

**МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. М.В. ЛОМОНОСОВА**

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

В.Д. Мятлев, Л.А. Панченко, А. Т. Терехин

ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Пособие по курсу

"Математические методы в биологии"

МОСКВА

МАКС Пресс

2002

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1. СЛУЧАЙНАЯ ВЫБОРКА И ЕЕ ОПИСАНИЕ

1.1. Понятие случайной выборки

1.2. Характеристики случайной выборки

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

2.1. Логика статистического оценивания

2.2. Доверительные интервалы

2.2.1. Доверительный интервал для математического ожидания нормально распределенной случайной величины с известной дисперсией

2.2.2. Доверительный интервал для математического ожидания нормально распределенной случайной величины с неизвестной дисперсией

2.2.3. Доверительный интервал для неизвестной дисперсии нормально распределенной случайной величины (при неизвестном математическом ожидании)

2.2.4. Доверительный интервал для неизвестного параметра p биномиального распределения

2.2.5. Доверительный интервал для неизвестного параметра l пуассоновского распределения

2.2.6. Приближенный доверительный интервал для неизвестного коэффициента корреляции двумерного нормального распределения

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Логика проверки статистических гипотез

3.2. Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с известной дисперсией

3.3. Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с неизвестной дисперсией (одновыборочный t -критерий)

3.4. Проверка гипотезы о равенстве заданному числу дисперсии нормально распределенной случайной величины (одновыборочный χ^2 -критерий)

3.5. Проверка гипотезы о равенстве математических ожиданий двух независимых нормально распределенных случайных величин (двухвыборочный t -критерий)

3.6. Проверка гипотезы о равенстве дисперсий двух независимых нормально распределенных случайных величин (двухвыборочный F -критерий)

3.7. Приближенный критерий для проверки гипотезы о равенстве параметров двух независимых биномиальных случайных величин (критерий для сравнения двух вероятностей)

3.8. Приближенный критерий для проверки гипотезы о равенстве параметров двух независимых пуассоновских случайных величин

3.9. Приближенный критерий для проверки гипотезы о равенстве нулю коэффициента корреляции между компонентами двумерной нормально распределенной случайной величины

3.10.. Критерии согласия

3.11. Непараметрические критерии

3.11.1. Одновыборочные критерии

3.11.2. Проверка гипотезы об отсутствии сдвига

3.11.3. Критерии однородности

3.11.4. Проверка гипотезы о независимости

ЛИТЕРАТУРА

ВВЕДЕНИЕ

Задача *математической статистики*, в строгом понимании этого термина, состоит в разработке и применении методов описания реальных явлений вероятностными моделями, исходя из данных, полученных в результате наблюдений за этими явлениями.

В более широком смысле математическая статистика понимается как совокупность методов планирования экспериментов и обработки данных, полученных в результате экспериментов, причем эти методы могут не основываться на вероятностных моделях. При таком широком понимании вместо термина «математическая статистика» часто используют термин «*анализ данных*».

Исторически вначале сформировались методы обработки данных, не связанные тесно с теорией вероятности, так называемая дескриптивная, *описательная статистика*. С начала этого века начали интенсивно развиваться методы анализа данных, основанные на вероятностных моделях, - это, прежде всего, методы *статистического оценивания* и *статистической проверки гипотез*, о которых будет идти речь в данной книге.

Бурное развитие вычислительной техники вызвало к жизни ряд новых методов анализа. Некоторые из этих методов разработаны на основе подходов, отличных от теоретико-вероятностного (геометрические, оптимизационные и др.). Вероятностное их обоснование либо отсутствует, либо недостаточно, что затрудняет количественную оценку степени достоверности выводов и исследование аналитическими средствами классической математической статистики. Однако в последние годы, также в связи с быстрым ростом производительности вычислительных машин, начали получать распространение процедуры так называемого *случайного моделирования* (пермутационные методы, бут-стрэп), позволяющие оценить статистические свойства получаемых решений без аналитических методов.

Начнем рассмотрение методов математической статистики с ее исходного понятия - понятия случайной выборки.

1. СЛУЧАЙНАЯ ВЫБОРКА И ЕЕ ОПИСАНИЕ

1.1. Понятие случайной выборки

Понятие случайной выборки тесно связано с понятием *случайного испытания* и случайной величины, о которых шла речь в предыдущей главе. *Случайная выборка* представляет собой совокупность наблюдений x_1, x_2, \dots, x_n случайной величины ξ , полученных в n независимых случайных испытаниях. Число полученных наблюдений n называется *объемом выборки*. Образно можно представить процесс получения случайной выборки как извлечение наудачу значений из гипотетической бесконечной *генеральной совокупности*, где разные значения содержатся в пропорциях, соответствующих распределению случайной величины, и тщательно перемешаны между собой. Случайная величина может быть не только одномерной, но и многомерной - $\vec{\xi}$, тогда каждая из компонент выборки $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ будет вектором.

Конкретная случайная выборка - это просто набор значений случайной величины (скалярных или векторных). Однако при оценке информации, которую несет эта выборка, мы должны рассматривать ее как одну из бесконечного числа потенциально возможных выборок объема n , т. е. как векторную n -мерную случайную величину

$$\mathbf{x} = \{x_1, x_2, \dots, x_n\}$$

(очевидно, что в случае векторной случайной величины $\vec{\xi}$ компоненты \mathbf{x} сами будут векторными случайными величинами, однако во избежание чрезмерного усложнения обозначений мы пока ограничимся одномерным случаем). Из условий получения выборки следует, что случайные величины x_1, x_2, \dots, x_n имеют одинаковые функции распределения $F(x)$, совпадающие с функцией распределения исходной случайной величины ξ . Кроме того, случайные величины x_1, x_2, \dots, x_n по определению случайной выборки независимы, поэтому их совместная функция распределения равна произведению одномерных функций распределения

$$F(x_1, x_2, \dots, x_n) = F(x_1) \times F(x_2) \times \dots \times F(x_n)$$

Как правило, информация, содержащаяся в выборке, интересует нас не столько сама по себе, сколько как информация обо всей генеральной совокупности. Однако чтобы отвечать этой цели, выборка должна быть правильно организованной и представительной. Существует специальный раздел математической статистики - *планирование выборочных обследований*. Мы будем рассматривать только один способ получения выборки - *простой случайный выбор*. В принципе схема его проста: из тщательно перемешанной генеральной совокупности извлекается наудачу n значений. На практике, однако, дело обстоит сложнее. Предположим, что

мы отловили n взрослых животных определенного вида и измерили их массу. Какую генеральную совокупность представляет эта выборка? Всех животных данного вида, т. е. живущих в разных местах, прошлых и будущих? Или популяцию, обитающую в данной местности? Или популяцию, обитающую в данной местности в данный год? Конечно, чем уже мы будем понимать генеральную совокупность, описываемую данной выборкой, тем ближе мы будем к истине, но, возможно, тем меньший интерес для нас она будет представлять.

1.2. Характеристики случайной выборки

Итак, мы имеем случайную выборку x_1, \dots, x_n значений случайной величины ξ с неизвестным распределением $F(x)$. Как нам разумно распорядиться этими значениями, чтобы получить представление о распределении $F(x)$, т.е. о генеральной совокупности, из которой извлечена эта выборка?

Можно использовать следующий эвристический принцип - будем считать, что исследуемая нами генеральная совокупность близка к гипотетической генеральной совокупности, состоящей только из значений x_1, \dots, x_n , содержащихся в ней в равной пропорции, т.е. случайная величина ξ близка к случайной величине ξ^* , принимающей n значений x_1, \dots, x_n с вероятностями $1/n$ (это, действительно, максимум информации о значениях случайной величины и их вероятностях, которую можно извлечь из выборки). Распределение случайной величины ξ^* называется эмпирическим распределением случайной величины ξ , а ее функция распределения $\hat{F}(x)$ - **эмпирической функцией распределения**. Очевидно, что каждой выборке соответствует своя эмпирическая функция распределения, т.е. можно сказать, что $\hat{F}(x)$ - случайная функция. $\hat{F}(x)$ представляет собой ступенчатую функцию, возрастающую от 0 до 1 со скачками высотой $1/n$ в точках x_1, \dots, x_n (очевидно, если некоторое значение повторяется k раз, то ему будет соответствовать один скачок величиной k/n). Можно определить эмпирическую функцию формулой $\hat{F}(x) = n_x / n$, где n_x - число значений выборки, не превосходящих x .

Пример. Пусть случайная величина ξ - это длина лепестка случайно выбранного цветка ириса разноцветного (*Iris versicolor*). Следующий ряд чисел содержит значения длин (в см) пятидесяти случайно выбранных лепестков:

Таблица 1

4.6	4.5	4.8	4.6	5.1	3.9	4.5	4.7	4.7	4.5
4	3.8	4.5	4	4.4	4.4	4.5	4.1	4.5	3.5
4.2	3.3	4.2	4.2	4.6	3.9	4.5	3.5	3.7	3.9
4.3	4.2	4	4.7	4.4	4.1	4.9	4.7	4.3	3
4.1	4.7	3.6	4.9	4	4	4.4	4.8	5	3.3

Упорядочим эти значения по величине, т.е. представим их в виде так называемого *вариационного ряда*:

Таблица 2

3	3.3	3.3	3.5	3.5	3.6	3.7	3.8	3.9	3.9
3.9	4	4	4	4	4	4.1	4.1	4.1	4.2
4.2	4.2	4.2	4.3	4.3	4.4	4.4	4.4	4.4	4.5
4.5	4.5	4.5	4.5	4.5	4.5	4.6	4.6	4.6	4.7
4.7	4.7	4.7	4.7	4.8	4.8	4.9	4.9	5	5.1

На рис. 1 представлено эмпирическое распределение случайной величины ξ , для этой выборки, т.е. распределение случайной величины ξ , а на рис. 2 - соответствующая эмпирическая функция распределения $\hat{F}(x)$.

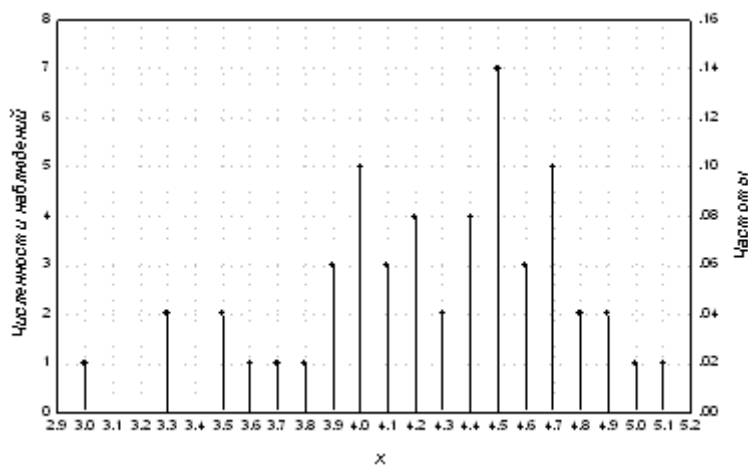


Рис.1. Пример эмпирического распределения.

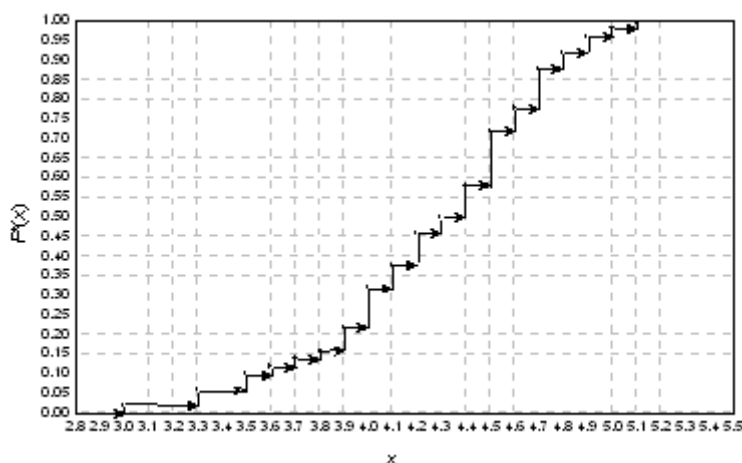


Рис.2. Пример эмпирической функции распределения.

Поскольку эмпирическая функция распределения $\hat{F}(x)$ является оценкой для $F(x)$ (можно доказать, что при $n \rightarrow \infty$ вероятность того, что максимальное расхождение между $\hat{F}(x)$ и $F(x)$ не превзойдет заданного малого числа ε , стремится к единице), можно взять характеристики $\hat{F}(x)$ в качестве оценок характеристик генерального распределения.

Ниже мы приводим полученные таким образом формулы для некоторых выборочных характеристик.

Название характеристики	Формула
Выборочный момент порядка k	$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$
Выборочный центральный момент Порядка k	$\hat{\mu}_{k,0} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$
Выборочное среднее - первый нецентральный момент	$\bar{x} = \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$
Выборочная дисперсия - (см. в главе 2 обоснование деления на $n-1$ вместо деления на n)	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Выборочный коэффициент асимметрии	$\hat{A}_1 = \frac{\hat{A}_{3p}}{\hat{A}_{2p}^3}$
Выборочный коэффициент эксцесса	$\hat{A}_2 = \frac{\hat{A}_{4p}}{\hat{A}_{2p}^2}$

Зная эмпирическую функцию распределения, можно найти эмпирические квантили, квартили и интерквартильную широту точно так же, как в случае обычной (теоретической) функции распределения. А именно, выборочная квантиль \hat{C}_p порядка p получается как абсцисса точки пересечения горизонтальной прямой, пересекающей ось ординат в точке p , с эмпирической функцией распределения $\hat{F}(x)$ (если пересечение не точка, а отрезок, то в качестве квантили можно взять абсциссу середины этого отрезка). Выборочные квантили $\hat{C}_{0.25}, \hat{C}_{0.50}, \hat{C}_{0.75}$ дают соответственно выборочную нижнюю квартиль \hat{Q}_1 , выборочную медиану \hat{Q}_2 и выборочную верхнюю квартиль \hat{Q}_3 , а разность $\hat{C}_{0.75} - \hat{C}_{0.25}$ - выборочную интерквартильную широту. Еще одна характеристика разброса значений случайной величины - размах R определяется как разность между максимальным и минимальным значением в выборке.

Если наблюдается не одномерная, а двумерная случайная величина (ξ, η) , т. е. выборка состоит из n пар значений $\{x_1, y_1\}, \{x_2, y_2\}, \dots, \{x_n, y_n\}$, то можно вычислить выборочный коэффициент ковариации для ξ и η

$$\hat{cov}(\xi, \eta) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{и} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Выборочный коэффициент корреляции определяется формулой

$$r_{\xi\eta} = \frac{\hat{cov}(\xi, \eta)}{s_{\xi} s_{\eta}}$$

где

$$s_{\xi}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{и} \quad s_{\eta}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

Выборочную моду x_{mod} целесообразно оценивать для дискретного и непрерывного генерального распределения различными способами. В дискретном случае оценкой моды будет значение, встретившееся в выборке наибольшее число раз.

Ниже даны значения характеристик, вычисленные для приведенного выше примера данных о длинах лепестков ириса.

Название характеристики	Значение характеристики
Выборочное среднее	$\bar{x} = 4.26$
Выборочная дисперсия	$s^2 = 0.2164$
Выборочный коэффициент асимметрии	$\hat{A}_1 = -0.61$
Выборочный коэффициент эксцесса	$\hat{A}_2 = 0.05$
Минимум	$x_{\min} = 3.0$
Максимум	$x_{\max} = 5.1$
Размах	$R = 2.1$
Нижняя квартиль	$\hat{Q}_1 = 4.00$
Верхняя квартиль	$\hat{Q}_3 = 4.60$
Интерквартильный размах	$\hat{Q}_3 - \hat{Q}_1 = 0.60$
Медиана	$\hat{Q}_2 = 4.35$
Мода	$x_{\text{mod}} = 4.50$

В непрерывном случае повторяющиеся значения редки или вообще отсутствуют, поэтому следует разбить диапазон изменения наблюдаемых значений точками a_0, a_1, \dots, a_k на k равных промежутков $[a_0, a_1], (a_1, a_2], \dots, (a_{k-1}, a_k]$ длиной $h = (a_k - a_0)/k$ и в качестве выборочной моды взять середину интервала, в который попало наибольшее число значений.

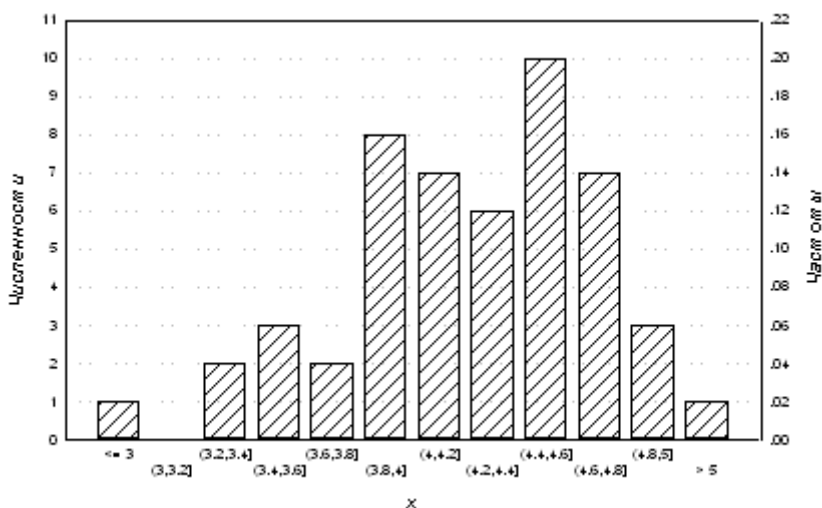


Рис.3. Пример гистограммы.

Результаты разбиения диапазона изменения значений выборки на интервалы и последующего подсчета числа значений n_i , попавших в интервалы $i=1, 2, \dots, k$, можно представить графически. Построив над каждым интервалом i столбик высотой n_i , мы получим так называемую **гистограмму**. Если при построении гистограммы оперировать не числом значений, попавших в интервалы, а их относительной частотой в выборке $\nu_i = n_i / n$ и откладывать столбики высотой ν_i , то полученная гистограмма будет выборочным аналогом плотности непрерывного распределения. В частности, сумма площадей всех столбиков гистограммы будет равна единице, как и для плотности распределения. Очевидно, что поскольку частоты ν_i пропорциональны численностям n_i , гистограммы для численностей и частот различаются только масштабом шкалы по оси ординат. На рис. 3 приведен пример гистограммы для рассмотренной выше выборки, содержащей 50 длин лепестков ириса (заметим, что рис. 3 можно рассматривать как результат преобразования рис. 1 путем группировки значений по оси ординат).

При построении гистограммы возникает трудно формализуемая проблема выбора оптимальных длин интервалов разбиения h . Обычно число интервалов разбиения выбирают из расчета, чтобы в каждый интервал попало в среднем не менее десяти наблюдений (при очень малых объемах выборки это число уменьшают). Таким образом, при увеличении объема выборки n можно уменьшать длины интервалов разбиения и более детально характеризовать выборку, а следовательно, и порождающее ее генеральное распределение. По сравнению с эмпирической функцией распределения гистограмма более наглядна, однако при ее построении привносится элемент субъективизма.

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

2.1. Логика статистического оценивания

Допустим, что у нас имеется случайная выборка x_1, x_2, \dots, x_n значений некоторой случайной величины ξ . Распределение этой случайной величины может быть либо полностью неизвестным, либо частично известным, например, может быть известна функциональная форма его распределения (вид функции распределения или плотности в непрерывном случае и формулы, определяющей вероятности отдельных значений, - в дискретном). Если распределение неизвестно, то нас могут интересовать различные его характеристики - математическое ожидание, мода, медиана, дисперсия, интерквартильная широта, моменты, асимметрия, эксцесс и т.д. Если вид распределения известен, а неизвестны лишь значения определяющих его параметров, нас могут интересовать также (или даже прежде всего) значения этих параметров, например, параметров μ и σ^2 в случае нормального распределения, параметра λ - в случае пуассоновского, параметра p - в случае биномиального. Задача оценивания неизвестного параметра или характеристики θ распределения случайной величины ξ состоит в вычислении на основе значений выборки x_1, x_2, \dots, x_n величины $\hat{\theta}(x_1, x_2, \dots, x_n)$, в каком-то смысле близкой к оцениваемому параметру или характеристике и называемой *точечной оценкой* θ .

Фактически мы решали задачу оценивания в предыдущем разделе, поскольку рассматривали выборочные математическое ожидание, дисперсию и т.д. как оценки неизвестных генеральных характеристик. Однако этот подход к оцениванию был скорее интуитивным, и сейчас мы попробуем его формализовать. Эта формализация касается уточнения смысла близости *выборочной оценки* $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ к оцениваемому параметру θ . Заметим, прежде всего, что $\hat{\theta}$ является случайной величиной, значения которой меняются от выборки к выборке. Для некоторых выборок значение $\hat{\theta}$ может оказаться очень близким к оцениваемому параметру θ , для других - довольно далеким. Можно, однако, потребовать, что хотя бы в среднем оценки для разных выборок группировались вокруг оцениваемого параметра, т.е. чтобы выполнялось условие $M\hat{\theta} = \theta$. Оценки, удовлетворяющие этому условию называются *несмещенными*. Отметим, что в зависимости от контекста мы рассматриваем $\hat{\theta}(x_1, x_2, \dots, x_n)$ либо как случайную величину (и только в этом случае мы имеем право говорить о математическом ожидании $\hat{\theta}$ - как это было в приведенном определении несмещенности), либо как алгебраическую формулу для вычисления значения оценки по выборке, либо просто как конкретное значение, полученное для конкретной выборки.

Можно доказать, пользуясь свойствами математического ожидания и определением выборки, что выборочное математическое ожидание (чаще используется термин **выборочное среднее**) \bar{x} является несмещенной оценкой генерального среднего $M\xi$. Действительно, имеем

$$M\bar{x} = M\left\{\frac{1}{n}\sum_{i=1}^n x_i\right\} = \frac{1}{n}\sum_{i=1}^n M x_i = \frac{1}{n}\sum_{i=1}^n M\xi = M\xi$$

(мы воспользовались также тем очевидным фактом, что распределение любой компоненты выборки x_i , совпадает с распределением анализируемой случайной величины ξ).

Однако, вопреки нашей интуиции, математическое ожидание второго центрального момента

$$\hat{A}_{2p} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$$

не равно генеральной дисперсии. Для получения несмещенной оценки дисперсии надо разделить сумму квадратов на $n-1$, а не на n . Действительно, пользуясь свойствами математического ожидания, в частности тем, что математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий, получаем

$$\begin{aligned} M\{s^2\} &= M\left\{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2\right\} = M\left\{\frac{1}{n-1}\sum_{i=1}^n [(x_i - M\xi) + (M\xi - \bar{x})]^2\right\} = \\ &= M\left\{\frac{1}{n-1}\left[\sum_{i=1}^n (x_i - M\xi)^2 + 2(M\xi - \bar{x})\sum_{i=1}^n (x_i - M\xi) + n(M\xi - \bar{x})^2\right]\right\} = \\ &= M\left\{\frac{1}{n-1}\left[\sum_{i=1}^n (x_i - M\xi)^2 - 2n(M\xi - \bar{x})^2 + n(M\xi - \bar{x})^2\right]\right\} = \\ &= M\left\{\frac{1}{n-1}\left[\sum_{i=1}^n (x_i - M\xi)^2 - n(\bar{x} - M\xi)^2\right]\right\} = \\ &= \frac{1}{n-1}\left[\sum_{i=1}^n M\{(x_i - M\xi)^2\} - nM\{(\bar{x} - M\xi)^2\}\right] = \frac{1}{n-1}\left[nD\xi - \frac{n}{n}D\xi\right] = D\xi \end{aligned}$$

Таким образом, несмещенной оценкой для $D\xi$ будет s^2 - сумма квадратов отклонений от среднего, деленная на $n-1$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$$

Следует, однако, отметить, что хотя выборочный центральный момент второго порядка не является несмещенной оценкой для дисперсии $D\xi$, его смещение (отклонение математического ожидания от оцениваемого

параметра $D\xi$), равное $\frac{D\xi}{n}$, стремится к нулю при $n \rightarrow \infty$. Оценки, удовлетворяющие этому свойству, называются **асимптотически несмещенными**, и часто, когда не удается найти несмещенной оценки для

оцениваемого параметра, довольствуются асимптотически несмещенными оценками. Конечно, нас не очень интересует, что происходит со смещением при $n \rightarrow \infty$, когда мы имеем дело с конкретной выборкой фиксированного объема n . Скорее, мы предпочитаем асимптотически несмещенные оценки из-за того, что, как правило, их смещение относительно невелико при не очень малых объемах выборки. Например, смещение $\hat{\mu}_{2.0}$ при $n=50$ равно 2%.

Конечно, коль скоро s^2 есть несмещенная оценка для $D\xi$, то ей и следует пользоваться. Однако нас чаще интересует не столько оценка дисперсии $D\xi$ случайной величины x , сколько оценка ее среднеквадратичного отклонения $\sqrt{D\xi}$. Естественно взять в качестве оценки среднеквадратичного отклонения квадратный корень несмещенной оценки, т.е. s , однако следует помнить, что s не будет несмещенной оценкой для $\sqrt{D\xi}$, а будет лишь асимптотически несмещенной.

Итак, на примере с получением несмещенной оценки дисперсии мы убедились, что "естественный" способ получения оценок, состоящий в использовании характеристик выборочного распределения в качестве оценок характеристик генерального распределения не всегда приводит к наилучшим результатам. Существуют ряд регулярных приемов получения оценок (метод моментов, метод максимального правдоподобия, метод минимума хи-квадрата), но они не всегда приводят к наилучшим оценкам (например, с точки зрения несмещенности). Поэтому общий подход к задаче оценивания состоит в том, что в качестве претендента на оценку характеристики или параметра (для краткости мы иногда будем использовать только термин "параметр") θ генерального распределения по случайной выборке x_1, x_2, \dots, x_n в принципе может претендовать любая функция $\hat{\theta}(x_1, x_2, \dots, x_n)$ от компонент выборки (часто вместо словосочетания "функция от выборки" используют более краткий термин "*статистика*"), среди которых должна быть выбрана наилучшая.

У нас уже есть один критерий для сравнения оценок - это наличие или отсутствие несмещенности (хотя бы асимптотической). Однако этого недостаточно. Действительно, было показано, что выборочное среднее \bar{x} - несмещенная оценка для математического ожидания $M\xi$. Но можно предложить другие несмещенные оценки для математического ожидания. Например, полусумма первого и последнего значений выборки $(x_1+x_n)/2$ как легко показать, будет также несмещенной оценкой для $M\xi$. Конечно, интуитивно мы чувствуем, что \bar{x} - более предпочтительная оценка, поскольку она более полно учитывает информацию, содержащуюся в выборке. Однако необходим какой-то критерий, позволяющий чисто формально показать, что \bar{x} лучше оценок типа $(x_1+x_n)/2$. Таким критерием, может служить состоятельность оценки. Оценка $\hat{\theta}(x_1, x_2, \dots, x_n)$ называется

состоятельной, если при $n \rightarrow \infty$ она сходится по вероятности к оцениваемому параметру θ , т.е. если для любого положительного ε выполняется условие

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

Достаточным условием состоятельности оценки $\hat{\theta}$ является ее несмещенность и стремление дисперсии оценки $\hat{\theta}$ к нулю при увеличении объема выборки, т.е. оценка будет состоятельной, если $M\hat{\theta} = \theta$ и $\lim_{n \rightarrow \infty} D\hat{\theta} = 0$. Справедливость этого утверждения непосредственно следует из неравенства Чебышева, которое в данном случае имеет следующий вид

$$P(|\hat{\theta} - M\hat{\theta}| > \varepsilon) < \frac{D\hat{\theta}}{\varepsilon^2}$$

Пользуясь этим достаточным условием, покажем, что выборочное среднее \bar{x} является состоятельной оценкой математического ожидания. Поскольку несмещенность \bar{x} уже была доказана, осталось показать, что $D\bar{x} \rightarrow 0$ при $n \rightarrow \infty$. Действительно, имеем

$$D\bar{x} = D\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n D x_i = \frac{1}{n^2} n D \xi = \frac{1}{n} D \xi \xrightarrow{n \rightarrow \infty} 0$$

Таким образом \bar{x} - состоятельная оценка для математического ожидания $M\xi$ случайной величины ξ . Однако легко видеть, что полусумма $(x_1 + x_n)/2$ первого и последнего значений выборки, будучи несмещенной, не является состоятельной, поскольку ее дисперсия не стремится к 0 при неограниченном увеличении n

$$D\left(\frac{x_1 + x_n}{2}\right) = \frac{1}{4} (D x_1 + D x_n) = \frac{1}{2} D \xi$$

Аналогично можно доказать, что несмещенная оценка дисперсии s^2 является состоятельной оценкой для генеральной дисперсии $D\xi$.

Еще одним желательным свойством оценки является ее *эффективность*. Несмещенная оценка $\hat{\theta}(x_1, x_2, \dots, x_n)$ параметра θ называется *эффективной* оценкой, если $D\hat{\theta} < D\theta^*$ для любой другой несмещенной оценки $\theta^*(x_1, x_2, \dots, x_n)$. Дело в том, что две оценки, будучи обе несмещенными и состоятельными, могут различаться своими дисперсиями. Например, следующая оценка для математического ожидания

$$\bar{x} = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i$$

будет, как легко проверить, несмещенной и состоятельной. Однако она не будет эффективной, т.к. ее дисперсия $\frac{D\xi}{n-1}$ больше дисперсии \bar{X} , которая равна $\frac{D\xi}{n}$.

2.2. Доверительные интервалы

Оценки, которые рассматривались в предыдущем параграфе, принято называть *точечными*, поскольку за оценку неизвестного параметра θ принимается конкретное значение (точка), вычисляемое по выборке (например, значение выборочного среднего в качестве оценки для математического ожидания). Однако часто нас интересует не только конкретное значение, но и такие свойства оценки, которые ассоциируются с ее точностью и надежностью. Этим требованиям отвечают так называемые интервальные оценки. *Интервальная оценка* - это некоторый интервал $(\theta_{\text{нижн}}, \theta_{\text{верхн}})$, где $\theta_{\text{нижн}}$ и $\theta_{\text{верхн}}$ есть функции от выборочных значений и γ : $\theta_{\text{нижн}} = \theta_{\text{нижн}}(x_1, x_2, \dots, x_n, \gamma)$, $\theta_{\text{верхн}} = \theta_{\text{верхн}}(x_1, x_2, \dots, x_n, \gamma)$, называемый *доверительным*, который с заданной (достаточно высокой) вероятностью γ , называемой доверительной, содержит истинное значение θ оцениваемого неизвестного параметра (случайные величины $\theta_{\text{нижн}}$ и $\theta_{\text{верхн}}$ называются, соответственно, верхним и нижним доверительными пределами), т.е.

$$P(\theta_{\text{нижн}} < \theta < \theta_{\text{верхн}}) = \gamma$$

(дополнение γ до 1 будем обозначать ε , т.е. $\gamma = 1 - \varepsilon$). Наиболее часто используются значения доверительной вероятности γ равные 0.95 или 0.99 (95%-ный и 99%-ный доверительные интервалы).

2.2.1. Доверительный интервал для математического ожидания нормально распределенной случайной величины с известной дисперсией

Проще всего понять логику интервального оценивания на примере построения доверительного интервала для математического ожидания нормально распределенной случайной величины с известной дисперсией. Пусть ξ - нормально распределенная случайная величина с неизвестным математическим ожиданием $M\xi = \mu$ и дисперсией $D\xi = \sigma^2$, т.е. в наших обозначениях $\xi \sim N(\mu, \sigma^2)$, и имеется выборка значений этой случайной величины x_1, x_2, \dots, x_n объема n . Требуется найти доверительный интервал для μ с доверительной вероятностью γ .

Выше было показано, что выборочное среднее (для любого распределения, в том числе и нормального) имеет математическое ожидание, равное математическому ожиданию исходной случайной величины, т.е. μ , а дисперсия - дисперсии исходной случайной величины, деленной на n , т.е. σ^2/n . Следовательно, статистика

$$u = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad (1)$$

полученная путем стандартизации выборочного среднего \bar{x} , будет иметь нулевое математическое ожидание и единичную дисперсию. Поскольку, как мы знаем, линейные комбинации нормально распределенных случайных величин имеют также нормальное распределение, а случайная величина u фактически является линейной комбинацией нормально распределенных случайных величин x_1, x_2, \dots, x_n , то u будет стандартно распределенной случайной величиной, т.е. $u \sim N(0,1)$. Стандартное нормальное распределение - это конкретное, полностью заданное распределение, квантили которого можно найти в соответствующих таблицах (или вычислить путем численного интегрирования). В частности, можно найти симметричные относительно центра распределения границы, внутрь которых u попадает с заданной вероятностью γ

$$P(u_{\varepsilon/2} < u < u_{1-\varepsilon/2}) = \gamma$$

или, с учетом симметрии,

$$P(-u_{1-\varepsilon/2} < u < u_{1-\varepsilon/2}) = \gamma \quad (2)$$

(через $u_{\varepsilon/2}$ и $u_{1-\varepsilon/2}$ обозначены квантили стандартного нормального распределения порядка $\varepsilon/2$ и $1-\varepsilon/2$). В частности, справедливы следующие неравенства

$$P(-1.96 < u < 1.96) = 0.95 \quad \text{и} \quad P(-2.58 < u < 2.58) = 0.99$$

Подставляя в (2) выражение для u из (1), получаем

$$P\left(-u_{1-\varepsilon/2} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < u_{1-\varepsilon/2}\right) = \gamma \quad (3)$$

или, после преобразований,

$$P\left(\bar{x} - u_{1-\varepsilon/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{1-\varepsilon/2} \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

Это означает, что интервал $\left(\bar{x} - u_{1-\varepsilon/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + u_{1-\varepsilon/2} \frac{\sigma}{\sqrt{n}}\right)$ будет $100\gamma\%$ -ным доверительным интервалом для неизвестного математического ожидания μ нормального распределения с известной дисперсией σ^2 . В частности, 95%-ным доверительным интервалом будет интервал

$(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$, а 99%-ным - $(\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}})$. Мы видим, что доверительный интервал уменьшается при уменьшении σ , увеличении объема выборки и снижении доверительной вероятности.

2.2.2. Доверительный интервал для математического ожидания нормально распределенной случайной величины с неизвестной дисперсией

В случае неизвестной дисперсии постановка задачи и ход рассуждений при построении доверительного интервала аналогичны случаю известной дисперсии, рассмотренному в предыдущем параграфе. Разница состоит в том, что в выражении (1) неизвестное среднеквадратичное отклонение σ заменяется на его выборочную оценку s

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad (4)$$

Полученная таким путем статистика t , будучи довольно сложной функцией от нормально распределенных случайных величин x_1, x_2, \dots, x_n , уже не будет нормально распределенной. Можно доказать, что t имеет t -распределение Стьюдента с $n-1$ степенями свободы. Отсюда следует, что справедливо равенство

$$P(-t_{n-1, 1-\alpha/2} < \frac{\bar{x} - \mu}{s / \sqrt{n}} < t_{n-1, 1-\alpha/2}) = \gamma$$

аналогичное уравнению (3) и отличающееся от него заменой σ на s и квантилей нормального распределения на соответствующие квантили t -распределения с $n-1$ степенями свободы. Соответственно 100 γ %-ный доверительный интервал для неизвестного математического ожидания μ нормального распределения с неизвестной дисперсией σ^2 будет иметь следующий вид

$$(\bar{x} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}), \quad \text{или} \quad (\bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}})$$

Известно, что этот доверительный интервал и доверительный интервал из предыдущего раздела являются **робастными**, т.е. они нечувствительны к умеренным отклонениям от предположения о нормальности распределения. Во всяком случае, как отмечается в пособиях по математической статистике, при объеме выборки не менее 15 становится целесообразно использовать приведенные доверительные интервалы для математического ожидания и в случае умеренного отклонения от предположения о нормальности.

Заметим, что при $n \rightarrow \infty$ t -распределение приближается к нормальному распределению, а его квантили - к квантилям нормального распределения. Например, при $n-1=60$ квантиль $t_{60, 0.975}$ равна 2.00, что не

очень сильно отличается от аналогичного значения $t_{0,975} = 1.96$ для нормального распределения (особенно на фоне выборочных флуктуаций \bar{x} и s). Поэтому при числе наблюдений порядка нескольких десятков можно пользоваться нормальным приближением для t -распределения. Однако при небольшом числе степеней свободы различие между квантилями t -распределения и нормального распределения довольно значительно. Например, для $n-1=1$ имеем $t_{1,0,975} = 12.71$, для $n-1=2$ - $t_{2,0,975} = 4.30$, для $n-1=5$ - $t_{5,0,975} = 2.57$. При $n-1=9$ (выборка из 10 наблюдений) получаем значение $t_{9,0,975} = 2.26$, что уже не очень сильно отличается от 1.96.

Возвращаясь к примеру с длинами лепестков ириса и учитывая, что $\bar{x} = 4.26$, $s = 0.47$, $t_{49,0,975} = 2.01$ (при числе степеней свободы $n-1=49$), а также предполагая, что распределение длин лепестков нормально (в следующем разделе мы рассмотрим процедуру проверки этого предположения), получаем, что 95%-ным доверительным интервалом для математического ожидания длины лепестка будет интервал (4.13, 4.39). Т.е. мы можем утверждать, что с вероятностью 0.95 неизвестное μ находится между 4.13 и 4.39 (точнее следует сказать, что найденный доверительный интервал с вероятностью 0.95 накроет неизвестное значение μ).

2.2.3. Доверительный интервал для неизвестной дисперсии нормально распределенной случайной величины (при неизвестном математическом ожидании)

Для нахождения доверительного интервала для неизвестной дисперсии нормально распределенной случайной величины рассмотрим статистику

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Можно показать, что эта статистика имеет χ^2 -распределение с $n-1$ степенями свободы. Следовательно, справедливо равенство

$$P\left(\chi_{n-1, \varepsilon/2}^2 < \frac{(n-1)s^2}{\sigma^2} < \chi_{n-1, 1-\varepsilon/2}^2\right) = \gamma$$

которое можно переписать в виде

$$P\left(\frac{(n-1)s^2}{\chi_{n-1, 1-\varepsilon/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, \varepsilon/2}^2}\right) = \gamma$$

Таким образом, 100 γ %-ный доверительный интервал для неизвестной дисперсии σ^2 нормального распределения с неизвестным математическим ожиданием μ будет иметь следующий вид

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1, 1-\varepsilon/2}}, \frac{(n-1)s^2}{\chi^2_{n-1, \varepsilon/2}} \right)$$

где $\chi^2_{n-1, \varepsilon/2}$ и $\chi^2_{n-1, 1-\varepsilon/2}$ - квантили распределения χ^2 с $n-1$ степенями свободы. В частности, для длины лепестков ириса, учитывая, что $s^2 = 0.22$, $n-1=49$, $\chi^2_{49, \varepsilon/2} = 70.24$ и $\chi^2_{n-1, 1-\varepsilon/2} = 31.55$, получаем, что 95%-ным доверительным интервалом для дисперсии (в предположении нормальности распределения) будет интервал (0.15, 0.34).

Заметим, что полученный доверительный интервал для дисперсии, в отличие от доверительного интервала для математического ожидания, чувствителен к отклонениям от исходного предположения о нормальности распределения.

2.2.4. Доверительный интервал для неизвестного параметра p биномиального распределения

Пусть произведено n независимых испытаний, в которых некоторое событие A произошло k раз. Требуется найти точечную и интервальную оценку неизвестной вероятности p появления этого события.

Эту задачу можно рассматривать в двух эквивалентных формулировках. В первой формулировке считается, что получено n наблюдений случайной величины ξ_1 , принимающей с вероятностью p значение 1 в случае появления события A и с вероятностью $1-p$ значение 0 - в случае непоявления события A . Во второй формулировке считается, что имеется только одно наблюдение случайной величины ξ - числа появлений события A в одном сложном испытании.

Поскольку математическое ожидание ξ_1 , как мы знаем, равно p , то получение оценки для p равносильно получению оценки для математического ожидания ξ_1 . Несмещенной, состоятельной и эффективной оценкой для математического ожидания является выборочное среднее, которое в данном случае совпадает с частотой $\nu = k/n$ появления события A в выборке. Таким образом, выборочная частота является несмещенной, состоятельной и эффективной оценкой для неизвестной вероятности. Во второй формулировке математическое ожидание ξ равно np , а выборочное среднее для одного наблюдения равно самому наблюдению, т.е. k . И поскольку оценкой для np служит k (среднее по одному наблюдению), то оценкой для p снова будет k/n .

Построение доверительных интервалов несколько проще обсуждать в терминах биномиального распределения. Можно построить как точные доверительные интервалы для p , так и приближенные. Математическая

техника нахождения точных доверительных интервалов довольно громоздка и мы приведем здесь лишь окончательные формулы для доверительных пределов. Напротив, выражения для приближенных доверительных пределов легко получаются на основе применения центральной предельной теоремы, однако они применимы лишь при достаточно большом n - ориентировочно при $n > \frac{25}{v(v-1)}$.

Точный доверительный интервал

Точный доверительный интервал для параметра биномиального распределения имеет вид

$$\left(\frac{k}{k + (n-k+1) F_{2(n-k+1), 2k, 1-\alpha/2}^2}, \frac{(k+1) F_{2(k+1), 2(n-k), 1-\alpha/2}^2}{n-k + (k+1) F_{2(k+1), 2(n-k), 1-\alpha/2}^2} \right)$$

где n - число испытаний, k - число появлений события A , а $F_{f, g, \alpha}^2$ обозначает квантиль порядка α распределения F с f, g степенями свободы.

Рассмотрим в качестве примера приведенные в п. 1.2 данные о длине лепестков ириса с точки зрения оценки вероятности появления в случайной выборке "длинных" лепестков, а именно, лепестков длиной 4.5 см и более. В данном случае $n=50$ и $k=21$. Следовательно, точечной оценкой для p будет значение $v = k/n = 0.42$, а 95%-ным доверительным интервалом - интервал

$$\left(\frac{21}{21 + 30 F_{60, 42, 0.975}^2}, \frac{22 F_{42, 58, 0.975}^2}{29 + 22 F_{42, 58, 0.975}^2} \right)$$

Учитывая, что $F_{60, 42, 0.975} = 1.78$ и $F_{42, 58, 0.975} = 1.74$, получаем окончательно интервал (0.28, 0.57).

Приближенный доверительный интервал

Поскольку число появлений события A в n испытаниях равно сумме чисел появлений этого события в отдельных испытаниях (0 или 1), то в соответствии с центральной предельной теоремой при больших n распределение биномиальной случайной величины будет близко к нормальному. Учитывая, что дисперсия биномиальной случайной величины равна $np(1-p)$, получаем для ее математического ожидания np

приближенные доверительные пределы $k \pm u_{1-\alpha/2} \sqrt{np(1-p)}$, а для

параметра p - соответственно $v \pm u_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$. Заменяя параметр p его выборочной оценкой v , получаем окончательно для приближенного доверительного интервала параметра p следующее выражение

$$v \pm u_{1-\alpha/2} \sqrt{\frac{v(1-v)}{n}}$$

В частности, для предыдущего примера, подставляя $\nu=0.42$, $z_{0.975}=1.96$ и $n=50$, получаем приближенный доверительный интервал (0.28, 0.56), не слишком отличающийся от точного доверительного интервала.

2.2.5. Доверительный интервал для неизвестного параметра λ пуассоновского распределения

Пусть число появлений некоторого события A является случайной величиной, имеющей пуассоновское распределение, и пусть в результате наблюдения событие A произошло k раз. Требуется найти точечную и интервальную оценку неизвестного параметра λ пуассоновского распределения. Поскольку λ является математическим ожиданием пуассоновской случайной величины, то несмещенной, состоятельной и эффективной оценкой для λ будет выборочное среднее, которое для рассматриваемой ситуации единственного наблюдения совпадает с k . Что касается доверительных пределов для параметра λ , то, как и в случае биномиального распределения можно предложить точное и приближенное решения. Выражения для приближенных доверительных пределов также основаны на применении центральной предельной теоремы и применимы лишь при достаточно большом k - ориентировочно при $k > 25$.

Точный доверительный интервал

Точный доверительный интервал для параметра λ пуассоновского распределения имеет вид

$$\left(\frac{1}{2} \chi_{2k, \alpha/2}^2, \frac{1}{2} \chi_{2k+2, 1-\alpha/2}^2 \right)$$

где k - число появлений события A , а $\chi_{f, \alpha}^2$ обозначает квантиль порядка α распределения χ^2 с f степенями свободы.

Рассмотрим следующий пример. Пусть после фильтрования 1 мл воды на фильтровальной бумаге обнаружено 100 клеток фитопланктона. Предполагая, что число клеток в заданном объеме воды имеет пуассоновское распределение, найти 95%-ные доверительные пределы для параметра λ этого распределения. Подставляя $k=100$, $\chi_{200, 0.025}^2 = 162.7$ и $\chi_{202, 0.975}^2 = 243.4$ в приведенные выше выражения, получаем 95%-ный доверительный интервал: (81.3, 121.7).

Приближенный доверительный интервал

Поскольку число появлений события A при наблюдении пуассоновской случайной величины X можно аппроксимировать биномиальным распределением, а последнее - нормальным, то при не слишком малых k в соответствии с центральной предельной теоремой

распределение случайной величины ξ будет близко к нормальному. Учитывая, что дисперсия пуассоновской случайной величины равна λ , получаем для ее математического ожидания λ приближенные доверительные пределы $k \pm u_{1-\alpha/2} \sqrt{k}$. Заменяя параметр λ его выборочной оценкой k , получаем окончательно для приближенного доверительного интервала параметра λ следующее выражение

$$k \pm u_{1-\alpha/2} \sqrt{k}.$$

В частности, для предыдущего примера, подставляя $k=100$ и $u_{0.975} = 1.96$, получаем приближенный доверительный интервал (80.4, 119.6), не слишком отличающийся от точного доверительного интервала.

2.2.6. Приближенный доверительный интервал для неизвестного коэффициента корреляции двумерного нормального распределения

Рассмотрим теперь вопрос построения доверительного интервала для коэффициента корреляции. Пусть $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - случайная выборка объема n из двумерного нормального распределения. Пусть ρ - коэффициент корреляции случайных величин ξ и η , а r - выборочный коэффициент корреляции. Распределение самого коэффициента корреляции r , особенно при значениях ρ близких к 0 или 1 может сильно отличаться от нормального. Однако следующая функция от r , называемая **преобразованием Фишера**, довольно хорошо аппроксимируется нормальным распределением

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

со средним $Mz = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ и дисперсией $Dz = \frac{1}{n-3}$. Соответственно, стандартизованная случайная величина будет иметь стандартное нормальное распределение

$$u = \sqrt{n-3} \left(z - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \right) \sim N(0,1)$$

и с вероятностью γ будет заключена в пределах $\pm u_{1-\alpha/2}$, т.е.

$$P(-u_{1-\alpha/2} < \sqrt{n-3} \left(z - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \right) < u_{1-\alpha/2}) = \gamma$$

Решая неравенство под знаком вероятности относительно неизвестного коэффициента корреляции ρ , получаем окончательно

$$P \left(\frac{e^{2(z - \sqrt{\frac{1}{n-3}} u_{1-\varepsilon/2})} - 1}{e^{2(z - \sqrt{\frac{1}{n-3}} u_{1-\varepsilon/2})} + 1} < \rho < \frac{e^{2(z + \sqrt{\frac{1}{n-3}} u_{1-\varepsilon/2})} - 1}{e^{2(z + \sqrt{\frac{1}{n-3}} u_{1-\varepsilon/2})} + 1} \right) = \gamma$$

Заметим, что полученный доверительный интервал для коэффициента корреляции чувствителен к отклонениям от исходного предположения о двумерной нормальности случайных величин ξ и η .

Пример. Пусть объем выборки $n=10$, а вычисленное по выборке значение $r=0.6$, тогда 95%-ным доверительным интервалом для неизвестного коэффициента корреляции будет интервал (0.05, 0.88).

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Логика проверки статистических гипотез

Пусть x_1, x_2, \dots, x_n - случайная выборка значений случайной величины ξ , имеющей некоторое полностью или частично неизвестное распределение $F(x)$. В предыдущем разделе рассматривались методы получения оценок параметров или характеристик этого неизвестного распределения. Однако часто нас интересуют не столько конкретные количественные оценки, сколько правильность или ошибочность некоторых утверждений, относящихся к распределению наблюдаемой случайной величины. Например, является ли это распределение нормальным или нет? Или, равно математическое ожидание заданному значению или нет? Если кроме выборки x_1, x_2, \dots, x_n имеется выборка y_1, y_2, \dots, y_n значений другой случайной величины η , то можно поставить вопрос о том, равны или нет математические ожидания случайных величин ξ и η ? Если имеется выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ двумерной случайной величины (ξ, η) , то может возникнуть вопрос о том, равен нулю или нет коэффициент корреляции между ξ и η ?

Решению задач проверки гипотез о генеральном распределении по выборке из этого распределения посвящен специальный раздел математической статистики - проверка статистических гипотез. Логика проверки гипотез в математической статистике (она напоминает логику доказательства от противного) состоит в следующем. Вначале предполагается, что проверяемая гипотеза (ее принято называть *нулевой гипотезой* и обозначать H_0) верна. В предположении, что H_0 верна, ищется распределение вероятностей некоторой функции $g(x_1, x_2, \dots, x_n)$ от значений выборки, называемой *статистикой критерия* (правило проверки гипотезы

принято называть **критерием**), и в области значений этой статистики выделяется некоторая область W , называемая **критической областью**, такая, что вероятность $P(g \in W)$ попадания выборочного значения статистики g в эту область не превосходит заданного малого значения α , называемого **уровнем значимости критерия** (обычно полагают α равным 0.05 или 0.01). Если для данной конкретной выборки g попадает в критическую область W , то гипотеза H_0 **отвергается** (говорят - "отвергается на уровне значимости α "), поскольку вероятность этого события при верной H_0 мала. Если же g не попадает в критическую область W , то говорят, что "гипотеза H_0 **не отвергается** на уровне значимости α " (или - "полученные данные не дают оснований отвергнуть гипотезу H_0 на уровне значимости α ").

Очевидно, однако, что можно разными способами задать статистику критерия $g(x_1, x_2, \dots, x_n)$, а для заданной статистики можно разными способами выбрать критическую область W , удовлетворяющую условию $P(g \in W) = \alpha$. Поэтому следует выбирать g и W в некотором смысле наилучшими из возможных, а именно такими, чтобы полученный критерий был **наиболее мощным**.

Для определения понятия **мощности критерия** введем понятие **альтернативной гипотезы** H_1 , т.е. гипотезы, которая выполняется, если не выполняется нулевая гипотеза H_0 . Тогда в терминах правильности или ошибочности принятия H_0 и H_1 можно указать четыре потенциально возможных результата применения критерия к выборке, представленные в табл. 3. Как мы видим мощность критерия - это вероятность принятия при применении данного критерия альтернативной гипотезы H_1 при условии, что она верна. Очевидно, что при фиксированной ошибке 1-го рода (ее мы задаем сами, и она не зависит от свойств критерия) критерий будет тем лучше, чем больше его мощность (т.е. чем меньше ошибка 2-го рода).

Таблица 3

		Принята гипотеза	
		H_0	H_1
H_0	H_0	$1 - \alpha = P(H_0 H_0)$ - вероятность правильно принять H_0 , когда верна H_0	$\alpha = P(H_1 H_0)$ - вероятность ошибочно принять H_1 , когда верна H_0 (ошибка 1-го рода, уровень значимости)
	H_1	$\beta = P(H_0 H_1)$ - вероятность ошибочно принять H_0 , когда верна H_1 (ошибка 2-го рода)	$1 - \beta = P(H_1 H_1)$ - вероятность правильно принять H_1 , когда верна H_1

Верна Гипотеза	H_1	$\beta = P(H_0 H_1)$ - вероятность ошибочно принять H_0 , когда верна H_1 (ошибка 2-го рода)	$1 - \beta = P(H_1 H_1)$ - вероятность правильно принять H_1 , когда верна H_1 (мощность критерия)
-------------------	-------	---	---

Проиллюстрируем основные понятия рассмотренной методологии на простом примере проверки гипотезы о равенстве математического ожидания нормально распределенной случайной величины с известной дисперсией заданному числу (с точки зрения практического применения этот пример несколько искусственен, поскольку дисперсия наблюдаемой случайной величины обычно неизвестна).

3.2. Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной случайной величины с известной дисперсией

Итак, пусть $\xi \sim N(M\xi, \sigma^2)$, $H_0: M\xi = \mu_0$ и $H_1: M\xi = \mu_1$, и пусть имеется выборка x_1, x_2, \dots, x_n значений случайной величины ξ объема n . Предположим, что H_0 верна и выберем в качестве статистики критерия стандартизованное выборочное среднее

$$u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

При верной H_0 статистика u имеет стандартное нормальное распределение, $u \sim N(0,1)$, представленное на рис. 4. На этом рисунке также указана критическая область уровня $\alpha = 0.05$, состоящая из двух бесконечных полуинтервалов $(-\infty, -1.96]$ и $[1.96, \infty)$, вероятность попадания в каждый из которых статистики u равна $\alpha/2 = 0.025$.

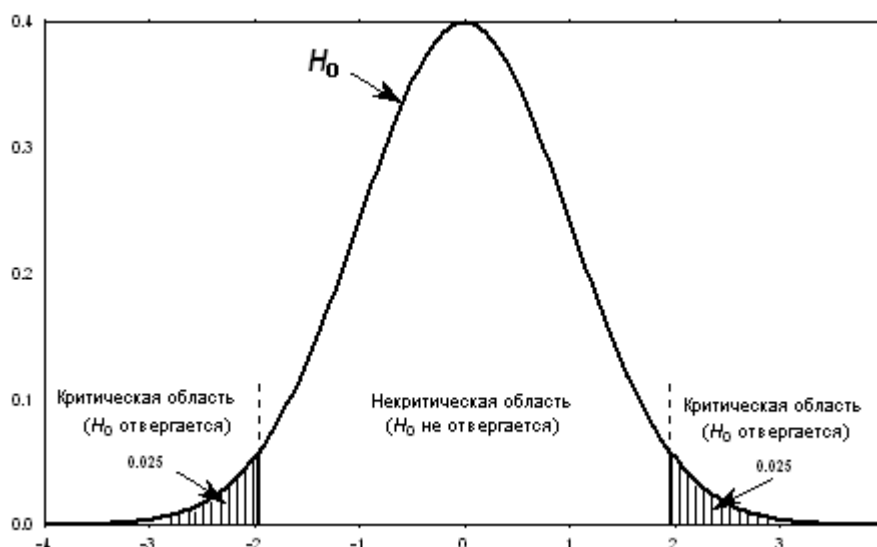


Рис. 4. Пример критической области.

Имеются формализованные подходы к выбору статистик критериев и построению критических областей, приводящие к наиболее мощным критериям, но мы их здесь не рассматриваем. Неформальное же правило состоит в том, чтобы выбирать в качестве статистики величину, характеризующую степень отклонения от нулевой гипотезы. Очевидно, разность $\bar{x} - \mu_0$ удовлетворяет этому условию, а деление на константу σ/\sqrt{n} , сохраняя это качество, приводит к величине u с полностью заданным распределением, что позволяет выбрать критическую область с требуемым уровнем значимости. Неформальное правило выбора критической области состоит в том, чтобы она включала значения статистики, соответствующие наибольшим отклонениям от нулевой гипотезы - на рис. 4 эта рекомендация соблюдена.

До сих пор мы говорили о свойствах критерия в предположении, что верна гипотеза H_0 . А что происходит, когда верна альтернативная гипотеза H_1 ? В этом случае распределение статистики критерия u изменится. Чтобы его найти, произведем преобразование

$$u = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu_0 - \mu_1 + \mu_1}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

из которого следует, что при гипотезе H_1 распределение статистики u

отличается от стандартного нормального сдвигом на величину $\Delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$, т.е. $u \sim N(\Delta, 1)$ при выполнении H_1 .

На рис. 5 взаимное расположение плотностей распределения статистики u при гипотезах H_0 и H_1 показано для случая $\mu_0 = 0$, $\mu_1 = 2$, $\sigma^2 = 1$, $\alpha = 0.05$ и $n=1$. Вероятности ошибки 2-го рода β соответствует площадь под кривой функции плотности при H_1 на промежутке от -1.96 до 1.96 , где не отвергается гипотеза H_0 , а следовательно, ошибочно не принимается гипотеза H_1 . В данном случае ошибка 2-го рода, $\beta = 0.48$, довольно велика. Это произошло, главным образом, потому, что мал объем выборки - имеется всего одно наблюдение, $n=1$. При увеличении n распределение, соответствующее альтернативной гипотезе H_1 , будет

сдвигаться вправо, поскольку величина $\Delta = \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}}$ будет увеличиваться, что приведет, как легко понять по рис. 5, к уменьшению ошибки β . Очевидно также, что большей величине разности $\mu_1 - \mu_0$ соответствует большая величина Δ , и следовательно меньшая ошибка 2-го рода. Ошибка 2-го рода уменьшается также при уменьшении дисперсии σ^2 наблюдаемой случайной величины. Кроме того β уменьшается при увеличении α , однако не принято брать α больше 0.05. При уменьшении α ошибка β , напротив, растет, поэтому не следует брать α слишком малым, если число наблюдений n мало, разность между μ_1 и μ_0 невелика, а дисперсия σ^2 - большая.

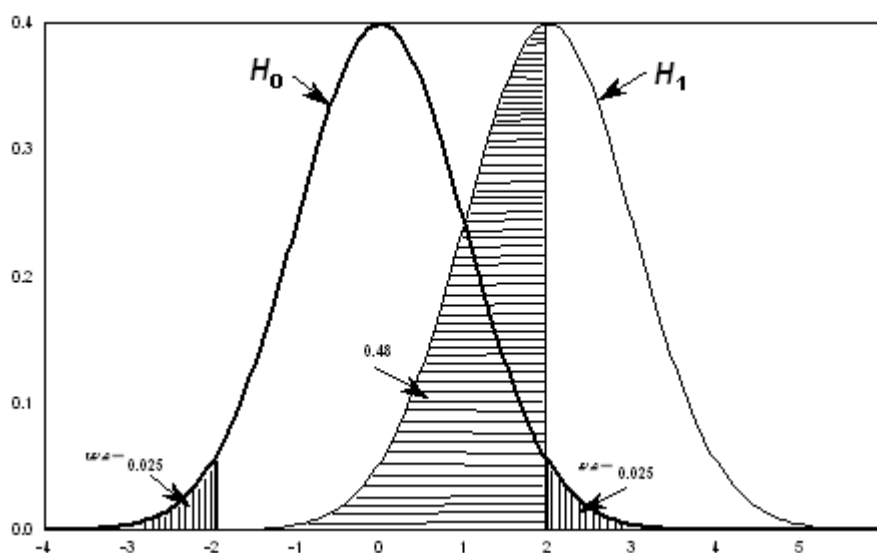


Рис. 5. Взаимосвязь между ошибками 1-го и 2-го рода при двусторонней альтернативе.

Содержательно, ошибка 1-го рода - это ошибка ложного обнаружения несуществующего отклонения от нулевой гипотезы (ложного обнаружения

несуществующего эффекта). Ошибка же 2-го рода - это ошибка ложного необнаружения существующего отклонения от нулевой гипотезы (ложного необнаружения существующего эффекта). Мощность критерия - это его способность обнаружить имеющееся отклонение от нулевой гипотезы.

В приведенном примере мы предполагали, что альтернативной гипотезе H_1 соответствует вполне определенное распределение $\xi \sim N(2, 1)$, что позволило нам найти конкретное значение ошибки 2-го рода. Такого рода альтернативные гипотезы называются *простыми альтернативами*. Однако на практике чаще встречается ситуация, когда конкретной нулевой гипотезе противопоставляется целый спектр альтернатив. Например, $H_1: M\xi \neq \mu_0$ или $H_1: M\xi > \mu_0$. Такого рода альтернативные гипотезы называются *сложными альтернативами*. В случае сложной альтернативной гипотезы мы не можем определить величину ошибки второго рода. Например, в рассматриваемой ситуации она может быть значительной даже при очень большом числе наблюдений если различие между μ_1 и μ_0 мало. Поэтому в ситуации, когда статистика критерия не попадает в критическую область, не утверждают категорично, что "нулевая гипотеза принимается", а формулируют вывод более осторожно: "нулевая гипотеза не отвергается". Тем самым подчеркивается, что хотя мы и не обнаружили отклонения от нулевой гипотезы, мы могли его при верной H_1 ошибочно не обнаружить с вероятностью β , которую мы не знаем и которая, возможно, довольно значительна. Если же ошибка β действительно велика, то утверждение "нулевая гипотеза принимается" не представляет большой ценности. Например, положив равной нулю ошибку 1-го рода, мы, независимо от результатов наблюдений, всегда будем принимать гипотезу H_0 , поскольку критическая область будет включать всю область определения статистики критерия. Однако при этом ошибка 2-го рода будет равна единице, т.е. если даже отклонение от нулевой гипотезы имеется, то мы его с вероятностью единица не обнаружим.

Сложные альтернативы могут быть *двусторонними* ($H_1: M\xi \neq \mu_0$) и *односторонними* ($H_1: M\xi > \mu_0$ или $H_1: M\xi < \mu_0$). Если имеется достоверная информация о направлении отклонения от нулевой гипотезы, то использование односторонней альтернативы предпочтительнее двусторонней, поскольку это повышает мощность критерия. Если, например, известно, что отклонение математического ожидания $M\xi$ от гипотетического значения μ_0 может произойти только в большую сторону, то в качестве альтернативы следует взять гипотезу $H_1: M\xi > \mu_0$. Критическая область уровня α в этом случае будет состоять не из двух бесконечных полуинтервалов $(-\infty, u_{\alpha/2}]$ и $[u_{1-\alpha/2}, \infty)$, из одного - $[u_{1-\alpha}, \infty)$.

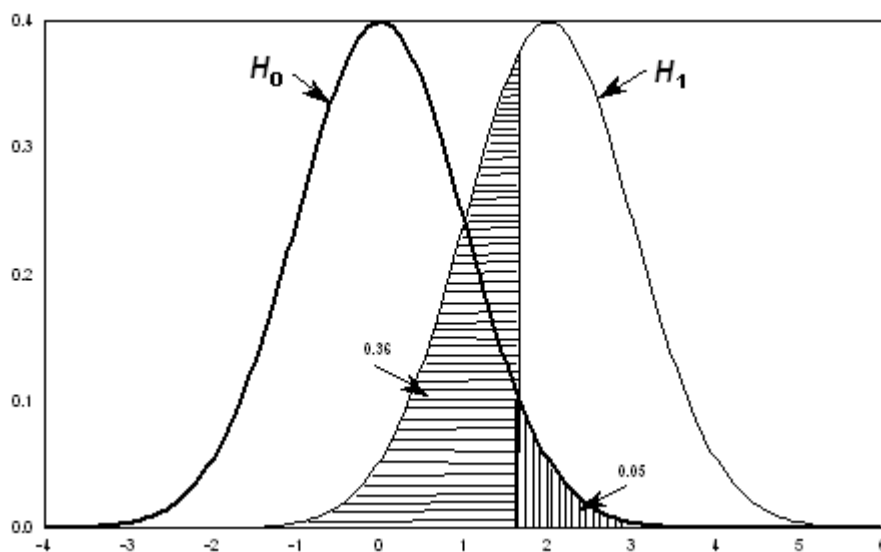


Рис. 6. Взаимосвязь между ошибками 1-го и 2-го рода при односторонней альтернативе.

На рис. 6 ситуация с односторонней альтернативой представлена для случая $\mu_0 = 0$, $\mu_1 = 2$, $\sigma^2 = 1$, $\alpha = 0.05$ и $n=1$. Вероятности ошибки 2-го рода β соответствует площадь под кривой плотности статистики критерия g при условии, что верна гипотеза H_1 на промежутке от $-\infty$ до 1.64. $\beta = 0.36$, что меньше, чем для аналогичной двусторонней альтернативы, представленной на рис. 5 (строго говоря, в ситуации рис. 5 тоже предпочтительнее было бы использовать альтернативу $H_1: M\xi > \mu_0$, поскольку направление отклонения от нулевой гипотезы было известно).

Пример. Известно, что датчик генерирует случайные числа, нормально распределенные с дисперсией 1, но есть сомнения в том, что математическое ожидание равно 0. Требуется проверить гипотезу о равенстве математического ожидания нулю по следующей случайной выборке объема $n=25$:

0.830 0.177 -0.294 0.471 -0.044 0.635 2.209 -0.394 -0.404
 1.257 1.137 -0.839 1.668 0.751 0.416 -0.922 1.473 -0.317
 0.220 0.414 0.428 1.088 -1.130 -0.015 0.142

Выборочное среднее равно $\bar{x} = 0.358$, следовательно, для статистики критерия получаем

$$u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{0.358 - 0}{1 / \sqrt{25}} = 1.79$$

Значение 1.79 не выходит за двусторонние 5%-ные критические пределы ± 1.96 , поэтому гипотеза не отвергается.

На самом деле математическое ожидание датчика было положено равным 0.25, т.е. отклонение ошибочно не было обнаружено - при проверке гипотезы была сделана ошибка 2-го рода. Очевидно, мощность критерия при данном числе наблюдений $n=25$, данной разности между гипотетическим и истинным математическими ожиданиями $\mu_1 - \mu_0 = 0.25$ и данной дисперсии $\sigma^2 = 1$ недостаточна.

В другом эксперименте с этим же датчиком была получена выборка значений объема $n=100$. Выборочное среднее оказалось равным $\bar{x}=0.218$, а значение статистики - равным $u=2.18$, что дало основание отвергнуть нулевую гипотезу.

3.3. Проверка гипотезы о равенстве заданному числу математического ожидания нормально распределенной

случайной величины с неизвестной дисперсией

(одновыборочный t -критерий)

Аналогично случаю построения доверительного интервала для неизвестного математического ожидания нормально распределенной случайной величины, в случае неизвестной дисперсии мы возьмем в качестве статистики критерия проверки гипотезы о равенстве математического ожидания заданному числу ту же статистику, что и в случае с известной дисперсией, но с заменой неизвестного среднеквадратичного отклонения σ на его выборочную оценку s

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Статистика t имеет t -распределение с $n-1$ степенями свободы. Соответственно, критическая область для проверки гипотезы $H_0: M\xi = \mu_0$ против двусторонней альтернативы $H_1: M\xi \neq \mu_0$ будет состоять из двух бесконечных полуинтервалов $(-\infty, t_{n-1, \alpha/2}]$ и $[t_{n-1, 1-\alpha/2}, \infty)$, против односторонней альтернативы $H_1: M\xi > \mu_0$ - из одного полуинтервала $[t_{n-1, 1-\alpha}, \infty)$ и против односторонней альтернативы $H_1: M\xi < \mu_0$ - также из одного полуинтервала

$(-\infty, t_{n-1-\alpha}]$, где $t_{\alpha/2}, t_{1-\alpha/2}, t_{1-\alpha}, t_{\alpha}$ обозначают квантили t -распределения с $n-1$ степенями свободы соответствующего уровня значимости (в силу симметричности t -распределения справедливы равенства $t_{\alpha/2} = -t_{1-\alpha/2}$ и $t_{\alpha} = -t_{1-\alpha}$).

Пример. Рассмотрим пример предыдущего параграфа с 25 случайными числами в предположении, что дисперсия неизвестна. В этом случае необходимо вычислить оценку среднеквадратичного отклонения, которая оказывается равной $s=0.830$. Выборочное значение статистики критерия, соответственно, равно

$$u = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{0.358 - 0}{0.830 / \sqrt{25}} = 2.15$$

Это значение должно быть сравнено с 5%-ными двусторонними критическими пределами, равными $\pm t_{24; 0.025} = \pm 2.06$. Выборочное значение статистики выходит за эти пределы, следовательно, гипотеза о равенстве математического ожидания нулю должна быть отвергнута на уровне значимости 5%.

Заметим, что хотя применение t -критерия требует нормальности исходной случайной величины, он может применяться и при умеренных отклонениях от нормальности и не слишком малых n ($n \geq 15$).

3.4. Проверка гипотезы о равенстве заданному числу дисперсии нормально распределенной случайной величины

(одновыборочный χ^2 -критерий)

Для проверки гипотезы $H_0: D\xi = \sigma_0^2$ о равенстве дисперсии $D\xi$ нормально распределенной случайной величины ξ заданному числу σ_0^2 рекомендуется использовать статистику

$$\chi^2 = (n-1) \frac{s^2}{\sigma_0^2}$$

Можно показать, что эта статистика при условии, что верна гипотеза H_0 , распределена по закону χ^2 с $n-1$ степенями свободы. Критическая область уровня α при двусторонней альтернативе $H_1: D\xi \neq \sigma_0^2$ состоит из двух промежутков: $[0, \chi_{n-1-\alpha/2}^2]$ и $[\chi_{n-1-\alpha/2}^2, \infty)$, где $\chi_{n-1-\alpha/2}^2$ и $\chi_{n-1-\alpha/2}^2$ - квантили порядка

$\alpha/2$ и $1-\alpha/2$ распределения χ^2 с $n-1$ степенями свободы. Для односторонней альтернативы $H_1: D\xi > \sigma_0^2$ критическая область имеет вид $[\chi_{n-1-\alpha}^2, \infty)$, а для альтернативы $H_1: D\xi < \sigma_0^2$ - соответственно, $[0, \chi_{n-1-\alpha}^2]$.

3.5. Проверка гипотезы о равенстве математических ожиданий двух независимых нормально распределенных случайных величин (двухвыборочный t -критерий)

Предположим, что имеются случайные выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m значений двух независимых нормально распределенных случайных величин $\xi \sim N(M\xi, D\xi)$ и $\eta \sim N(M\eta, D\eta)$ и требуется проверить гипотезу $H_0: M\xi = M\eta$ о равенстве математических ожиданий этих случайных величин.

(а) Если известно, что дисперсии случайных величин x и h равны, $D\xi = D\eta = \sigma^2$ (значение σ^2 неизвестно), то можно получить следующую объединенную несмещенную оценку для σ^2

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{j=1}^m (y_j - \bar{y})^2}{n+m-2}$$

В этом случае s^2/n и s^2/m будут несмещенными оценками для дисперсии выборочных средних \bar{x} и \bar{y} , а сумма $s^2/n + s^2/m$ - несмещенной оценкой для дисперсии разности средних $\bar{x} - \bar{y}$. Соответственно, статистика

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s^2}{n} + \frac{s^2}{m}}}$$

как можно показать, будет иметь t -распределение с $n+m-2$ степенями свободы. Критическая область уровня α для проверки гипотезы $H_0: M\xi = M\eta$ против двусторонней альтернативы $H_1: M\xi \neq M\eta$ будет состоять из двух бесконечных полуинтервалов $(-\infty, t_{n+m-2, \alpha/2}]$ и $[t_{n+m-2, 1-\alpha/2}, \infty)$, против односторонней альтернативы $H_1: M\xi > M\eta$ - из полуинтервала $[t_{n+m-2, 1-\alpha}, \infty)$ и против альтернативы $H_1: M\xi < M\eta$ - из полуинтервала $(-\infty, t_{n+m-2, \alpha}]$, где $t_{n+m-2, \alpha/2}$, $t_{n+m-2, 1-\alpha/2}$, $t_{n+m-2, \alpha}$ обозначают соответствующие квантили t -распределения с $n+m-2$ степенями свободы.

(б) Если нет оснований считать, что дисперсии случайных величин x и h равны, то для каждой из дисперсий $D\xi$ и $D\eta$ вычисляется своя оценка

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad \text{и} \quad s_y^2 = \frac{\sum_{j=1}^m (y_j - \bar{y})^2}{m-1}$$

и соответственно модифицируется статистика критерия

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

которая, как можно показать, имеет t -распределение с числом степеней свободы, равным целой части от $1/k$, где k выражается следующей формулой

$$k = \frac{\left(\frac{s_x^2/n}{s_x^2/n + s_y^2/m} \right)^2}{n-1} + \frac{\left(1 - \frac{s_x^2/n}{s_x^2/n + s_y^2/m} \right)^2}{m-1}$$

3.6. Проверка гипотезы о равенстве дисперсий двух независимых нормально распределенных случайных величин

(двухвыборочный F -критерий)

В предыдущем параграфе мы видели, что процедура проверки гипотезы о равенстве двух математических ожиданий двух нормально распределенных случайных величин упрощается, если их дисперсии одинаковы. Следующий критерий позволяет проверить нулевую гипотезу $H_0: D\xi = D\eta$ о равенстве дисперсий двух нормально распределенных случайных величин. В качестве статистики критерия используется отношение несмещенных оценок дисперсий этих случайных величин

$$F = \frac{s_x^2}{s_y^2}$$

При условии, что верна гипотеза H_0 , можно доказать, что статистика критерия имеет F -распределение с $n-1$ и $m-1$ степенями свободы. Соответственно, критическая область уровня α для проверки гипотезы $H_0: D\xi = D\eta$ против двусторонней альтернативы $H_1: D\xi \neq D\eta$ будет состоять из двух промежутков: $(0, F_{\alpha/2, n-1, m-1, \alpha/2}]$ и $[F_{\alpha/2, n-1, m-1, \alpha/2}, \infty)$, где $F_{\alpha/2, n-1, m-1, \alpha/2}$ -

квантили порядка $\alpha/2$ и $1-\alpha/2$ F -распределения с $n-1$ и $m-1$ степенями свободы. Для односторонней альтернативы $H_1: D\xi > D\eta$ критическая область имеет вид $[F_{n-1, m-1, 1-\alpha}, \infty)$, а для альтернативы $H_1: D\xi < D\eta$ - соответственно $[0, F_{n-1, m-1, \alpha}]$. Если в качестве статистики использовать отношение большей оценки дисперсии к меньшей, то в качестве критической области при двусторонней альтернативе используется односторонняя критическая область $[F_{n-1, m-1, 1-\alpha}, \infty)$ - это позволяет ограничиться таблицами F -распределения, содержащими значения функции распределения только для аргументов больших единиц.

Заметим, что в отличие от t -критерия F -критерий чувствителен к отклонениям исходных случайных величин от нормальности. При значительных отклонениях от нормальности, особенно при небольшом числе наблюдений его не следует применять.

3.7. Приближенный критерий для проверки гипотезы о равенстве параметров двух независимых биномиальных случайных величин (критерий для сравнения двух вероятностей)

Пусть две независимые биномиально распределенные случайные величины ξ и η с параметрами n, p_ξ и m, p_η , соответственно, при проведении независимых испытаний приняли значения k и l . Требуется проверить гипотезу $H_0: p_\xi = p_\eta$ о равенстве параметров p_ξ и p_η . Для этого можно использовать статистику

$$u = \frac{v_\xi - v_\eta - \frac{1}{2(n+m)}}{\sqrt{v(1-v)\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

где v_ξ, v_η и v - выборочные частоты, вычисленные по первой, второй и объединенной выборкам: $v_\xi = k/n, v_\eta = l/m$ и $v = (k+l)/(n+m)$. Если верна гипотеза H_0 , то для v_ξ, v_η , не очень близких к 0 или 1, и при достаточно больших n, m эта статистика имеет приближенно стандартное нормальное распределение. Практически приближение применимо, если каждая из четырех численностей $k, l, n-k$ и $m-l$ больше пяти.

Критическая область уровня значимости α для проверки гипотезы $H_0: p_\xi = p_\eta$ против двусторонней альтернативы $H_1: p_\xi \neq p_\eta$ будет состоять из двух бесконечных полуинтервалов $(-\infty, u_{\alpha/2}]$ и $[u_{1-\alpha/2}, \infty)$, против

односторонней альтернативы $H_1: P_\xi > P_\eta$ - из одного полуинтервала $[u_{1-\alpha}, \infty)$ и против односторонней альтернативы $H_1: P_\xi < P_\eta$ - также из одного полуинтервала $(-\infty, u_\alpha]$, где $u_{\alpha/2}$, $u_{1-\alpha/2}$, $u_{1-\alpha}$, и u_α обозначают квантили соответствующего порядка стандартного нормального распределения.

Имеется также точный критерий для проверки этой гипотезы (см., напр., [3]).

3.8. Приближенный критерий для проверки гипотезы о равенстве параметров двух независимых пуассоновских случайных величин

Пусть две независимые случайные величины ξ и η , имеющие пуассоновское распределение с параметрами λ_ξ и λ_η , соответственно, при проведении испытаний приняли значения k и l . Требуется проверить гипотезу $H_0: \lambda_\xi = \lambda_\eta$ о равенстве параметров λ_ξ и λ_η распределений этих случайных величин. Для этого можно использовать статистику

$$u = \frac{k - l}{\sqrt{k + l}}$$

распределение которой при выполнении H_0 и при $k+l > 5$ довольно точно приближается стандартным нормальным распределением. Соответственно, как и в предыдущем параграфе, критическая область уровня значимости α для проверки гипотезы $H_0: P_\xi = P_\eta$ против двусторонней альтернативы $H_1: P_\xi \neq P_\eta$ будет состоять из двух бесконечных полуинтервалов $(-\infty, u_{\alpha/2}]$ и $[u_{1-\alpha/2}, \infty)$, против односторонней альтернативы $H_1: P_\xi > P_\eta$ - из одного полуинтервала $[u_{1-\alpha}, \infty)$ и против односторонней альтернативы $H_1: P_\xi < P_\eta$ - также из одного полуинтервала $(-\infty, u_\alpha]$.

3.9. Приближенный критерий для проверки гипотезы о равенстве нулю коэффициента корреляции между компонентами двумерной нормально распределенной случайной величины

Пусть $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - случайная выборка пар значений двумерной случайной величины (ξ, η) , имеющей двумерное нормальное распределение. Требуется проверить гипотезу $H_0: \rho = \rho_0$ о равенстве коэффициента корреляции ρ этого двумерного распределения заданному числу ρ_0 . Для проверки этой гипотезы можно использовать статистику

$$u = \frac{\sqrt{n-3}}{2} \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} \right)$$

распределение которой при выполнении H_0 и при достаточно большом n довольно точно приближается стандартным нормальным распределением. Соответственно, как и в предыдущих двух параграфах, критическая область уровня значимости α для проверки гипотезы $H_0: \rho = \rho_0$ против двусторонней альтернативы $H_1: \rho \neq \rho_0$ будет состоять из двух бесконечных полуинтервалов $(-\infty, u_{\alpha/2}]$ и $[u_{1-\alpha/2}, \infty)$, против односторонней альтернативы $H_1: \rho > \rho_0$ - из одного полуинтервала $[u_{1-\alpha}, \infty)$ и против односторонней альтернативы $H_1: \rho < \rho_0$ - также из одного полуинтервала $(-\infty, u_{\alpha}]$.

Обычно проверяется гипотеза о равенстве коэффициента корреляции нулю, что в случае двумерного нормального распределения, как ранее отмечалось, эквивалентно проверке гипотезы о независимости ξ и η . В этом случае приведенное выше выражение для статистики критерия упрощается

$$u = \frac{\sqrt{n-3}}{2} \left(\ln \frac{1+r}{1-r} \right) \quad (5)$$

Пример. Пусть объем выборки $n=10$, вычисленное по выборке значение $r=0.6$ и требуется проверить гипотезу $H_0: \rho = 0$ против альтернативы $H_1: \rho \neq 0$.

Выборочное значение статистики u , вычисленное по формуле (5), равно 1.83. Поскольку оно не выходит за двусторонние 5%-ные критические пределы стандартного нормального распределения ± 1.96 , то у нас нет оснований отвергнуть нулевую гипотезу об отсутствии корреляции. Если бы у нас были основания предполагать, что корреляционная зависимость в случае ее наличия может быть только положительной, то следовало бы использовать для проверки H_0 , одностороннюю критическую область, которая для $\alpha = 0.05$ представляет собой бесконечный полуинтервал $[1.64, \infty)$. Значение 1.83 попадает в эту критическую область и, следовательно, гипотеза об отсутствии корреляции должна бы была быть отвергнута. Заметим, однако, что число наблюдений в данном примере недостаточно велико для уверенного использования данного приближенного критерия. Если к этому добавить тот факт, что выборочное значение статистики критерия находится вблизи границы критической области, то следует заключить, что по имеющимся данным нельзя сделать надежного вывода ни о наличии, ни об отсутствии корреляции.

Отметим, что если бы, скажем, значение $r=0.6$ было получено для $n=50$, то выборочное значение статистики u было бы равно 4.75, и гипотеза

однозначно должна бы была быть отвергнута не только на уровне значимости 5%, но и 1% (и даже более высоком, т.к. вероятность того, что стандартно распределенная случайная величина примет значение большее 4.75 равна 0.000001).

3.10. Критерии согласия

Все рассмотренные до сих пор критерии принято относить к группе так называемых *параметрических критериев*. Применение этих критериев требует знания типа распределения наблюдаемых случайных величин (нормальное, биномиальное, пуассоновское, двумерное нормальное или какое-либо иное) и проверяемая гипотеза касается параметров данных распределений. Прежде чем применять параметрические методы, необходимо убедиться в том, что мы действительно имеем дело с распределением требуемого типа.

Предположение о виде распределения случайной величины – это статистическая гипотеза, которую можно проверить с помощью экспериментальных данных. Критерии, позволяющие решать такого рода задачи, называются критериями согласия – согласия выборочных данных некоторому наперед заданному теоретическому распределению.

Пусть имеется выборка x_1, x_2, \dots, x_n значений случайной величины ξ с неизвестной функцией распределения $F(x)$. Требуется проверить гипотезу $H_0: F(x) = F_0(x)$ о том, что случайная величина имеет некоторое заданное распределение $F_0(x)$ против альтернативной гипотезы $H_1: F(x) \neq F_0(x)$. Распределение $F_0(x)$ может быть либо задано полностью (простая нулевая гипотеза), либо с точностью до параметров (сложная нулевая гипотеза). Во втором случае фактически проверяется принадлежность распределения к заданному типу, например, проверяется гипотеза о нормальности. Часто это делается с целью обоснования применения для обработки полученных данных методов, требующих принадлежности распределения к заданному типу (например, при применении t -критерия предполагается, что выборка извлечена из нормальной генеральной совокупности). Следует однако помнить, что неотвержение гипотезы H_0 не является убедительным доводом в пользу ее справедливости при неизвестной ошибке второго рода, которая может быть довольно высокой при небольшом числе наблюдений.

Для проверки гипотезы о принадлежности распределения к заданному типу часто используется так называемый *критерий согласия χ^2* . Относительно распределения $F(x)$ не делается никаких предположений, оно может быть как непрерывным, так и дискретным. Статистика критерия χ^2

вычисляется следующим образом. Область изменения значений выборки разбивается на k интервалов с таким расчетом, чтобы число наблюдений n_i (наблюдённая частота) в большинстве из интервалов $i, i=1, \dots, k$, было не менее 10. Для каждого из интервалов вычисляется также вероятность p_i попадания в этот интервал случайной величины при условии выполнения гипотезы H_0 . Статистика χ^2 равна нормированной сумме квадратов отклонений числа наблюдений n_i от гипотетической частоты np_i

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

Для легкости запоминания эту формулу можно рассматривать как сумму квадратов пуассоновских случайных величин n_i , стандартизованных путем вычитания из них гипотетических средних np_i и деления разности на их среднеквадратические отклонения $(np_i)^{1/2}$.

Если верна гипотеза H_0 и при достаточно большом n (не менее 50) распределение данной статистики хорошо приближается распределением χ^2 с $k-l$ степенями свободы, где l - число параметров гипотетического распределения $F_0(x)$, оцененных по выборке (одна степень свободы вычитается даже при полностью заданном $F_0(x)$, поскольку наблюдаемые частоты связаны соотношением $n_1+n_2+\dots+n_k=n$). Следовательно, критическое множество уровня значимости α состоит из одного полуинтервала $[\chi_{k-l-1-\alpha}^2, +\infty)$ где $\chi_{k-l-1-\alpha}^2$ - квантиль χ^2 -распределения с числом степеней свободы $k-l-1$ порядка $1-\alpha$.

Необходимость в проверке простых гипотез возникает относительно редко. Гораздо чаще $F_0(x)$ бывает известна с точностью до r параметров, $F_0(x) = F_0(x, \theta_1, \dots, \theta_r)$, где $\theta_1, \dots, \theta_r$ - неизвестные параметры. В этом случае теоретические вероятности p_i не удастся вычислить непосредственно, поэтому находим $\hat{p}_i = p_i(\hat{\theta}_1, \dots, \hat{\theta}_r)$, $i=1, \dots, k$, где $\hat{\theta}_1, \dots, \hat{\theta}_r$ - оценки параметров $\theta_1, \dots, \theta_r$, определяемые через наблюдаемые частоты n_1, \dots, n_k . Статистика критерия имеет вид

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

Если нулевая гипотеза H_0 верна, статистика критерия χ^2 при $n \rightarrow \infty$ распределена асимптотически как χ^2 с числом степеней свободы $k-r-1$. Следовательно, критическое множество уровня значимости α состоит из

полуинтервала $[x_{k-r-1,1-\alpha}^2, +\infty)$ где $x_{k-r-1,1-\alpha}^2$ - квантиль χ^2 -распределения с числом степеней свободы $k-r-1$ порядка $1-\alpha$.

Часто оценки неизвестных параметров $\hat{\theta}_1, \dots, \hat{\theta}_r$ определяются не по наблюдаемым частотам n_i , а по всей выборке. Например, при проверке нормальности ожидаемую частоту в i -ом интервале, $n\hat{p}_i$, находят, используя выборочное среднее \bar{x} и выборочную дисперсию s^2 , определенные по всей выборке. В этом случае статистика критерия χ^2 при справедливости H_0 уже не имеет асимптотически распределения χ^2 , ее распределение заключено между χ_{k-1}^2 и χ_{k-r-1}^2 . Различием между ними можно пренебречь при больших k . Но для малых k при определении критического множества полезно убедиться, что выборочное значение статистики критерия $\chi^2 > \chi_{k-r-1,1-\alpha}^2$ и $\chi^2 > \chi_{k-1,1-\alpha}^2$ [3].

Для проверки соответствия непрерывного распределения $F(x)$ заданному $F_0(x)$ используются также **одновыборочные критерии Колмогорова и Смирнова**. Статистика Колмогорова для проверки гипотезы H_0 против двусторонней альтернативы $H_1: F(x) \neq F_0(x)$ определяется как максимум модуля отклонения эмпирической функции распределения $\hat{F}(x)$ от гипотетической $F_0(x)$

$$D_n = \max_x |\hat{F}(x) - F_0(x)|$$

Статистика Смирнова, D_n^+ , для проверки гипотезы H_0 против правосторонней альтернативы $H_1: F(x) > F_0(x)$ имеет вид

$$D_n^+ = \max_x [\hat{F}(x) - F_0(x)]$$

Для случая простой нулевой гипотезы распределения статистик D_n и D_n^+ при справедливости H_0 не зависят от типа $F_0(x)$. Если верна нулевая гипотеза, предельным распределением статистики $\sqrt{n}D_n$ при $n \rightarrow \infty$ является распределение Колмогорова, а Н.В. Смирнов получил точное и предельное распределение статистики D_n^+ . Соответственно, критическое множество уровня значимости α для проверки гипотезы H_0 против двусторонней альтернативы H_1 состоит из полуинтервала $[D_{n,1-\alpha}, +\infty)$, и против правосторонней альтернативы H_1 - $[D_{n,1-\alpha}^+, +\infty)$, где $D_{n,1-\alpha}$, $D_{n,1-\alpha}^+$ - критические значения статистик D_n и D_n^+ , соответственно, уровня значимости α . При

$\alpha \leq 0,2$ $D_{n,1-\alpha} \approx D_{n,1-\alpha}^+$ с большой точностью (большей 0,00005). Поэтому критические значения статистики D_n^+ могут быть заменены критическими значениями статистики D_n .

В случае сложной нулевой гипотезы, когда $F_0(x)$ известна с точностью до параметров, $F_0(x) = F_0(x, \theta_1, \dots, \theta_r)$, где $\theta_1, \dots, \theta_r$ – неизвестные параметры, статистика критерия для проверки гипотезы H_0 против двусторонней альтернативы H_1 имеет вид

$$\hat{D}_n = \max_x |\hat{F}(x) - F_0(x, \hat{\theta}_1, \dots, \hat{\theta}_r)|$$

где $\hat{\theta}_1, \dots, \hat{\theta}_r$ – оценки неизвестных параметров.

При условии, что нулевая гипотеза H_0 верна, распределение статистики \hat{D}_n (и \hat{D}_n^+) уже зависит от конкретного вида распределения $F_0(x, \theta_1, \dots, \theta_r)$. Для некоторых типов распределений – нормального, показательного, логистического – Лиллифорсом получены таблицы критических значений статистики \hat{D}_n при условии, что гипотеза H_0 верна [8, 9]. Соответственно, критическое множество уровня значимости α для проверки гипотезы H_0 против двусторонней альтернативы H_1 состоит из одного полуинтервала $[\hat{D}_{n,1-\alpha}^+, +\infty)$, где $\hat{D}_{n,1-\alpha}^+$ – критическое значение статистики \hat{D}_n для заданных α , n и $F_0(x)$.

Статистика \hat{D}_n может быть преобразована к виду, практически не зависящему от n . Например, для нормального распределения Стефенсом получено следующее выражение для модифицированной формы статистики Колмогорова \hat{D}^* [5]:

$$\hat{D}^* = \hat{D}_n(\sqrt{n} - 0,001 + 0,85/\sqrt{n})$$

Это дает возможность проводить проверку гипотезы практически при всех n , зная значения $\hat{D}_{1-\alpha}^*$ для небольшого набора значений α . В частности, для $\alpha = 0.05$ имеем $\hat{D}_{0.95}^* = 0.895$.

При проверке гипотезы о нормальности распределения с неизвестными средним и дисперсией критерий Колмогорова-Смирнова является более мощным, чем критерий χ^2 .

Заметим, что в англоязычной литературе и в ППП статистики D_n и D_n^* называют одновыборочными статистиками Колмогорова-Смирнова, двусторонней и односторонней, соответственно.

Среди других критериев согласия отметим критерий Шапиро - Уилка для проверки нормальности [6].

Если для конкретной выборки мы отклоняем гипотезу о нормальности, и, следовательно, не имеем права пользоваться методами, основанными на нормальности, то для получения статистических выводов можно поступать разными способами. Например, если объем выборки достаточно велик, можно предпочесть использовать параметрические критерии как приближенные. Другой путь состоит в подборе замены переменной, приводящей к нормальному распределению. Третий путь - применение непараметрических критериев.

Пример. Пусть получена следующая выборка 50 значений случайной величины ξ с неизвестным распределением:

45	89	93	40	91	60	2	59	87	78
57	39	50	0	35	91	67	62	25	93
19	98	55	78	34	45	86	31	15	95
50	52	35	66	0	44	93	36	29	44
17	85	17	63	34	43	100	75	84	9

Проверим гипотезу о том, что эта случайная величина имеет нормальное распределение. После разбиения области изменения выборочных значений на 5 равных интервалов получаем следующие наблюдаемые и гипотетические частоты:

Интервал	$(-\infty, 20]$	$(20, 40]$	$(40, 60]$	$(60, 80]$	$(-80, \infty)$
Наблюдённая частота, n_i	8	10	12	7	13
Гипотетическая Частота, np_i	6.1	9.7	13.4	11.6	9.2

Гипотетические частоты вычислялись для нормального распределения $N(\mu = 53.3, \sigma^2 = 840.4)$ с параметрами, оцененными по выборке - соответственно, число степеней свободы статистики критерия равно $5-1-2=2$. Выборочное значение статистики равно $\chi^2 = 4.2$, что не выходит за критический 5%-ный предел, равный $\chi_{2,0.95}^2 = 6.0$. Следовательно, у нас нет оснований отвергнуть гипотезу о нормальности.

В действительности, выборка была получена с помощью датчика случайных чисел, равномерно распределенных на отрезке $[0, 100]$. Т.е. мы видим, что при данном числе наблюдений (в общем-то, конечно, небольшом для проверки гипотезы о типе распределения) критерий χ^2 не обнаруживает отклонения от нормальности в направлении равномерности.

Величина статистики одновыборочного критерия Колмогорова - Смирнова равна $D=0.11$, что также не выходит за 5%-ный предел этого критерия в предположении, что гипотетические средние равны выборочным. Однако в случае неизвестных параметров гипотетического нормального распределения лучше пользоваться модификацией критерия Колмогорова - Смирнова, предложенной Стефенсом (Лиллифорсом). Но в этом случае значение $\hat{D}_S = 0.11 < D_{S,0.95}^* = 0.895$, т.е. нет оснований отвергнуть гипотезу и по этому критерию.

Пример. Расчеты, аналогичные предыдущим, проведенные для выборки объема 150 значений случайной величины, равномерно распределенной на отрезке $[0, 100]$, дали значение $\chi^2 = 7.6$, что позволило отвергнуть гипотезу о нормальности на уровне значимости 5%. По критерию Колмогорова - Смирнова гипотеза отвергалась лишь на уровне 10%, а по критерию Лиллифорса - на уровне 1%, что показывает неправомочность применения критерия Колмогорова - Смирнова в данной ситуации.

Пример. Расчеты статистик критериев согласия для данных таблицы 1, содержащей 50 выборочных значений длины лепестка ириса разноцветного, приводят к значению статистики χ^2 равному 2.1, и значению статистики \hat{D}_* , равному 0.117. В этом случае гипотеза о нормальности не отвергается ни критерием χ^2 , ни критерием Колмогорова - Смирнова - Лиллифорса.

Пример. В некоторых классических экспериментах с селекцией гороха Мендель наблюдал частоты различных видов семян, получаемых при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Они приводятся ниже вместе с теоретическими вероятностями, вычисленными в соответствии с теорией наследственности Менделя.

Семена	Наблюденная численность	Ожидаемая численность
Круглые и желтые	315	$556 \times \frac{9}{16} = 313$
Морщинистые и желтые	101	$556 \times \frac{3}{16} = 104$
Круглые и зеленые	108	$556 \times \frac{3}{16} = 104$
Морщинистые и зеленые	32	$556 \times \frac{1}{16} = 35$
Всего	556	556

В этом случае теоретическое распределение дискретно и известно полностью. Для проверки согласия экспериментальных данных теоретическому распределению используем критерий χ^2 для простой гипотезы. Значение статистики, вычисленное по выборке равно $\sum_{i=1}^4 \frac{(n_i - np_i)^2}{np_i} = 0.47$, что меньше 5%-ного критического значения $\chi_{3095}^2 = 7.81$. Следовательно, теория наследственности Менделя не противоречит полученным экспериментальным данным.

Наряду с количественными статистическими критериями для определения типа распределения по выборочным данным используются *графические методы*.

Простейший способ - построение по имеющейся выборке гистограммы относительных частот и на том же графике и в том же масштабе, - кривой плотности нормального распределения с выборочным средним и выборочной дисперсией в качестве параметров. Значительные отклонения от нормальности (сильная асимметрия, бимодальность) легко обнаруживаются на графике.

Пример. Применим этот прием к рассмотренной выше модельной выборке объема $n=50$, извлеченной из равномерного распределения. На рис. 7 приведена гистограмма и кривая нормальной плотности. Можно сказать, что визуально отклонение от нормальности в пользу равномерности заметно (хотя, как мы видели, статистически значимо при таком числе наблюдений оно не подтверждается).

С точки зрения визуального обнаружения отклонений от нормальности сравнение эмпирической и гипотетической функций распределения гораздо менее наглядно, чем сравнение гистограммы с графиком плотности. Однако обычно сравнивают не сами функции распределения, а обратные нормальные преобразования от них, так называемые **пробит-графики**. Пробит-график от теоретической нормальной функции распределения представляет собой прямую, а пробит-график эмпирической функции распределения тем ближе к прямой, чем ближе она к нормальной. Этот прием позволяет на первом этапе анализа данных выявить их особенности, выдвинуть гипотезы о характере распределения, решить вопрос о целесообразности замены переменной.

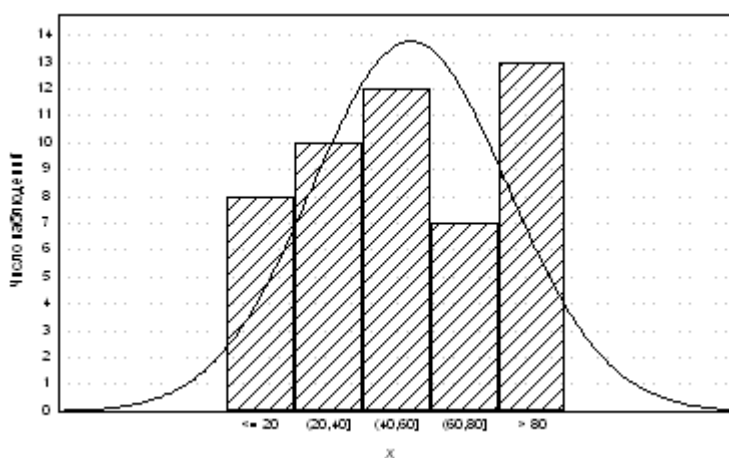


Рис. 7. Пример сравнения гистограммы и кривой нормальной плотности.

3.11. Непараметрические критерии

В большинстве случаев надежная априорная информация о типе распределения отсутствует, а имеющиеся выборочные данные слишком малочисленны для определения типа распределения. В этих ситуациях применяются так называемые **непараметрические критерии**, характеризующиеся тем, что в качестве их статистик используются такие функции от наблюдений, распределение которых не зависит от вида распределения наблюдаемых случайных величин.

Часто статистики непараметрических критериев основаны не непосредственно на численных значениях наблюдений выборки x_1, x_2, \dots, x_n , а на их **рангах**, т.е. на порядковых номерах $R(x_1), R(x_2), \dots, R(x_n)$ наблюдений при их упорядочении по возрастанию (в их вариационном ряду). Ранги наблюдений, будучи функциями выборочных значений, являются случайными величинами с возможными значениями $1, 2, \dots, n$. Оказалось,

что набор рангов $R(x_1), R(x_2), \dots, R(x_n)$ содержит значительную долю информации о распределении наблюдаемой случайной величины, что обеспечивает этим методам высокую эффективность.

Если статистика рангового критерия, $g(R_1, R_2, \dots, R_n)$, – дискретная случайная величина, то для заданного уровня значимости α может не существовать значения квантили распределения статистики критерия при справедливости нулевой гипотезы порядка α . Поэтому для определения критического множества используется верхнее критическое значение статистики критерия $g_{1-\alpha}$, равное наименьшему значению g , такому, что $P(g > g_{1-\alpha}) < \alpha$, и нижнее критическое значение g_α , равное наибольшему значению g , такому, что $P(g < g_\alpha) < \alpha$. Значения g_α и $g_{1-\alpha}$ находятся по таблицам. Для всех рассматриваемых критериев существуют таблицы критических значений статистики, например, в [1, 2, 7].

Важной особенностью ранговых критериев является их применимость и в тех случаях, когда наблюдения не являются количественными, но допускают упорядочение, что часто имеет место в исследованиях по биологии, медицине, психологии и социологии. Рассмотрим некоторые непараметрические критерии.

3.11.1. Одновыборочные критерии

Ряд одновыборочных критериев предназначен для проверки гипотезы о равенстве медианы заданному значению. Пусть имеется выборка x_1, x_2, \dots, x_n значений случайной величины ξ с неизвестной функцией распределения $F(x, M)$ и неизвестной медианой M . Требуется проверить гипотезу $H_0: M = M_0$ о равенстве медианы M заданному числу M_0 .

Для решения этой задачи можно воспользоваться *критерием знаков*. Возьмем в качестве статистики критерия число n^+ положительных разностей среди n разностей $x_i - M_0, i=1, \dots, n$. Если верна нулевая гипотеза H_0 , то $P(x_i > M_0) = P(x_i < M_0) = 1/2$ и, следовательно, статистика критерия n^+ – дискретная случайная величина, распределенная по биномиальному закону с параметрами n и $p=1/2$.

Критическая область уровня значимости α для проверки гипотезы $H_0: M = M_0$ против двусторонней альтернативы $H_1: M \neq M_0$ будет состоять из двух интервалов $[0, n_{\alpha/2}^+]$ и $[n_{1-\alpha/2}^+, n]$, причем $n_{\alpha/2}^+ = n - n_{1-\alpha/2}^+$, так как распределение статистики критерия n^+ при H_0 симметрично относительно своего среднего $n/2$, а $n_{\alpha/2}^+, n_{1-\alpha/2}^+$ – нижнее и верхнее критические значения статистики n^+ , соответственно. Критическая область против правосторонней

альтернативы $H_1: M > M_0$ состоит из одного интервала $[T_{1-\alpha}^+, n]$ и против левосторонней альтернативы $H_1: M < M_0$ - также из одного интервала $[0, T_\alpha^+]$.

При малых n критические значения можно вычислить точно с помощью непосредственного перебора равновозможных 2^n последовательностей с $n^+ = 0, 1, \dots, n$. При $n \rightarrow \infty$ случайная величина n^+ распределена асимптотически нормально, $n^+ \sim N(n/2, n/4)$, и для нахождения критических значений можно воспользоваться нормальным приближением.

Критерий знаков обладает недостаточной чувствительностью к различению нулевой и альтернативной гипотез (его асимптотическая эффективность по отношению к одновыборочному t -критерию равна 0,637), но из-за простоты и наглядности часто используется для предварительного анализа данных.

Большой мощностью обладает **критерий знаковых рангов** (асимптотическая эффективность по отношению к одновыборочному t -критерию равна 0,955). Статистика знаковых рангов Вилкоксона равна сумме рангов положительных разностей

$$T^+ = \sum_{R(x_i - M_0) > 0} R(x_i - M_0)$$

где $R(x_i - M_0)$ ранг разности $x_i - M_0$ и суммирование рангов ведется по положительным разностям.

Если нулевая гипотеза H_0 верна, вероятность каждого из возможных 2^n исходов для набора рангов положительных разностей равна $(1/2)^n$, что и определяет распределение статистики критерия для заданного n , симметричного относительно среднего $n(n+1)/4$, откуда следует, что $T_{1-\alpha}^+ = n(n+1)/2 - T_\alpha^+$. где $T_\alpha^+, T_{1-\alpha}^+$ - верхнее и нижнее критические значения статистики критерия при заданных α и n , соответственно. Критическая область уровня значимости α для проверки гипотезы $H_0: M = M_0$ против двусторонней альтернативы $H_1: M \neq M_0$ будет состоять из двух интервалов $[0, T_{\alpha/2}^+]$ и $[T_{1-\alpha/2}^+, n(n+1)/2]$. Критическая область против правосторонней альтернативы $H_2: M > M_0$ - из одного интервала $[T_{1-\alpha}^+, n(n+1)/2]$ и против левосторонней альтернативы $H_3: M < M_0$ - также из одного интервала $[0, T_\alpha^+]$.

Если верна H_0 , то при $n \rightarrow \infty$ распределение статистики критерия стремится к нормальному, $T^+ \sim N(n(n+1)/4, n(n+1)(2n+1)/24)$. При $n > 25$ этим

приближением можно воспользоваться для определения критических значений статистики.

3.11.2. Проверка гипотезы об отсутствии сдвига

Пусть имеются выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m значений случайных величин ξ и η с неизвестными функциями распределения $F(x)$ и $G(x)$. Известно однако, что $F(x)$ и $G(x)$ имеют одинаковую форму и различаются лишь сдвигом, т.е. $G(x) = F(x - \Delta)$. Требуется проверить гипотезу $H_0: \Delta = 0$ об отсутствии сдвига между распределениями случайных величин ξ и η .

Случай независимых выборок

Пусть x_1, \dots, x_n и y_1, \dots, y_m – независимые выборки из непрерывных распределений $F(x)$ и $G(x)$, соответственно, причем $G(x) = F(x - \Delta)$. Для решения задачи об отсутствии сдвига между $F(x)$ и $G(x)$ можно применить **критерий Вилкоксона** или **критерий Манна - Уитни**. Пусть $n \leq m$ – в противном случае выборки поменяем местами. Упорядочим $n+m$ наблюдений по возрастанию и обозначим через R_i ранг i -ого наблюдения в объединенном ряду, $i=1, \dots, n+m$. Если есть совпадающие значения внутри какой-либо из выборок, то ранги их можно взять в произвольном порядке. Если же совпадают значения, принадлежащие разным выборкам, то их ранги заменяются средним арифметическим рангов, которые бы получились, если бы наблюдения различались.

В качестве статистики Манна-Уитни U используется общее число случаев (инверсий) в упорядоченной по возрастанию последовательности из x и y , в которых x появляется позднее некоторого y :

$$U = \sum_{i=1}^n \sum_{j=1}^m \delta_{ij}, \quad \delta_{ij} = \begin{cases} 1, & x_i > y_j \\ 0, & x_i < y_j \end{cases}$$

Если $x_i = y_j$, к значению U прибавляется $1/2$. Статистика U – дискретная случайная величина, принимающая значения от 0 до nm .

Если нулевая гипотеза H_0 верна, C_{n+m}^m последовательностей из x и y являются равновероятными, что и определяет распределение статистики U , симметричное относительно своего среднего $nm/2$. Критическая область уровня значимости α для проверки гипотезы $H_0: M = M_0$ против двусторонней альтернативы $H_1: M \neq M_0$ будет состоять из двух интервалов $[0, U_{\alpha/2}]$ и $[U_{1-\alpha/2}, nm]$, где U_{α} и $U_{1-\alpha}$ – нижнее и верхнее критические значения

статистики критерия U , связанные соотношением $U_{1-\alpha, n, m} = m - U_{\alpha, n, m}$. Критическая область против правосторонней альтернативы $H_2: M > M_0$ - из одного интервала $[U_{1-\alpha, n, m}]$ и против левосторонней альтернативы $H_3: M < M_0$ - из одного интервала $[0, U_{\alpha}]$.

При малых n и m значение U_{α} определяется непосредственным подсчетом последовательностей с наименьшим количеством инверсий. При больших n и m распределение U можно аппроксимировать нормальным распределением. Если нулевая гипотеза H_0 верна, то при $n \rightarrow \infty, m \rightarrow \infty$ $U \sim N(m/2, m(m+n+1)/12)$, и для вычисления критических значений можно воспользоваться нормальным приближением.

Статистику критерия U можно также вычислить по формуле

$$U = W - \frac{n(n+1)}{2}$$

где $W = \sum_{j=1}^n R_j$, сумма рангов наблюдений $x_j, j=1, \dots, n$, есть статистика критерия Вилкоксона. Следовательно критерии, основанные на статистиках U и W эквивалентны.

Пример. Пусть получены выборки значений двух случайных величин ξ и η объема $n=4$ и $m=5$:

ξ : 174 175 183 174

η : 187 185 185 179 181

Составим из них общий вариационный ряд (т.е. расположим в порядке возрастания), сохранив информацию о принадлежности к выборке:

Ранг	1	2	3	4	5	6	7	8	9
ξ	174	174	175			183			
η				179	181		185	185	187

Сумма рангов выборки значений случайной величины ξ равна $W=1+2+3+6=12$. Это значение не выходит за двусторонние критические пределы $W_{0.025}=11$ и $W_{0.975}=34$ уровня значимости 5%. Выборочное значение статистики $U=2$ и соответствующее ему $p=0.032$ меньше 0.05 и, следовательно, у нас нет оснований отвергнуть нулевую гипотезу о том, что сдвиг между распределениями $F(x)$ и $G(x)$ отсутствует.

Заметим, что приведенные данные были получены с помощью датчика нормально распределенных случайных чисел $\xi \sim N(\mu = 175, \sigma^2 = 25)$ и $\eta \sim N(\mu = 180, \sigma^2 = 25)$. Приведенные выше значения могли бы быть, например, значениями роста четырех случайно выбранных французов и пяти случайно выбранных норвежцев (средний рост взрослых мужчин Франции и Норвегии равен 175 и 180 см, соответственно). Т.е. в действительности сдвиг Δ между распределениями отличен от нуля (он равен $\Delta = 175 - 180 = -5$) и гипотеза $H_0: \Delta = 0$ неверна, но критерии Вилкоксона и Манна - Уитни не обнаружили различия между распределениями $F(x)$ и $G(x)$. Если применить к приведенным данным t -критерий Стьюдента для сравнения математических ожиданий двух нормально распределенных случайных величин с неизвестными дисперсиями, то получим выборочное значение $t = -2.72$ для статистики критерия. Поскольку это значение выходит за 5%-ные критические пределы $t_{7,0.025} = -2.36$ и $t_{7,0.975} = 2.36$ t -распределения с $4 + 5 - 2 = 7$ степенями свободы, то гипотеза о равенстве математических ожиданий должна быть отвергнута. Это типичная ситуация - непараметрические критерии обладают меньшей мощностью по сравнению с аналогичными параметрическими критериями, использующими дополнительную информацию о наблюдаемых случайных величинах. Поэтому, если имеется достоверная дополнительная информация, то предпочтительнее использовать критерий, учитывающий эту информацию.

Гипотезу об отсутствии сдвига можно проверить также с помощью **критерия Ван-дер-Вардена**. Обозначим $N = n + m$. Статистика критерия имеет вид

$$X = \sum_{i=1}^n \Psi \left(\frac{R(x_i)}{N+1} \right)$$

где $R(x_i)$ - ранг наблюдения x_i , а $\Psi(p)$ - p -квантиль стандартного нормального распределения.

Если нулевая гипотеза H_0 верна, то C_{n+m}^m последовательностей длиной N из x_i и y_i являются равновероятными. При малых n и m критические значения можно вычислить точно с помощью непосредственного перебора равновозможных последовательностей из x и y . Верхнее, $X_{1-\alpha, n, m}$, и нижнее, $X_{\alpha, n, m}$, критические значения, соответствующие уровню значимости α , при заданных n и m связаны соотношением $X_{1-\alpha} = -X_{\alpha}$. Критическая область уровня значимости α для проверки гипотезы $H_0: \Delta = 0$ против двусторонней альтернативы $H_1: \Delta \neq 0$ будет состоять из двух бесконечных полуинтервалов $(-\infty, X_{\alpha/2}]$ и $[X_{1-\alpha/2}, \infty)$. Критическая область против правосторонней альтернативы $H_1: \Delta > 0$ - из одного полуинтервала $[X_{1-\alpha}, \infty)$ и против

левосторонней альтернативы $H_1: \Delta < 0$ - также из одного полуинтервала $(-\infty, X_{\alpha}]$. При $N \rightarrow \infty$, независимо от поведения n и m по отдельности, статистика X распределена асимптотически нормально, $X \sim N(0, nm/(N-1))$.

Критерий Ван-дер-Вардена является наиболее мощным непараметрическим критерием для решения задачи двух выборок, если функции распределений $F(x)$ и $G(x)$ отличаются лишь параметром сдвига. Если обе выборки извлечены из нормальных совокупностей, то при постоянном n и $m \rightarrow \infty$ критерий Ван-дер-Вардена имеет такую же мощность, как и двухвыборочный t -критерий.

Случай связанных выборок

Пусть x_1, \dots, x_n и y_1, \dots, y_n - связанные выборки из непрерывных распределений $F(x)$ и $G(x)$, соответственно, причем $G(x) = F(x - \Delta)$. Например, каждая пара наблюдений (x_i, y_i) , $i=1, \dots, n$, принадлежит одному объекту, либо (x_i, y_i) попарно связаны тем, что условия проведения наблюдений менялись от опыта к опыту, но для каждой пары (x_i, y_i) оставались постоянными, что в практике биологического эксперимента встречается очень часто.

Обозначим через $z_i = x_i - y_i$. Тогда задача об отсутствии сдвига между $F(x)$ и $G(x)$ сводится к одновыборочной задаче, рассмотренной в 3.11.1. И для проверки гипотезы H_0 можно применить критерий знаков или критерий знаковых рангов.

Асимптотическая относительная эффективность критерия знаков для связанных выборок по отношению к двухвыборочному t -критерию для связанных выборок равна 0.637, а критерия знаковых рангов Вилкоксона - 0.955.

3.11.3. Критерии однородности

Критерии Манна - Уитни (Вилкоксона) и Ван-дер-Вардена позволяют обнаруживать лишь различия в центральных тенденциях распределений двух случайных величин. Если важно обнаружить любые расхождения в форме распределений, то пользуются **критериями однородности**, например, **двухвыборочным критерием Смирнова**. С помощью этого критерия проверяется гипотеза $H_0: F_{\xi}(x) = F_{\eta}(x)$ о том, что функции распределения $F_{\xi}(x)$ и $F_{\eta}(x)$ случайных величин ξ и η идентичны против альтернативной гипотезы $H_1: F_{\xi}(x) \neq F_{\eta}(x)$ о том, что они различны.

Статистика критерия Смирнова $D_{m,n}$ определяется как максимум модуля разности между эмпирической функцией $\hat{F}_\xi(x)$, построенной по выборке x_1, x_2, \dots, x_n , и эмпирической функцией $\hat{F}_\eta(x)$, построенной по выборке y_1, y_2, \dots, y_m

$$D_{m,n} = \max_x |\hat{F}_\xi(x) - \hat{F}_\eta(x)|$$

При справедливости гипотезы H_0 статистика $\sqrt{\frac{mn}{m+n}} D_{m,n}$ имеет асимптотическое (при $\min(m, n) \rightarrow \infty$ так, что отношение m/n остается постоянным) распределение Колмогорова. Критическая область уровня значимости α для проверки гипотезы H_0 против двусторонней альтернативы H_1 будет состоять из одного полуинтервала $[D_{m,n,1-\alpha}, \infty)$, где $D_{m,n,1-\alpha}$ - квантиль распределения статистики $D_{m,n}$ при H_0 порядка $1-\alpha$.

Заметим, что в англоязычной литературе и в ППП критерий однородности двух выборок Смирнова называют **двухвыборочным критерием Колмогорова-Смирнова**.

3.11.4. Проверка гипотезы о независимости

Пусть имеется двумерная выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ из неизвестного двумерного распределения. Причем наблюдаемые признаки могут быть как количественными, так и порядковыми. Найдем ранги, $R(x_i)$ и $R(y_i)$, в последовательностях x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_n , упорядоченных по отдельности. Мерой зависимости двух случайных величин, каждая из которых может быть как количественной, так и порядковой, является **коэффициент ранговой корреляции Спирмена**, определяемый формулой

$$r_s = 1 - 6 \sum_{i=1}^n d_i^2 / n(n^2 - 1)$$

где $d_i = R(x_i) - R(y_i)$. Как и обычный коэффициент корреляции, коэффициент ранговой корреляции r_s принимает значения $-1 \leq r_s \leq +1$, причем $r_s = +1$, когда $R(x_i) = R(y_i)$, $i=1, \dots, n$, и $r_s = -1$, когда последовательности рангов полностью противоположны, $R(x_i) = (n+1) - R(y_i)$, $i=1, \dots, n$. Коэффициент r_s используется для проверки гипотезы о независимости признаков. Нулевая гипотеза формулируется как $H_0: r_s = 0$. Чаще всего H_0 проверяется против альтернативы $H_1: r_s \neq 0$. Статистикой критерия является r_s . Если нулевая гипотеза H_0 верна, то распределение r_s симметрично относительно 0 с $M r_s = 0$ и $D r_s = 1/(n-1)$. Следовательно, множество принятия нулевой гипотезы имеет

вид $|r_s| < r_{s, 1-\alpha/2}$, где $r_{s, 1-\alpha/2}$ – верхнее критическое значение статистики критерия r_s , соответствующее уровню значимости α при заданном n .

Если верна нулевая гипотеза, случайная величина $\sqrt{n-1}r_s$ при $n \rightarrow \infty$ распределена асимптотически нормально с параметрами $(0, 1)$. При $4 \leq n \leq 10$ критические значения статистики критерия находят по таблицам точного распределения r_s при H_0 , а при $n > 10$ пользуются нормальной аппроксимацией.

Асимптотическая относительная эффективность критерия, основанного на r_s , по отношению к критерию, основанному на выборочном коэффициенте корреляции, равна 0.912.

ЛИТЕРАТУРА

1. Благовещенский Ю.Н., Самсонова В.П., Дмитриев Е.А. Непараметрические методы в почвенных исследованиях. М.: Наука, 1987.
2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983.
3. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973
4. Компьютерная биометрика. М.: Изд-во МГУ, 1990
5. Тюрин Ю.Н. Непараметрические методы статистики. М.: Знание, 1978.
6. Хан Г., Шапиро С. Статистические модели в инженерных задачах. М.: Статистика, 1980.
7. Холлендер М., Вульф Д. Непараметрические методы статистики. М.: Финансы и статистика, 1983.
8. Lilliefors H.W. The Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. J. Amer. Stat. Assn. v.62: 399-402.
9. Sokal R.R., Rohlf F.J. Biometry. The Principles and Practice of Statistics in Biological Research. N-Y, 1995.