистических критериев и области их применения. Все критерии по отношению к законам распределения *исходных выборок*, делятся на два типа:

- *параметрические,* когда проверяемые выборки соответствуют нормальной совокупности (которая определяется средним и дисперсией, поэтому критериально достаточно сравнить только эти два параметра); это, прежде всего, наиболее совершенные параметрические критерии *t*-Стьюдента и *F*-Фишера;
- *непараметрические*, справедливые для *любых видов исходных* распределений (в том числе и нормального закона).

Непараметрические критерии имеют более широкую область применения, т.к. применимы к любым выборкам, но зато уступают параметрическим по своей мощности. Если об исходных законах СВ ничего не известно, то надо использовать непараметрические тесты. Хотя решения будут получаться менее надежными, чем в случае нормального распределения исходной СВ, но зато они будут статистически полностью обоснованны и поэтому дают вполне строгие решения. В то же время, наиболее совершенные непараметрические критерии лишь не намного уступают самым совершенным параметрическим тестам, таким как *t*-критерий и *F*-критерий. Относительная мощность лучших непараметрических критериев достигает 95% от мощности параметрических *t* и *F* тестов. Поэтому качество получаемых решений при их использовании снижается несущественно.

Кроме того, по возможному способу построения критических областей и параметрические и непараметрические критерии делятся на *односторонние (критические области 1 и 2)* и *двухсторонние (критическая область 3)*.

Далее будут рассмотрены наиболее употребительные критерии обоих типов применительно к решению задач 1-5, перечисленных в п. 5.1.1.

Глава 5.2. ПРОВЕРКА ГИПОТЕЗ ОДНОРОДНОСТИ ВЫБОРОК С ПОМОЩЬЮ ПАРАМЕТРИЧЕСКИХ И НЕПАРАМЕТРИЧЕСКИХ КРИТЕРИЕВ

Метеорологический ряд наблюдений называется однородным, если он получен в одинаковых погодно-климатических условиях и с одинаковой точностью измерений. Следовательно, к неоднородности ряда могут привести изменение климатических условий

территории за счет общего изменения климата или локального влияния человеческой деятельности (например, постепенная застройка территории), а также изменение технологии измерений (смена приборов и методик наблюдений).

Можно проверять гипотезы однородности с разных подходов:

- относительно средних значений двух или более метеорологических рядов;
- относительно средних значений различных частей одного и того же длинного ряда;
- относительно дисперсий двух или более метеорологических рядов;
- относительно дисперсий различных частей одного и того же ряда;
- относительно законов распределений двух или более метеорологических рядов, т.е. равенства выборок в целом.

При этом применяемые критерии в зависимости от распределения исходных величин могут быть параметрическими (исходные СВ распределены нормально или приближенно нормально) или непараметрическими, когда исходные СВ распределены не нормально или закон их распределения неизвестен.

В настоящей главе сначала рассматривается решение всех этих задач с использованием двух самых совершенных параметрических критериев, а затем с помощью одного их лучших непарамеррических критерия. При этом сначала рассмотрим критерии для сравнения выборочных дисперсий, а затем критерии для сравнения выборочных средних. Такой порядок необходим вследствие того, что при сравнении средних по *t*-критерию необходимо знать равенство дисперсий выборок, т.е. первоначально требуется сравнить выборочные дисперсии.

5.2.1. Проверка гипотезы равенства дисперсий с помощью параметрических критериев F-Фишера и Бартлета

Парамеметрический критерий *F*-Фирера используется для проверки равенства двух выборочных дисперсий, а параметрический критерий Бартлета позволяет сравнивать дисперсии двух и более выборок.

Проверка гипотезы равенства двух дисперсий с помощью F-критерия Фмшера. Пусть имеются две выборки СВ x_1 и x_2, имеющие объемы n_1 и n_2, которые соответствуют *нормальной* генеральной совокупности. Оценки дисперсий этих выборок равны s_1^2 и s_2^2. Ставится задача: сравнить две дисперсии, т.е. проверить нулевую гипотезу

$$H_0: \; s_1^2 = s_2^2 . \qquad (5.10)$$

Эта гипотеза проверяется с помощью *F*-распределения Фишера, для чего ее надо сформулировать в виде:

$$H_0: \frac{s_1^2}{s_2^2} = 1. \qquad (5.11)$$

Эмпирическое значение критерия *F(эм)*, равно отношению

$$F(эм) = \frac{s_1^2}{s_2^2}, \qquad (5.12)$$

И распределено как *F*-распределение Фишера с (n_1-1) для числителя и (n_2-1) для знаменателя степенями свободы.

Здесь для определенности в числителе всегда стоит большая из двух дисперсий (это не снижает общности решения, т.к. выборки всегда можно поменять местами). Тогда используется односторонний критерий с альтернативой

$$\overline{H}_1 : s_1^2 > s_2^2. \qquad (5.13)$$

Если на уровне доверительной вероятности *p* (уровне значимости *q=1-p*), эмпирическое значение $F_{эм}$ по (5.12) меньше критического $F_{эм} < F_{кр}$, найденного по программе Excel «F,ОБР.ПХ» при заданном *q* и $CC_1=(n_1-1)$ и $CC_2=(n_2-1)$, т.е. как

$$F_{кр} = F_{(n_1-1);(n_2-1);q}, \qquad (5.14)$$

то нулевая гипотеза *H_0 (5.11) принимается* на уровне значимости *q*

В противном случае нулевая гипотеза о равенстве дисперсий отклоняется на уровне значимости *q и принимается альтернативная (5.13).*

Пример. На двух близко расположенных станциях Байтык (1579 м) и Чон-Арык (1110 м) за счет разницы в их высотах наблюдается заметное различие годовых сумм осадков, на Байтыке она равна 593 мм, а в Чон-Арыке 515 мм. Покажем, что для этих станций дисперсии осадков одинаковы. По климатическим данным имеем:

$s_Б^2$ =9216 мм^2, $n_Б$=49 лет; $s_{ЧА}^2$=10609 мм^2, $n_{ЧА}$=47 лет.

Рассчитаем по (5.12) $F_{эм}$=10609/9216=1,15 и, задав уровень значимости $q_{0,05}$ по программе «F,ОБР.ПХ» находим $F_{кр}=F_{46;48;0,05}$=1,63.

Так как

$$F_{эм}=1,15<F_{кр}=1,63,$$

то на уровне значимости *q*=0,05 (с доверительной вероятностью *p*=0,95) нулевая гипотеза о равенстве дисперсий принимается. Климтически это означает, что, хотя на обеих станциях годовые суммы осадков существенно различаются, их межгодовая изменчивость может быть принята одинаковой и, с этой точки зрения, режимы осадков на обеих станциях идентичны.

Проверка гипотезы равенства двух и более дисперсий с помощью критерия Бартлет. Пусть теперь имеется k-выборок разного объема из нормальной генеральной совокупности. Введем обозначения: n_i – объем i-той выборки, имеющей дисперсию s_i^2; $(n_i\text{-}1)$ – число степеней свободы i-той выборки; $(n\text{-}k)$ – общее число степеней свободы; s^2 – среднее взвешенное значение дисперсии; n – объем объединенной выборки; $n=\Sigma n_i$. Тогда эмпирической статистикой критерия Бартлета будет величина:

$$\chi^2_{Б} = \frac{1}{c}\left[2{,}3026\left((n-k)\lg s^2 - \sum_{i=1}^{k}(n_i-1)\lg s_i^2\right)\right], \qquad (5.15)$$

$$c = \frac{\sum_{i=1}^{k}\frac{1}{(n_i-1)} - \frac{1}{n}}{3(k-1)} + 1, \qquad (5.16)$$

$$s^2 = \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{(n-k)}. \qquad (5.17)$$

Распределение статистики $\chi^2_{Б}$ по (5.5) подчиняется χ^2-распреде-лению с $(k\text{-}1)$ степенями свободы. Если $\chi^2_{Б.эм}$ по (5.5) меньше $\chi^2_{кр}$, найденного по χ^2-распределению для уровня значимости q и числа степеней свободы $(k\text{-}1)$, то нулевая гипотеза

$$H_0 : s_1^2 = s_2^2 = \ldots = s_k^2 = s^2, \qquad (5.18)$$

принимается на уровне значимости q против альтернативы

$$\overline{H}_3 : s_i^2 \neq s^2. \qquad (5.19)$$

Пример из[14]. Даны три выборки (k=3) объема n_1=9, n_2=6 и n_3=5 с дисперсиями, приведенными в табличке. Проверим нулевую гипотезу (5.8) о равенстве трех дисперсий на уровне значимости q=0,05.

Выборка	s_i^2	(n_i-1)	$(n_i-1)\ s_i^2$	$\lg s_i^2$	$(n_i-1)\lg s_i^2$
1	8,00	8	64,00	0,9031	7,2248
2	4,67	5	23,35	0,6693	3,3465
3	4,00	4	16,00	0,6021	2,4084
		17	103,35		12,9797

$$s^2 = \frac{103{,}35}{17} = 6{,}079, \qquad \lg s^2 = 0{,}7838,$$

$$\chi^2 = \frac{1}{c}\left[2{,}3026\,(17\cdot 0{,}7838 - 12{,}9797)\right] = \frac{1}{c}0{,}794,$$

$$c = \frac{\left[\frac{1}{8}+\frac{1}{5}+\frac{1}{4}\right]-17}{3(3-1)}+\frac{1}{17}=1,086,$$

$$\chi^2_{эмпир}=\frac{0,794}{1,086}=0,731.$$

По программе «ХИ2.ОБР.ПХ» для степеней свободы *CC*=(3-1)=2 и уровня значимости *q*=0,05 находим, что

$$\chi^2_{кр}=5,99.$$

Так как $\chi^2_{эм}$=0,731< $\chi^2_{кр}$=5,99, то H_0 по (5.8) принимается на уровне значимости *q*=0,05, т.е. сравниваемые дисперсии статистически значимо не различаются между собой. Заметим, что чисто визуально кажется что их различия диспрсий велики, т.к. наибольшая из них равна 8,0, а наименьшая 4,0. Полученное критериальное решение – дисперсии равны – есть следствие двух факторов: малых объемов всех выборок, а так же тем, что дисперсия описывается формулой, в которую переменная входит в квдрате.

Теперь за оценку общей дисперсии s^2 можно теперь принять ее среднее взвешенное значение s^2=6,079, полученное по трем дисперсиям.

Следует заметить, что критерии Фишера и Бартлета достаточно чувствительныны к отклонениям от нормальности выборок, что надо учитывать на практике.

5.2.2. Проверка гипотезы равенства двух средних значений с помощью параметрического t-критерия

Параметрический *t*-критерий Стьюдента используется для сравнения двух средних значения $\bar{x}_1$ и $\bar{x}_2$, полученных по выборокам разного объема n_1 и n_2 при условии: выборки сделаны из нормально распределенных генеральных совокупностей, которые имеют пусть неизвестные, но *равные* дисперсии, т.е. $s_1^2 = s_2^2$.

При этих условиях эмпирической статистикой *t*-критерия является выражение:

$$t(эм)=\frac{|\bar{x}_1-\bar{x}_2|}{[n_1s_1^2+n_2s_2^2]^{0,5}}\cdot\left[\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}\right]^{0,5}. \qquad (5.20)$$

В этой формуле условно принято, что среднее $\bar{x}_1 > \bar{x}_2$, это не снижает общность решения, т.к. безразлично что брать за первую и вторую выборку. Поэтому величина *t* принимает только положительные значения. С этой целью для разности $\bar{x}_1$ - $\bar{x}_2$ в (5.12) использованы прямые модульные скобки, чтобы подчеркнуть, что используется модуль разности..

Статистика критерия по (5.) распределена по *t*-распределению Стьюдента с (n_1+n_2-2) степенями свободы. Поэтому его критическое значение находится по программе Excel «СТЬЮДЕНТ.ОБР.2Х». При этом надо учесть, что программа составлена для двухстороннего критерия. Это значит, что если, например, в программу вводится вероятость q=0,05, то она строит две критические области слева и справа с вероятностями соответственно равными α=0,025 и β=0,025.

Следовательно, если использовать эту прогамму для одностороннеего критерия (именно это мы и будем делать в настоящем пункте), то надо задавать в программе значение вероятности не q, а равное 2q. Однако относить полученное $t_{кр}$(одност) надо к значению *q*. Так, например, для двусторонней альтернативы звдавая q=0,05 и СС=77, получим $t_{кр}$(двухст)=1,991, а для односоронней альтернативы, задавая теперь 2*0,05=0,10 и СС=77, получим $t_{кр}$(одност) =1,665. Напомним, что для односторонней альтернативы $t_{кр}$ всегда меньше, чем для двусторонней. Но полученное для односторонней альтернативы $t_{кр}$=1,665 теперь надо отнести как соответствующее одностороннему уровню значимости q=0,05.

Кроме того, перед тем как использовать *t*-критерий Стьюдента надо методами предыдущего пункта 5.2.1 убедиться в равенстве дисперсий обеих выборок, $s_1^2 = s_2^2$. При их неравенстве критерий применять нельзя.

Покажем процедуру расчета *t*-критерия и проверки нулевой гипотезы на примере сравнения средних годовых темтператур воздуха на метеостанции Токмак для двух частей временного ряда: 1)1932-1975 гг., когда средние по земному шару температуры менялись мало (стабильный мировой климат), 2)1976-2009 гг., когда наблюдался быстрый рост этих температур (потепление мирового климата). Необходимые для проверки нулевой гипотезы значения статистик оказались равными:

1932-1975 гг.		1976-2009 гг.	
$\bar{x}_1$ (°C)	9.88	$\bar{x}_2$ (°C)	11.14
s_1^2 (°C)^2	0.695	s_2^2 (°C)^2	0.424
(n_1-1	43	n_2-1	34

Из этих данных видно, что от первого ко второму периоду произошло потепление в поле средних годовых температур в среднем на 1,26°C. Встает вопрос: можно ли считать это потепление закономерным, как местный климатический отклик, на наблюдаевшееся в это время потепление глобального климата. Сформулируем нулевую гипотезу как

H_0: $\bar{x}_1 = \bar{x}_2$,

против односторонней альтернативы: $\bar{x}_2 > \bar{x}_1$.

В начале проверим нулевую гипотезу о равенстве дисперсий выборок по *F*-критерию (5.12) на уровне *q*=0,05, чтобы можно было обоснованно применять *t*-критерий для сравнения средних значений годовых температур за два эти периода. По (5.12) получим, что F(эм)=1,64, а по прогамме «F.ОБР.ПХ» для q=0,05, $СС_1$=43 и $СС_2$=34 найдем, что F(кр)=1,73. Так как F(эм)=1,64< F(кр)=1,73, что на уровне доверительной вероятности p=0,95 гипотезу о равенсте дисперсий средних годовых температур в эти два периода следует принять верной. Это дает право использовать *t*-критерий для сравнения средних значений темпрарур за периоды - 1932-1975 и 1976-2009 гг.

Вычислим теперь эмпирическое значение *t*-критерия по формуле (5.20). Оно оказалось равным *t(эм)=9,17. По* «СТЬЮДЕНТ.ОБР.2Х»
для q=2*0,05=0,10 и СС=79-2=77, получим что для односторонней альтернативы *t(кр)=1,665.* Так как *t(эм)=9,17>t(кр)=1,665,* то нулевая гипотеза равенства температур в эти два периода на уровне значимости q= 0,05 *отклоняется и принимается* односторонняя альтернатива: среднее значенмие температуры во второй период 1976-2009 гг. (11,14°С) было значимо выше, чем в период 1932-1975 гг. (9,88°С), т.е. наблюдавшееся потепление климата, равное 1,27°С, было не случайным, а статистически значимым на уровне доверительной вероятности p=0,95.

Приведем еще пример по режиму осадков. В п. 4.2.1 было показано,что на достаточно близко расположенных метеостанциях Байтык (1579 м, нижняя часть склона) и Чон-Арык (1110 м, подножье склона) годовые суммы осадков *r* и их СКО, полученные соответственно за 49 и 47 лет, равны:

$$r_Б=593 \text{ мм}; s_Б=96 \text{ мм}; r_{ЧА}=515 \text{ мм}; s_{ЧА}=103 \text{ мм}.$$

Оказалось, что можно принять гипотезу о равенстве дисперсий на этих станциях. Сравним, существенно ли отличаются средние годовые количества осадков на обеих станциях по *t*-критерию, т.е. проверим

$$H_0: r_{ЧА}-r_Б=0 \text{ (или } r_{ЧА}=r_Б),$$

против односторонней альтернативы

$$\bar{H}_2: r_{ЧБ}>r_{ЧА}.$$

Подставив исходные данные в (5.20) получим:

$$t_{эм} = \frac{593-515}{[48\cdot 9216+46\cdot 10609]^{0,5}}\cdot\left[\frac{49\cdot 47\cdot(96-2)}{96}\right]^{0,5} = 3,84 .$$

Примем в качестве доверительной вероятности *p*=0,95 (уровень значимости *q*=0,05). Тогда, по программе «СТЬЮДЕНТ.ОБР.2Х» для 2*q*=0,10 и числа степеней свободы 96-2=94 для односторонней альтернативы имеем *t(кр)*=1,66. В результате, имеем

$t(эм)=3{,}84>t(кр)=1{,}66$

и поэтому H_0 на уровне доверительной вероятности $p=0{,}95$ *следует отвергнуть*, приняв альтернативу: годовая сумма осадков на вышерасположенной станции Байтык существенно больше, чем на МС Чон-Арык.

Следует еще раз подчеркнуть, что рассмотренные критерии *F*-Фишера и *t*-Стьюдента являются самыми совершенными параметрическими тестами.

5.2.3. Проверка гипотезы равенства выборок с помощью непараметрического критерия Крускаля-Уоллиса

До сих пор рассматривались только парные параметрические критерии, пригодные для сравнения выборок из норально распределенных (или близких к ним) генеральных совокупностей. Непараметрический ранговый *X*-критерий Крускаля-Уоллиса позволяет сравнивать две и более выборки для любых видов их распределений, в том числе и нормального. По-существу, критерий представляет собой непараметрическую модель однофакторного дисперсионного анализа. Это один из самых эффективных непараметрических критериев группового сравнения выборок, когда их число две и больше. Эмпирической статистикой критерия служит величина *X*, определяемая формулой:

$$X(эм)=\frac{12}{n(n+1)}\sum_{i=1}^{m}\frac{R_i^2}{n_i}-3(n+1), \qquad (5.21)$$

где R_i – сумма рангов *i*-той выборки в общем ранжированном ряду, состоящем из *m* выборок ($m\geq 2$); n_i – объем *i*-той выборки; *n* – общий объем выборок ($n=\Sigma n_i$).

При этом, при выполнении нулевой гипотезы, средние и максимальные значения *X* определяются выражениями:

$$\overline{X}=m-1, \qquad X_{макс}=\frac{(n^3-\sum n_i^3)}{(n^2+n)}. \qquad (5.22)$$

Для контроля правильности расчетов R_i служит выражение

$$\sum_{i=1}^{m}R_i=\frac{n(n+1)}{2}. \qquad (5.23)$$

При $n>15$ распределение *X*(эм) хорошо аппроксимируется χ^2-распределением с (m-1) степенями свободы. Для $m=3$ и $n\geq 6$, точные критические вероятности *q* для *X* приведены в табл. 4.1.

Свойства и применение критерия:

1. Критерий наиболее чувствителен для сравнения средних значений выборок, но т.к. сравнивается и их взаимное расположение то, следовательно, сравниваются и другие свой-

ства распределений. Поэтому H_0 формулируется в предположении, что все m выборок принадлежат одной генеральной совокупности, т.е. имеют одинаковые законы распределений:

$$H_0: F_1(x_1)=F_2(x_2)= \ldots =F_m(x_m). \quad (5.24)$$

Альтернативная односторонняя гипотеза утверждает, что группа m-выборок не принадлежит одной генеральной совокупности, т.е. их законы распределения не одинаковы

$$\overline{H} : F_i(x_i) \text{ - различны}. \quad (5.25)$$

2. Критерий имеет относительную мощность 95% по отношению F-критерию Фишера, являющегося одним из самых совершенных параметрических критериев. Он начинает работать уже при малых объемах выборок, когда n_i=1-2, а $n\geq5$. Это исключительно важно, когда мы имеем дело с выборками *очень малых объемов* (3-15 случаев), например, при экспедиционных наблюдениях или отдельных замерах. По существу, параметрические критерии в этих случаях не применимы, а другие, непараметрические, менее пригодны, чем X-критерий.

3. Для применения критерия ранжируют совокупную выборку (состоящую из m отдельных выборок) в возрастающем порядке, присваивая каждому члену ранг, соответствующий его номеру в общем ряду. Совпадающим членам разных выборок присваивают средний ранг (совпадение членов одной частной выборки значения не имеет). При работе в Excel ячейки каждой выборки следует выделить *отдельным цветом*. Тогда в общей ранжированной выборке *члены каждой отдельной* выборки будет хорошо отличимы от других. Это значительно облегчит вычисления суммы рангов каждой выборки в общем ранжированном ряду.

4. Подсчитывается *сумма рангов каждой выборки*, которая в (5.13) обозначена как R_i.

5. По формуле (5.13) рассчитывается эмпирическое значение статистики критерия $X(эм)$.

6. В табл. 4.1 для m=3 и $n\leq15$ даны точные значения вероятностей $p_{табл.}$, соответствующие приведенным в нем значениям критерия $X_{табл}$ и комбинациям выборок n_1, n_2, n_3. Задав уровень значимости q, по табл.4.1 интерполяцией по $X_{табл}$ и $p_{табл.}$ можно вычислить $X_{крит.,q.}$ Если $X(эм)<X_{табл.,q}$, то нулевая гипотеза на уровне доверительной вероятности $p=1-q$ – принимается. При $X(эм)>X_{табл.,q}$ – H_0 отвергается.

Таблица 4.1

Точные значения *X*-критерия Крускаля-Уоллиса и соответствующие им вероятности *p* при *m*=3, *n*≤15 и различных комбинациях n_1, n_2 и n_3

n	n_1	n_2	n_3	X_p	*P*	*n*	n_1	n_2	n_3	X_p	*P*
6	2	2	2	4,571	0,067	12	5	4	3	7,445	0,010
	3	2	1	4,286	0,100					5,631	0,050
										4,549	0,099
7	3	2	2	4,714	0,048	9	4	4	1	6,667	0,010
				4,464	0,105					4,967	0,048
										4,067	0,102
7	3	3	1	5,143	0,043	9	5	2	2	6,553	0,008
				4,571	0,100					5,040	0,056
										4,293	0,122
7	4	2	1	4,821	0,057	9	5	3	1	6,400	0,012
				4,018	0,114					4,960	0,048
										4,018	0,095

	n_1	n_2	n_3	X_p	*P*	*n*	n_1	n_2	n_3	X_p	*P*
8	3	3	2	6,250	0,011	10	4	4	2	6,873	0,011
				5,139	0,061					5,236	0,052
										4,445	0,103
8	4	2	2	6,000	0,014	10	4	3	3	6,746	0,010
				5,125	0,052					5,727	0,050
				4,458	0,100					4,700	0,101
8	4	3	1	5,208	0,050	10	5	3	2	6,822	0,010
				4,056	0,093					2,251	0,049
										4,495	0,101
8	5	2	1	5,000	0,048	10	5	4	1	6,955	0,008
				4,200	0,095					4,986	0,044
										3,987	0,098
9	3	3	3	6,489	0,011	11	4	4	3	7,144	0,010
				5,600	0,050					5,576	0,051
				4,622	0,100					4,477	0,102
9	4	3	2	6,444	0,008	12	5	5	2	7,269	0,010

				5,400	0,051					5,246	0,051
				4,444	0,102					4,508	0,100
11	5	3	3	7,079	0,009	13	5	4	4	7,760	0,010
				5,649	0,049					5,618	0,050
				4,533	0,097					4,619	0,100
11	5	4	2	7,118	0,010	13	5	5	3	7,543	0,009
				5,268	0,050					5,626	0,051
				4,518	0,101					4,545	0,100
11	5	5	1	7,309	0,009	14	5	5	4	7,791	0,010
				5,127	0,046					5,643	0,050
				4,036	0,105					4,520	0,101
12	4	4	4	7,654	0,008	15	5	5	5	7,980	0,010
				5,692	0,049					5,780	0,049
				4,500	0,104					4,560	0,100

7. При $n>15$ $X(кр)$ находится по таблицам χ^2-распределения по заданному уровню значимости q и (m-1) степеням свободы. Точно также, если $X(эм)<X(кр)$, то значения критерия попадает в область допустимых значений и нулевая гипотеза однородности выборок принимается на уровне доверительной вероятности p.

8. При наличии совпадений в общем ранжированном ряду n, которые относятся к разным выборкам, вычисленную по (5.13) статистику X надо разделить на величину c

$$c = 1 - \frac{1}{n^3 - n}\sum_{i=1}^{x}(t^3 - t), \qquad (5.26)$$

где x – число групп совпадений, относящихся к разным выборкам; t – число совпадений в каждой группе (число совпадений, не превышающее 25% объема выборки, существенно не сказывается на работу критерия).

В этом случае берется исправленное эмпирическое значение критерия (5.13):

$$X_{испр.эмпир.} = \frac{1}{c}X, \qquad (5.27)$$

и для него оценивается выполнение неравенства

$$X_{испр.эмпир.} < X_{крит.}, \qquad (5.28)$$

на основании которого решается вопрос о принятии H_0.

Пример.

Для определения гололедных нагрузок на проектируемую ЛЭП 500 кВт «Токтогульская ГЭС-п/ст Фрунзе» были проведены в течение одного осеннего сезона очень трудные

экспедиционные измерения гололедных отложений на тех перевалах Киргизского и Таласского хребтов, где нет жилья, дорог и каких-либо других коммуникаций. Проверим по критерию Крускаля–Уоллиса нулевую гипотезу об однородности гололедных выборок на этих перевалах – Чунгур (3,6 км), Джаргарт (3,64 км) и Ак-Таш (3,1 км). Перевалы расположены на максимальном удалении до 65 км. Число гололедных процессов за одну осень на них колебалось от 6 до 9 которые дали различные значения осадков на 1 погонный метр провода (кг/м). Чтобы можно было, как можно полнее использовать совокупные данные всех измерений для расчета гололедных нагрузок на перевальных участках на ЛЭП 500, необходимо проверить на однородность полученные измерения (выборки) на всех трех перевалах.

Значения гололедных осадков на этих пунктах за осенний период 1972 г. было следующим (кг/м):

1. Чунгур: 0,10; 0,12; 0,14; 0,28; 0,30; 0,32; 0,33; 0,40; 0,48;
2. Ак-Таш: 0,08; 0,22; 0,36; 0,40; 0,44; 0,66;
3. Джаргарт: 0,06; 0,10; 0,22; 0,28; 0,38; 0,42; 0,98.

Составим совокупный ранжированный ряд, обозначая с помощью индексов 1, 2 и 3 члены соответствующих выборок:

Ряд:	$0{,}06_3$	$0{,}08_2$	$0{,}10_1$	$0{,}10_3$	$0{,}12_1$	$0{,}14_1$	$0{,}22_2$	$0{,}22_3$
Ранг:	1	2	3,5	3,5	5	6	7,5	7,5
Ряд:	$0{,}28_1$	$0{,}28_3$	$0{,}30_1$	$0{,}32_1$	$0{,}33_1$	$0{,}36_2$	$0{,}38_3$	$0{,}40_1$
Ранг:	9,5	9,5	11	12	13	14	15	16,5
Ряд:	$0{,}40_2$	$0{,}42_3$	$0{,}44_2$	$0{,}48_1$	$0{,}66_2$	$0{,}98_3$		
Ранг:	16,5	18	19	20	21	22		

Ранги каждой из выборок 1, 2 и 3 и их суммы R_i равны:

1. 3,5 5 6 9,5 11 12 13 16,520; R_1=96,5
2. 2 7,5 14 16,519 21; R_2=80
3. 1 3,5 7,5 9,5 15 18 22; R_i=76,5

Проверка: по (5.23) $\Sigma R_i = 22 \cdot 23/2 = 253$; ΣR_i(факт)=253.

Подставляя найденные значения R_i в (5.21) получим:

$$X_{эмпир} = \frac{12}{22 \cdot 23}\left[\frac{96{,}5^2}{9} + \frac{80^2}{6} + \frac{76{,}5^2}{7}\right] - 3 \cdot 23 = 0{,}66 .$$

Исправим $X_{эм}$=0,66 на совпадения, относящиеся к разным выборкам. Имеются 4 группы совпадений по 2 члена в каждой. Поэтому

$$c = 1 - \frac{1}{22^3 - 22}\left[(2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (2^3 - 2)\right] = 1 - 0{,}0001 \approx 1{,}0 .$$

Величиной *c*, таким образом, можно пренебречь.

По программе «ХИ2.ОБР.ПХ» для q=0,05 и числа степеней свободы СС=(3-1)=2 имеем:

$$X_{крит}=X_{2;0,05}=5,99.$$

В результате, $X_{эмп}=0,66<X_{крит}$ и нулевая гипотеза о равенстве масс гололедных осадков на трех перевалах принимается на уровне доверительной вероятности p=0,95. Из этого следуют два очень важных вывода: 1)все три выборки можно объединить в одну с объемом 22 измерения, что даст по ней гораздо боле надежные результаты статистических расчетов, 2)на все три перевальных участка следует распространить одну оценку максимальной гололедной нагрузки (вероятную 1 раз в 15 лет, как это требуют «Правила электроустановок»), равную 6,7 кг/м. Эта оценка была найдена по полученной таким путем совокупной выборке с учетом дополнительных 5-летних наблюдений на пер. Чунгур. Дальнейшая многолетняя эксплуатация ЛЭП показала правильность такого решения.

Кроме X-критерия имеется и много других непараметрических тестов. К сожалению, их рассмотрение выходит за рамки настоящего учебника. Основные критерии можно найти в работе [23], а весьма полный перечень содержится в обстоятельной работе [14].

Глава 5.3. ПРОВЕРКА ГИПОТЕЗ СОГЛАСИЯ ЭМПИРИЧЕСКИХ И ТЕОРЕТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ, ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ СТАТИСТИК, ПРОВЕРКА ГИПОТЕЗ О ЗНАЧИМОСТИ КОРРЕЛЯЦИИ И РЕГРЕССИИ

С вопросами, перечисленными в заголовкеэтой главы , нам уже *по необходимости* приходилось сталкиваться при изложении материала предыдущих тем 1-4. Теперь рассмотрим их более полно с использованием общего подхода и материала лекций 5.1-5.2. При этом, в ряде случаев неизбежны некоторые повторы, однако постараемся свести их к минимуму, используя принципиальный подход, уже имеющиеся сведения и ссылку на приведенные ранее расчетные примеры. Для решения этих задач используются как параметрические, так и непараметрические критери. Эти критерии построены на использовании нрмального закона, t- распределения Стьюдента, F-распределения Фишера и χ^2 – Пирсона.

5.3.1. Непараметрические критерии согласия эмпирических и теоретических распределений: χ^2-Пирсона, λ-Колмогорова-Смирнова и nw^2

После того, как по данным выборки тем или иным методом подобран конкретный теоретический закон распределения в виде $F(x)$ или $f(x)$ необходимо убедиться, что эти теоретические законы хорошо описывают выборку. Это можно сделать с помощью различных *критериев согласия*, наиболее употребительными из которых являются χ^2-Пирсона, λ-Колмогорова–Смирнова и nw^2. Если между эмпирическим и теоретическим распределением имеется согласие, то можно заключить, что эмпирическое распределение *вызвано теми же причинами*, которые лежат в основе теоретического распределения. Это значительно повышает *обоснованность* практического использования теоретического закона, который аппроксимирует выборку.

Нулевая гипотеза при использовании всех критериев согласия формулируется одинаково

$$H_0: F_{эм.}=F_{теор.}, \tag{5.29}$$

против односторонней альтернативы

$$\overline{H}_3: F_{эм.}\neq F_{теор.}. \tag{5.30}$$

Обычно используются уровни значимости q=0,10, q=0,05 или q=0,01, которым соответствуют доверительные вероятности p=0,90, p=0,95 и p=0,99. Будем, как и ранее, задавать q=0,05, а p=0,95.

Принять H_0 означает, что имеющиеся различия между теоретическим и эмпирическим распределениями признаются чисто случайным, на самом деле ими можно пренебречь, и на практике вполне обоснованно использовать $F_{теор.}$ вместо выборки. Это не только технически более удобно, но гораздо шире по возможностям применения. Например, по выборке нельзя обычно оценить квантили на концах распределения при F<0,05 и F>0,95, которые в большинстве случаев наиболее важны в практических приложениях. Эти квантили без труда находятся по найденному теоретическому распределению, обоснованность применения которого дополнительно «санкционируется», в том числе и фактором принятия H_0.

1. Критерий согласия χ^2-Пирсона

Непараметрический критерий согласия χ^2-Пирсона основан на использовании χ^2-распределения и применяется только *для сгруппированных* выборок, принадлежащих *к любым законам* распределений. Он был подробно описан в п. 2.1.4, где показан его расчет и использование для примера аппроксимации нормальным законом сгруппированной выборки средних годовых температур на метеостанции Байтык.

2. Критерий согласия λ-Колмогорова–Смирнова

Непараметрический критерий согласия Колмогорова–Смирнова (λ-критерий) основан на оценке модуля максимальной разности между интегральными функциями эмпирического $F_{эм.}$ и теоретического $F_{теор.}$ распределений, имея статистику

$$\lambda = \left| F_{теор.} - F_{эм.} \right|_{макс}. \qquad (5.31)$$

Критерий может использоваться в силу своей структуры как для сгруппированных, так и несгруппированных выборок *для любых законов* распределений. По сравнению с χ^2 он лучше устанавливает различия в форме кривых *F(x)*, тогда как χ^2 лучше оценивает нерегулярность отклонений *F(x)* в распределении сгруппированных классов.

Практическое использование λ-критерия выглядит следующим образом.

1. Рассчитывают *F(x)* для эмпирического и теоретического законов в форме обеспеченностей $F_{эм}$ и $F_{теор}$. При этом $F_{эм}$ для несгрупптрованной выборке может быть рассчитана, как это было показано в п.1.1.4 для средних годовых температур на станции Байтык (табл.1.1), или для сгруппированной выборки, как например, в табл. 2.3 п. 2.1.3.

2. Определяют модуль максимальной разности между $F_{эм}$ и $F_{теор}$.

3. По табл. 5.2 для выборок (n=3 ..., 100) определяют критическое значение $\lambda_{кр}$ для уровней значимости q=0,10 и 0,05. При объеме выборок n>35 границу критической области $\lambda_{кр}$ можно найти, пользуясь аппроксимацией:

уровень значимости q	0,20	0,10	0,05	0,01
значения $\lambda_{кр}$	$1{,}07/\sqrt{n}$	$1{,}22/\sqrt{n}$	$1{,}36/\sqrt{n}$	$1{,}63/\sqrt{n}$

4. Если для заданного уровня значимости q:

$$\lambda_{эм}(5.23) < \lambda_{крит}, \qquad (5.32)$$

то нулевая гипотеза (5.29) принимается на уровне доверительной вероятности $p=1-q$. (В противном случае принимается альтернатива (5.30)).

Таблица 5.2

Критические значения ($\lambda_{крит}$) критерия Колмогорова-Смирнова

n	$q_{0,10}$	$q_{0,05}$	n	$q_{0,10}$	$q_{0,05}$
3	0,636	0,708	23	0,247	0,275
4	0,565	0,624	24	0,242	0,269
5	0,509	0,563	25	0,238	0,264
6	0,468	0,519	26	0,233	0,259
7	0,436	0,483	27	0,229	0,254

8	0,410	0,454	28	0,225	0,250
9	0,387	0,430	29	0,221	0,246
10	0,369	0,409	30	0,218	0,242
11	0,352	0,391	31	0,214	0,238
12	0,338	0,375	32	0,211	0,234
13	0,325	0,361	33	0,208	0,231
14	0,314	0,349	34	0,205	0,227
15	0,304	0,338	35	0,202	0,224
16	0,295	0,327	36	0,199	0,221
17	0,286	0,318	37	0,196	0,218
18	0,278	0,309	38	0,194	0,215
19	0,271	0,301	39	0,191	0,213
20	0,265	0,294	40	0,189	0,210
21	0,259	0,287	50	0,170	0,177
22	0,253	0,281	100	0,121	0,134

Проверим по критерию λ степень согласованности сгруппированной выборки летних сумм осадков по 500 метеостанциям России [15], которая была аппроксимирована нормальным законом Классы осадков *r*, мм и соответствующие им обеспеченности верхних границ классов - $F_{теор}$ и $F_{эм}$ - приведены в табл. 5.3

Таблица 5.3

Расчет $|\Delta F| = |F_{теор.} - F_{эмпир.}|$

r, мм	145-150	150-155	155-160	160-165	165-170	170-175	175-180	180-185	185-190	190-195	195-200
$F_{теор}$	0,0008	0,0068	0,0374	0,1378	0,3454	0,6160	0,8382	0,9532	0,9906	0,9884	0,9994
$F_{эм}$	0	0,0008	0,0208	0,1428	0,3428	0,6028	0,8308	0,9548	0,9768	0,9928	0,9928
$\|\Delta F\|$	0,0008	0,0060	0,0166	0,0050	0,0026	0,0132	0,0074	0,0016	0,0138	0,0044	0,0066

Как видно, $\Delta F_{макс}$=0,0166 и соответствует классу 155-160 мм. Зададим уровень значимости *q*=0,05 и, используя аппроксимацию $\lambda_{крит} = 1{,}36/\sqrt{n}$ (где *n*=500) получим, что $\lambda_{крит}$=0,0608. Таким образом, на уровне доверительной вероятности *p*=0,95 имеем, что

$$\lambda_{эм}=0{,}01666<\lambda_{крит}=0{,}0608$$

и, следовательно, H_0: $F_{теор}=F_{эм}$ принимается.

3. Критерий согласия nw^2 (читается –эн омега квадрат).

Непараметрический критерий согласия nw^2 наиболее полно использует информацию исходных рядов, т.к. применяется *к несгруппированным* выборкам и оценивает квадрат средней разности эмпирического $F_{эм}$ и теоретического $F_{теор}$ распределений. Статистикой критерия является функция:

$$nw^2(эм) = \frac{1}{12n} + \sum_{r_i}[F_{теор}(x_{r_i}) - F_{эмпир}(x_{r_i})]^2 , \qquad (5.33)$$

где n – объем выборки; r_i – ранг члена выборки x_{r_i} в ранжированном по возрастанию ряду; $F_{теор}(x_{r_i})$ – значение теоретической функции распределения для x_{r_i}; $F_{эмпир}(x_{r_i})$ – эмпирическая функция распределения для x_{r_i}, рассчитываемая по формуле:

$$F_{эмпир}(x_{r_i}) = \frac{r_i - 0{,}5}{n} . \qquad (5.34)$$

Математическое ожидание и дисперсия величины nw^2 определяются выражениями

$$мо\left(nw^2\right) = \frac{1}{6n}, \quad \sigma^2(nw^2) = \frac{(4n-3)}{180n^3} . \qquad (5.35)$$

Распределение статистики критерия быстро сходится к своему предельному распределению. При $n \geq 40$ аппроксимация верхних критических значений $nw^2_{крит.}$ в зависимости от уровня значимости q приведена в табл. 5.4.

Таблица 5.4

Верхние критические значения $(nw^2)_{крит.}$ в зависимости от q

q	0,30	0,20	0,10	0,05	0,03	0,02	0,01	0,001
$(nw^2)_{крит.}$	0,1843	0,2412	0,3473	0,4614	0,5489	0,6198	0,7435	1,1679

Применение критерия элементарно:

1. Исходный ряд ранжируется в возрастающем порядке и по (5.34) рассчитывается эмпирическая функция распределения по значениям рангов r_i.

2. Для значений переменной x_{r_i} по тому или иному закону рассчитывается теоретическая функция распределения.

3. По формуле (5.33) определяется эмпирическая статистика критерия $(nw^2)_{эм.}$, а по табл. 5.4 его критическое значение $(nw^2)_{крит.}$ для заданного уровня значимости q.

4. Если

$$(nw^2)_{эм.} < (nw^2)_{крит.}, \qquad (5.36)$$

то нулевая гипотеза принимается на уровне доверительной вероятности $p=1-q$. В пртивном случае приимается односторонняя альтернативная гипотеза.

5.3.2. Построение доверительных интервалов статистик и их использование

В математической статистике разработаны практические методы по оценке доверительных интервалов для таких статистик как среднее значение, дисперсия и среднее квадратическое отклонение, коэффициенты корреляции и коэффициенты уравнений регрессии для парной и множественной линейной корреляции и регрессии. Во всех случаях построение доверительных интервалов выполняется по единой схеме. Пусть, по выборке найдено точечная оценка статистики x_m, ее средняя квадратичеческая (стандартная ошибка) s_{cm} и требуется найти интервальную оценку статистики $x_{ин}$ с заданным уровнем значимости q, т.е. с уровнем доверительной вероятности $p=1-q$. Исходным выражением является задание отношения точечной оценки самой статистики x_m к ее стандарной ошибке $s_{ст}$, т.е.

$$y_{эм} = x_m/s_{cm}, \qquad (5.37)$$

где $y_{эм}$ является случайной величиной.

Это выражение для $y_{эм}$ подчинается одному из следующих законов: нормальному $N(0,1)$, t-распределению Стьюдента, χ^2-Пирсона или F-Фишера. Тогда, интервальное значение статистики $y_{ин}$ может быть определено как квантили этих законов для заданного значимости q

$$x_m - \alpha_{(q/2)}\, s_{cm} < y_{ин} < x_m + \beta_{(1-q/2)}\, s_{cm} \quad , \qquad (5.38)$$

где $\alpha_{(q/2)}$ и $\beta_{(1-q/2)}$ – квантили известного распределения $y_{эм} = x_m/s_{cm}$.

В случае применения симметричных закона $N(0,1)$ и t-распре-деления Стьюдента выражение (5.30) можно записать также в виде

$$y_{ин} = x_m \pm \alpha_{(q/2)}\, s_{cm} \qquad (5.39)$$

Исключение представляет построение доверительного интервала для дисперсии и среднего квадратического отклонения, для которых эта схема, принципиально оставаясь верной, технически имеет несколько иной вид.

Уравень значимости q это есть ошибка первого рода. Увеличивая значение p мы будем расширять область доверительных значений $y_{ин}$ и сужать область ее критических значений. Обычно p задается значениями 0,90, 0,95, 0,99 и 0,999, тогда q соответственно равно 0,10, 0,05, 0,01 и 0,001. Но увеличивая p и уменьшая q, мы одновремнно увеличиваем риск совершения ошибки второго рода. На пратике, как уже отмечалось, во многих случаях принимается q= 0,05 и p=0,95, что обеспечивает некоторый средний баланс вероятностей обеих ошибок. Одновременно для снижения ошибки второго рода выбирается наиболее мощный из возможных критериев. Рассмотрим конкретную технику построения доверительных интервалов применительно к названным статистикам.

1. Доверительный интервал для среднего.

Построение доверительного интервала для среднего значения с использованием нормального распределения (при больших выбоках, когда n>30-50) и *t*-распределения Стьюдента применимого для выборок любого объема, показано соответственно в п. 2.2.1 и 2.2.2.

2. Доверительный интервал для дисперсии и СКО.

Доверительные интервалы для этих статистик находятся с помощью χ^2-распределения Пирсона. Технология их построения подробно описана в п. 2.2.3.

3. Доверительные интервалы для коэффициентов парной линейной регрессии.

Для углового коэффициента b_1 и остаточного члена b_0 парной линейной регрессии доверительные интервалы этих статистик при выборках любого объема могут быть найдены с использованием *t*-распределения Стьюдента. При больших выборках ($n \geq 30$) аналогичным образом можно применять также нормальное распределение.

Формулы для стандартных ошибок b_0 *и* b_1 имеют вид:

$$s_{b_0}^2 = \frac{s_2^2 \sum x_i}{n\sum (x_i - \bar{x})^2} = \frac{s_2^2[(n-1)s_x^2 + n\bar{x}^2]}{n(n-1)s_x^2}, \qquad (5.40)$$

$$s_{b_1}^2 = \frac{s_2^2}{\sum (x_i - \bar{x})^2} = \frac{s_2^2}{(n-1)s_x^2}. \qquad (5.41)$$

Напомним, что значения этих стандартных ошибок в Excel вычисляются при использовании программы ЛИНЕЙН как для парной, так и для множественной линейной регрессии.

Можно считать, что отношение (5.37) для среднего значения имеет *t*-распределение Стьюдента с (n-2) степенями свободы, которое при больших n сходится к нормальному закону. Поэтому формулы для построения симметричных двухсторонних доверительных интервалов по (5.39) будут иметь вид

$$b_1(ин) = b_1 \pm s_{b_1} \cdot t_{(n-2),1-q/2}, \qquad (5.42)$$

$$b_0(ин) = b_0 \pm s_{b_0} \cdot t_{(n-2),1-q/2}. \qquad (5.43)$$

где b_0 и b_1 – точечные оценки коэффициентов регрессии, $t_{(n-2),1-q/2}$ – квантиль *t*-распределения с (n-2) степенями свободы и уровнем значимости $q=1-p$ (p – доверительная вероятность). При больших выборках (n>30) или квантиль *t*-распределения может быть заменен на квантиль стандартного нормального закона $z_{(1-q/2)}$.

Пример оценки интервала для b_1. Пусть, например, b_1=0,45, s_{b_1}=0,12 и n=30. По программе «СТЬЮДЕНТ.ОБР.2Х» (которая составлена для двустороннего критеря) для уровня значимости q=0,05 и CC=30-2=28 получим квантильное значение $t_{28;0,05}$ =2,05. Тогда,

b_1*(ин)*=0,45±0,12·2,05⇒b_1*(ин)*=0,20 ..., 0,80. Это значит, что с доверительной вероятностью p=0,95 истинное значение углового коэффициента регрессии b_1 лежит в интервале от 0,20 до 0,80. В принципе мы вправе *взять любое значение* b_1 из этого интервала, т.к. они равновероятны. Но обычно разумно использовать найденное по выборке значение b_1, лежащее в середине этого интервала и равное b_1=0,45. При этом дополнительно можно сделать еще один из двух следующих важных выводов:

а) если b_1*(ин*) *не включает ноль* (как в нашем случае), то регрессиию следуе также признать значимой на уровне q =0.05.

б) если доверительный интервал b_1(ин) *включает ноль* (например, если бы было получено *b(ин)*= -0,15 ..., 0,55), то теперь следует признать, что b_1=0; это означает принятие незначимыми на уровне q =0.05, как регрессии, так и корреляции, что следует из того, что $b_1=rs_y/s_x$.

Таким образом, построением доверительного интервала для b_1 на самом деле одновременно решаются две задачи: получение численных границ доверительного интервала для b_1 и оценка значимости регрессии и корреляции.

Остаточный член парной регрессии b_0. Техника построения для него доверительного интервала полностью аналогична описанной для b_1. Но дополнительные выводы теперь иные. По смыслу b_0 есть отрезок, отсекаемый графиком регрессии на оси абцис. Поэтому, если доверительный инттервал для b_0 ключает ноль, то это означает, что следует приять нулевую гипотезу: $b_0=0$. Важность принятия такого решения может проистекать из физической сути задачи, когда требуется, чтобы при нулевом значении предиктора *x,* предиктант *y* также соответствовал нулю. Например, при сравнении показаний двух ветроизмерительных приборов желательно, чтобы при нулевых показаниях эталонного прибора поверяемый также показывал ноль.

4. Доверительные интервалы для коэффициентов множественной линейной регрессии.

Технология построения доверительных интервалов для угловых коэффициентов множественной линейной регрессии b_k ничем не отличается от их технологии для парной регрессии, если учесть, что по формуле (5.34) надо в отдельности для каждого углового коэффициента b_k использовать свою стандартную ошибку s_{b_k}. При этом значения коэффициентов b_k и их ошибок s_{b_k} вычисляются по программе ЛИНЕЙН. Разница состоит лишь в том, что теперь при нахождении квантиля *t*-распределения для формулы (5.42) по программе «СТЮДНЕТ,ОБР.2Х» надо взять число степеней не (*n-2*), а равное (*n-k-1*), где *k* есть число независимых переменных в множественном уравнении регрессии. Так, например, при *n*=52 и числе переменных *k*=3 число *CC*=52-3-1=48.

Однако выводы, которые можно сделать из оценки доверительного интервала для каждого b_k более узкие, чем для парной регрессии, и состоят в следующем: 1)доверительный интервал для каждого b_k показывает область значений, в которой с доверительной вероятностью p находится истинное значение b_k; 2)если доверительный интервал включает ноль, то это *только означает*, что можно считать, что этот коэффициент b_k=0 (т.е. принять H_0: b_k=0) и, следовательно, предиктор x_k может быть исключен из уравнения регрессии; важно, что сама множественная регрессия не перестает от этого быть значимой, если значимы в ней остальные (хотя бы один из них) предикторы.

Процедура построения доверительного интервала для b_0 с использованием формулы (5.43) как для парной, так и множественной регрессии полностью аналогичны.

6. *Доверительный интервал для парного линейного коэффициента корреляци r.*

Для нахождения доверительного интервала для коэфициента парной линейной корреляции r используются два метода расчета – приближенный и более точный, который лучше применять всегда.

Способ 1 (приближенный). Для больших выборок (n>30-50) и не очень больших по модулю r ($|r|\le$0,7-0,8) выборочный коэффициент корреляции распределен приблизительно нормально со средней квадратической ошибкой s_r

$$s_r = \frac{1-r^2}{\sqrt{n-2}}. \qquad (5.44)$$

Тогда, доверительный интервал для r с уровнем значимости q и доверительной вероятностью $p=1-q$ будет равен

$$r(ин.) = r \pm t_{(n-2);1-q/2} \cdot s_{r.} \qquad (5.45)$$

где $t_{(n-2);\ 1-q/2}$ – квантиль t-распределения с (n-2) степенями свободы для q=0,05, который находится по программе «СТЬЮДЕНТ.ОБР.2Х», где задается вероятность q (при n>30 квантиль t-распределения можно заменить на $z_{q/2}$ – квантиль N(0,1)).

Способ 2 (более точный, который лучше применять всегда). Здесь надо перейти от рассчитанного по выборке значения r к величине z, по так называемому z-преобразованию Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}, \quad s_z = \frac{1}{\sqrt{n-3}}. \qquad (5.46)$$

Величина z распределена нормально уже при малых n и любых r. Поэтому, сначала надо построить доверительный интервал для z по соотношению

$$z(ин.) = z \pm t_{(n-2);1-q/2} \cdot s_z, \qquad (5.47)$$

а затем обратным решением левой формулы (5.46) найти концы доверительного интервала для *r* по концам для z.

Для облегчения этой процедуры используются программы Excel для прямого («ФИШЕР», находится z по r) по и обратного («ФИШЕРОБР», находится r по z) преобразования Фишера. Имеются также специальные таблицы этих преобразований.

Кроме получения самого интервала для r, можно также сделать заключение о значимости *r* и линейной регрессии: если доверительный интервал для *r* включает ноль, то и корреляция и регрессия – статистически *незначимы*, если доверительный интервал для *r* не включает ноль, то корреляцию (а, следовательно, и регрессию) надо с уровнем вероятности *p* признать *значимой*. При этом для практического использования можно брать любое значение *r* из доверительного интервала, но обычно предпочитают использовать *r* выборочное.

Пример 1. Пусть значение *r* не велико: *r*=0,50, *n*=71 и по (5.44) ошибка s_r=0,0903. Тогда при *p*=0,95 и *q*=0,05 имеем:

$$r(ин)=0,50\pm1,96\cdot0,0903=0,32\ ...,\ 0,68.$$

Доверительный интервал для r равен 0,32…,0,68 и не включает ноль, поэтому корреляция, а также регрессия значимы на уровне q=0.05.

Пример 2. Пусть значение велико: *r*=0,90, *n*=71. Тогда, по программе «ФИШЕР» находим, что для *r*=0,90 значение *z*=1,4722 и по (5.46) ошибку s_z =0,121. По (5.47) для *p*=0,95 и *q*=0,05 имеем:

$$z(ин)=1,4722\pm1,96\cdot0,121=1,2345\ ...,\ 1,7099.$$

Переводя концы интервала для *z* 1,2345 и 1,7099 по программе «ФИШЕРОБР» в концы интервала для *r*, получим: *r*(ин)=0,84 ..., 0,94.

Таким образом, доверительный интервал, найденный для точечного значения r=0.90 с использование z-преобразования Фишера равен 0,84…, 94; он не включает ноль, что так же означает, что линейная корреляция и соответствуюшая ей регрессия значимы на уровне q=0.05.

7. *Доверительный интервал для множественного линейного коэффициента корреляции R.*

Напомним, что стандартная ошибка коэффициента множественной линейной коррклляции s_R определяется формулой

$$s_R = \frac{1-R^2}{\sqrt{n-k-1}}, \qquad (5.48)$$

где *k* – число независимых переменных в выборке.

Поэтому доверительный интервал *R(ин)* будет выражаться формулой

$$R(ин) = R \pm t_{q;(n-k-1)}\, s_R, \qquad (5.49)$$

где $t_{q;(n-k-1)}$ – квантиль t–распределения с уровнем значимоси q и степенью свободы $CC=n-k-1$.

В формуле (5.49) квантиль $t_{q;(n-k-1)}$ определяется по программе «СТЬЮДЕНТ.ОБР.2Х», где сразу предусмотрено задание 2 задание двустороннего критерия со сначением вероятности q. При больших выборках (n>30) квантиль t–распределения в (5.49) можно заменить на квантиль стандартного нормального закона $z_{(1-q/2)}$.

Пример. Пусть R=0.90, число независимых переменных k=3 и n=52. Тогда, значение s_R=0,0274. По программе Excel «СТЬЮДЕНТ.ОБР.2Х» для q=0.05 и CC=52-3-1=48 получим $t_{0.05;\ 48}$=2,01. Подставляя эти значения в (5.49) найдем границы доверительного интервала - *R(ин)=0,85…,0,96.* Это значит, что с доверительной вероятностью p=0,95 неизвесное нам истинное значение R лежит в интервале от 0,85 до 0,96. Как и для парного r справедливыми также будут два дополнительных вывода: если доверительный интервал для R включает ноль, то и корреляция и регрессия – статистически *незначимы*, если доверительный интервал для R не включает ноль, то корреляцию (а, следовательно, и регрессию) надо с уровнем вероятности p признать *значимой*. При этом для практического использования можно брать любое значение R из доверительного интервала, но обычно предпочитают использовать R выборочное.

5.3.3 Проверка гипотез значимости корреляции и регрессии

Задачи проверки гипотез значимости как парной, так и множественной корреляции и регресии попутно уже решались в предыдущем пункте при остроении доверительных интервалов для угловых коэфициентов регресии и коэффициентов корреляции. Рассмотрим теперь этот вопрос более подробно.

1. *Проверка гипотез значимости парной и множественной линейной корреляции и регресии с помощью F- критерия Фишера.*

Описываемая ниже процедура идентично применима для парной и множественной линейной регрессии и корреляции. Для парной регрессии и корреляции этот вопрос уже был рассмотрен в п. 4.1.3, а для множественной в п.4.4.1. Приведем теперь основные положения и возможности использования двух вариантов F-критерия Фишера. В обоих случаях нулевая гипотеза H_0 формулируется как: *регрессия (корреляция) отсутствует, т.е.*

$$H_0: b_1=0\ (H_0: r=0); \quad H_0: b_1=b_2=\ldots,=b_k=0\ (H_0: R=0), \qquad (5.50)$$

против одноторонней альтернативы $\overline{H}_3$

$\overline{H}_3$: *регессия (корреляция) статистически значима.* (5.51)

Вариант 1(реализован в Excel). Эмпирическое значение критерия *F(эм)* строится как отношение:

$$F(эм) = \frac{\sum_1 : k_1}{\sum_2 : k_2}, \tag{5.52}$$

где k_1 – есть число независимых переменных в уравнении регрессии, $k_2=(n-k_1-1)$ и Σ_1 – есть закономерная регрессионная сумма (см. п. 4.1.2),

$$\Sigma_1 = \Sigma_1(\tilde{y}_i - \bar{y})^2, \tag{5.53}$$

а $\sum_2$ – есть остаточная сумма, равная

$$\Sigma_2 = \Sigma_2(y_i - \tilde{y}_i)^2 \tag{4.54}$$

Величина *F* (эм) по (5.44) распределена по *F*-распределению Фишера с числом степеней свободы k_1 (CC_1) для числителя и k_2 (CC_2) – для знаменателя. Поэтому критическое значения *F*-критерия определяется выражением

$$F_{крит} = F_{k_1,(n-k1-1),q}. \tag{5.55}$$

Значение $F_{крит}$ находится по прогамме «F.ОБР.ПХ», для чего надо задать уровень значимости *q* и последовательно две степени свободы: $CC_1=k_1$ *и* $CC_2= k_2=(n-k_1-1)$.

Если $F_{эм} < F_{крит}$, то на уровне доверительной вероятностью *p* (на уровне значимости *q*) нулевая гипотеза принимается и регрессию (корреляцию) следует признать статистически *незначимыми*. В противном случае нулевая гипотеза отвергается и регрессия, а также корреляция принимаются *значимыми*.

В Excel использован именно этот вариант *F*- критерия в программах ЛИНЕЙН и ЛГРФПРИБЛ. При этом на печать, вместе со значением *F*(эм), выдается число степеней свободы $CC_2=k_2=(n-k_1-1)$, а также две суммы - Σ_1 и Σ_2. Порядок работы с программой ЛИНЕЙН на примере линейной регрессии (k_1=1) приведен в п. 4.1.3.

Вариант 2 (в Excel отсутствует). Эмпирическое значение критерия задается отношением

$$F = \frac{\sum_y : (n-1)}{\sum_2 : (n-k-1)} = \frac{s_y^2}{s_2^2}, \tag{5.56}$$

где $\sum_y$ и s_y^2 – полная сумма квадратов и полная дисперсия, а $\sum_2$ и s_2^2 – остаточная сумма квадратов и остаточная дисперсия, равные (см. п. 4.1.2):

$$\Sigma_y = \Sigma_y(y_i - \bar{y})^2, \quad s_y^2 = \frac{1}{n-1}\Sigma_y(y_i - \bar{y})^2, \tag{5.57}$$

$$\Sigma_2 = \Sigma_2 (y_i - \tilde{y}_i)^2 , \qquad s_2^2 = \frac{1}{n - k_1 - 1} \Sigma_2 (y_i - \tilde{y}_i)^2 . \qquad (5.58)$$

Величина F по (5.56) распределена по F-распределению Фишера с (n–1) степенями свободы для числителя и (n–k_1–1) степенями для знаменателя (n – объем выборки, а k_1 – число независимых переменных в уравнении регрессии), т.е.

$$F_{крит} = F_{(n-1),(n-k_1-1),q} , \qquad (5.59)$$

где q – уровень значимости.

Таким образом, критическое значение F-критерия в варианте 2 также, как и в варианте 1, определяется по программе «F.ОБР.ПХ», но оба числа степеней свободы теперь другие.

Если $F_{эмп}$ (5.57)<$F_{крит}$ (5.59), то нулевая гипотеза принимается с доверительной вероятностью $p=1-q$, т.е. корреляция и регрессия статистически *незначимы*. В противном случае нулевая гипотеза отвергается и регрессия, а также корреляция принимаются *значимыми*.

Одновременно величина *F(эм)* по (5.56) показывает - во сколько раз $\tilde{y}$ предсказывается по регрессии лучше, по сравнению с его таким простым прогнозом как $\tilde{y} = \bar{y}$ (среднее значение по выборке). Численный пример использования F -критерия по варианту 2 дается в п. 4.1.3.

При применеии F-критерия в обоих вариантах для проверки гипотезы значимости *нелинейной парной* корреляции и регресии технология его расчетов и использования полностью аналогична технологии, которая применяется *для парной линейной* корреляции и регрессии. Однако значение *F(эмп)* в этих случаях приходится рассчитывать самостоятельно. Исключением является программа нелинейной регрессии ЛГРФПРИБЛ, в которой предусмотрен расчет *F(эмп)*.

2. *Проверка гипотезы значимости парной и множественной линейной корреляции и регрессии с помощью t-критерия Стьюдента.*

Статистические нулевые гипотезы относительно угловых коэффициентов регрессии остаточного члена регрессии b_0, коэффициента парной r и множественной R линейной корреляции могут быть так же проверены с помощью t-распределения Стьюдента. Нулевые гипотезы H_0 формулируются однотипно по формулам (5.52) против *двухсторонней* альтернативы - $\overline{H}_3$: рассчитанные по выборке эмпирические значения этих статистик значимо отличаются от нуля.

Эмпрические значения критериев во всех случаях строятся как отношения найденных точечных оценок статистик к их стандартным ошибкам. Эти отношения распределены по t-распределению Стьюдента с различными для разных статистик, но известными,

степенями свободы. Задавая уровень значимости q и извесные степени свободы CC, по программе «СТЬЮДЕНТ.ОБР.2Х» находятся критические значения критерия, равные

$$t(кр) = t_{q;cc} \qquad (5.6\ 0)$$

Если $t(эм) < t(кр)$, то нулевая гипотеза принимается на уровне значимости q и значение статистики считается *незначимо отличающимися от нуля (кратко – не значимы)* на уровне доверительной вероятности $p=1-q$. В противном случае принимается альтернативная гипотеза – статистики значимы на уровне $p=1-q$. Покажем эти процедуры для каждой статистики отдельно.

Угловой коэфициент парной линейной регресси b_1. Эмпирическое значение критерия находится по формуле

$$t(эм) = b_1/s_{(b1)}, \qquad (5.61)$$

b_1 и $s_{(b1)}$ – найденные по выборке угловой коэффициент регрессии (всегда берется по модулю) и его стандартная ошибка, которые рассчитываются по программе «ЛИНЕЙН».

Критические значения критерия – $t(кр)$ находятся по программе «СТЬЮДЕНТ.ОБР.2Х», для чего задаются уровень значимости q и числло степеней свободы $CC=(n-2)$. Если $t(эм) < t(кр)$, то нулевая гипотеза *принимается* на уровне значимости q и угловой коэффициент b_1 считается *не значимыми* (равным нулю, что означает незначимость регрессии) на уровне доверительной вероятности $p=1-q$. В противном случае *принимается альтернативная гипотеза* – b_1 значимо отличается от нуля и регресия значима на уровне $p=1-q$.

Пример. Значение углового коэффициента парной регрессии и его стандартной ошибки оказались равными: b_1= -0,426, $s_{(b1)}$=0,190, объем выборки n=80. Находим $t(эм)$=2.242. По программе «СТЬЮДЕНТ.ОБР.2Х» для q=0,05 и СС=78 получим $t(кр) = t_{0,05,78}=1,99$. Так как $t(эм)>t(кр)$, то нулевая гипотеза отвергается и на уровне значимости 0,05 принимается альтернативная - регрессия статистически значима ($b_1 \neq 0$). Одновременно это означает и значимость корреляции ($r \neq 0$), так как $b_1 = r s_y/s_x$.

Угловой коэфициент множественной линейной регресси b_k. В случае множественной линйной регресии с числом независимых переменных (предикторов), равным k, процедура проверки идентична парной регрессии, только надо учесть, что теперь проверяется статистическая значимость каждого углового коэффициента b_k в отдельности, а число степеней свободы для них определяется по формуле - $CC=(n-k-1)$. Изменится и формулировка выодов результатов проверки, как это показано в следующем примере - они будут относиться только к значимости каждого коэффициента b_k.

Пример. Пусть значение углового коэффициента b_2 множественной регрессии (k=4) и его стандартной ошибки, как и в предыдущем примере, оказались равными: b_2= -0,426,

$s_{(b2)}$=0,191, объем выборки n= 80. Находим $t(эм)$=2.242. По программе «СТЬЮДЕНТ.ОБР.2Х» для q=0,05 и CC=75 получим $t(кр)= t_{0,05,75}=1,992$. Так как $t(эм)>t(кр)$, то нулевая гипотеза отвергается и на уровне значимости 0,05 принимается альтернативная – угловой коэффициент множественной регрессии b_2 статистически значим ($b_2 \neq 0$) и этот фактор следует оставить в уравнении регрессии как статистически значимый. Как видно, с учетом того, что регрессия теперь множественная, изменилась и формулировка вывода, который касается только значимости коэффициента b_2.

Обращаем внимание на полученную очень малую зависимость значений $t(кр)$ от числа степеней сводобы при объемах выборок более 50. Это является следствием достаточно быстрого схождения по вероятности t-распределения к $N(0,1)$. Так для примеров выше квантиль $z_{(1-q/2)}=z_{0.975}$ =1,96. Однако при малых объемах выборок (n<30) эта зависимость выражена боле сильно и тем сильнее, чем меньше n.

Коэффициент парной линейной корреляции r. Эмпирическое значение критерия находится по формуле

$$t(эм) = r/s_{(R)}, \qquad (5.62)$$

где r и $s_{(r)}$ – найденные по выборки значения коэффициента корреляции и его стандартной ошибки по формуле (4.23)

$$s_r = \frac{1-r^2}{\sqrt{n-2}}.$$

Критическое значение $t(кр)$ находится по программе «СТЬЮДЕНТ.ОБР.2Х» для уровня значимости q и степени свободы $CC=(n-2)$. Если $t(эм)< t(кр)$, то нулевая гипотеза *принимается* на уровне значимости q и коэффициент корррреляции r считается *незначимыми* (равным нулю, что означает одновременно означает и незначимость регрессии) на уровне доверительной вероятности $p=1- q$. В противном случае *принимается двусторонняя альтернативная гипотеза – r* значимо отличается от нуля на уровне $p=1- q$ (одновременно значима также и регресия т.к. $b_1=rs_y/s_x$).

Пример. Пусть найденное значение r= -0,50, n=71 и $s_{(r)}$ =0,0903. Тогда $t(эм) =0,50/0,0903=5,54$.

По программе «СТЬЮДЕНТ.ОБР.2Х» для q=0,05 и CC=69 получим $t(кр)= t_{0,05,69}=1,995$. Так как $t(эм)>t(кр)$, то нулевая гипотеза отвергается и на уровне значимости q=0,05 принимается двусторонняя альтернатива – коэффициент корреляции r статистически значим (r≠0). Одновременно это означает и значимость парной линейной регрессии.

Коэффициент множественной линейной корреляции R. Эмпирическое значение критерия находится по формуле

$$t(эм) = R/s_{(R)}, \qquad (5.63)$$

где R и $s_{(R)}$ – найденные по выборки значения коэффициента корреляции и его стандартной ошибки, определяемой по формуле (4.55)

$$s_R = \frac{1-R^2}{\sqrt{n-k-1}},$$

где k – число независимых переменных в выборке.

Критическое значение *t(кр)* находится по программе «СТЬЮДЕНТ.ОБР.2Х» для уровня значимости q и степени свободы $CC=(n-k-1)$. Если $t(эм)< t(кр)$, то нулевая гипотеза *принимается* на уровне значимости q и коэффициент корррреляции R считается *не значимыми* (равным нулю, что означает одновременно означает и незначимость регрессии). В противном случае *принимается двусторонняя альтернативная гипотеза* – R значимо отличается от нуля на уровне $p=1- q$ (одновременно значима также и множественная линейная регресия).

Пример. Пример. Пусть найденное значение R= 0,80, n=59, число независимых переменных k=4 и $s_{(r)}$=0,0490. Тогда $t(эм) =0,80/0,0490=16,33$.

По программе «СТЬЮДЕНТ.ОБР.2Х» для q=0,05 и CC=54 получим $t(кр)= t_{0,05,54}=2,01$. Так как $t(эм)>t(кр)$, то нулевая гипотеза отвергается и на уровне значимости q=0,05 принимается альтернатива – коэффициент корреляции R статистически значим (R>0). Одновременно это означает и значимость множественной регрессии.

Таким образом, рассмотренные методы главы 5.3 позволяют оценить значимость параметров корреляции и регрессии по двум типам критериев, основанных на использовании F-распределении Фишера и t-распределении Стьюдента. Причем, так как корреляция и регрессия связаны между собой, то значимость/(не значимость) одной из них автоматически влечет за собой также значимость/(не значимость) другой. Поэтому на практике достаточно выполнять оценку значимости только для одной их видов зависимостей, например, регрессии. Однако формально следует иметь собственный метод оценивать значимость другой характеристики, например, корреляции, когда регессия нас вообще не интересует.

Так как в Excel в программах ЛИНЕЙН и ЛГРФПРИБЛ делаются вычисления $F(эм)$ по варианту 1 (формула (5.52)), то обычно достаточно использовать только один этот метод оценки значимости регрессии и корреляции, прибегая дополнительно к другим в случае какой-либо неоходимости.

К сожалению, подобные простые методы оценки значимости статстик нелинейной парной корреляции и регрессии, а также нелинейной множественной корреляции и регрессии пока не разработаны.

ЛИТЕРАТУРА

1. *Алексеев Г.А.* Объективные методы выравнивания и нормализации корреляционных связей. – Л.: Гидрометеоиздат, 1971. – 363 с.
2. *Афифи А.А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982. – 488 с.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Вычислительный центр АН СССР, 1968. – 464 с.
4. *Боровиков В.* Statistica: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Изд-во Дом ПИТЕР, 2001. – 650 с.
5. *Брукс К., Карузерс Н.* Применение статистических методов в метеорологии /Пер. с англ. Ивановой Е.Ф. и Френкеля Л.Л. – Л.: Гидрометеоиздат, 1963. – 416 с.
6. *Вентцель Е.С.* Теория вероятностей. – М.: Наука, 1969. – 576 с.
7. *Верещагин М.А., Наумов Э.П., Шаталинский К.М.* Статистические методы в метеорологии. – Казань: Казанск. госуд. ун-т, 1990. – 107 с.
8. *Гмурман В.Е.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1977.
9. *Горяинов В.Т., Журавлев А.Г., Тихонов В.И.* Статистическая радиотехника. Примеры и задачи. – М.: Советское радио, 1980. – 543 с.
10. *Гумбель Э.* Статистика экстремальных значений /Пер. с англ. В.Ю. Татарского. – М.: Мир, 1965. – 450 с.
11. *Давлетгалиев С.К.* Математические методы обработки гидрологических данных. – Алматы: КазГНУ, 1998. – 166 с.
12. *Зажигаев Л.С., Кишьян А.А., Романиков Ю.И.* Методы планирования и обработки результатов физического эксперимента. – М.: Атомиздат, 1978. – 231 с.
13. *Заварина М.В.* Строительная климатология. – Л.: Гидрометеоиздат, 1976. – 312 с.
14. *Закс Л.* Статистическое оценивание /Пер. с нем. В.Н. Варыгиной. – М.: Статистика, 1976. – 598 с.
15. *Исаев А.А.* Статистика в метеорологии и гидрологии. – М.: МГУ, 1988. – 245 с.
16. *Кендал М.Дж., Стьюарт А.* Теория распределений /Пер. с англ. В.В. Сазонова, А.Н. Ширяева. – М.: Наука, 1966. – 587 с.
17. *Кобышева Н.В.* Косвенные расчеты климатических характеристик. – Л.: Гидрометеоиздат, 1971.

18. *Комелев В.А., Староверов О.В., Турундаевский В.Б.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1971. – 400 с.
19. *Линник Ю.В.* Метод наименьших квадратов и основы теории обработки наблюдений. – М.: Изд-во физ.-мат. лит-ры, 1962. – 350 с.
20. *Львовский Е.Н.* Статистические методы построения эмпирических формул. Учебное пособие для вузов. – М.: Высшая школа, 1988. – 239 с.
21. *Митропольский А.К.* Техника статистических вычислений. – М.: Наука, 1971.
22. *Пановский Г.А., Брайер Г.В.* Статистические методы в метеорологии /Пер. с англ. И.П. Гейбера, В.А. Шнайдмана. – Л.: Гидрометеоиздат, 1972. – 209 с.
23. *Подрезов О.А.* Методы статистической обработки и анализа гидрометеорологических наблюдений. - Бишкек: Изд-во КРСУ, 2009. – 262 с.
24. *Романовский В.И.* Математическая статистика. Том 1 и 2. – Ташкент: Изд-во АН Узб.ССР, 1961, 1963.
25. *Рождественский А.В., Чеботарев А.И.* Статистические методы в гидрологии. – Л.: Гидрометеоиздат, 1974. – 424 с.
26. *Рунион Р.* Справочник по непараметрической статистике /Пер. с англ. Е.З. Демиденко. – М.: Финансы и статистика, 1982. – 198 с.
27. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1969.
28. Справочник по теории вероятностей и математической статистике. – М.: Изд-во физ.-мат. лит-ры, 1985. – 640 с.
29. *Уланова Е.С., Забелин В.Н.* Методы корреляционного и регрессионного анализа в агрометеорологии. – Л.: Гидрометеоиздат, 1990. – 207 с.
30. *Шор Я.Б., Кузмин Ф.И.* Таблицы для анализа и контроля надежности. – М.: Светское радио, 1968. – 284 с.

Подрезов Олег Андреевич

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ И АНАЛИЗА ГИДРОМЕТЕОРОЛОГИЧЕСКИХ НАБЛЮДЕНИЙ

Часть 1

Методы анализа с использованием статистик, аппроксимации распределений, регрессии, корреляции и проверки гипотез

Редактор И.С. Волоскова
Технический редактор О.А. Матвеева
Компьютерная верстка Д.Р. Зайнулиной, *Ю.Ю. Юдаковой*

Подписано к печати 4.06.2003. Формат $60 \times 84\,{}^{1}/_{16}$.
Офсетная печать. Объем 16,5 п.л.
Тираж 100 экз. Заказ 7.

Издательство Кыргызско-Российского
Славянского университета
720000, Бишкек, Киевская, 44

Отпечатано в типографии КРСУ
720048, Бишкек, Горького, 2