Московский инженерно-физический институт
(Государственный университет)

Учебно-методический комплекс

по дисциплине
«Математическая статистика в радиационной физике»

Автор: ст. преподаватель каф.№1 Морозова Н.И.
(для групп Т7-01, Т7-01М)

Москва
2008

Содержание

ЛЕКЦИЯ 1.

ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ И ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Целью любого физического эксперимента является ***определение неизвестных параметров изучаемого объекта на основании полученных результатов измерений.***

Полученные результаты физических экспериментов, представляют собой сумму двух составляющих: $X = X + \sigma_{j=1}^{n} X_j$, где X - определяет истинное значение параметра изучаемого или измеряемого объекта, а X_j - вклад j-го внешнего мешающего воздействия на результат. Влияющие на точность измерения факторы можно разделить на ***две группы***.

Если группа факторов вносит в измерение ***случайные по знаку и по величине, но небольшие искажения, не смещающие результат в сторону увеличения или уменьшения***, то экспериментатор имеет дело со ***случайной ошибкой***.

Если появляются факторы, которые ***систематически смещают центр тяжести результатов измерений в какую- либо одну сторону***, то появляется ***систематическая ошибка.***

Мы будем рассматривать лишь случаи, когда математическое ожидание случайной величины совпадает со значением неизвестного параметра : M =

Стоит задача определения - неизвестного параметра исследуемого объекта по данным независимых измерений этого параметра: $x_1, x_2, \ldots x_n$.

Оценка параметра - некоторая функция от результатов измерений = (...), приближенно равная истинному значению.

Любая функция от результатов измерений называется в современной терминологии ***описательной*** или ***дескриптивной статистикой.***

Рассмотрим каждую из этих описательных статистик и ее свойства.

Пусть у нас есть случайная величина ξ и практическая реализация этой случайной величины при измерении:

$$X_1, X_2, \ldots X_n,$$

Минимум и максимум – соответственно минимальное и максимальное значения полученных при данной выборке значений случайной величины.

Среднее (оценка среднего $\bar{a}$) вычисляется как среднее результатов наблюдений

$$\bar{a} = \frac{\sigma \; = \;}{}$$

Оценку среднего часто называют ***выборочным средним***. Выборочное среднее обладает следующим ***свойством***: ***сумма отклонений наблюдаемых значений от среднего арифметического равна нулю.*** Эта статистика единственная, которая обладает этим свойством.

Дисперсия выборки или ***выборочная дисперсия*** определяется следующим образом:

$$\sigma^2 = \frac{\sum (x_i - a)^2}{n - 1}.$$

Эту описательную статистику очень часто называют ***мерой рассеяния выборки случайной величины.*** Эта статистика была впервые введена Фишером в 1882 г.

Стандартное отклонение (standard deviation) вычисляется как ***корень квадратный из выборочной дисперсии.*** На практике более удобно использовать, поскольку измеряется в тех же единицах, что и измеряемая величина, т.е. более наглядно показывает как сильно разбросаны значения выборки относительного выборочного среднего.

Содержание

ЛЕКЦИЯ 2.

ОЦЕНКИ И ИХ СВОЙСТВА

Точечные оценки и их свойства.

Для наглядности и простоты проанализируем следующие результаты независимых измерений случайной величины

ξ: 2.15, 2.32, 1.99, 2.26, 2.04, 2.19, 2.21, 2.14, 2.24.

Нужно ***найти точечную оценку*** неизвестного параметра [illegible].

Виды оценок неизвестного параметра [illegible]:

1. [illegible] = σ [illegible] / [illegible] … = [illegible]. [illegible] ***(среднее арифметическое)***

2. [illegible] = [illegible] ([illegible], [illegible], … [illegible]) = [illegible] ([illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible], [illegible]. [illegible]) = [illegible]. [illegible].

3. [illegible] = ([illegible] + [illegible]) / [illegible] = [illegible]. [illegible] ***(оценка по двум крайним точкам)***

4. [illegible] = [illegible] ς [illegible] = [illegible]. [illegible] ***(среднее геометрическое)***

На основании одних и тех же данных может быть получено по разным формулам несколько оценок, близких по величине между собой. Чтобы определить наилучшую, рассмотрим следующие свойства точечных оценок.

Состоятельность.

Оценка θ_n является ***состоятельной***, если при увеличении объема выборки n ее значение сходится к истинному значению параметра $\lim_{n \to \infty} \theta_n = \theta$

Здесь подразумевается ***выполнение двух условий***:

1. Сходимость математического ожидания оценки к истинному значению параметра:
   $\lim_{n \to \infty} M[\theta_n] = \theta$.
2. Сходимость дисперсии оценки к 0:
   $\lim_{n \to \infty} D[\theta_n] = 0$.

Если объем выборки мал, то не обязательно требовать, чтобы используемые оценки были состоятельны. При малых n сходимость не наблюдается для любых оценок.

***Несмещенность*.**

Оценка $\hat{\alpha}_n$ является ***несмещенной***, если для любого объема выборки n ее математическое ожидание совпадает с истинным значением параметра:

$$M[\hat{\alpha}_n] = \alpha$$

Эффективность.

Оценка $\hat{\alpha}_n$ называется ***эффективной для заданного объема выборки*** если ***ее дисперсия минимальна среди других видов оценок для заданного n.***

Оценка $\hat{\theta}_n$ называется ***ассимптотически эффективной, если она эффективна при*** $n \to \infty$.

Таким образом, для того чтобы исследовать ту или иную оценку на эффективность ***нужно иметь несколько видов описательных статистик***, а ***эффективная оценка*** $\hat{\theta}_{эф}$ будет выбираться из условия:

$$\frac{D[\hat{\theta}_{эф}]}{D[\hat{\theta}_n]} \leq 1$$

Кроме того, как будет показано далее для нахождения минимума дисперсии необходима информация о виде распределения случайной величины, согласно неравенства Крамера-Рао.

Достаточность. Для определения достаточности оценки θ_n ***необходимо знать закон распределения случайной величины ξ.***

Пусть ***p(x, θ) - плотность распределения***, зависящая от параметра θ. Тогда ***вероятность появления выборки*** $x_1, x_2, \ldots x_n$ определится соотношением:

$$P(x_1, x_2 \ldots x_n) = P(x_1, x_2 \ldots x_n) \cdot P(x_1, x_2 \ldots x_n) =$$

$$= \prod_{i=1}^{n} p(x_i, \theta) \cdot \prod_{i=1}^{n} p(x_i)$$

Второй сомножитель в этом выражении не зависит от параметра θ и не оказывает влияния на величину вероятности. Первый же сомножитель зависит и от параметра θ, и от исходных данных и называется ***функцией правдоподобия*** $L(x_1, x_2, \ldots x_n, \theta)$.

Оценка θ_n называется ***достаточной, если*** функцию правдоподобия можно представить в виде двух сомножителей:

$$L(x_1, x_2, \ldots x_n, \theta) = L_1(\theta_n, \theta) * L_2(x_1, x_2, \ldots, x_n)$$

Пример. Попытаемся ***исследовать на достаточность оценку в виде среднеарифметического*** $\bar{x} = \frac{\sum x_i}{n}$. Далее предположим, что случайная величина ξ. распределена по нормальному закону $N(a,1)$. ***Функция правдоподобия*** для выборки $x_1, x_2, \ldots, x_n$ будет выглядеть следующим образом:

$$L(x_1, x_2 \ldots x_n, a) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - a)^2}{2}\right),$$

тогда ***для оценки*** $\bar{x}$, ***получаем следующую функцию правдоподобия***:

$$L(x_1, x_2 \ldots x_n, \bar{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(-2x_i\bar{x} + \bar{x}^2)}{2}\right) \exp\left(-\frac{\sum_{i=1}^{n}x_i^2}{2}\right)$$

Принимая во внимание что $\frac{\sigma \sum_{i=1}^{n}}{n}$ есть $x_{ср}$, первую экспоненту можно представить в следующем виде: $\exp\left[-\frac{n x_{ср}^2}{2} + \ldots\right]$ и в результате получаем, что ***функция правдоподобия для случайной величины распределенной по нормальному закону N(x,1) с учетом того, что для оценивания среднего используем среднеарифметическую оценку можно представить в виде двух сомножителей:***

$$x_1, x_2 \ldots x_n, x = \frac{1}{\sqrt{2\pi}}^{n} \exp\left[-\frac{n x_{ср}^2}{2} + x_{ср} \sum_{i=1}^{n} - \frac{\sigma \sum_{i=1}^{n} x_i^2}{2}\right]$$

таким образом среднее арифметическое является достаточной оценкой в случае если случайная величина распределена по нормальному закону.

Если аналогичные выкладки провести для оценки из примера 1.1 $x_{ср} = \frac{x_{max} + x_{min}}{2}$, то достаточно легко можно убедиться что эта оценка не является достаточной.

Робастность.

Оценка, ***свойства которой не зависят от конкретного вида распределения***, называется ***устойчивой или робастной***. Робастные оценки устойчивы к выбросам, присутствие которых в выборках обусловлено наличием так называемой «шумовой» составляющей и наложением ее распределения на основное.

На ***примере*** нашей выборки и выбранных оценок это свойство можно достаточно просто пояснить следующим образом. Заменим значение 2.24 на значение 10.23 (ситуация, когда имеет место поломка прибора или кратковременный скачок напряжения в сети). И среди результатов выборки появляется значение, резко отличающееся от других – так называемый ***«выброс»***. Из-за этого оценка изменилась и стала равной 3.06, но оценка осталась равной 2.19, т.е. не изменилась. Таким образом вторая оценка является ***более устойчивой*** по отношению к другим видам статистик из нашего примера.

Вычислительная простота.

Это свойство точечных оценок также имеет право на существование, но с развитием персональных компьютеров это свойство оценок уходит на задний план. ***Из приведенных в примере оценок наиболее простой с точки зрения вычислений является***  .

В итоге, понятно что найти наилучшую оценку, удовлетворяющую всем перечисленным свойствам одновременно нельзя.

Надежность оценок. (Понятие интервального оценивания.)

Точечные оценки неизвестных параметров, как правило, являются первоначальным этапом статистического оценивания результатов измерений. Поскольку далее результат конкретного эксперимента представляется в виде ▨ ± ▨ возникает вопрос о том, какова вероятность нахождения истинного значения ▨ в этом интервале. В этой связи вводится понятие ***надежности оценок***.

Надежность оценки определяется как ***вероятность нахождения истинного значения в указанном интервале с заданной вероятностью.***

Пусть задана ***выборка случайной величины*** : $x_1, x_2,\ldots x_n$ ***распределенная с плотностью вероятности*** $p(x, a_1, a_2,\ldots,a_k)$ Стоит ***задача*** оценивания неизвестных параметров распределения.

Интервальной оценкой неизвестного параметра называется ***интервал (a, b) такой, что истинное значение попадает в этот интервал с заданной вероятностью (1-)***. Вспоминаем, приведенные в первой лекции ***обозначения и определения*** связанные с интервальным оцениванием:

(a, b) - интервальная оценка параметра ;

a – нижний доверительный предел;

b – верхний доверительный предел;

– доверительный уровень или уровень значимости;

(1-e) – доверительная вероятность.

В некоторых прикладных задача x наоборот требуется найти для заданной интервальной оценки неизвестную доверительную вероятность. Более подробно для нескольких частных случаев вопросы, связанные с интервальным оценивание будут рассмотрены далее в лекции 7.

Содержание

ЛЕКЦИЯ 3.

ОЦЕНИВАНИЕ ПРИ ОТСУТСТВИИ ИНФОРМАЦИИ О ВИДЕ РАСПРЕДЕЛЕНИЯ

Метод подстановки.

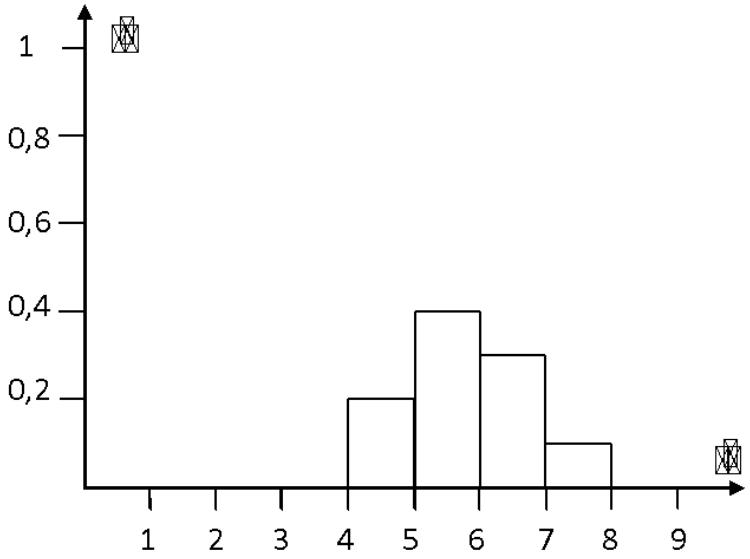
Поставим ***задачу*** определения оценок среднего и дисперсии по данным независимых измерений случайной ***величины*** x_1, $x_2,\ldots x_n$. Из теории вероятности известны ***определения среднего*** $\bar{x}$ ***и дисперсии*** $D\ (\sigma^2)$ **:**

$$\bar{x} = \int_{-\infty}^{\infty} x\,p(x)\,dx \qquad D = \sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 p(x)\,dx \qquad (Л3.1)$$

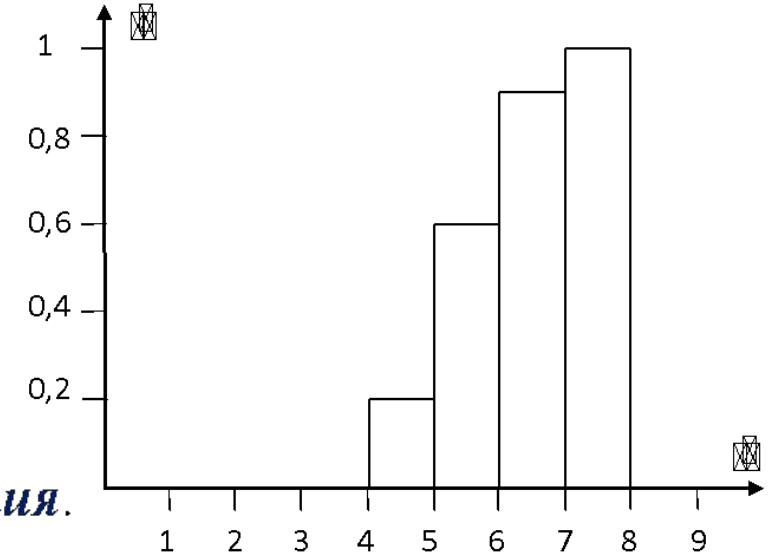
где $p(x)$ - плотность вероятности распределения случайной величины x

Принцип подстановки формулируется следующим образом. ***Для нахождения оценки величины*** θ, ***которая зависит от функции распределения случайной величины или плотности распределения, в формулу описывающую эту зависимость вместо неизвестных*** $p(x)$ ***или*** $F(x)$ ***подставляются их эмпирические аналоги.***

Эмпирическая функция распределения $F(x)$ в каждой точке x равна ***отношению числа выборочных значений меньших x к размеру выборки n*** и ***представляет собой ступенчатую функцию,*** имеющую значения от 0 до 1 и размер ступенек кратен 1/n, и определяется числом выборочных значений попавших в выделенный интервал.



Эмпирическая плотность распределения $f(x)$ равна ***производной эмпирической функции распределения***. На рис. 1 приведены эмпирические плотность и функция распределения для следующей выборки:



Если ввести определение ***единичной функции U(x)*** и ***дельта-функции*** $\delta(x)$:

$$U(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

(Л3.2)

То ***эмпирические функция и плотность распределения*** запишутся:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} U(x - x_i) \qquad p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$$

(Л3.3)

Воспользуемся ***принципом подстановки*** для получения ***оценок среднего и дисперсии***:

$$\bar{x}_n = \int_{-\infty}^{\infty} x\, p_n(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\, dx = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (Л3.4)$$

$$s_n^2 = \int_{-\infty}^{\infty} (x - \bar{x}_n)^2 p_n(x)\, dx = \int_{-\infty}^{\infty} (x - \bar{x}_n)^2 \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\, dx$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n} \qquad (Л3.5)$$

Таким образом, мы получили, что хорошо известные оценки ***среднее арифметическое и дисперсия являются оценками, полученными методом подстановки.***

Свойства оценок среднего и дисперсии, полученных этим методом.

Проверим - является ли оценка среднего ***состоятельной и несмещенной.***

Согласно определению ***состоятельности*** оценки мы найдем математическое ожидание и дисперсию этой оценки. При этом считаем, что все результаты измерений x_i ***независимы*** и ***распределены по тому же закону, что и сама случайная величина*** , другими словами $M[x_i] =$, ***а*** $D[x_i] = M[(x_i -)^2] = \sigma^2$.

Прикладывая оператор среднего и дисперсии к оценке (средне арифметической), и вспоминая свойства оператора математического ожидания и дисперсии получаем:

$$M[\quad] = M\left[\frac{}{}\sigma_{=}\quad\right] = \frac{}{}\sigma_{=}\quad [x_i] = \frac{}{} = \quad \text{(ЛЗ.6)}$$

$$D[\quad] = D\left[\frac{}{}\sigma_{=}\quad\right] = \frac{}{}\sigma_{=}\quad [\quad] = \frac{}{} = \frac{}{} \quad \text{(ЛЗ.7)}$$

Из (ЛЗ.5) следует, что оценка среднего, полученная на основе метода подстановки, является ***несмещенной и состоятельной***.

Теперь проверим ***состоятельность и несмещенность оценки дисперсии,*** которая получена на основе метода подстановки. Для начала найдем математическое ожидание этой оценки.

$$M[\hat{\sigma}^2] = M\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n} M[(x_i - \bar{x})^2]$$

Рассмотрим отдельно каждое i-ое слагаемое суммы:

$$M[(x_i - \bar{x})^2] = M[(x_i - a + a - \bar{x})^2] = M[(x_i - a)^2] - M[(\bar{x} - a)^2] + M[(x_i - a)(\bar{x} - a)] = \ldots = \sigma^2 - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Подставляя полученное выражение для каждого слагаемого в сумму (ЛЗ.7) получим:

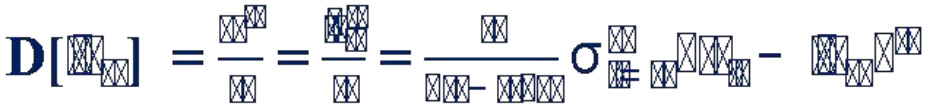
$$M[\sigma^2] = \frac{1}{n}\sum_{i=1}^{n}\frac{n-1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2. \qquad (ЛЗ.8)$$

Этот результат говорит о том, что ***для оценки дисперсии, полученной методом подстановки***, ***выполняется первое условие состоятельности*** ($\lim_{n\to\infty} M[\hat{\sigma}^2] = \sigma^2$). Но условие ***несмещенности не выполняется. Второе условие состоятельности*** ($\lim_{n\to\infty} M[\hat{\sigma}^2] = \sigma^2$) также выполняется.

Для того, ***чтобы избавиться от смещения в оценке для дисперсии*** необходимо в выражение Л3.5 вместо n подставить (n-1). Тогда получим ***выражение для несмещенной и состоятельной оценки*** дисперсии:

(Л3.9)

Еще раз воспользуемся ***методом подстановки*** для нахождения квадрата погрешности оценки среднего(оценка дисперсии оценки среднего):



(Л3.10)

Очевидно, что ***погрешность оценки среднего также обладает своей погрешностью.*** В теории математической статистики доказывается, что ***отношение погрешности погрешности среднего к погрешности среднего есть величина обратная пропорциональная*** Тем обстоятельством, что каждая вычисленная погрешность в свою очередь обладает своей погрешностью, и обусловлен принцип отбрасывания знаков записи результатов физического эксперимента в виде 

Содержание

ЛЕКЦИЯ 4.

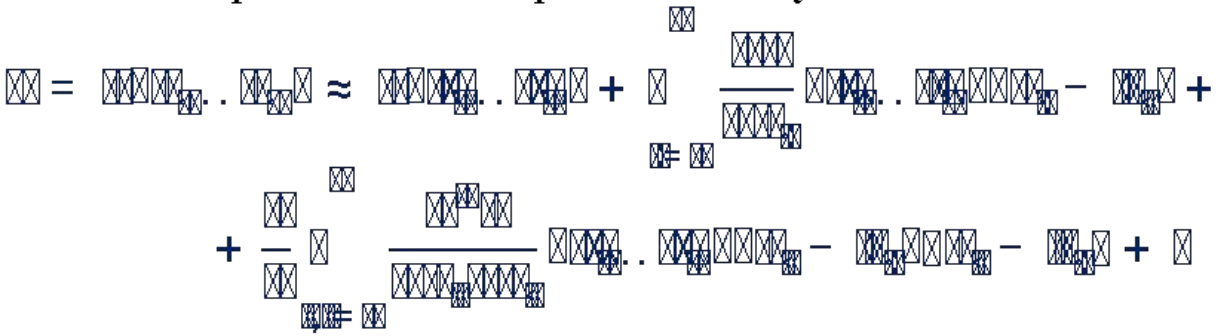
ПЕРЕНОС ОШИБОК

Л4.1 Погрешность косвенных измерений (Одномерный случай).

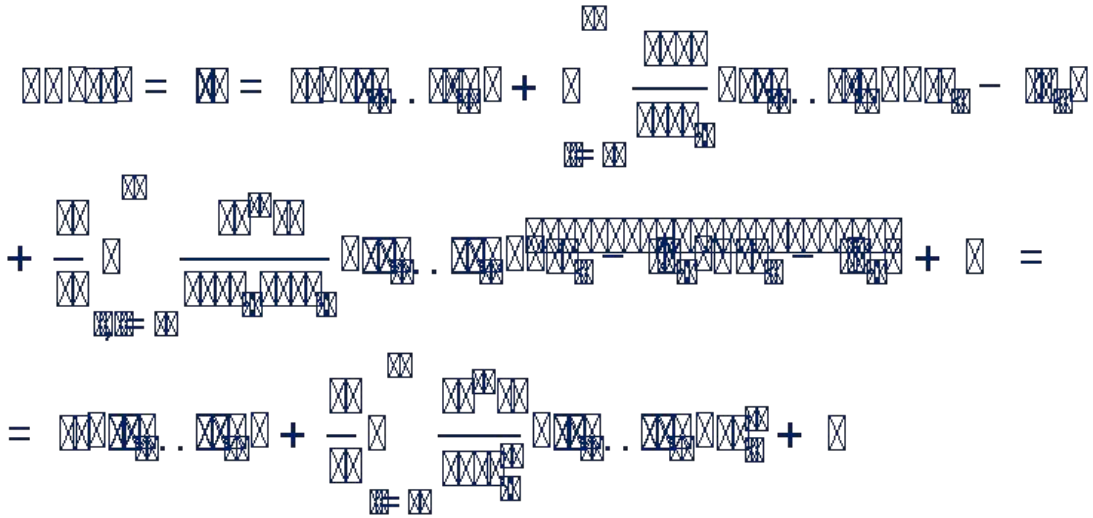
Пусть ***есть некоторая случайная величина Y, которая является функцией случайных величин*** $x_1, x_2, .. x_n$. Стоит ***задача найти оценку среднего величины Y.***

Например, рассмотрим случай $Y = x^2$.

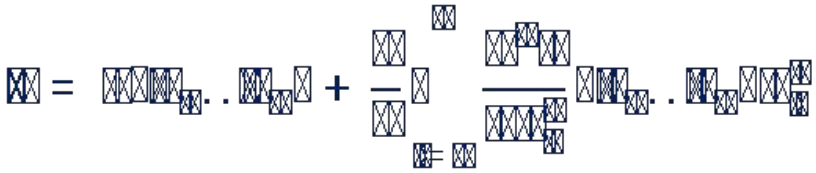
Как правильно построить оценку величины Y.



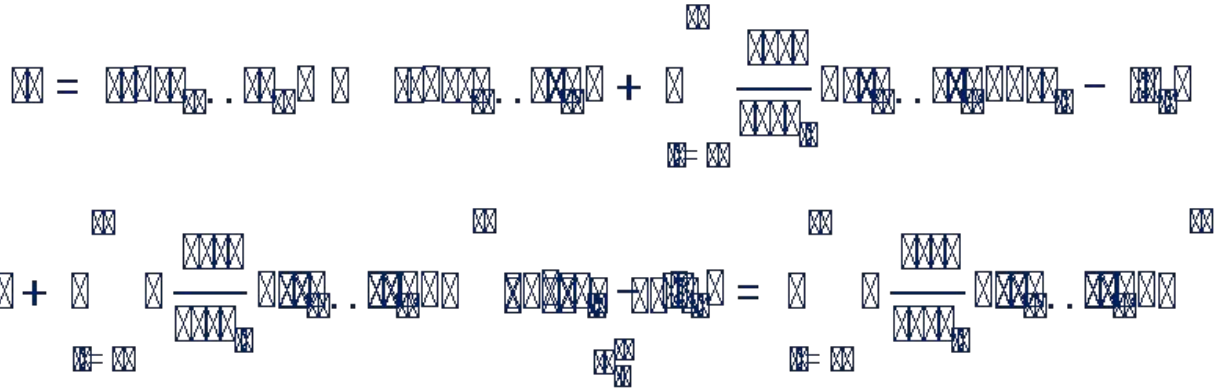
Тогда математическое ожидание *y* будет равно:



Воспользовавшись ***методом подстановки***, можем подставить в разложение оценки $\theta_1 \dots \theta_k$:



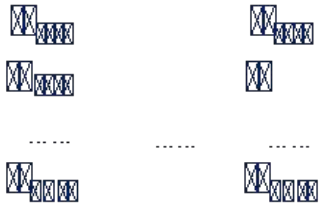
Данная формула является ***более точной*** по сравнению с ранее предложенной и ***содержит квадратичную поправку***. В случае, если значение этой поправки мало, простая подстановка на начальном этапе может быть удовлетворительной. Ограничимся теперь линейным членом разложения ряда и вычислим ***дисперсию*** D:



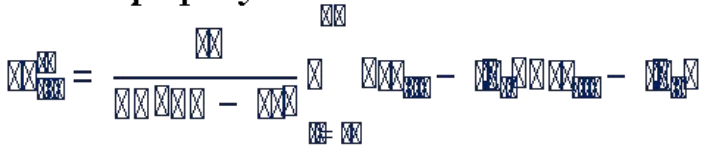
Полученная формула также ***приближенная***, так как нелинейные члены ряда отброшены. Но т.к. θ_0 уже определена неточно, то учет членов ряда более высоких порядков не приведет к существенному увеличению точности.

Л4.2. Формула переноса в случае зависимых измерений.

Рассмотрим более ***общий случай***. Пусть ***случайные величины*** .. ***не являются зависимыми***, в этом случае должна быть известна = - ***ковариационная матрица*** .. . Эта матрица, также как и оценки , … может быть получена на основе экспериментальных данных. Например, если оценки , … получены на основе выборок одинакового размера:

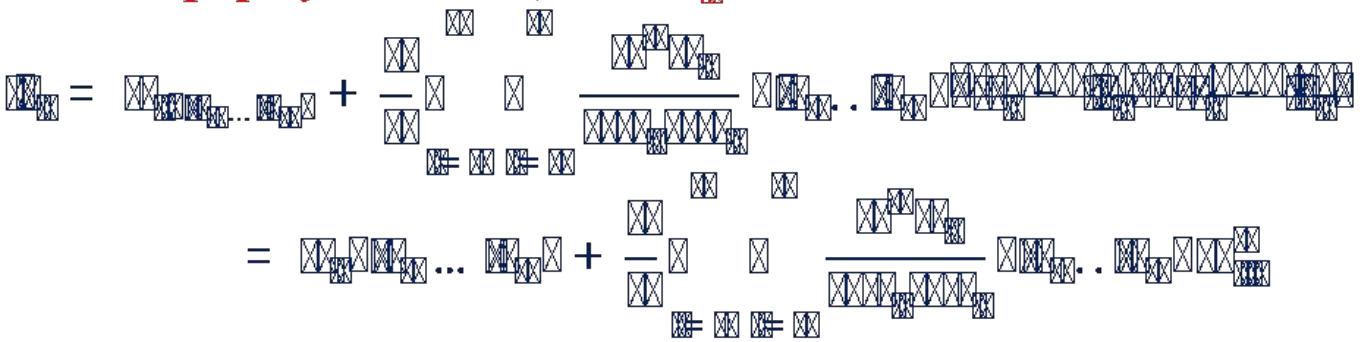


где первый индекс указывает на номер измерения, а второй – на номер случайной величины, то ***оценка (ij)-ого элемента*** ковариантной матрицы определиться по формуле:

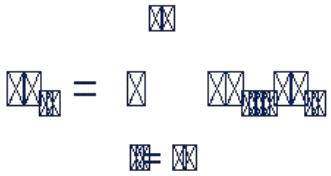


где = σ , = . . - ***оценка среднего для каждого столбца***. Стоит ***задача*** определения **и** - ***ковариационной матрицы*** … , где = ... зависимые функции n случайных переменных.

Разложение в ряд Тейлора в окрестности точки (...) и последующее применение оператора математического ожидания приводит к ***формуле для оценки*** ... аналогичной:

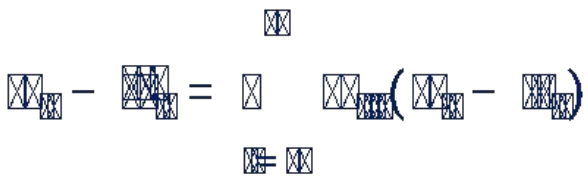


Для нахождения ... рассмотрим вначале ***частный случай***. Пусть функция ... = ..., ***j=1,k, линейны относительно*** ...:

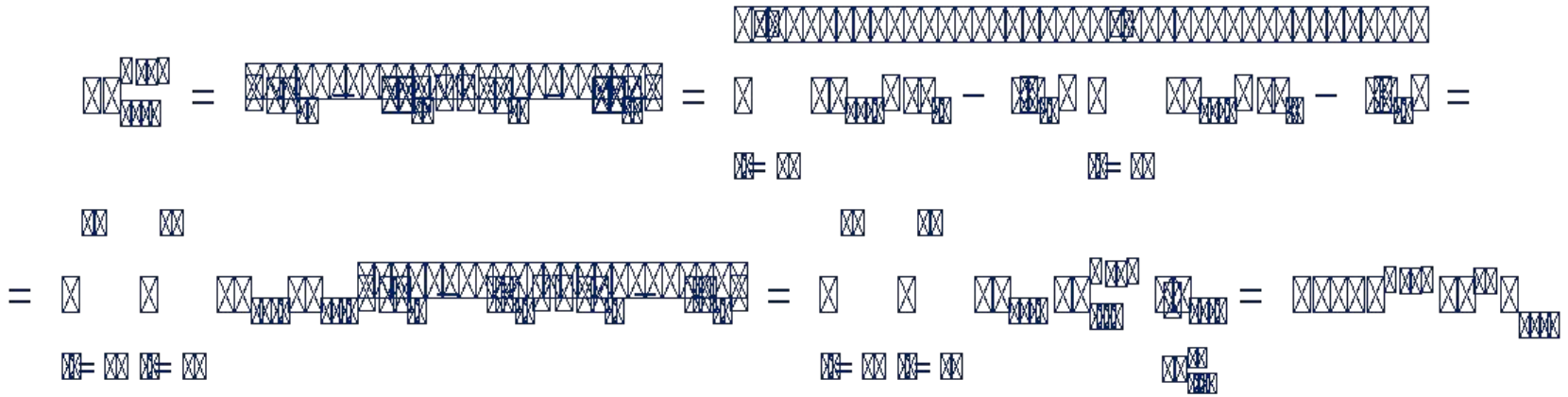


Применим оператор математического ожидания к обеим частям равенства: ... = σ ...

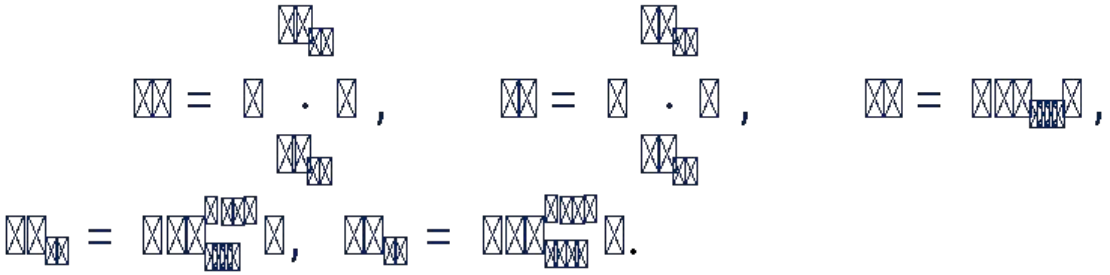
Вычтем из первого равенства второе. Получим:



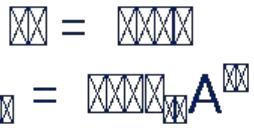
Подставим в формулу для элемента ковариационной матрицы σ_{ij}, который по определению равен:



Введем ***векторные обозначения***:



Тогда в векторной форме:



Откуда, пользуясь принципом подстановки, получаем:

Обратимся к ***общему случаю***, когда функции $y_m = f_m(x_1, x_2, \ldots, x_n)$, $x_i = \varphi_i(z)$, ***нелинейны***. Разложим каждую функцию в ряд Тейлора вокруг оценки среднего (в идеале нужно раскладывать вокруг истинного среднего, но оно все равно неизвестно, поэтому еще раз применим ***принцип подстановки***) с точностью до линейных членов разложения:

$$y_m = f_m(x_1, x_2, \ldots, x_n) + \frac{\partial f_m}{\partial x_1}(x_1 - x_1) + \frac{\partial f_m}{\partial x_2}(x_2 - x_2) + \ldots + \frac{\partial f_m}{\partial x_n}(x_n - x_n)$$

Для такой линейной функции можно применить результат, при этом матрица A определиться следующим образом:

$$a_{mn} = \frac{\partial f_m}{\partial x_n}(x_1, x_2, \ldots, x_n)$$

Сдвиг на const не влияет на величину дисперсии, т.к. D[const]=0.

Итак, ***в общем случае получим***:

$$D_y = A D_x A^T,$$

где элементы матрицы определяются из выражения

$$a_{mn} = \frac{\partial f_m}{\partial x_n}(x_1, x_2, \ldots, x_n)$$

Содержание

ЛЕКЦИЯ 5.

МЕТОДЫ ПОЛУЧЕНИЯ ОЦЕНОК ПРИ ИЗВЕСТНОМ ВИДЕ РАСПРЕДЕЛЕНИЯ

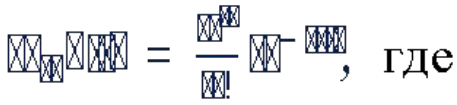
Л5.1. Виды распределений случайных величин, встречающиеся в радиационной физике

1. ***Случайная величина*** $\xi = \xi_0 + \sigma \sum_{i=1}^{n} \xi_i$, ($n$ – велико, ξ_i - случайные факторы) – есть ***результат действия большого числа случайных факторов на исходную постоянную величину.*** В этом случае ***применима центральная предельная теорема,*** и величина ξ распределена нормально, т.е. $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$, где a, σ – неизвестные параметры нормального распределения.

2. ***Имеется N образцов, для которых событие A имеет место с вероятностью p (или не имеет места с вероятностью 1-p)***. В эксперименте ***подсчитывается случайная величина ν – количество исходов в эксперименте, когда событие A имело место.*** В этом случае ***действует схема независимых испытаний или схема Бернулли***, при этом распределение случайной величины будет биноминальным: $P_m = C_n^m p^m q^{n-m} = \frac{n!}{m!(n-m)!} p^m q^{n-m}$, где p – неизвестный параметр биноминального распределения.

3. Эксперимент состоит в том, что ***в течение некоторого фиксированного промежутка времени t подсчитывается число появлений некоторого события*** (например, подсчет гамма-квантов от радиоактивного источника детектирующей системой). Если можно считать, что

- ***для любых двух неперекрывающихся промежутков времени число зарегистрированных гамма-квантов***  ***за время***  ***и число зарегистрированных гамма-квантов***  ***за время***  ***независимы;***
- ***вероятность одновременной регистрация двух гамма-квантов мала***

то ***случайное число зарегистрированных гамма-квантов за фиксированный промежуток времени подчиняется закону Пуассона:***  , где λ – параметр распределения.

4. ***Если число зарегистрированных гамма-квантов за фиксированный промежуток времени будет больше 20-30***, то распределение зарегистрированных гамма-квантов за фиксированный интервал времени будет ближе к ***нормальному распределению***.

Таким образом, ***задача статистической обработки данных*** в этом случае будет ***заключаться в нахождении оценок параметров априорно известных видов распределений.***

Неравенство Крамера-Рао.

Рассмотрим ***неравенство***, которое ***позволяет установить нижнюю границу дисперсии оценки случайной величины с известным видом распределения.***

Для любого метода оценивания всегда существует предел точности. Дисперсия оценки, найденная любым методом, ***не может быть меньше значения, установленного этим неравенством.***

Вспомним, что ***оценка, которая имеет при заданном размере выборки наименьшую дисперсию*** называется ***эффективной***. (таким образом, зная вид распределения всегда можно задать нижний предел оценивания для дисперсии).

Пусть измерены $x_1, x_2, \ldots x_n$ - ***независимые реализации случайной величины*** ξ. Считаем, что ***случайная величина*** ξ ***распределена в общем случае по закону*** $p(x, a_1, a_2, \ldots a_k)$, где $a_1, a_2, \ldots a_к$ – неизвестные параметры распределения.

Для того, чтобы упростить дальнейшее изложение ***рассмотрим простейший случай***, когда ***закон распределения случайной величины зависит только от одного параметра***, а затем обобщим полученный результат.

Точечная оценка параметра $\theta = f(x_1, x_2, \ldots x_n)$. Закон распределения каждого выборочного значения, тот же, что и у случайной величины ξ. Если все выборочные значения независимы, то закон распределения оценки θ определится произведением:

$$P_{\theta} = \prod_{i=1}^{n} p(x_i, a) \qquad \textbf{(Л4.1)}$$

Будем предполагать, что оценка θ - ***несмещенная, т.е.*** $M\theta = a$. Для закона распределения p(x, a) выполняется следующее ***условие нормировки***:

$$\int_{-\infty}^{+\infty} p(x, a)\, dx = 1 \qquad \text{(Л4.2)}$$

Продифференцируем обе части (Л4.2) по a. Получим:

$$\int_{-\infty}^{+\infty} p'_a(x, a)\, dx = 0 \qquad \text{(Л4.3)}$$

Поделим и умножим обе части равенства ((Л4.3) на p(x,a).

$$\int_{-\infty}^{+\infty} \frac{p(x,a)}{p(x,a)}\, p'_a(x, a)\, dx = 0$$

Получим ***для случая одного неизвестного параметра распределения неравенство Крамера-Рао в виде:***

$$D_{\theta} \geq \frac{1}{\int_{-\infty}^{+\infty} \frac{1}{p(x,a)} \left[p'_a(x,a)\right]^2 dx} \qquad \text{(Л4.4)}$$

Пример. Рассмотрим, какой вид приобретает неравенство Крамера-Рао для случайной величины, нормально распределенной. Пусть:

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

$$\ln(f(x, a)) = -\frac{(x-a)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)$$

$$\frac{\partial \ln(f(x, a))}{\partial a} = \frac{x-a}{\sigma^2}$$

$$I = \int_{-\infty}^{+\infty} \left(\frac{x-a}{\sigma^2}\right)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sigma^2}$$

Таким образом, ***каким бы способом не была найдена оценка, ее дисперсия не может быть меньше величины, определяемой неравенством:***

$$D[\hat{a}] \geq \frac{\sigma^2}{n} \qquad \text{(Л4.5)}$$

Ранее методом подстановки была получена ***оценка среднего и дисперсия оценки среднего*** $D[\hat{a}] = D\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{\sigma^2}{n}$. На основании сравнения (Л4.5) и этой оценки дисперсии оценки среднего можно сделать вывод о том, что ***для нормального распределения эта оценка будет эффективной.***

Содержание

ЛЕКЦИЯ 6.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Постановка задачи

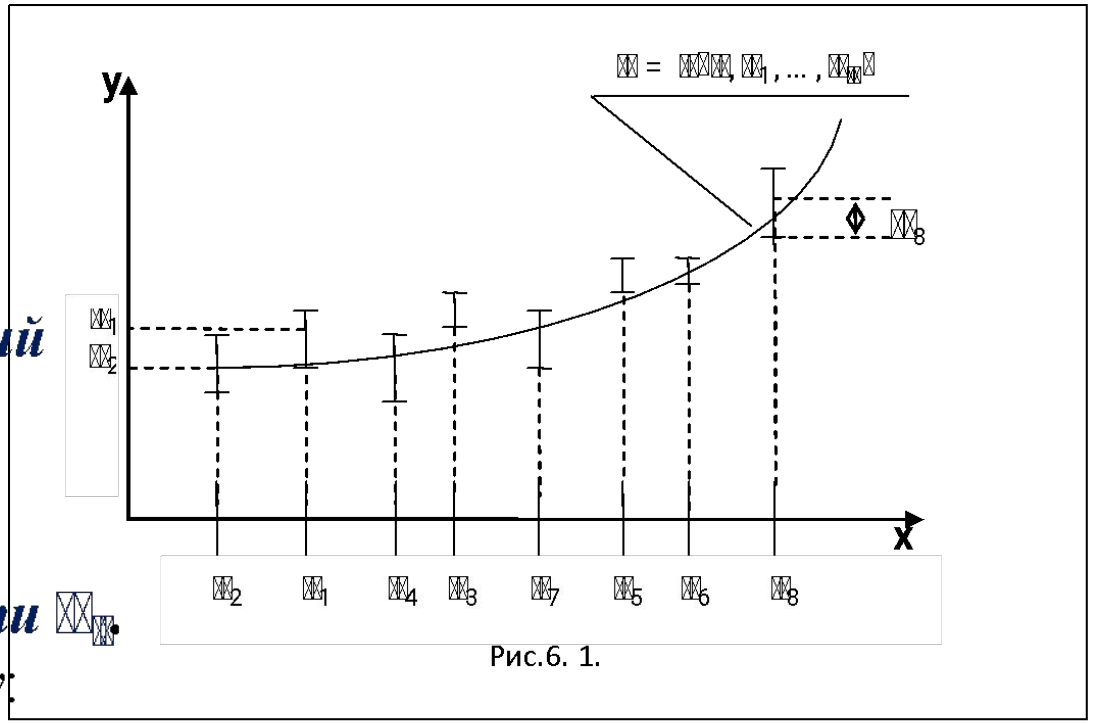


Рис.6. 1.

Имеются ***пары измерений*** (x_i, y_i), $i = 1, n$,
при этом x_i измерены
без погрешности,
а y_i ***имеют погрешности*** ε_i.
Зависимость между x и y:
$y = f(x, a_1, a_2, \dots a_m)$.

Задача состоит в отыскании оценок неизвестных параметров $a_1, \dots a_m$ ***и их погрешностей***. Аналитическая зависимость может быть известна из теоретических соображений или выбрана из практических соображений. Эту постановку задачи можно легко распространить и на многомерный случай, когда функция $y = f(x, x_1, \dots x_m)$ зависит не от одной переменной x, а от нескольких переменных.

Пусть ***случайные величины*** y_i ***подчиняются нормальному закону с параметрами*** $a, a_1, \ldots a_m, \sigma$. В этом случае задача отыскания неизвестных параметров может быть решена ***методом максимального правдоподобия.*** При этом максимизируется функция правдоподобия:

$$L(a, a_1, \ldots a_m) = \varsigma \frac{1}{\xi \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} p_i (y_i - f(x_i, a_1, \ldots a_m))^2}$$

(Л6.1)

где $p_i = \frac{1}{\sigma_i^2}$ ***статистический вес i-го измерения***.

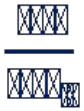
Очевидно, ***максимум функции правдоподобия соответствует минимуму функционала***:

$$\Phi(a, a_1, \ldots a_m) = \sum_{i=1}^{n} p_i (y_i - f(x_i, a_1, \ldots a_m))^2$$

(Л6.2)

Этот функционал представляет собой сумму взвешенных квадратов отклонений измеренных значений y_i от их ожидаемых значений. При этом, чем больше погрешность измерения, тем меньше вклад этого измерения в сумму.

Оценки параметров  находятся из ***условия минимума суммы квадратов уклонений теоретических значений от экспериментальных в точках***  ***взятых с весами обратно пропорциональными квадрату среднеквадратичной погрешности измерений.*** В результате ***оценки параметров*** могут быть найдены как решение системы уравнений:

 ***= 0, при j =1,…, k*** (Л6.3)

В этом суть метода, получившего в дальнейшем название ***метода наименьших квадратов (МНК).***

Метод был впервые предложен Лежандром в 1805 году, дальнейшее развитие получил в работах Гаусса в 1809 году. При выводе метода было сделано предположение, что измеренные величины распределены по нормальному закону. Как показала практика и теория, этот ***метод может быть применим и для других видов распределения, но только в случае нормального распределения измеренных величин***, оценки, полученные МНК совпадают с оценками полученными методом максимального правдоподобия, т.е. обладают свойством асимптотической эффективности.

Метод наименьших квадратов для случая линейной зависимости от неизвестных параметров.

Большое количество задач, решаемых МНК, может быть сведено к случаю, когда ***функция*** $y = y(x_1, x_2, \ldots x_n)$, ***зависит от неизвестных параметров линейно***. Функцию можно представить в виде:

$$y(x_1, x_2, \ldots x_n) = \sum_{j=1}^{k} a_j x_j f_j(x_i) \qquad (Л6.4)$$

Введем следующее обозначение:

$$A_{ij} = x_j(x_i) \qquad (Л6.5)$$

Матрица A размерности nxk, элементы которой определяются выражением (Л6.5), называется конструкционной матрицей. Так, например, если стоит задача аппроксимации зависимости полиномом второй степени, а число замеров равно пяти, то конструкционная матрица A размерности 5x3 запишется следующим образом.

$$\begin{vmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{vmatrix}$$

С учетом введенных обозначений (Л6.4) и (Л6.5) ***минимизируемый функционал*** перепишется в виде:

$$S(a_0, a_1, \dots a_m) = \sum_{i=1}^{n} w_i y_i^2 - \sum_{i=1}^{n} w_i y_i \left(\sum_{j=0}^{m} a_j x_i^j \right)^2 \qquad \text{(Л6.6)}$$

Минимум этого функционала находится из системы уравнений:

$$\frac{\partial}{\partial a_k} \left(\sum_{i=1}^{n} w_i y_i \left(\sum_{j=0}^{m} a_j x_i^j \right)^2 \right) = \sum_{i=1}^{n} w_i y_i x_i^k \qquad \text{(Л6.7)}$$

Линейный метод наименьших квадратов очень удобно представлять в матричной записи. Введем следующие ***обозначения*** векторов и матриц:

$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \qquad a = \begin{pmatrix} a_0 \\ \dots \\ a_m \end{pmatrix} \qquad A = a_{ij} = x_i^j \qquad W = \begin{pmatrix} w_1 & & 0 \\ & \dots & \\ 0 & & w_n \end{pmatrix}$$

(Л6.7)

Тогда систему (Л6.7) можно переписать в следующем виде:

$$A^T W A a = A^T W Y \qquad \text{(Л6.8)}$$

Решение системы (Л6.8) запишется в виде:

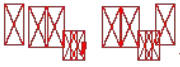
$$a = (A^T W A)^{-1} A^T W Y \qquad \text{(Л6.9)}$$

Ковариационную матрицу B_a для нахождения погрешности параметров a ***найдем с помощью матричной записи формулы переноса ошибок***. Промежуточные выкладки опускаем и получаем:

$$B_a = (A^T W A)^{-1} \qquad \text{(Л6.10)}$$

Теорема Гаусса- Маркова.

Теорема позволяет оценить эффективность оценок, полученных с помощью линейного метода.

Формулируется следующим образом, ***пары измерений***  , ***при этом*** ***измерены без погрешности (т.е. не являются случайными величинами), а***  ***имеют погрешности*** Известно, что ***зависимость между x и y описывается формулой:*** . При определении оценок неизвестных параметров аппроксимирующих ***зависимостей метод наименьших квадратов является оптимальным в случае линейной зависимости от параметров при любом виде распределения.***

Метод наименьших квадратов в нелинейном случае.

Метод наименьших квадратов может ***применяться и в случае, когда функция*** $y = f(x, a_1, \dots a_m)$ ***в функционале (Л6.2) зависит от неизвестных параметров нелинейно***.

Существует огромное число методов, позволяющих решить ***задачу минимизации функционала*** (Л6.2). В частности, иногда встречается случай, когда ***можно подобрать преобразование обеих частей равенства*** $y = f(x, a_1, \dots a_m)$ ***таким образом, что с вводом новых переменных*** $Y = \varphi(y)$, $X = \psi(x)$, $b_0 = \chi_0(a_0, \dots a_m)$, $b_1 = \chi_1(a_0, \dots a_m)$, ... $b_m = \chi_m(a_0, \dots a_m)$ ***вновь полученное равенство*** $Y = g(X, b_1, \dots b_m)$ будет ***содержать функцию g, зависящую от параметров линейно.*** Для некоторых частных функций f при минимизации функционала (Л6.2) ***удается получить нелинейную систему уравнений***, аналогичную (Л6.2), которая может быть ***решена аналитически***.

Однако, в ряде случаев нелинейности функции нельзя получить решение аналитически. В этом случае применяются специальные численные методы.

Метод линеаризации Ньютона-Рафсона.

Достоинство этого метода заключается в том, что он является универсальным и не зависит от конкретного вида функции.

Пусть так же, как и ранее ***имеются пары независимых измерений*** x_i, y_i, $i = 1, n$, при этом x_i **измерены без погрешности, а** y_i **имеют погрешность** ε_i. Известно, что зависимость между x и y описывается нелинейной относительно параметров функцией $y = f(x, a_1, \ldots a_m)$.

1. Выбирается ***начальное приближение*** $a_1^0, \ldots a_m^0$, на основе конкретного вида функции f. Например, для функции

$$f(x, a_1, a_2) = \frac{a_1}{\xi\sqrt{a_2}} e^{-\frac{x}{a_2}} - \frac{a_1}{\xi\sqrt{a_2}} x - a_2^{a_1}$$

Начальное приближение a_1^0 - ***площадь под ломаной кривой***, соединяющей точки x_i, y_i, $i = 1, n$. Начальное приближение $a_2^0 = \frac{\Delta_{1/2}}{2,355}$, где $\Delta_{1/2}$ – ширина на половине максимальной высоты над осью X графика ломаной соединяющей точки x_i, y_i, $i = 1, n$. Начальное приближения $a_3^0 = x_k$, соответствующее максимальному значению y_i, $i = 1, n$.

2. Функция $y = f(x_1, x_2, \dots x_n)$ ***раскладывается в ряд Тейлора в окрестности точки*** $x_1^0, \dots x_n^0$ с точностью до линейных членов разложения:

$$f(x_1, x_2, \dots x_n) \approx f(x_1^0, x_2^0, \dots x_n^0) + \sum_{i=1}^{n} \frac{\partial f(x_1^0, x_2^0, \dots x_n^0)}{\partial x_i} \Delta x_i,$$

где $\Delta x_i = x_i - x_i^0$.

3.***Полученная функция линейна по отношению к параметрам*** Δx_i. Следовательно, можно применить ***метод наименьших квадратов*** для линейного случая. ***Матричные обозначения***, аналогичные (Л6.8), будут выглядеть следующим образом:

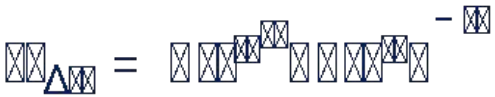
$$\Delta Y = \begin{pmatrix} y_1 - f(x_1^0, x_2^0, \dots x_n^0) \\ \dots \\ y_m - f(x_1^0, x_2^0, \dots x_n^0) \end{pmatrix} \qquad \Delta Y = \begin{pmatrix} y_1 - y_1^0 \\ \dots \\ y_m - y_m^0 \end{pmatrix}$$

$$A = \left(\frac{\partial f(x_1^0, x_2^0, \dots x_n^0)}{\partial x_j}\right) = \frac{\partial f(x_1^0, x_2^0, \dots x_n^0)}{\partial x_j} \qquad A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

(Л6.9)

4. ***Решение линейной задачи*** запишется в соответствии с (Л6.10) и обозначениями (Л6.11) следующим образом:

Ковариационная матрица параметров  :



5. ***Следующее приближение значение параметров*** запишется:

6. Возвращаемся к пункту 2 и повторяем процедуру относительно первого приближения и так далее. ***Итерационная формула*** будет выглядеть следующим образом:

Ковариационная матрица параметров:

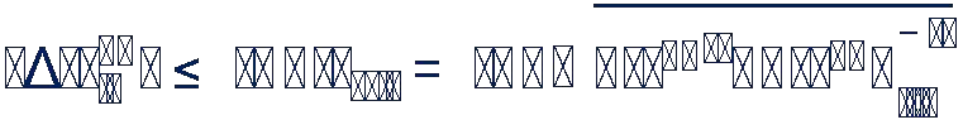
7. ***Очередное приближение значений параметров*** запишется:

8. Итерационный процесс продолжается до тех пор, ***пока не выполниться одно из следующих предварительно выбранных условий:***

a) Достигается заранее ***назначенная точность вычисления параметров*** , то есть для всех *j* выполняется:

b) Приращение минимизируемого функционала достигает заранее назначенной точности :

c) ***Модуль приращения*** каждого из параметров на данной итерации будет ***составлять заданную долю*** = ÷ ***вычислительной погрешности:***



Если ***процесс расходится, или достигается локальный минимум вместо глобального***, то существуют различные способы выхода из подобных ситуаций:

1. ***Решение задачи для нескольких начальных приближений и выбор решения, которое соответствует глобальному минимуму функционала*** (Л6.2). ***(если функционал имеет несколько локальных минимумов).***

2. ***Замена параметра*** $x_j' = \frac{x_j}{x_j}$, ***если минимум находится на бесконечности***, то есть начиная с некоторой итерации приращение некоторого *j*–ого параметра превышает заданное значение: $\Delta x_j > x_j$ (наблюдается ***расходимость*** по этому параметру).

В случае переопределенности, т.е. ***если два (или более) параметра*** x_i ***и*** x_j ***сильно коррелированны между собой***, наблюдается расходимость. Чтобы ***проверить***, действительно ли это случай переопределенности, вычисляются элементы корреляционной матрицы: $r_{ij} = \frac{\Delta x_i \Delta x_j}{\Delta x_i^2 \Delta x_j^2}$. Если для каких-нибудь *i и j* $r_{ij} > x_i \div x_j$, то ***между параметрами*** x_i ***и*** x_j ***сущесвует сильная зависимость***. В этом случае рекомендуется ***замена***: $x_i = x_i x_j$, обоснованная конкретными условиями задачи.

Метод наименьших квадратов в случае, когда обе переменные имеют погрешность.

Пусть заданы ***пары независимых измерений*** (x_i, y_i), $i = \overline{1, n}$, при этом x_i **имеют погрешность** Δx_i, **а** y_i **имеют погрешность** Δy_i. Зависимость между x и y описывается нелинейной относительно параметров функцией $y = f(x, a_0, a_1, \ldots a_m)$.

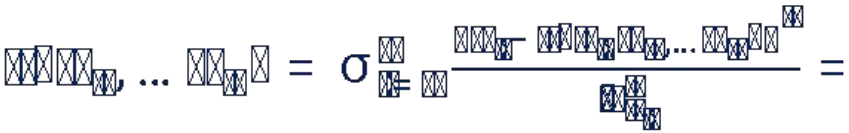
Поскольку значения x_i измерены с ошибкой, то это само по себе вносит погрешность в значения $y(x_i, a_0, a_1, \ldots a_m)$. ***По формуле переноса ошибок*** эта погрешность будет определяться:

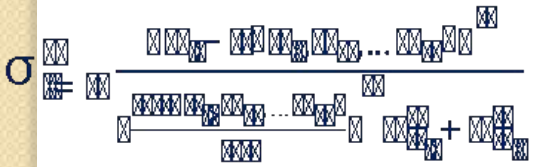
$$\Delta y_i = \left| \frac{\partial y(x_i, a_0, a_1, \ldots a_m)}{\partial x_i} \right| \Delta x_i$$

Перенесем эту погрешность на значения y_i, а x_i будем считать измеренными без погрешности. Таким образом, ***полная погрешность y*** определиться равенством:

$$\Delta y_i = \left| \frac{\partial y(x_i, a_0, a_1, \ldots a_m)}{\partial x_i} \right| \Delta x_i + \Delta y_i$$

С учетом полной погрешности функционал (Л6.2) перепишется следующим образом:

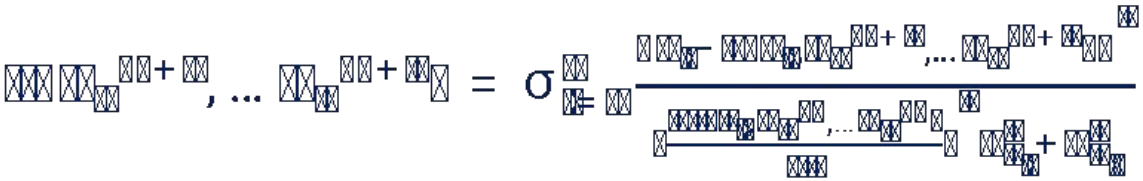




(Л6.12)

Если  , то дополнительной погрешностью, вносимой , ***можно пренебречь***, и задача сводиться к одному из случаев, рассмотренных ранее.

В противном случае минимум функционала (Л6.12) ищется с помощью ***итерационной процедуры***, аналогичной рассмотренной выше. При этом ***на m–ой итерации будет минимизироваться функционал:***



(Л6.13)

где  – искомое *(m+1)*–ое приближение, по которому производится минимизация функционала (Л6.12),  – *m* –ое приближение, уже найденное по предыдущей итерации.

Содержание

ЛЕКЦИЯ 7. I.

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Метод доверительных интервалов.

Пусть ***задана выборка случайной величины*** $x_1, x_2,\ldots x_n$ ***распределенная с плотностью вероятности*** $p(x, a_1,a_2,\ldots,a_k)$ Стоит ***задача*** оценивания неизвестных параметров распределения.

Интервальной оценкой неизвестного параметра θ называется ***интервал (a, b) такой, что истинное значение*** θ ***попадает в этот интервал с заданной вероятностью (1-*** α***).*** Вспоминаем, приведенные в первой лекции ***обозначения и определения*** связанные с интервальным оцениванием:

(a, b) - интервальная оценка параметра θ***;***

a – нижний доверительный предел;

b – верхний доверительный предел;

α ***– доверительный уровень или уровень значимости;***

(1-e) – доверительная вероятность.

В некоторых прикладных задачах наоборот требуется найти для заданной интервальной оценки неизвестную доверительную вероятность.

Рассмотрим решение поставленной задачи для двух частных случаев, а затем обобщим решение для общего случая.

Частный случай 1.

Пусть случайная величина x распределена по нормальному закону:

$$f_{a,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$$

При этом ***стандартное отклонение σ заранее задано***. Таким образом закон распределения имеет только один ***неизвестных параметр a***. Воспользуемся ***точечной оценкой среднего значения,*** полученной для нормального распределения методом максимального правдоподобия:

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Случайная величина $\hat{a}$, как сумма нормально распределенных случайных величин, также подчиняется нормальному распределению с параметрами $(a, \sigma/\sqrt{n})$:

$$f_{a,\sigma}(x) = \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{n})} exp\left(-\frac{(x-a)^2}{2(\frac{\sigma}{\sqrt{n}})}\right) \qquad (Л7.1.1)$$

Введем ***нормированную случайную величину*** $t = \frac{(\bar{x} - a)}{\sigma / \sqrt{n}}$, которая согласно ЦПТ будет распределена по нормальному закону N(0,1). Тогда по таблицам нормального распределения ***можно подобрать такое значение*** t_0, ***что вероятность того, что значение*** t ***не превысит*** t_0 ***равно заданной доверительной вероятности:***

$$P(|t| \le t_0) = \frac{1}{\sqrt{2\pi}} \int_{-t_0}^{t_0} \exp\left(-\frac{t^2}{2}\right) dt = 1 - \alpha \qquad (Л7.1.2)$$

С учетом того, что $t = \frac{(\bar{x} - a)}{\sigma / \sqrt{n}}$

$$P\left(\left|\frac{\bar{x} - a}{\sigma / \sqrt{n}}\right| < t_0\right) = P\left(\bar{x} - \frac{t_0 \sigma}{\sqrt{n}} \le a < \bar{x} + \frac{t_0 \sigma}{\sqrt{n}}\right) = 1 - \alpha$$

(Л7.1.3)

Таким образом, ***найден доверительный интервал (a, b):*** $a = \bar{x} - \frac{t_0 \sigma}{\sqrt{n}}$ и $b = \bar{x} + \frac{t_0 \sigma}{\sqrt{n}}$, такой, что неизвестное истинное значение параметра a попадает в него с вероятностью $1-\alpha$

Частный случай 2.

Пусть случайная величина распределена по нормальному закону: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$, ***но при этом значение a неизвестно***. Воспользуемся ***точечными оценками среднего значения и дисперсии, полученными методом максимального правдоподобия:***

$$\bar{a} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad s^2 = \frac{n}{n-1}\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{a})^2$$

Случайная величина $\bar{a}$, также как и предыдущем случае подчиняется нормальному распределению, но нормированная величина $t = \frac{(\bar{a}-a)}{s/\sqrt{n}}$, уже ***не будет распределена по закону N(0,1)***, ***так как в знаменателе выражения стоит величина, зависящая от исходных данных***. Согласно закона распределения функции случайных величин $t = F(x_1, x_2, ..x_n)$, в этом случае t будет распределено ***по закону распределения Стьюдента***:

$$f_m(t) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \cdot \left(1 + \frac{t^2}{m}\right)^{\frac{m+1}{2}} \qquad (Л7.1.4)$$

где m – число степеней свободы, а Г - гамма-функция, определяемая выражениями:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt; \quad \Gamma(x) = (x-1)*\Gamma(x-1); \qquad (Л7.1.5)$$

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}; \quad \Gamma(0) = \Gamma(1) = 1.$$

В данном случае ***число степеней свободы m равно (n-1),*** т.к. на исходные данные накладывается одна дополнительная связь при подстановке в выражение для t вместо истинного значения его оценки .

По таблицам распределения Стьюдента для заданного числа степеней свободы = − можно подобрать такое значение t_0, что вероятность того, что значение t не превысит t_0, равна заданной доверительной вероятности:

$$P(|t| \le t_0) = \int_{-t_0}^{t_0} S(t)\,dt = 1 - \alpha \qquad (Л7.1.6)$$

С учетом того, что $t = \frac{(\bar{x} - x)}{s/\sqrt{n}}$ соответствующее равенство для запишется:

$$P\left(\left|\frac{\bar{x} - x}{s/\sqrt{n}}\right| < t_0\right) = P\left(\bar{x} - \frac{t_0 s}{\sqrt{n}} \le x < \bar{x} + \frac{t_0 s}{\sqrt{n}}\right) = 1 - \alpha$$

(Л7.1.7)

Таким образом, для второго частного случая ***найден доверительный интервал (a, b), a =*** $\bar{x} - \frac{t_0 s}{\sqrt{n}}$ и ***b =*** $\bar{x} + \frac{t_0 s}{\sqrt{n}}$, такой, что неизвестное истинное значение параметра попадает в него с вероятностью 1-

Содержание

ЛЕКЦИЯ 7.2.

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Критерий хи-квадрат.

Пусть на заключительном этапе обработки результатов измерений ***стоят следующие задачи***:

1. имеется выборка, состоящая из реализации случайной величины X X_1, ... X_N; и необходимо сделать обоснованное заключение о виде распределения случайной величины
2. измерения случайной величины X_N в точках X_N с погрешностью X_N аппроксимирующей отношение между величинами зависимости
3. несколько выборок произвольного (может быть одинакового) размера одной или нескольких случайных величин с известной или неизвестной дисперсией.

Тогда ***любое из перечисленных выше предположений экспериментатора, проверяемое на основании выборочных экспериментальных данных, будет называться статистической гипотезой***.

В первом случае это будет гипотеза о виде распределения случайной величины, во втором случае это будет гипотеза о виде аппроксимирующей зависимости между двумя случайными или в ряде случаев зависимыми величинами и третьем случае гипотеза о статистиках случайных величин о наличии между ними зависимости.

На данном занятии рассмотрим несколько ***частных случаев*** проверки статистических гипотез.

Частный случай 1. ***Имеются измерения случайной величины*** y_i ***в точках*** x_i ***с погрешностью*** σ_i, $i = 1..n$. ***Известно, что*** y_i ***распределены по нормальному закону*** $N(f(x_i), \sigma_i)$. Требуется ***проверить статистическую гипотезу о том, что данная зависимость описывается кривой*** $f(x)$ ***на уровне значимости*** α. Отсутствие параметров в формуле для $f(x)$ означает, что они заранее известны (фиксированы).

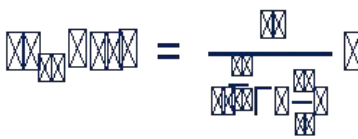
Например: $f(x) = ax + b = 2x + 1$. Параметры линейной зависимости здесь фиксированы. Отнормируем значения y_i:

$$u_i = \frac{y_i - f(x_i)}{\sigma_i}$$

Отнормированные значения будут распределяться по нормальному закону N(0,1). Найдем ***сумму квадратов***:

$$\chi^2 = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 \qquad (Л7.2.1)$$

Случайная величина Z распределена по закону (хи-квадрат с степенями свободы) с плотностью распределения :



(Л7.2.2)

где Г – гамма функция

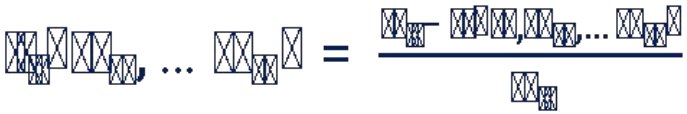
Таким образом, для заданных значений *n* и ***по таблице хи-квадрат распределения находим значение*** , соответствующее решению уравнения:

(Л7.2.3)

В случае, ***если*** найденное значение < , то ***исходная гипотеза*** о том, что данные описываются кривой , ***отвергается***. В противном случае гипотеза не отвергается. Однако, это еще не означает, что она бесспорно верна, поскольку ***существует бесконечно много аналогичных гипотез,*** которые не будут отвергнуты на том же уровне значимости. Поэтому, в результате проверки такой статистической гипотезы не заключают, что гипотеза верна, а ***делают вывод о том, что гипотеза не противоречит данным с доверительной вероятностью*** – При этом – вероятность отвергнуть правильную гипотезу.

Частный случай 2.

Имеются ***измерения случайной величины***  ***в точках*** ***с погрешностью*** , = , . . , ***известно, что*** ***распределены по нормальному закону.*** Требуется ***проверить статистическую гипотезу о том, что данная зависимость описывается кривой*** , , ... ***на уровне значимости*** ***при некоторых значениях неизвестных параметров*** , ... ***Отнормированные значения*** будут функциями неизвестных параметров:

 (Л7.2.4)

и будут как и в предыдущем случае распределены по нормальному закону. ***Сумма квадратов***

(Л7.2.5) будет теперь ***распределена по закону*** ***(хи-квадрат с n-k степенями свободы) с плотностью распределения*** . (Л7.2.2).

Для каких значений параметров $a_1, \ldots a_m$ следует проверять исходную гипотезу? Если найти фиксированные значения параметров $A_1, \ldots A_m$ ***методом наименьших квадратов***, ***то они будут минимизировать значение критерия*** (Л7.2.5). ***Если гипотеза будет отвергнута при минимальном значении критерия, то она будет отвергнута и при любом другом его значении.***

Число степеней свободы определяется размером выборки минус число неизвестных параметров. Для проверки исходной гипотезы при найденных значениях параметров $A_1, \ldots A_m$ ***определим значение критерия*** $\chi^2_{эксп}, \ldots \chi^2_{эксп}$ (Л7.2.5) и по таблице распределения хи-квадрат для числа степеней свободы $n - m$ и заданного уровня значимости α ***найдем*** $\chi^2_{табл}$, соответствующее решению уравнения:

$$P(\chi^2_{табл}, n - m) = 1 - \alpha \qquad \text{(Л7.2.6)}$$

В случае, если найденное значение $\chi^2_{эксп} < \chi^2_{табл}$, то исходная гипотеза о том, что данные описываются кривой $f(x, A_1, \ldots A_m)$, ***отвергается*** (данная модельная кривая неадекватно описывает выборочные данные на заданном уровне значимости). В случае $\chi^2_{эксп} > \chi^2_{табл}$ гипотеза ***не отвергается,*** т.е. предлагаемая модель не противоречит данным. Но не исключено, что ***возможно существуют другие модели***, также не противоречащие данным. Поэтому ***при проверке таких гипотез однозначен только отрицательный ответ.***

Частный случай 3

Пусть имеется ***выборка, состоящая из независимых реализаций случайной величины*** $x_1, x_2, \ldots x_n$. Требуется ***проверить гипотезу о том, что случайная величина*** x ***подчиняется закону с плотностью распределения*** $f(x)$ ***на уровне значимости*** α. Так же, как и в первом частном случае, плотность распределения может иметь параметры, но значение их заранее известно (фиксировано). Например, $f(x)$ может соответствовать плотности распределения N(1,0).

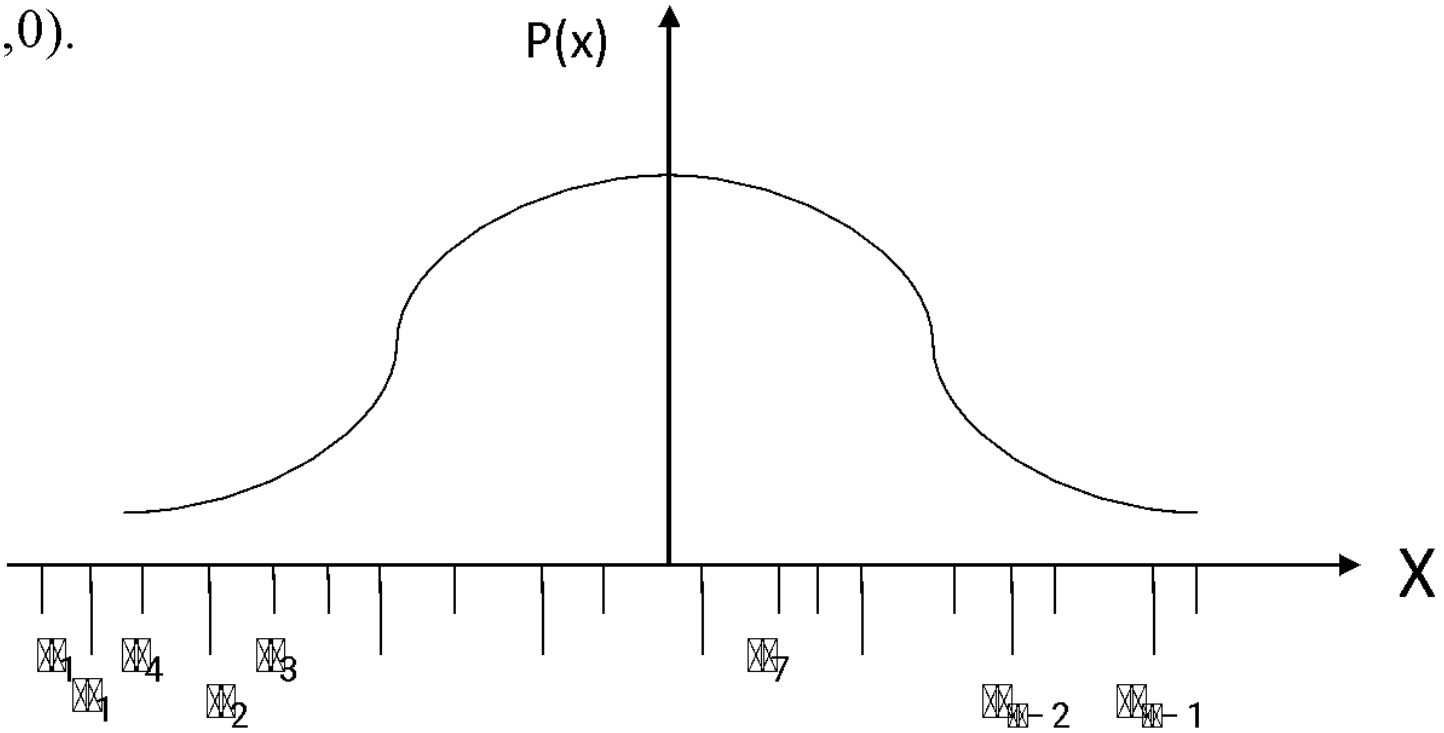


Рис. Л7.2.1

Поделим область значений случайной величины (на рис. Л7.2.1) величина определена от − ∞до +∞) на *r* участков:

: − ∞, ; : , ; … . : , + ∞

При этом каждое выборочное значение , … попадает в один из образовавшихся интервалов. Теоретическая вероятность попадания значения случайной величины ξ в отрезок равна

= (Л7.2.7)

– наблюдаемое количество выборочных значений, попавших в *i*–ый интервал. Случайные величины подчиняются биноминальному закону распределения с параметрами *n*, . ***Математическое ожидание*** = а ***дисперсия*** = − . При ∞ биноминальное распределение стремиться к нормальному . Вычислим ***сумму квадратов нормированных величин***:

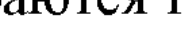
= σ (Л7.2.8)

При ∞ случайная величина Z будет распределена по закону распределения (хи-квадрат с *r-1* степенями свободы) с плотностью распределения (см Л7.2.1). Одна дополнительная связь при определении числа степеней свободы присутствует благодаря ограничению σ =

Таким образом, ***процедура проверки гипотезы о соответствии выборки заданному распределению сводится к следующим шагам:***

1. Разбиваем область значений случайной величины на *r* интервалов.
2. Подсчитываем значения n_i и теоретические вероятности p_i (Л7.2.7).
3. Подсчитываем Z по формуле (Л7.2.8).
4. По таблице распределения хи-квадрат для заданного числа степеней свободы $(r - 1)$ и уровня значимости α находим z_0, как решение уравнения $\int_0^{z_0} f_{r-1}(x)\,dx = 1 - \alpha$
5. Если $z_0 < z$, то исходная гипотеза отвергается, при этом вероятность отвергнуть правильную гипотезу равняется α, в противном случае гипотеза не отвергается.

Способ разбиения и количество интервалов могут повлиять на результат проверки. При разных разбиениях и одних и тех же данных одна и та же гипотеза может быть принята и отвергнута на одном и том же уровне значимости зависимости от способа разбиения. ***На практике при использовании критерия хи-квадрат для проверки рассматриваемой гипотезы для разбиения области значений случайной величины на отрезки пользуются следующими эмпирическими правилами при значениях*** 

1. Разбиение должно быть таким, чтобы 
2. Границы отрезков выбираются таким образом, чтобы  ***для любого i.***

При выполнении этих двух правил ***сходимость распределения случайной величины Z (Л7.2.8) к распределению хи-квадрат самая быстрая,*** то есть получение достоверных выводов возможно при меньшем размере выборки.

Частный случай 4

Пусть имеется выборка, состоящая из независимых реализаций случайной величины X $x_1, \dots x_n$. Требуется ***проверить гипотезу о том, что случайная величина*** X ***подчиняется закону с плотностью распределения*** $f(x, \theta_1, \dots \theta_k)$ ***на уровне значимости*** α ***при некоторых значениях k неизвестных параметров.***

Процедура проверки гипотезы о соответствии выборки распределения сведется к следующим шагам:

1. Разбиваем область значений случайной величины на *r* интервалов.
2. Подсчитываем значения n_i, теоретические вероятности p_i в данном случае будут функциями от неизвестных параметров $\theta_1, \dots \theta_k$.
3. В данном случае Z также будет функцией неизвестных параметров:

$$Z(\theta_1, \dots \theta_k) = \sigma \left(\frac{n_i - np_i(\theta_1, \dots \theta_k)}{\sqrt{np_i(\theta_1, \dots \theta_k)}} \right)^2 \quad \text{(Л7.2.9)}$$

5. Если проверяемая гипотеза будет отвергнута при минимальном значении критерия (Л7.2.9), то она будет отвергнута и при любом другом его значении. Тогда, набор значений параметров находим при ***решении задачи минимизации***: $\alpha_1^*, \ldots \alpha_k^* = \arg\min_{\alpha_1, \ldots \alpha_k} Z^2$, а минимальное значение Z^2 находим при найденных значениях параметров по формуле (Л7.2.9).
6. По таблицам распределения хи-квадрат для заданного числа степеней свободы $r - k - 1$ и уровня значимости α находим χ_0, как решение уравнения $\int_{\chi_0}^{\infty} f(r-k-1, x)\,dx = \alpha$.
7. Если $\chi^2 < \chi_0$, то исходная гипотеза ***отвергается***, при этом вероятность ошибки α, ***в противном случае*** гипотеза о соответствии выборки заданному распределению (с точностью до значений параметров) ***не отвергается.***

Замечания о разбиении на интервалы остаются справедливыми и для данного случая. Задача осложняется тем, что выполнение второго эмпирического правила разбиения оказывается невозможным, по крайней мере, в общем случае. Поэтому при разбиении следует руководствоваться тем ***правилом***, что ***при значениях*** $x > x_{max} \div x_{min}$, $x = x_1, x_{max} \div x_1, x_{min}$ ***число точек, попадающих в каждый интервал*** $n_i > 5 \div 10$

Частный случай 5

Критерий хи-квадрат применяется и в том случае, когда ***нужно выяснить***, являются ли значения двух величин ***коррелированными***.

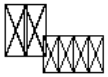
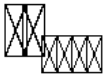
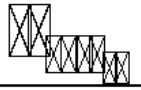
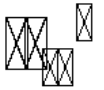
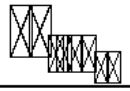
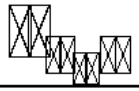
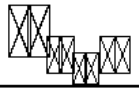
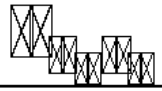
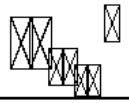
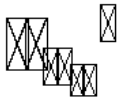
Пусть ***имеются пары измерений двух случайных величин X и Y, проведенные в последовательные моменты времени:*** (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), ... (x_n, y_n), (x_{n+1}, y_{n+1}). Требуется ***проверить гипотезу о том, что эти две случайные величины независимыми.***

В основание проверки этой гипотезы положено то обстоятельство, что для ***независимых случайных величин X и Y выполняется***:

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j).$$

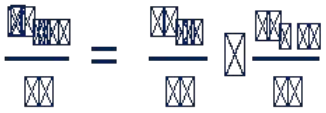
Разобьем область значений X на r_1 интервалов, а область Y значений на r_2 интервалов (используя первое эмпирическое правило, упомянутое при рассмотрении третьего частного случая).

Составим следующую таблицу, содержащую в клетке (k, j) число n_{kj} таких пар *x*, значения которых попадают в *k*–ый интервал разбиения области значений X, а значения *y* которых попадают в *j*–ый интервал разбиения области значений Y:

	1	…	k	…		
1				…		
…		…		….		
j						
…		…		….		
		…		….		
		…		…		n

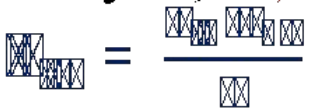
В ***последнем столбце*** таблицы стоит ***количество***  ***таких пар, значения y которых попадают в j–ый интервал разбиения области значений Y, при этом x может быть любым.*** Аналогично, в ***последней строке*** таблицы стоит ***количество***  ***таких пар, значения x которых попадают в k–ый интервал разбиения области значений X, при этом y может быть любым.***

Поделив каждое значение в таблице на n, получим ***оценки вероятностей попадания значений пар в соответствующие интервалы.*** В предположении о независимости оценки вероятности ***одновременного*** попадания x в k–ый интервал, а y в j–ый определиться по формуле:



(Л7.2.10)

Откуда, ***оценочное значение количества таких пар***:



(Л7.2.11)

Поскольку величины  подчиняются биномиальному распределению, то при  случайная величина

 (Л7.2.12)

будет распределена ***по закону распределения***  (хи-квадрат с *s* степенями свободы) с плотностью распределения  (см. Л7.2.2). ***Число степеней свободы*** в этом случае подсчитывается с учетом  дополнительных независимых связей. Таким образом

Если для заданного уровня значимости и числа степеней свободы найденное по таблице хи-квадрат распределения значение  , соответствующее решению уравнения, аналогичного (Л7.2.6), будет ***меньше Z***, найденного по формуле (Л7.2.12), то гипотеза о независимости случайных величин X и Y ***отвергается***. Вероятность отвергнуть правильную гипотезу при этом равняется величине уровня значимости

Пример 2

Заданные пары измерений помещены в таблице:

X	1	2	3	4	5	6	7	8	9	10
Y	1	2	3	4	5	5	4	3	2	1

Проверим гипотезу о независимости. Сами интервалы запишем в таблицу для подсчета пар, значения которых попадают в соответствующие интервалы. Оценочные значения помещены в ту же таблицу рядом с эмпирическими значениями под косой чертой:

	(0,3.5)	(3.5,7.5)	(7.5,10.5)	
(0,2.5)	2 /1.2	0 /1.6	2 /1.2	4
(2.5,4.5)	1 /1.2	2 /1.6	1 /1.2	4
(4.5,5.5)	0 /0.6	2 /0.8	0 /0.6	2
	3	4	3	1

По формуле (Л7.2.12) получим Z=5.826. Для числа степеней s=4 свободы гипотеза о независимости отвергается на уровне значимости α = 0.25 (χ^2_0 = 5.39) и не отвергается на уровне значимости α = 0.1. (χ^2_0 = 7.78) . Т.о. с 75% уверенностью можно утверждать, что X и Y зависимы, а с 90% - нельзя.

Содержание

ЛЕКЦИЯ 8

ЭЛЕМЕНТЫ ДИСПЕРСИОННОГО АНАЛИЗА

Дисперсионный анализ широко используется при обработке данных в радиационной физике и биофизике. ***Основное назначение метода – выявление факта однородности или неоднородности выборки вследствие разных причин***.

Рассмотрим совокупность результатов измерений (выборку размера n): $x_1, x_2, \ldots x_n$. Понятно, что все выборочные значения отличаются друг от друга. Возможны ***две причины отличия***:

- ***статистическая***, т.е. значения различаются как практическая реализация одной и той же физической случайной величины;
- ***детерминистическая***, т.е. значения являются реализацией разных случайных величин, различающихся между собой.

Выборка, ***состоящая из одной и той же случайной величины***, называется ***однородной***. Выборка, ***состоящая из разных случайных величин,*** называется ***неоднородной***.

Целью дисперсионного анализа является ***выяснение характера различий между отдельными значениями выборки***, т. е. необходимость решения следующей задачи: носят ли различия случайный(стохастический) характер или детерминированы.

В некоторых случаях ***детерминированные различия*** носят ***мешающий*** характер. По существу ***в этом случае имеем дело со смесью распределений с различными средними значениями***. Наличие детерминированных различий в такой ситуации затрудняет интерпретацию результатов измерений. В других случаях ***детерминированные различия вызываются специально,*** с ***целью изучения эффекта каких-то воз***действий. ***Дисперсионный анализ*** представляет собой существенное ***обобщение проверки гипотезы о равенстве средних двух выборок.***

В дальнейшем будем считать, что ***все измеряемые величины подчиняются нормальному закону с плотностью вероятности***:

$$\varphi(x, a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \qquad (Л8.1)$$

где параметры a и σ неизвестны. Если выборка ***неоднородна***, то считается, что ***она состоит из подвыборок, подчиняющихся нормальным распределениям с разными средними значениями, но с одинаковым среднеквадратическим отклонением*** σ.

Проверка однородности выборки.

Рассмотрим следующий ***пример из биофизической практики.*** Для проведения исследования биолог выращивает культуры клеток в нескольких чашках Петри. Затем он берет несколько проб из каждой чашки и проводит необходимые измерения. Он предполагает, что пробы, взятые из разных чашек, одинаковы в том смысле, что соответствующие результаты измерений являются реализациями одной и той же случайной величины. На самом же деле условия выращивания культур в разных чашках могут различаться. В свою очередь это различие приводит к различию значений средних и может исказить конечный результат исследования. Поэтому, ***прежде всего следует убедиться, что выборка однородна.*** Сделать это можно с помощью дисперсионного анализа, который излагается далее применительно к данной задаче.

Пусть выборка объемом n состоит из r подвыборок (групп): x_{ij}, где i=1,…,r – номер подвыборки (чашки), j =1,…,n_i - номер измерения внутри подвыборки, $\sigma_{i=1}$ = n. Величины x_{ij} подчиняются нормальному распределению (Л8.1).

Имеется подозрение, что средние значения $\bar{x}_i$ ***могут различаться между собой***. Таким образом, следует проверить гипотезу о равенстве средних значений $\bar{x}_i$ разных подвыборок друг другу. Если эта гипотеза будет отвергнута, то это будет означать, что выборка неоднородна.

Введем следующие обозначения:

$\bar{x}_i = \frac{1}{n_i}\sigma_{j=1}^{n_i} x_{ij}$ (Л8.2) - ***выборочное среднее*** (оценка среднего значения) ***в i-ой подвыборке***;

$\bar{x} = \frac{1}{n}\sigma_{i=1}^{r}\sigma_{j=1}^{n_i} x_{ij}$ (Л8.3) - ***выборочное среднее*** (оценка среднего значения) ***по всей выборке.***

Рассмотрим ***величину*** $\square = \sigma_{\square=\square}\sigma_{\square-\square\square\square} - \square^{\square}$ (Л8.4)

Если выборка ***однородна***, то величина $\frac{\square}{\square-\square}$ является ***оценкой величины дисперсии распределения*** (Л8.1). Преобразуем величину (Л8.4) следующим образом:

$$\square = \sigma_{\square=\square}\sigma_{\square-\square\square\square} - \square^{\square} = \sigma_{\square=\square}\sigma_{\square-\square\square\square} - \square_{\square} + \square_{\square} - \square^{\square} =$$

$$\sigma_{\square=\square}\sigma_{\square-\square\square\square} - \square_{\square}^{\square} + 2\sigma_{\square=\square}\square\square - \square\sigma_{\square=\square}\square\square - \square_{\square}\square +$$

$$+\sigma_{\square=\square}\sigma_{\square=\square}\square\square - \square^{\square}$$

Второе слагаемое в этой сумме равно нулю, т.к. $\sigma_{\square=\square}\square\square - \square_{\square} = 0$ в соответствии с определением (Л8.2). В третьем слагаемом нет зависимости от индекса j, поэтому $\sigma_{\square=\square}\square\square - \square^{\square} = \square_{\square}\square_{\square} - \square^{\square}$. Окончательно имеем: $Q = Q_1 + Q_2$, где $Q_1 = \sigma_{\square=\square}\sigma_{\square-\square\square\square} - \square_{\square}^{\square}$, $\square_{\square} = \sigma_{\square=\square}\square_{\square}\square_{\square} - \square^{\square}$ (Л8.5)

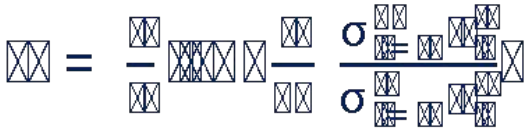
Как видно, величина Q_1 характеризует ***отклонение экспериментальных значений от подвыборочных средних***, а Q_2 - ***отклонение подвыборочных средних от общего выборочного среднего.***

Выдвигаем гипотезу о том, что выборка ***однородна***, т.е. все x_{ij} ***подчиняются распределению*** (Л8.1). В этом случае можно показать, что величины $\frac{\square}{\square}$, $\frac{\square}{\square}$, $\frac{\square}{\square}$ ***независимы*** и ***подчиняются распределению хи-квадрат*** с числом степеней свободы (n-1), (n-r), (r-1) соответственно. Можно также показать, что величины: $\square = \frac{\square}{\square}$, $\square = \frac{\square}{\square}$, $\square = \frac{\square}{\square}$ (Л8.6), являются ***несмещенными оценками величины*** $\square$. Однако т.к. истинное значение $\square^2$ неизвестно, применить критерий хи-квадрат для проверки гипотезы непосредственно нельзя. Поэтому в качестве статистики для проверки гипотезы выбирается соотношение:

$$\square = \frac{\square}{\square} = \frac{\square}{\square} \cdot \frac{\sigma \square}{\sigma \square \cdot \sigma \square} \qquad \text{(Л8.7)},$$

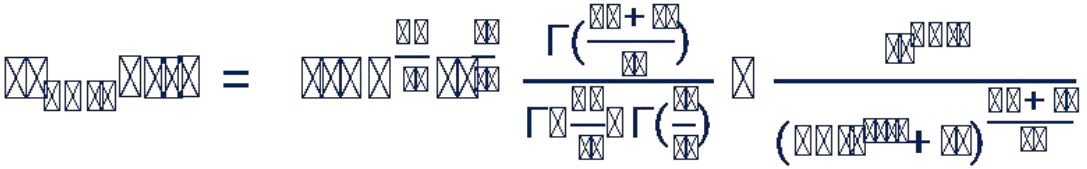
которое называется ***дисперсионным отношением***. Если умножить числитель и знаменатель дисперсионного отношения на масштабный множитель $\square^2$, то получим отношение двух величин, распределенных по закону хи-квадрат с соответствующим числом степеней свободы.

Английским математиков Р.Фишером показано, что величина:



(Л8.8),

где две независимые выборки: $x_1, x_2, \ldots x_m$ и $y_1, y_2, \ldots y_k$ являются реализациями одной и той же случайной величины, распределенной по нормальному закону с параметрами $N(0, \sigma^2)$, ***подчиняются закону распределения с плотностью вероятности:***



(Л8.9),

где Г – гамма-функция. Это распределение называется ***распределением Фишера с числом степеней свободы m и k***. Значения интегралов табулированы для многих значений z_0, m, k.

Если фактическое значение Z, вычисленное на основании экспериментальных данных по формулам (Л8,1), (Л8.3), (Л8.7), (Л8.8), удовлетворяет неравенству $\alpha \leq$, то ***гипотеза о равенстве всех средних друг другу не отвергается***. В противном случае, когда $\alpha >$ ***гипотеза отвергается, и выборка признается неоднородной.*** При этом вероятность ошибки равна Заметим, что ***гипотеза отвергается в двух случаях:***

- ***Дисперсионное отношение велико ( )***
- ***Дисперсионное отношение мало ( )***

Первый случай понятен; он соответствует ситуации, когда разброс значений подвыборочных средних относительно общего среднего велик по сравнению с разбросом внутри каждой подвыборки. Это как раз и свидетельствует о различии истинного среднего значения в каждой подвыборке по сравнению с другими подвыборочными средними.

Второй случай требует комментария.

Исходную выборку можно представить в виде:

$x_{11}, x_{12}, \ldots x_{1n1}$

$x_{21}, x_{22}, \ldots x_{2n2}$

$\ldots, \quad \ldots, \quad \ldots.$

$x_{r1}, x_{22}, \ldots x_{rnr}$

Здесь строки являются ***подвыборками***, число элементов в каждой строке в общем случае ***различно***. ***Формально можно считать подвыборками не строки, а столбцы.*** Тогда число элементов в каждом столбце в первых подвыборках будет равно r, а в последних оно будет меньше, т.к. различны. Может оказаться так, что средние в различных новых, соответствующих столбцам, подвыборках существенно отличаются друг от друга, в то время как построчные средние, напротив, могут быть близки по значениям. Легко показать, что ***именно этой ситуации соответствует малое значение дисперсионного отношения R***(Л8.7).

На практике подобную ситуацию можно представить следующим образом.

Обратимся вновь к примеру выращивания клеток в чашках Петри. Допустим, последовательность действий биологи следующая: выращивается культура в одной чашке, берется готовая проба, проводится измерение, берется вторая проба, проводится измерение и т.д. Далее выращивается культура во второй чашке, и вновь пробы и измерения проводятся по той же схеме. Затем все то же проделывается с третьей чашкой и так далее. Допустим теперь, что культура клеток неустойчива (например, она постепенно разлагается). Тогда свойства пробы зависят от времени между моментом окончания процесса выращивания и моментом измерения. Тогда истинные значения средних проб, взятых из одной чашки в разные моменты времени, могут различаться, и наоборот, можно ожидать равенства истинных средних у всех первых проб, у всех вторых проб и т.д.

Понятно, что во втором случае, (R мало) неприятие гипотезы об однородности выборки также физически обосновано, как и в первом случае, когда R велико.

И в заключении, ***если гипотеза об однородности выборки отвергается, то можно взять любые две подвыборки (например, с наиболее различающимися подвыборочными средними) и получить интервальную оценку разности средних.***

Содержание