

Министерство образования и науки Украины
Одесский национальный университет имени И.И. Мечникова
Институт математики, экономики и механики
Кафедра методов математической физики

Ю.С. Процеров

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Одесса – 2016

Печатается по решению Научно-методического совета ОНУ имени И.И. Мечникова (протокол № 1 от 20 октября 2016 года).

Составитель: Ю.С. Процеров, канд. физ. – мат. наук, доцент кафедры методов математической физики

Рецензенты: Н.Д. Вайсфельд, докт. физ-мат. наук, профессор, заведующая кафедрой Методов математической физики

В.Г. Попов, докт. физ-мат. наук, профессор, заведующий кафедрой Высшей математики Национального университета «Одесская морская академия»

А.В. Усов, докт. техн. наук, профессор, заведующий кафедрой Высшей математики и моделирования систем Одесского национального политехнического университета, лауреат Государственной премии Украины

Данное учебно-методическое пособие ставит своей целью оказать помощь студентам III-го курса направления 6.040301 «Прикладная математика» в изучении курса «Математическая статистика». В нем подробно и систематично изложены основные темы программы курса:

статистическое распределение выборки, эмпирическая функция распределения и эмпирические моменты, точечные оценки параметров распределений и методы их нахождения, интервальные оценки параметров распределений, статистическая проверка гипотез о значении параметров распределений и о виде закона распределения, теория корреляции.

Изложение материала сопровождается многочисленными примерами. Также приведено большое число заданий, предназначенных для самостоятельной работы студентов. Все задания сопровождаются ответами.

Оглавление

§ 1. Основные задачи математической статистики	4
§ 2. Статистическое распределение выборки. Полигон и гистограмма частот	5
§ 3. Эмпирическая функция распределения	7
§ 4. Эмпирические моменты	8
§ 5. Точечные оценки параметров распределений	10
§ 6. Основные виды распределений случайных величин и их числовые характеристики	12
§ 7. Метод моментов нахождения точечных оценок параметров распределений	14
§ 8. Метод максимального правдоподобия нахождения точечных оценок параметров распределений	17
§ 9. Теорема о свойствах логарифмической функции правдоподобия. Неравенство Рао-Крамера	22
§ 10. Интервальные оценки параметров распределений	28
§ 10.1. Построение доверительных интервалов для параметров нормального распределения	29
§ 10.2. Построение доверительного интервала для параметра λ распределения Пуассона	33
§ 10.3. Построение доверительного интервала для вероятности p биномиального распределения	34
§ 11. Статистическая проверка гипотез о значениях параметров распределений	37
§ 11.1. Проверка гипотез о значении параметров нормального распределения	40
§ 11.2. Проверка гипотез о значении вероятности p биномиального распределения ...	52
§ 12. Статистическая проверка гипотез о виде закона распределения	59
§ 12.1. Критерий Колмогорова	59
§ 12.2. Критерий Пирсона или χ^2	61
§ 13. Элементы теории корреляции	73
Таблица 1 значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	79
Таблица 2 значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$	80
Таблица 3 значений $t_\gamma = t(\gamma, n)$	81
Таблица 4 значений $q = q(\gamma, n)$	81
Таблица 5 критических точек распределения Стьюдента	82
Таблица 6 критических точек распределения χ^2	83
Таблица 7 значений функции распределения Колмогорова	
$K(x) = 1 + 2 \sum_{m=1}^{\infty} (-1)^m e^{-2m^2 x^2}$	84
Ответы	85
Литература	87

§ 1. Основные задачи математической статистики.

Теория вероятностей занимается изучением математических моделей случайных явлений. Имея подходящую математическую модель какого-либо случайного явления, мы можем вычислять вероятности тех или иных событий; по вероятностям одних случайных событий находить вероятности других, с ними связанных; по числовым характеристикам и функциям распределения одних случайных величин находить функции распределения и числовые характеристики других. Но естественно возникает вопрос: как найти эти исходные вероятности, функции распределения и числовые характеристики? Это как раз и является предметом исследования другой науки о массовых случайных явлениях, которая получила название математической статистики. Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении методами теории вероятности статистических данных – результатов наблюдений или экспериментов.

Как наука с оформившейся тематикой и методами исследования математическая статистика возникла в конце XIX, начале XX веков, хотя отдельные задачи, как, например, сбор и обработка статистических сведений о народонаселении, рассматривались еще в древнем Китае, Египте, Римской империи.

Задачи математической статистики являются в определенной мере обратными к задачам теории вероятности. Если в теории вероятностей мы считаем заданной модель явления и, исходя из нее, производим расчет возможного реального течения этого явления, то в математической статистике мы исходим из известных реализаций каких-либо случайных событий, то есть из статистических данных, и уже по ним подбираем подходящую теоретико-вероятностную модель.

Исходным материалом для статистического исследования реального явления служит набор результатов наблюдений над ним или же результатов специально поставленных испытаний. При этом возникают следующие основные задачи.

Первая задача математической статистики состоит в определении числа необходимых данных и разработке способа их сбора.

Пусть, например, требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака. Так, в партии некоторых изделий качественным признаком может служить их стандартность, а количественным – контролируемый размер изделия. Изучаемую совокупность объектов принято называть генеральной совокупностью. Из нее некоторым способом отбирают для исследования n объектов. Эти отобранные объекты принято называть выборочной совокупностью или выборкой объема n . Выборка может быть повторной, когда отобранный объект после изучения возвращается в генеральную совокупность, бесповторной – когда отобранный объект в генеральную совокупность не возвращается. При этом способы самого отбора могут быть различными. От сплошной выборки, когда исследуются все объекты, до разбиения генеральной совокупности на части и отбора одного или нескольких объектов из каждой части. Для того, чтобы по данной выборке можно было достаточно уверенно судить о всей генеральной совокупности, необходимо, чтобы выборка была репрезентативной, то есть представительной. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществлять случайно и ее объем n не должен быть мал.

Вторая задача математической статистики – разработка методов анализа статистических данных в зависимости от целей исследования. Сюда относятся:

- оценивание вероятностей некоторых событий исходя из статистических данных;
- определение, хотя бы и приближенно, закона распределения исследуемой случайной величины;
- нахождение неизвестных параметров закона распределения случайной величины, тип которого известен из других соображений;

Например, во многих случаях можно считать, что некоторая случайная величина распределена по нормальному закону, но его параметры μ и σ неизвестны.

- статистическая проверка гипотез о законе распределения случайной величины или его параметрах;

Например, на основании некоторых соображений мы выдвигаем гипотезу о конкретном законе распределения или о значении его параметра, если этот закон известен, и на основании имеющихся статистических данных мы должны подтвердить или опровергнуть выдвинутую гипотезу.

- нахождение корреляционной зависимости между двумя или большим числом случайных величин.

§ 2. Статистическое распределение выборки. Полигон и гистограмма частот.

Пусть имеется случайная величина ξ , заданная на вероятностном пространстве $\langle \Omega, F, P \rangle$. Пусть в результате n наблюдений (экспериментов) получено n ее значений $\xi_1, \xi_2, \dots, \xi_n$ - выборка объема n из генеральной совокупности всех возможных значений случайной величины ξ . Значения ξ_i в выборке могут повторяться. Обозначим через $x_1, x_2, \dots, x_k$ их различные значения (их принято называть вариантами), расположенные в порядке возрастания, а через $n_1, n_2, \dots, n_k$ частоты этих значений, то есть варианта x_1 встретилась n_1 раз, варианта x_2 - n_2 раза и так далее. При этом $n_1 + n_2 + \dots + n_k = n$. Если теперь каждой варианте x_i поставить в соответствие ее частоту n_i , то получим так называемое статистическое распределение выборки

x_i	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k

где $n_1 + n_2 + \dots + n_k = n$.

Наряду с частотами n_i рассматривают еще относительные частоты $\omega_i = \frac{n_i}{n}$. В этом случае статистическое распределение выборки записывают в виде

x_i	x_1	x_2	$\dots$	x_k
ω_i	ω_1	ω_2	$\dots$	ω_k

где $\omega_1 + \omega_2 + \dots + \omega_k = 1$.

Статистическое распределение выборки также задают в виде последовательности интервалов и соответствующих им частот, то есть сумму частот вариантов, попавших в этот интервал

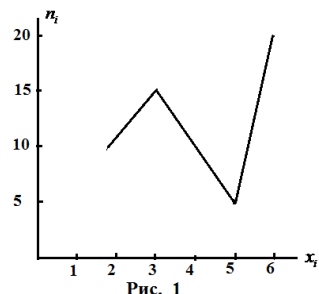
$(x_{i-1}; x_i)$	$(x_0; x_1)$	$(x_1; x_2)$	$\dots$	$(x_{k-1}; x_k)$
n_i	n_1	n_2	$\dots$	n_k

где $n_1 + n_2 + \dots + n_k = n$.

Здесь также вместо частот n_i могут стоять относительные частоты ω_i .

Для наглядного изображения данных выборки используют графическую иллюстрацию в виде полигонов и гистограмм частот. По статистическому распределению выборки можно построить полигон частот или полигон относительных частот, то есть ломаную с вершинами в точках $(x_i; n_i)$ или в точках $(x_i; \omega_i)$ соответственно. На рис. 1 изображен полигон частот следующего распределения выборки объема $n = 50$

x_i	2	3	5	6
n_i	10	15	5	20



В случае, когда статистическое распределение выборки задается в виде интервалов одинаковой длины h , то строят гистограмму частот или гистограмму относительных частот. Гистограмма частот это ступенчатая фигура, состоящая из прямоугольников с основаниями длиной h и высотой $\frac{n_i}{h}$. Площадь i -го прямоугольника будет равна частоте

$h \cdot \frac{n_i}{h} = n_i$, а площадь всей гистограммы равна $n_1 + n_2 + \dots + n_k = n$, то есть объему выборки.

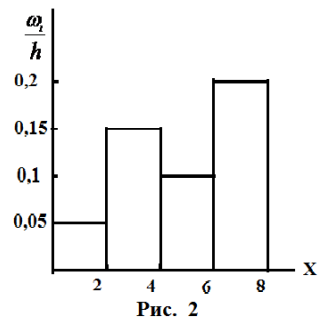
При построении гистограмм относительных частот высоты прямоугольников берут равными $\frac{\omega_i}{h}$, тогда площадь одного прямоугольника равна относительной частоте ω_i , а площадь всей гистограммы равна единице.

Например, построим гистограмму относительных частот для следующего распределения выборки объема $n = 50$

$(x_{i-1}; x_i)$	(0;2)	(2;4)	(4;6)	(6;8)
n_i	5	15	10	20

Найдем сначала относительные частоты $\omega_1 = 0,1$; $\omega_2 = 0,3$; $\omega_3 = 0,2$; $\omega_4 = 0,4$, а затем высоты прямоугольников $\frac{\omega_i}{h} : 0,05; 0,15; 0,1; 0,2$. На рис. 2

изображена требуемая гистограмма относительных частот



§ 3. Эмпирическая функция распределения.

В курсе теории вероятности при изучении свойств случайных величин основную роль играют их функции распределения $F_{\xi}(x) = P\{\omega: \xi(\omega) < x\}$, которые в статистике принято называть теоретическими функциями распределения.

Пусть дана выборка объема n

x_i	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k

где $n_1 + n_2 + \dots + n_k = n$.

Эмпирической функцией распределения или функцией распределения выборки называют функцию $F_n(x) = \frac{n_x}{n}$, где n_x число вариантов, меньших x .

Она обладает теми же свойствами, что и теоретическая функция распределения, а именно: ее значения лежат от 0 до 1; она не убывает; непрерывна слева; $F_n(x) = 0$ при $x \leq x_1$ и $F_n(x) = 1$ при $x > x_k$.

Отличие в том, что $F_{\xi}(x)$ это вероятность тех событий, для которых $\xi(\omega) < x$, а $F_n(x)$ есть сумма относительных частот тех вариантов x_i , которые меньше x

$$F_n(x) = \begin{cases} 0, & x \leq x_1 \\ \frac{n_1}{n}, & x_1 < x \leq x_2 \\ \frac{n_1}{n} + \frac{n_2}{n}, & x_2 < x \leq x_3 \\ \dots\dots\dots \\ 1, & x > x_k \end{cases}$$

Ее графиком будет ступенчатая кусочно-постоянная кривая со скачками $\frac{n_i}{n}$ в точках x_i .

Имеет место следующая

Теорема Гливенко. При увеличении объема выборки n эмпирическая функция распределения $F_n(x)$ сходится по вероятности к теоретической функции распределения $F_{\xi}(x)$

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{-\infty < x < \infty} |F_n(x) - F_{\xi}(x)| < \varepsilon \right\} = 1.$$

Пример. По выборке объема $n = 50$

x_i	1	4	6
n_i	10	15	25

найти эмпирическую функцию распределения и построить ее график.

Эмпирическая функция распределения имеет вид
$$F_n(x) = \begin{cases} 0, & x \leq 1 \\ 0,2, & 1 < x \leq 4 \\ 0,5, & 4 < x \leq 6 \\ 1, & x > 6 \end{cases},$$

а ее график приведен на рис. 3.

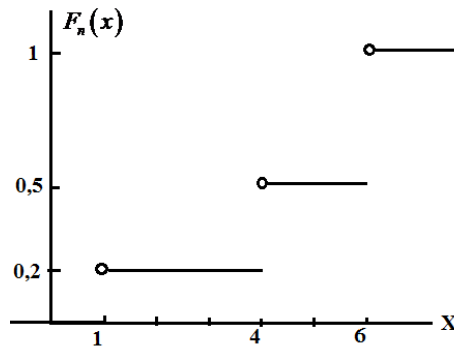


Рис. 3

Если выборка задана в виде интервалов, то при нахождении эмпирической функции распределения в качестве x_i берут середины интервалов $x_i^* = \frac{x_{i-1} + x_i}{2}$.

§ 4. Эмпирические моменты.

В курсе теории вероятности для случайной величины ξ были введены следующие моменты

$\nu_m = M(\xi^m) = \int_{-\infty}^{\infty} x^m dF_{\xi}(x)$ - начальный момент m -го порядка, в частности $\nu_1 = M\xi$ -

математическое ожидание случайной величины, и

$\mu_m = M(\xi - M\xi)^m = \int_{-\infty}^{\infty} (x - M\xi)^m dF_{\xi}(x)$ - центральный момент m -го порядка, в

частности $\mu_2 = M(\xi - M\xi)^2 = D\xi$ - дисперсия случайной величины.

Введем эмпирические или выборочные моменты

$\nu_m^* = \int_{-\infty}^{\infty} x^m dF_n(x)$ - начальный момент m -го порядка и

$\mu_m^* = \int_{-\infty}^{\infty} (x - \nu_1^*)^m dF_n(x)$ - центральный момент m -го порядка.

Так как эмпирическая функция распределения $F_n(x)$ кусочно-постоянная и терпит

скачки $\frac{n_i}{n}$ в точках x_i , то по свойству интеграла Стильеса

$$\nu_m^* = \sum_{i=1}^k x_i^m \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^k x_i^m n_i = \frac{1}{n} \sum_{i=1}^n \xi_i^m.$$

Начальный эмпирический момент первого порядка $\nu_1^* = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{1}{n} \sum_{i=1}^n \xi_i = \bar{x}_e$

называют выборочным средним. Выборочное среднее $\bar{x}_e$ есть среднее арифметическое значений выборки.

Аналогично вычисляется

$$\mu_m^* = \sum_{i=1}^k (x_i - \nu_1^*)^m \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_e)^m n_i = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x}_e)^m.$$

Центральный эмпирический момент второго порядка

$$\mu_2^* = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_e)^2 n_i = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x}_e)^2 = D_e$$

называют выборочной дисперсией. Выборочная дисперсия D_e есть среднеквадратичное отклонение значений выборки от их среднего значения $\bar{x}_e$.

Для вычисления выборочной дисперсии удобно пользоваться следующей формулой

$$\begin{aligned} D_e &= \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_e)^2 n_i = \frac{1}{n} \sum_{i=1}^k (x_i^2 - 2x_i \bar{x}_e + \bar{x}_e^2) n_i = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - 2\bar{x}_e \frac{1}{n} \sum_{i=1}^k x_i n_i + \bar{x}_e^2 \frac{1}{n} \sum_{i=1}^k n_i = \\ &= \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - 2\bar{x}_e^2 + \bar{x}_e^2 = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - \bar{x}_e^2, \end{aligned}$$

то есть выборочная дисперсия равна среднему из квадратов минус квадрат среднего.

Если выборка задана в виде интервалов, то при вычислении выборочного среднего и выборочной дисперсии в качестве x_i берут середины интервалов $x_i^* = \frac{x_{i-1} + x_i}{2}$.

Пример 1. По выборке объема $n = 20$

x_i	-2	1	3	6
n_i	3	8	5	4

найти выборочное среднее и выборочную дисперсию.

$$\text{Вычислим } \bar{x}_e = \frac{1}{20} (-2 \cdot 3 + 1 \cdot 8 + 3 \cdot 5 + 6 \cdot 4) = \frac{41}{20} = 2,05$$

$$D_e = \frac{1}{20} (4 \cdot 3 + 1 \cdot 8 + 9 \cdot 5 + 36 \cdot 4) - 2,05^2 = \frac{209}{20} - 2,05^2 = 10,45 - 4,2025 = 6,2475.$$

Во многих задачах варианты x_i являются довольно большими числами, и работать с ними по полученным формулам не очень удобно. В этом случае используется **метод условных вариантов**. Он заключается в том, что от вариантов x_i при помощи линейной замены переходят к условным вариантам u_i , которые уже малы. Например, если варианты

x_i равноотстоящие с шагом $h: x_{i+1} - x_i = h$, то положим $u_i = \frac{x_i - C}{h}$, где C некоторое значение из середины выборки (его обычно берут с наибольшей частотой),

Тогда $\bar{x}_e = hu_i + C$ и

$$\begin{aligned} \bar{x}_e &= \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{1}{n} \sum_{i=1}^k (hu_i + C) n_i = h \cdot \frac{1}{n} \sum_{i=1}^k u_i n_i + C \cdot \frac{1}{n} \sum_{i=1}^k n_i = h \cdot \bar{u}_e + C \\ D_x &= \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_e)^2 n_i = \frac{1}{n} \sum_{i=1}^k (hu_i + C - h\bar{u}_e - C)^2 n_i = h^2 \cdot \frac{1}{n} \sum_{i=1}^k (u_i - \bar{u}_e)^2 n_i = h^2 D_u. \end{aligned}$$

Пример 2. По выборке объема $n = 80$

$(x_{i-1}; x_i)$	$(-10; 0)$	$(0; 10)$	$(10; 20)$	$(20; 30)$	$(30; 40)$
n_i	10	15	25	20	10

методом условных вариантов найти выборочное среднее и выборочную дисперсию.

Отнесем частоты n_i к серединам интервалов $x_i^* = \frac{x_{i-1} + x_i}{2}$

x_i^*	-5	5	15	25	35
n_i	10	15	25	20	10

и перейдем к условным вариантам $u_i = \frac{x_i^* - 15}{10}$. Тогда

u_i	-2	-1	0	1	2
n_i	10	15	25	20	10

Вычислим $\bar{u}_e = \frac{1}{80}(-2 \cdot 10 - 1 \cdot 15 + 0 \cdot 25 + 1 \cdot 20 + 2 \cdot 10) = \frac{5}{80} = 0,0625$ и

$$D_u = \frac{1}{80}(4 \cdot 10 + 1 \cdot 15 + 0 \cdot 25 + 1 \cdot 20 + 4 \cdot 10) - 0,0625^2 = \frac{115}{80} - 0,0625^2 = 1,4375 - 0,0039 = 1,4336$$

Тогда $\bar{x}_e = 10 \cdot \bar{u}_e + 15 = 0,625 + 15 = 15,625$ и $D_e = 10^2 \cdot D_u = 143,36$.

§ 5. Точечные оценки параметров распределений.

Пусть для случайной величины ξ , заданной на вероятностном пространстве $\langle \Omega, F, P \rangle$, из каких-либо соображений известен закон ее распределения, но не известны параметры этого закона. Поставим задачу как по выборке найти приближенные значения этих параметров.

Пусть $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ выборка объема n из генеральной совокупности, а θ неизвестный параметр известного закона распределения $F_\xi(x)$. Нашей задачей является построение приближенной формулы

$$\theta \approx \theta^* = h(\zeta) = h(\xi_1, \xi_2, \dots, \xi_n), \quad (1)$$

где $h(\zeta)$ функция от выборки, называемая статистикой, а приближенное значение параметра $\theta^* = h(\zeta)$ называют точечной оценкой параметра θ . При этом предполагается, что функция $h(\zeta)$ не должна зависеть от оцениваемого параметра θ и от других параметров функции распределения $F_\xi(x)$. Кроме того функция $h(\zeta)$ должна быть одной и той же для всех распределений некоторого семейства $\{F_\xi(x)\}$. Если изменяется объем выборки, то будем обозначать $\theta \approx \theta_n^* = h_n(\zeta)$, а последовательность $\{h_n(\zeta)\}$ будем называть последовательностью порядковых статистик.

Следует отметить, что для каждой выборки $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ значение $\theta^* = h(\zeta)$ будет свое, а потому θ^* , как функция выборки, будет случайной величиной.

Здесь сразу возникает вопрос: как оценить точность приближенной формулы (1)? Будем оценивать величину погрешности при замене θ на θ^* среднеквадратичной погрешностью $M|\theta - \theta^*|^2$, то есть дисперсией оценки $D\theta^*$ - величиной рассеяния θ^* относительно точного значения θ . Среди оценок с одной и той же дисперсией $D\theta^*$ минимальную величину рассеяния относительно θ имеют оценки, для которых $M\theta^* = \theta$.

Определение. Точечная оценка θ^* называется несмещенной оценкой параметра θ , если $M\theta^* = \theta$.

Требование несмещенности оценки есть условие отсутствия «систематической ошибки». Его смысл состоит в том, что при многократном использовании формулы (1) для различных выборок среднее значение ошибки при замене θ на θ^* равно нулю.

Определение. Точечная оценка θ_n^* параметра θ называется асимптотически несмещенной, если $\lim_{n \rightarrow \infty} M\theta_n^* = \theta$.

Покажем, что выборочное среднее $\bar{x}_e$ является несмещенной точечной оценкой для математического ожидания $M\xi$. Рассмотрим

$$M\bar{x}_e = M\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n} \sum_{i=1}^n M\xi_i = \frac{1}{n} \sum_{i=1}^n M\xi = \frac{1}{n} \cdot n \cdot M\xi = M\xi,$$

так как ξ_i есть значение случайной величины ξ , полученное в i -том наблюдении, а значит все они одинаково распределены и $M\xi_i = M\xi$.

Покажем теперь, что выборочная дисперсия D_e не является несмещенной оценкой для дисперсии $D\xi$:

$$MD_e = M\left(\frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x}_e)^2\right) = \frac{1}{n} \sum_{i=1}^n M(\xi_i - \bar{x}_e)^2 = \frac{1}{n} \cdot n \cdot M(\xi_i - \bar{x}_e)^2 = M(\xi_i - \bar{x}_e)^2$$

Так как $M(\xi_i - \bar{x}_e) = M\xi_i - M\bar{x}_e = M\xi - M\xi = 0$, то

$$D(\xi_i - \bar{x}_e) = M(\xi_i - \bar{x}_e)^2 - (M(\xi_i - \bar{x}_e))^2 = M(\xi_i - \bar{x}_e)^2 \quad \text{и}$$

$$MD_e = D(\xi_i - \bar{x}_e) = D\left(\xi_i - \frac{1}{n} \sum_{j=1}^n \xi_j\right) = D\left(\left(1 - \frac{1}{n}\right)\xi_i - \frac{1}{n} \sum_{j=1, j \neq i}^n \xi_j\right) =$$

поскольку в последней сумме стоят независимые слагаемые, то

$$= \left(1 - \frac{1}{n}\right)^2 D\xi_i + \frac{1}{n^2} \sum_{j=1, j \neq i}^n D\xi_j = \left(1 - \frac{1}{n}\right)^2 D\xi + \frac{n-1}{n^2} D\xi = \frac{n^2 - 2n + 1 + n - 1}{n^2} D\xi = \frac{n-1}{n} D\xi,$$

учитывая, что все ξ_i одинаково распределены и $D\xi_i = D\xi$.

Таким образом, $MD_{\xi} = \frac{n-1}{n} D\xi \neq D\xi$ и выборочная дисперсия есть смещенная оценка для дисперсии $D\xi$. Однако эта оценка является асимптотически несмещенной, так как

$$\lim_{n \rightarrow \infty} MD_{\xi} = \lim_{n \rightarrow \infty} \frac{n-1}{n} D\xi = D\xi.$$

Несложно видоизменить выборочную дисперсию так, чтобы измененная оценка стала уже несмещенной. Для этого введем исправленную выборочную дисперсию

$$s^2 = \frac{n}{n-1} D_{\xi}$$

Тогда $Ms^2 = \frac{n}{n-1} MD_{\xi} = \frac{n}{n-1} \cdot \frac{n-1}{n} D\xi = D\xi$ и эта оценка является уже несмещенной.

Обычно при вычислениях, если объем n выборки невелик ($n \leq 50$), то в качестве оценки для дисперсии берут исправленную выборочную дисперсию s^2 , а при больших объемах выборки берут выборочную дисперсию D_{ξ} .

Определение. Оценка θ_n^* параметра θ называется состоятельной, если она сходится по вероятности к параметру θ при $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} P\{|\theta_n^* - \theta| < \varepsilon\} = 1.$$

Определение. Оценка θ_n^* параметра θ называется эффективной, если ее дисперсия с ростом n уменьшается:

$$D\theta_{n+1}^* \leq D\theta_n^*.$$

Покажем, что выборочное среднее $\bar{x}_{\xi}$ является состоятельной и эффективной оценкой для математического ожидания $M\xi$. Найдем

$$D\bar{x}_{\xi} = D\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D\xi_i = \frac{1}{n^2} \cdot n \cdot D\xi = \frac{D\xi}{n}.$$

Состоятельность оценки следует из II-го неравенства Чебышева

$$P\{|\bar{x}_{\xi} - M\xi| < \varepsilon\} = P\{|\bar{x}_{\xi} - M\bar{x}_{\xi}| < \varepsilon\} \geq 1 - \frac{D\bar{x}_{\xi}}{\varepsilon^2} = 1 - \frac{D\xi}{n\varepsilon^2} \rightarrow 1 \quad n \rightarrow \infty.$$

Эффективность оценки вытекает из неравенства

$$D\bar{x}_{\xi}^{-(n+1)} = \frac{D\xi}{n+1} < \frac{D\xi}{n} = D\bar{x}_{\xi}^{-(n)}.$$

§ 6. Основные виды распределений случайных величин и их числовые характеристики

В дальнейшем мы неоднократно будем пользоваться различными законами распределения случайных величин и их числовыми характеристиками, изученными ранее в курсе «Теории вероятности». Поэтому приведем здесь их основные характеристики.

Дискретные распределения.

1) Распределение Пуассона.

$$P\{\xi = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0, k = 0, 1, 2, \dots; \quad M\xi = \lambda, D\xi = \lambda.$$

2) Биномиальное распределение.

$$P\{\xi = k\} = C_n^k p^k q^{n-k}, k = 0, 1, 2, \dots, n; \quad M\xi = np, D\xi = npq.$$

3) Геометрическое распределение.

$$P\{\xi = k\} = q^{k-1} p, k = 1, 2, \dots; \quad M\xi = \frac{1}{p}, D\xi = \frac{q}{p^2}.$$

Абсолютно-непрерывные распределения.

1) Нормальное распределение или распределение Гаусса.

$$p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, F_\xi(x) = \frac{1}{2} + \Phi\left(\frac{x-a}{\sigma}\right), \text{ где } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

$$M\xi = a, D\xi = \sigma^2.$$

2) Равномерное на промежутке $(a;b)$ распределение.

$$p_\xi(x) = \begin{cases} \frac{1}{b-a}, & x \in (a;b) \\ 0, & x \notin (a;b) \end{cases}, F_\xi(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases}$$

$$M\xi = \frac{a+b}{2}, D\xi = \frac{(b-a)^2}{12}.$$

3) Показательное распределение.

$$p_\xi(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}, F_\xi(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \lambda > 0$$

$$M\xi = \frac{1}{\lambda}, D\xi = \frac{1}{\lambda^2}.$$

4) Гамма распределение

$$p_\xi(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \alpha > 0, \beta > 0$$

$$M\xi = \frac{\alpha}{\beta}, D\xi = \frac{\alpha}{\beta^2}.$$

5) Распределение Пирсона или χ^2 с n степенями свободы.

Пусть $\xi_1, \xi_2, \dots, \xi_n$ нормированные нормально распределенные случайные величины ($a = 0, \sigma = 1$). Тогда случайная величина

$$\chi^2 = \sum_{i=1}^n \xi_i^2$$

имеет распределение Пирсона или χ^2 с n степенями свободы. Его плотность

$$p_{\chi^2}(x) = k_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0, M\chi^2 = n, D\chi^2 = 2n. \\ 0, & x \leq 0 \end{cases}$$

С этим распределением тесно связаны следующие распределения

$$\chi = \sqrt{\sum_{i=1}^n \xi_i^2} \text{ с плотностью } p_{\chi}(x) = 2xk_n(x^2),$$

$$\frac{1}{n}\chi^2 = \frac{1}{n}\sum_{i=1}^n \xi_i^2 \text{ с плотностью } p(x) = nk_n(nx),$$

$$\frac{1}{\sqrt{n}}\chi = \sqrt{\frac{1}{n}\sum_{i=1}^n \xi_i^2} \text{ с плотностью } p(x) = 2nxk_n(nx^2).$$

6) Распределение Стьюдента с n степенями свободы.

Пусть η и $\xi_1, \xi_2, \dots, \xi_n$ нормированные нормально распределенные случайные величины ($a = 0, \sigma = 1$). Тогда случайная величина

$$t = \frac{\eta}{\sqrt{\frac{1}{n}\sum_{i=1}^n \xi_i^2}}$$

имеет распределение Стьюдента с n степенями свободы. Его плотность

$$p_t(x) = \frac{1}{\sqrt{\pi n}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad Mt = 0 \ (n > 1), \quad Dt = \frac{n}{n-2} \ (n > 2).$$

С ростом n распределения Пирсона и Стьюдента стремятся к нормальному распределению.

§ 7. Метод моментов нахождения точечных оценок параметров распределений.

Предположим, что известная функция распределения случайной величины ξ зависит от m неизвестных параметров $\theta_1, \theta_2, \dots, \theta_m$: $F_{\xi}(x, \theta_1, \theta_2, \dots, \theta_m)$. Запишем для нее выражения для теоретических моментов

$$\nu_k = \int_{-\infty}^{\infty} x^k dF_{\xi}(x, \theta_1, \dots, \theta_m) \text{ и } \mu_k = \int_{-\infty}^{\infty} (x - M\xi)^k dF_{\xi}(x, \theta_1, \dots, \theta_m),$$

которые будут содержать эти неизвестные параметры. С другой стороны по выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ подсчитаем эмпирические моменты

$$\nu_k^* = \frac{1}{n} \sum_{i=1}^n \xi_i^k \text{ и } \mu_k^* = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x}_{\xi})^k,$$

которые будут числовыми значениями. Приравнивая соответствующие теоретические и эмпирические моменты, получим систему уравнений для определения параметров $\theta_1, \theta_2, \dots, \theta_m$. Ее решение $\theta_1^*, \theta_2^*, \dots, \theta_m^*$ и берем в качестве точечных оценок искомых параметров распределения.

В частности, если функция распределения содержит один параметр, то берут $M\xi = \bar{x}_{\xi}$, а если два параметра, то берут $M\xi = \bar{x}_{\xi}$ и $D\xi = s^2$ или $D\xi = D_{\xi}$, если n велико.

Пример 1. Случайная величина ξ - продолжительность безотказной работы прибора

имеет показательное распределение с плотностью $p_{\xi}(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$. Найти

методом моментов точечную оценку параметра λ по выборке объема $n = 200$

$(x_{i-1}; x_i)$	(0;5)	(5;10)	(10;15)	(15;20)	(20;25)	(25;30)
n_i	133	45	15	4	2	1

Для показательного распределения $M\xi = \frac{1}{\lambda}$. Согласно методу моментов $M\xi = \frac{1}{\lambda} = \bar{x}_{\xi}$,

откуда $\lambda \approx \lambda^* = \frac{1}{\bar{x}_{\xi}}$. Для вычисления $\bar{x}_{\xi}$ рассмотрим середины интервалов

$$x_1^* = 2,5; x_2^* = 7,5; x_3^* = 12,5; x_4^* = 17,5; x_5^* = 22,5; x_6^* = 27,5;$$

и перейдем к условным вариантам $u_i = \frac{x_i^* - 2,5}{5}$. Для них выборка примет вид

u_i	0	1	2	3	4	5
n_i	133	45	15	4	2	1

Вычислим $\bar{u}_{\xi} = \frac{1}{200}(0 + 45 + 30 + 12 + 8 + 5) = \frac{100}{200} = 0,5$. Тогда $\bar{x}_{\xi} = 5 \cdot 0,5 + 2,5 = 5$ и

$$\lambda \approx \lambda^* = \frac{1}{\bar{x}_{\xi}} = \frac{1}{5} = 0,2.$$

Пример 2. Случайная величина ξ - число появлений события А в $m = 5$ независимых испытаниях подчинена биномиальному закону распределения с неизвестным параметром p - вероятностью наступления события А в одном испытании. Проведено $n = 50$ опытов по $m = 5$ испытаний в каждом опыте, результаты которых представлены выборкой

x_i	0	1	2	3	4	5
n_i	10	16	12	8	3	1

где x_i число появлений события А в одном опыте, а n_i количество опытов, в которых наблюдалось x_i появлений события А. Найти методом моментов точечную оценку вероятности p .

Для биномиального распределения $M\xi = mp$. Согласно методу моментов $M\xi = \bar{x}_e$, то есть $mp = \bar{x}_e$ и $p \approx p^* = \frac{1}{m} \bar{x}_e$. Вычислим

$$\bar{x}_e = \frac{1}{50}(0+16+24+24+12+5) = \frac{81}{50} = 1,62.$$

Тогда $p \approx p^* = \frac{1}{50} \cdot 1,62 = 0,0324$.

Задания.

1. Случайная величина ξ распределена по закону Пуассона. По выборке

x_i	0	1	2	3	4
n_i	140	40	15	4	1

методом моментов найти точечную оценку неизвестного параметра λ распределения Пуассона.

2. Случайная величина ξ имеет геометрическое распределение. По выборке

x_i	1	2	3	4	5	6
n_i	2	3	6	5	2	2

найти методом моментов точечную оценку вероятности p наступления события А в одном испытании. Здесь x_i номер испытания, в котором впервые появилось событие А, а n_i количество опытов, в которых событие А произошло на x_i испытании.

3. Случайная величина ξ распределена нормально. По выборке

x_i	0,2	0,4	0,6	0,8	1,0	1,2	1,4	1,6	1,8	2,0	2,2
n_i	6	9	25	26	30	26	24	21	20	8	5

найти методом моментов точечные оценки неизвестных параметров a и σ нормального распределения.

4. Случайная величина ξ распределена равномерно с неизвестными параметрами a и b . По выборке

x_i	3	5	7	9	11	13	15	17	19	21
n_i	21	16	15	26	22	14	21	22	18	25

методом моментов найти точечные оценки параметров a и b равномерного распределения.

5. Случайная величина ξ имеет гамма распределение с неизвестными параметрами α и β . По выборке

x_i	5	10	15	20	25	30	35	40	45
n_i	1	3	6	7	7	5	4	8	4

методом моментов найти точечные оценки параметров α и β гамма распределения.

§ 8. Метод максимального правдоподобия нахождения точечных оценок параметров распределений.

Пусть ξ дискретная случайная величина, для которой известен закон ее распределения, но неизвестен параметр θ этого закона распределения. Обозначим через

$$P\{\xi = \xi_i\} = p(\xi_i, \theta)$$

вероятность того, что случайная величина ξ приняла значение ξ_i .

По заданной выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ построим функцию правдоподобия

$$L(\zeta, \theta) = p(\xi_1, \theta) p(\xi_2, \theta) \dots p(\xi_n, \theta) = (p(x_1, \theta))^{n_1} (p(x_2, \theta))^{n_2} \dots (p(x_k, \theta))^{n_k},$$

которая является функцией от выборки ζ и неизвестного параметра θ . Так как значения ξ_i независимы, то функция правдоподобия $L(\zeta, \theta)$ есть вероятность появления исхода $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ в n независимых испытаниях, то есть как раз того результата, который дан выборкой.

Параметр θ будем искать из условия максимума функции правдоподобия $L(\zeta, \theta)$, то есть из условия максимума вероятности полученной выборки ζ . Иными словами, чтобы полученный в выборке результат был наиболее правдоподобным.

Для удобства вычислений будем искать максимум не самой функции правдоподобия $L(\zeta, \theta)$, а ее логарифма – логарифмической функции правдоподобия

$$\ln L(\zeta, \theta) = \sum_{i=1}^n \ln p(\xi_i, \theta) = \sum_{i=1}^k n_i \ln p(x_i, \theta).$$

Необходимое условие экстремума нас приводит к уравнению $\frac{d}{d\theta} \ln L(\zeta, \theta) = 0$

относительно θ . Если θ^* его решение и выполняется достаточное условие максимума

$\frac{d^2}{d\theta^2} \ln L(\zeta, \theta)|_{\theta=\theta^*} < 0$, то это значение θ^* и берем в качестве точечной оценки параметра θ

Пусть теперь случайная величина ξ имеет абсолютно-непрерывное распределение с плотностью $p_{\xi}(x, \theta) = p(x, \theta)$. Для выборки $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ функция правдоподобия имеет вид

$$L(\zeta, \theta) = p(\xi_1, \theta) p(\xi_2, \theta) \dots p(\xi_n, \theta) = (p(x_1, \theta))^{n_1} (p(x_2, \theta))^{n_2} \dots (p(x_k, \theta))^{n_k}.$$

В силу независимости величин ξ_i их совместная плотность распределения, то есть плотность распределения вектора $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$, равна произведению плотностей компонент. Таким образом, функция правдоподобия $L(\zeta, \theta)$ есть плотность распределения выборки ζ .

В этом случае параметр θ также ищется из условия максимума функции правдоподобия, точнее из условия максимума логарифмической функции правдоподобия

$$\frac{d}{d\theta} \ln L(\zeta, \theta) = 0 \text{ и } \frac{d^2}{d\theta^2} \ln L(\zeta, \theta) \big|_{\theta=\theta^*} < 0.$$

Как видно, формальная математическая формулировка обоих случаев ничем не отличается друг от друга. Только в первом случае $p(\xi, \theta)$ это вероятность, а во втором случае это плотность функции распределения.

Наконец, если функция распределения зависит не от одного, а от m параметров $\theta_1, \theta_2, \dots, \theta_m$, то и функция правдоподобия также будет зависеть от этих m параметров и задача сводится к их нахождению из условий максимума логарифмической функции

$$\ln L(\zeta, \theta_1, \theta_2, \dots, \theta_m) = \sum_{i=1}^n \ln p(\xi_i, \theta_1, \theta_2, \dots, \theta_m).$$

Необходимые условия экстремума нам дадут m уравнений

$$\frac{\partial}{\partial \theta_j} \ln L(\zeta, \theta_1, \theta_2, \dots, \theta_m) = 0, \quad j = \overline{1, m}.$$

Если $\bar{\theta}^* = (\theta_1^*, \theta_2^*, \dots, \theta_m^*)$ есть решение этой системы и выполняется достаточное условие максимума $d^2 \ln L(\zeta, \theta_1, \theta_2, \dots, \theta_m) \big|_{\bar{\theta}^*} < 0$, то найденные значения и берутся в качестве точечных оценок параметров распределения.

Пример 1. Случайная величина ξ - число появлений события А в $m = 10$ независимых испытаниях подчинена биномиальному закону распределения с неизвестным параметром p - вероятностью наступления события А в одном испытании. Проведено $n = 100$ опытов по $m = 10$ испытаний в каждом опыте, результаты которых представлены выборкой

x_i	0	1	2	3	4	5	6	7
n_i	2	3	10	22	26	20	12	5

где x_i число появлений события А в одном опыте, а n_i количество опытов, в которых наблюдалось x_i появлений события А. Найти методом максимального правдоподобия точечную оценку вероятности p .

В данном случае $P(\xi = x_i) = p(x_i, p) = C_m^{x_i} p^{x_i} (1-p)^{m-x_i}$. Составим функцию правдоподобия

$$L(\zeta, p) = \left(C_m^{x_1} p^{x_1} (1-p)^{m-x_1}\right)^{n_1} \cdot \left(C_m^{x_2} p^{x_2} (1-p)^{m-x_2}\right)^{n_2} \cdot \dots \cdot \left(C_m^{x_k} p^{x_k} (1-p)^{m-x_k}\right)^{n_k} = \\ = \left(C_m^{x_1}\right)^{n_1} \left(C_m^{x_2}\right)^{n_2} \dots \left(C_m^{x_k}\right)^{n_k} p^{x_1 n_1 + x_2 n_2 + \dots + x_k n_k} (1-p)^{(m-x_1)n_1 + (m-x_2)n_2 + \dots + (m-x_k)n_k}.$$

Учтем, что $x_1 n_1 + x_2 n_2 + \dots + x_k n_k = n \cdot \frac{1}{n} \sum_{i=1}^k x_i n_i = n \cdot \bar{x}_e$, $n_1 + n_2 + \dots + n_k = n$

$$(m-x_1)n_1 + (m-x_2)n_2 + \dots + (m-x_k)n_k = m(n_1 + n_2 + \dots + n_k) - (x_1 n_1 + x_2 n_2 + \dots + x_k n_k) = \\ = mn - n\bar{x}_e = n(n - \bar{x}_e)$$

и обозначим через $C = \left(C_m^{x_1}\right)^{n_1} \left(C_m^{x_2}\right)^{n_2} \dots \left(C_m^{x_k}\right)^{n_k}$ - величину, не зависящую от p .

Тогда $L(\zeta, p) = C \cdot p^{n\bar{x}_e} \cdot (1-p)^{n(m-\bar{x}_e)}$ и $\ln L(\zeta, p) = \ln C + n\bar{x}_e \ln p + n(m-\bar{x}_e) \ln(1-p)$.

Найдем

$$\frac{d}{dp} \ln L(\zeta, p) = \frac{n\bar{x}_e}{p} - \frac{n(m-\bar{x}_e)}{1-p} = n \frac{\bar{x}_e - p\bar{x}_e - mp + p\bar{x}_e}{p(1-p)} = n \frac{\bar{x}_e - mp}{p(1-p)}.$$

Это выражение будет равно нулю, если $\bar{x}_e - mp = 0$, то есть $p = \frac{1}{m} \bar{x}_e$.

Найдем теперь $\frac{d^2}{dp^2} L(\zeta, p) = -n \frac{\bar{x}_e}{p^2} - n \frac{m-\bar{x}_e}{(1-p)^2}$. Эта величина отрицательна, так как

$0 \leq \bar{x}_e \leq m$. Таким образом в точке $p = \frac{1}{m} \bar{x}_e$ функция правдоподобия достигает

максимума, и мы нашли точечную оценку вероятности $p \approx p^* = \frac{1}{m} \bar{x}_e$.

По заданной выборке найдем

$$\bar{x}_e = \frac{1}{100} (0 + 3 + 20 + 66 + 104 + 100 + 72 + 35) = \frac{400}{100} = 4.$$

Тогда $p^* = \frac{1}{10} \cdot 4 = 0,4$.

Пример 2. Случайная величина ξ распределена по нормальному закону с плотностью

$$p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}. \text{ По выборке объема } n=100$$

x_i	-4	-2	0	2	4	6	8	10
n_i	3	7	15	25	24	16	8	2

методом максимального правдоподобия найти точечные оценки параметров a и σ .

Составим функцию правдоподобия

$$L(\zeta, a, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-a)^2}{2\sigma^2}} \right)^{n_1} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2-a)^2}{2\sigma^2}} \right)^{n_2} \dots \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_k-a)^2}{2\sigma^2}} \right)^{n_k} =$$

$$= (\sigma\sqrt{2\pi})^{-(n_1+n_2+\dots+n_k)} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^k (x_i-a)^2 n_i}.$$

Логарифмическая функция правдоподобия имеет вид

$$\ln L(\zeta, a, \sigma) = -n(\ln \sigma + \ln \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^k (x_i - a)^2 n_i.$$

Найдем частные производные

$$\begin{aligned} \frac{\partial}{\partial a} \ln L(\zeta, a, \sigma) &= -\frac{1}{2\sigma^2} (-2) \sum_{i=1}^k (x_i - a) n_i = \frac{1}{\sigma^2} \left(\sum_{i=1}^k x_i n_i - a \sum_{i=1}^k n_i \right) = \\ &= \frac{1}{\sigma^2} (n\bar{x}_g - an) = \frac{n}{\sigma^2} (\bar{x}_g - a) \\ \frac{\partial}{\partial \sigma} \ln L(\zeta, a, \sigma) &= -\frac{n}{\sigma} + \frac{2}{2\sigma^3} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{1}{\sigma^3} \left[-n\sigma^2 + \sum_{i=1}^k (x_i - a)^2 n_i \right] \end{aligned}$$

Из системы уравнений

$$\begin{cases} \frac{n}{\sigma^2} (\bar{x}_g - a) = 0 \\ \frac{1}{\sigma^3} \left[-n\sigma^2 + \sum_{i=1}^k (x_i - a)^2 n_i \right] = 0 \end{cases}$$

находим $a = \bar{x}_g$ и $\sigma^2 = \frac{1}{n} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_g)^2 n_i = D_g$.

Для проверки выполнимости достаточного условия максимума вычислим вторые производные

$$\begin{aligned} \frac{\partial^2}{\partial a^2} \ln L(\zeta, a, \sigma) &= -\frac{n}{\sigma^2}; \quad \frac{\partial^2}{\partial a \partial \sigma} \ln L(\zeta, a, \sigma) = -\frac{2n}{\sigma^3} (\bar{x}_g - a) \\ \frac{\partial^2}{\partial \sigma^2} \ln L(\zeta, a, \sigma) &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^k (x_i - a)^2 n_i \end{aligned}$$

и найдем дифференциал второго порядка

$$d^2 \ln L(\zeta, a, \sigma) = -\frac{n}{\sigma^2} da^2 - \frac{4n}{\sigma^3} (\bar{x}_g - a) dad\sigma + \left[\frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^k (x_i - a)^2 n_i \right] d\sigma^2$$

При $a = \bar{x}_g$ и $\sigma^2 = D_g$ получаем, что

$$\begin{aligned} d^2 \ln L(\zeta, a, \sigma) &= -\frac{n}{D_g} da^2 + \left[\frac{n}{D_g} - \frac{3}{D_g^2} \sum_{i=1}^k (x_i - \bar{x}_g)^2 n_i \right] d\sigma^2 = \\ &= -\frac{n}{D_g} da^2 + \left[\frac{n}{D_g} - \frac{3}{D_g^2} \cdot nD_g \right] d\sigma^2 = -\frac{n}{D_g} da^2 - \frac{2n}{D_g} d\sigma^2 < 0. \end{aligned}$$

Таким образом, функция правдоподобия имеет максимум в найденной точке, и мы нашли точечные оценки параметров $a \approx a^* = \bar{x}_g$ и $\sigma \approx \sigma^* = \sqrt{D_g}$.

По заданной выборке вычислим

$$\bar{x}_g = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{1}{100} (-12 - 14 + 50 + 96 + 96 + 64 + 20) = \frac{300}{100} = 3$$

$$D_s = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_s)^2 n_i = \frac{1}{100} (49 \cdot 3 + 25 \cdot 7 + 3 \cdot 15 + 1 \cdot 25 + 1 \cdot 24 + 9 \cdot 16 + 25 \cdot 8 + 49 \cdot 2) = \frac{948}{100} = 9,48$$

Тогда $a^* = 3$ и $\sigma^* = \sqrt{9,48} = 3,08$.

Задания.

1. Случайная величина ξ распределена по закону Пуассона. По выборке

x_i	0	1	2	3	4
n_i	120	55	15	8	2

методом максимального правдоподобия найти точечную оценку неизвестного параметра λ распределения Пуассона.

2. Случайная величина ξ имеет геометрическое распределение. По выборке

x_i	2	3	4	5	6
n_i	2	4	5	5	4

найти методом максимального правдоподобия точечную оценку вероятности p наступления события А в одном испытании. Здесь x_i номер испытания, в котором впервые появилось событие А, а n_i количество опытов, в которых событие А произошло на x_i испытании.

3. Случайная величина ξ имеет показательное распределение. По выборке

x_i	5	10	15	20	25	30
n_i	150	80	40	20	8	2

методом максимального правдоподобия найти точечную оценку неизвестного параметра λ показательного распределения.

4. Случайная величина ξ распределена по нормальному закону с известным среднеквадратичным отклонением $\sigma = 4$. По выборке

x_i	2,5	5	7,5	10	12,5	15	17,5
n_i	10	20	35	60	40	25	15

найти методом максимального правдоподобия точечную оценку неизвестного параметра a нормального распределения.

5. Случайная величина ξ распределена по нормальному закону с известным математическим ожиданием $a = 25$. По выборке

x_i	16	19	22	25	28	31	34
n_i	4	15	25	45	20	10	1

найти методом максимального правдоподобия точечную оценку неизвестного параметра σ нормального распределения.

6. Случайная величина ξ имеет гамма распределение с известным параметром $\alpha = 1,5$. По выборке

x_i	1	5	9	13	17	21
n_i	40	25	15	10	6	4

найти методом максимального правдоподобия точечную оценку неизвестного параметра β гамма распределения.

§ 9. Теорема о свойствах логарифмической функции правдоподобия. Неравенство Рао-Крамера.

Рассмотрим здесь ряд вопросов, которые с одной стороны служат обоснованием рассмотренного выше метода максимального правдоподобия, а с другой стороны позволяют найти нижнюю границу дисперсии несмещенных точечных оценок параметра. Последнее позволяет сравнивать дисперсию полученной несмещенной оценки с нижней границы дисперсии и тем самым судить о том насколько полученная оценка близка к оптимально возможной. Кроме того будут получены условия достижения этой нижней границы дисперсии, а значит условия получения наиболее эффективной точечной несмещенной оценки.

Теорема. Пусть случайная величина ξ имеет абсолютно-непрерывное распределение с плотностью $p(x, \theta)$ и пусть для логарифмической функции правдоподобия

$$\ln L(\zeta, \theta) = \sum_{i=1}^n \ln p(\xi_i, \theta) \text{ выполняются следующие условия:}$$

1. Существуют производные $\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)$ и $\frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta)$;
2. Существуют и конечны математические ожидания $M\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)$,

$$M\left(\frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta)\right) \text{ и } M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right).$$

$$\text{Тогда } M\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right) = 0, \text{ а } M\left(\frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta)\right) = -M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right) < 0.$$

Доказательство. Так как функция правдоподобия $L(\zeta, \theta) = p(\xi_1, \theta) p(\xi_2, \theta) \dots p(\xi_n, \theta)$ есть совместная плотность распределения случайного вектора $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$, то справедливо тождество

$$\int_{R^n} L(\zeta, \theta) d\zeta = 1.$$

Продифференцируем этот интеграл по θ , что допустимо по условиям теоремы

$$\int_{R^n} \frac{\partial}{\partial \theta} L(\zeta, \theta) d\zeta = 0. \quad (1)$$

Преобразуем подынтегральное выражение

$$\int_{R^n} \frac{\frac{\partial}{\partial \theta} L(\zeta, \theta)}{L(\zeta, \theta)} L(\zeta, \theta) d\zeta = 0 \quad \text{или} \quad \int_{R^n} \frac{\partial}{\partial \theta} \ln L(\zeta, \theta) \cdot L(\zeta, \theta) d\zeta = 0,$$

а это и есть математическое ожидание от производной логарифмической функции правдоподобия

$$M\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right) = \int_{R^n} \frac{\partial}{\partial \theta} \ln L(\zeta, \theta) \cdot L(\zeta, \theta) d\zeta = 0.$$

Таким образом, первое утверждение теоремы доказано.

Рассмотрим теперь вторую производную от логарифмической функции правдоподобия

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta) &= \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta) \right) = \frac{\partial}{\partial \theta} \left(\frac{\frac{\partial}{\partial \theta} L(\zeta, \theta)}{L(\zeta, \theta)} \right) = \\ &= \frac{\frac{\partial^2}{\partial \theta^2} L(\zeta, \theta) \cdot L(\zeta, \theta) - \frac{\partial}{\partial \theta} L(\zeta, \theta) \cdot \frac{\partial}{\partial \theta} L(\zeta, \theta)}{(L(\zeta, \theta))^2} = \\ &= \frac{\frac{\partial^2}{\partial \theta^2} L(\zeta, \theta)}{L(\zeta, \theta)} - \left(\frac{\frac{\partial}{\partial \theta} L(\zeta, \theta)}{L(\zeta, \theta)} \right)^2 = \frac{\frac{\partial^2}{\partial \theta^2} L(\zeta, \theta)}{L(\zeta, \theta)} - \left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta) \right)^2 \end{aligned}$$

и найдем от нее математическое ожидание

$$\begin{aligned} M\left(\frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta)\right) &= \int_{R^n} \frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta) \cdot L(\zeta, \theta) d\zeta = \\ &= \int_{R^n} \frac{\partial^2}{\partial \theta^2} L(\zeta, \theta) d\zeta - \int_{R^n} \left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta) \right)^2 L(\zeta, \theta) d\zeta = \\ &= \int_{R^n} \frac{\partial^2}{\partial \theta^2} L(\zeta, \theta) d\zeta - M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right). \end{aligned}$$

Дифференцируя по θ полученное ранее равенство (1), получим

$$\int_{R^n} \frac{\partial^2}{\partial \theta^2} L(\zeta, \theta) d\zeta = 0.$$

Таким образом, $M\left(\frac{\partial^2}{\partial \theta^2} \ln L(\zeta, \theta)\right) = -M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right) < 0$ и теорема доказана.

Функцию $I(\theta) = M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right)$ называют количеством информации по Фишеру.

В определенном смысле ее можно рассматривать как количество информации, содержащемся в векторе $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ о значении параметра θ .

Теперь утверждения теоремы мы можем записать в виде

$$M\left(\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\right)=0 \text{ и } M\left(\frac{\partial^2}{\partial\theta^2}\ln L(\zeta,\theta)\right)=-I(\theta)<0.$$

Полученный результат может служить определенным обоснованием рассмотренного метода максимального правдоподобия, так как точечная оценка θ^* параметра θ находилась из условий максимума логарифмической функции правдоподобия $\ln L(\zeta,\theta)$, а именно из условий $\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)=0$ и $\frac{\partial^2}{\partial\theta^2}\ln L(\zeta,\theta)<0$.

Неравенство Рао-Крамера. Пусть выполнены оба условия предыдущей теоремы и пусть $\theta^*=h(\zeta)$ есть несмещенная точечная оценка параметра θ . Тогда

$$Dh(\zeta)=M\left(\left(h(\zeta)-\theta\right)^2\right)\geq\frac{1}{I(\theta)}.$$

Доказательство. Рассмотрим равенство

$$M\theta^*=Mh(\zeta)=\int_{R^n}h(\zeta)L(\zeta,\theta)d\zeta=\theta$$

и продифференцируем его по θ

$$\int_{R^n}h(\zeta)\frac{\partial}{\partial\theta}L(\zeta,\theta)d\theta=1.$$

Преобразуем левую часть этого равенства

$$\int_{R^n}h(\zeta)\frac{\partial}{\partial\theta}L(\zeta,\theta)d\zeta=\int_{R^n}h(\zeta)\frac{\frac{\partial}{\partial\theta}L(\zeta,\theta)}{L(\zeta,\theta)}L(\zeta,\theta)d\zeta=\int_{R^n}h(\zeta)\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\cdot L(\zeta,\theta)d\zeta.$$

Таким образом

$$\int_{R^n}h(\zeta)\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\cdot L(\zeta,\theta)d\zeta=1.$$

Из доказанной теоремы следует, что

$$M\left(\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\right)=\int_{R^n}\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\cdot L(\zeta,\theta)d\zeta=0.$$

Умножим последнее равенство на θ и вычтем из предыдущего

$$\int_{R^n}(h(\zeta)-\theta)\cdot\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\cdot L(\zeta,\theta)d\zeta=1, \text{ то есть}$$

$$M\left((h(\zeta)-\theta)\cdot\frac{\partial}{\partial\theta}\ln L(\zeta,\theta)\right)=1. \quad (2)$$

Воспользуемся неравенством Коши-Буняковского

$$\int_{R^n}f(\zeta)g(\zeta)d\zeta\leq\sqrt{\int_{R^n}f^2(\zeta)d\zeta}\cdot\sqrt{\int_{R^n}g^2(\zeta)d\zeta},$$

где равенство достигается только если $g(\zeta)=C\cdot f(\zeta)$, $C-\text{const}$.

Преобразуем его применительно к математическим ожиданиям

$$\begin{aligned}
M(f(\zeta) \cdot g(\zeta)) &= \int_{R^n} f(\zeta) g(\zeta) L(\zeta, \theta) d\zeta = \int_{R^n} \left(f(\zeta) \sqrt{L(\zeta, \theta)} \right) \cdot \left(g(\zeta) \sqrt{L(\zeta, \theta)} \right) d\zeta \leq \\
&\leq \sqrt{\int_{R^n} f^2(\zeta) L(\zeta, \theta) d\zeta} \cdot \sqrt{\int_{R^n} g^2(\zeta) L(\zeta, \theta) d\zeta} = \sqrt{M(f^2(\zeta))} \cdot \sqrt{M(g^2(\zeta))}.
\end{aligned}$$

Применим полученное неравенство к соотношению (2)

$$\begin{aligned}
1 &= M\left((h(\zeta) - \theta) \cdot \frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right) \leq \sqrt{M((h(\zeta) - \theta)^2)} \cdot \sqrt{M\left(\left(\frac{\partial}{\partial \theta} \ln L(\zeta, \theta)\right)^2\right)} = \\
&= \sqrt{Dh(\zeta)} \cdot \sqrt{I(\theta)},
\end{aligned}$$

откуда получаем неравенство Рао-Крамера $Dh(\zeta) \geq \frac{1}{I(\theta)}$.

При этом равенство в нем будет достигаться, если

$$\frac{\partial}{\partial \theta} \ln L(\zeta, \theta) = C(\theta)(h(\zeta) - \theta). \quad (3)$$

Доказательство теоремы и получение неравенства Рао-Крамера было проведено для абсолютно-непрерывных распределений, но все остается в силе и для дискретных распределений. Надо только при доказательстве заменить плотность $p(\xi, \theta)$ на вероятности $p(\xi_i, \theta)$, а интегралы на суммы.

Определение. Несмещенная точечная оценка θ^* параметра θ называется оптимальной или наиболее эффективной, если

$$D\theta^* = \inf_n D\theta_n^*.$$

Выполнение условия (3), при котором неравенство Рао-Крамера переходит в точное равенство, есть условие того, что полученная несмещенная оценка будет наиболее эффективной.

Пример 1. Пусть случайная величина ξ распределена по закону Пуассона

$$P\{\xi = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0, k = 0, 1, 2, \dots$$

Найдем по выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ методом максимального правдоподобия точечную оценку параметра λ и исследуем ее на эффективность.

Составим функцию правдоподобия

$$L(\zeta, \lambda) = \frac{\lambda^{\xi_1}}{\xi_1!} e^{-\lambda} \cdot \frac{\lambda^{\xi_2}}{\xi_2!} e^{-\lambda} \cdot \dots \cdot \frac{\lambda^{\xi_n}}{\xi_n!} e^{-\lambda} = A \cdot \lambda^{\xi_1 + \xi_2 + \dots + \xi_n} \cdot e^{-\lambda n} = A \cdot \lambda^{n\bar{x}_\theta} e^{-\lambda n},$$

где $A = (\xi_1! \cdot \xi_2! \cdot \dots \cdot \xi_n!)^{-1} - \text{const}$ и $\xi_1 + \xi_2 + \dots + \xi_n = n\bar{x}_\theta$.

Найдем $\ln L(\zeta, \lambda) = \ln A + n\bar{x}_\theta \ln \lambda - \lambda n$ и вычислим производную

$$\frac{\partial}{\partial \lambda} \ln L(\zeta, \lambda) = n\bar{x}_\theta \cdot \frac{1}{\lambda} - n = \frac{n}{\lambda} (\bar{x}_\theta - \lambda),$$

которая обращается в ноль при $\lambda = \lambda^* = \bar{x}_e$.

При этом $\frac{\partial^2}{\partial \lambda^2} \ln L(\zeta, \lambda) = -\frac{n\bar{x}_e}{\lambda^2} < 0$, а значит, при $\lambda = \lambda^* = \bar{x}_e$ функция правдоподобия

имеет максимум и мы нашли точечную оценку $\lambda^* = \bar{x}_e$. Эта оценка будет несмещенной $M\lambda^* = M\bar{x}_e = M\xi = \lambda$, поскольку для распределения Пуассона $M\xi = \lambda$.

Кроме того из полученного равенства $\frac{\partial}{\partial \lambda} \ln L(\zeta, \lambda) = \frac{n}{\lambda}(\bar{x}_e - \lambda)$ следует, что выполняется

условие (3) (при $C(\lambda) = \frac{n}{\lambda}$, $h(\zeta) = \bar{x}_e$) того, что полученная несмещенная оценка будет наиболее эффективной.

В этом случае неравенство Рао-Крамера переходит в точное равенство

$$Dh(\zeta) = D\bar{x}_e = \frac{1}{I(\lambda)}.$$

Найдем $D\bar{x}_e = D\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D\xi_i = \frac{1}{n^2} \cdot n D\xi = \frac{\lambda}{n}$, так как для распределения Пуассона $D\xi = \lambda$.

Таким образом, $D\bar{x}_e = \frac{\lambda}{n} = \frac{1}{I(\lambda)}$, откуда $I(\lambda) = \frac{n}{\lambda}$.

Пример 2. Пусть случайная величина ξ распределена по нормальному закону с

плотностью $p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$. Предположим, что параметр a известен, а

для параметра σ требуется по выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ методом максимального правдоподобия найти точечную оценку и исследовать ее на эффективность.

Составим функцию правдоподобия

$$L(\zeta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\xi_1-a)^2}{2\sigma^2}} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\xi_2-a)^2}{2\sigma^2}} \cdot \dots \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\xi_n-a)^2}{2\sigma^2}} = (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i-a)^2}.$$

Найдем $\ln L(\zeta, \sigma) = -n \ln \sigma - n \ln \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i - a)^2$ и вычислим

$$\frac{\partial}{\partial \sigma} \ln L(\zeta, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (\xi_i - a)^2 = \frac{n}{\sigma^3} \left[\frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2 - \sigma^2 \right].$$

Производная равна нулю, если $\sigma^2 = \sigma^{*2} = \frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2$. При этом

$$\frac{\partial^2}{\partial \sigma^2} \ln L(\zeta, \sigma^*) = \frac{n}{\sigma^{*2}} - \frac{3}{\sigma^{*4}} \sum_{i=1}^n (\xi_i - a)^2 = \frac{n}{\sigma^{*2}} - \frac{3}{\sigma^{*4}} n \sigma^{*2} = -\frac{2n}{\sigma^{*2}} < 0.$$

Таким образом, функция правдоподобия имеет максимум при $\sigma = \sigma^*$ и мы нашли

точечную оценку $\sigma^2 \approx \sigma^{*2} = \frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2$. Отметим сразу, что эта точечная оценка для

дисперсии $D\xi$ не является выборочной дисперсией $D_\varepsilon = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x}_\varepsilon)^2$, так как параметр a здесь известен.

Покажем, что полученная оценка является несмещенной

$$M(\sigma^{*2}) = M\left(\frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2\right) = \frac{1}{n} \sum_{i=1}^n M(\xi_i - a)^2 = \frac{1}{n} \cdot n M(\xi - a)^2 = D\xi = \sigma^2.$$

Кроме того выполнено условие (3)

$$\frac{\partial}{\partial \sigma} \ln L(\zeta, \sigma) = \frac{n}{\sigma^3} \left[\frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2 - \sigma^2 \right] \text{ при } C(\sigma) = \frac{n}{\sigma^3}, h(\zeta) = \frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2,$$

а значит, полученная несмещенная точечная оценка является наиболее эффективной и в неравенстве Рао-Крамера будет точное равенство.

Вычислим дисперсию полученной оценки

$$\begin{aligned} D\sigma^{*2} &= D\left(\frac{1}{n} \sum_{i=1}^n (\xi_i - a)^2\right) = \frac{1}{n^2} \sum_{i=1}^n D(\xi_i - a)^2 = \frac{1}{n^2} \cdot n D(\xi - a)^2 = \\ &= \frac{1}{n} \left[M(\xi - a)^4 - (M(\xi - a)^2)^2 \right] = \frac{1}{n} \left[M(\xi - a)^4 - (D\xi)^2 \right] = \frac{1}{n} \left[M(\xi - a)^4 - \sigma^4 \right]. \end{aligned}$$

$$\begin{aligned} \text{Найдем } M(\xi - a)^4 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - a)^4 e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \left[\begin{matrix} x - a = \sigma t, x = a + \sigma t \\ dx = \sigma dt \end{matrix} \right] = \\ &= \frac{\sigma^4}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 e^{-\frac{t^2}{2}} dt = \left[\begin{matrix} u = t^3, du = 3t^2 dt \\ dv = te^{-\frac{t^2}{2}} dt, v = -e^{-\frac{t^2}{2}} \end{matrix} \right] = \frac{\sigma^4}{\sqrt{2\pi}} \left[-t^3 e^{-\frac{t^2}{2}} \Big|_{-\infty}^{\infty} + 3 \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} dt \right] = \\ &= \left[\begin{matrix} u = t, du = dt \\ dv = te^{-\frac{t^2}{2}} dt, v = -e^{-\frac{t^2}{2}} \end{matrix} \right] = \frac{3\sigma^4}{\sqrt{2\pi}} \left[-te^{-\frac{t^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \right] = \frac{3\sigma^4}{\sqrt{2\pi}} \sqrt{2\pi} = 3\sigma^4. \end{aligned}$$

$$\text{Тогда } D\sigma^{*2} = \frac{1}{n} (3\sigma^4 - \sigma^4) = \frac{2\sigma^4}{n} \text{ и } D\sigma^{*2} = \frac{2\sigma^4}{n} = \frac{1}{I(\sigma)}, \text{ откуда } I(\sigma) = \frac{n}{2\sigma^4}.$$

Задание.

Пусть случайная величина ξ распределена по нормальному закону с

плотностью $p_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$. Предположим, что параметр σ известен, а

для параметра a требуется по выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ методом максимального правдоподобия найти точечную оценку и исследовать ее на эффективность.

§ 10. Интервальные оценки параметров распределений.

Понятие оценки параметра распределения будет неполным, если не указать величину ошибки, возникающей при использовании полученной оценки. Наглядное представление о точности оценок дают доверительные интервалы.

Доверительным интервалом для параметра распределения θ называется случайный интервал $(\theta'; \theta'')$, внутри которого с вероятностью γ , близкой к единице, содержится точное значение оцениваемого параметра

$$P\{\theta \in (\theta'; \theta'')\} = \gamma.$$

Величина γ называется доверительной вероятностью. Она выбирается в соответствии с готовностью мириться с возможностью ошибки. При заданной доверительной вероятности γ длина доверительного интервала характеризует точность, с которой локализовано значение параметра. Поэтому естественно выбирать среди различных доверительных интервалов при заданной доверительной вероятности γ интервал кратчайшей длины. Такой выбор обеспечивает наиболее точную локализацию параметра.

Отличие интервальной оценки от точечной состоит в следующем:

- доверительный интервал $(\theta'; \theta'')$ как оценка менее точен в том смысле, что он указывает целое множество возможных значений θ ;
- с другой стороны утверждение: « $\theta \in (\theta'; \theta'')$ с вероятностью γ » является истинным, в то время как событие $\theta = \theta^*$, как правило, имеет вероятность, равную нулю. Объявляя верным утверждение $\theta \in (\theta'; \theta'')$, мы используем тот факт, что если некоторое событие имеет вероятность γ и это γ близко к единице, то практически невозможно, чтобы это событие не произошло при единичном испытании.

Очевидно, что существуют различные способы построения доверительных интервалов. Рассмотрим один из них, основанный на использовании точечной оценки θ^* параметра. Построенные на основе этой оценки интервалы будут тем короче, чем меньше дисперсия оценки. Таким образом, эффективные и асимптотически эффективные оценки приводят к кратчайшим или асимптотически кратчайшим доверительным интервалам.

Закон распределения оценки θ^* в общем случае зависит от самих оцениваемых параметров и, как правило, не известен. Поэтому будем переходить от случайной величины θ^* к некоторой другой случайной величине, закон которой уже не зависит от неизвестных параметров, а зависит только от числа опытов n и от вида закона распределения случайной величины ξ . Как правило, это нормированное нормальное распределение, распределение Стьюдента или распределение Пирсона. Это позволит нам по доверительной вероятности γ найти доверительный интервал. Кроме того эти доверительные интервалы будем строить симметричными относительно оценки θ^* . Рассмотрим применение этой схемы для нахождения доверительных интервалов для параметров некоторых видов распределений.

§ 10.1. Построение доверительных интервалов для параметров нормального распределения.

Поскольку нормальное распределение имеет два параметра a и σ , то рассмотрим следующие случаи.

I. Нахождение доверительного интервала для параметра a , если дисперсия σ^2 известна.

Несмещенной точечной оценкой для параметра a является выборочное среднее

$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n \xi_i$. Оно имеет математическое ожидание $M\bar{x}_n = a$ и дисперсию

$$D\bar{x}_n = D\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D\xi_i = \frac{1}{n^2} \cdot n D\xi = \frac{\sigma^2}{n}.$$

Введем нормированную случайную величину

$$\eta = \frac{\bar{x}_n - M\bar{x}_n}{\sqrt{D\bar{x}_n}} = \frac{\bar{x}_n - a}{\sigma} \sqrt{n},$$

которая является нормированной суммой случайных величин, имеющих математическое ожидание и дисперсию. Тогда к ней применима центральная предельная теорема, согласно которой она будет асимптотически нормальной с параметрами $a = 0$ и $\sigma = 1$.

Тогда $P\{|\eta| < \delta\} = 2\Phi(\delta) = \gamma$, где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$.

По таблице значений функции $\Phi(x)$ находим решение $\delta = \delta_\gamma$ уравнения $\Phi(\delta) = \frac{\gamma}{2}$.

Тогда $\frac{|\bar{x}_n - a|}{\sigma} \sqrt{n} < \delta_\gamma$, $|\bar{x}_n - a| < \frac{\sigma}{\sqrt{n}} \delta_\gamma$ и $a \in \left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \delta_\gamma; \bar{x}_n + \frac{\sigma}{\sqrt{n}} \delta_\gamma\right)$ требуемый

доверительный интервал для параметра a .

Пример 1. Найти доверительный интервал для оценки с надежностью $\gamma = 0,95$ параметра a нормального распределения, если среднеквадратичное отклонение $\sigma = 5$, выборочное среднее $\bar{x}_n = 14$, а объем выборки $n = 64$.

Из уравнения $\Phi(\delta) = \frac{\gamma}{2} = 0,475$ по таблице значений функции $\Phi(x)$ (Таблица 2) находим $\delta = \delta_\gamma = 1,96$. Доверительный интервал

$$a \in \left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \delta_\gamma; \bar{x}_n + \frac{\sigma}{\sqrt{n}} \delta_\gamma\right) = \left(14 - 1,96 \cdot \frac{5}{8}; 14 + 1,96 \cdot \frac{5}{8}\right) = (12,775; 15,225).$$

Пример 2. Найти минимальный объем выборки, при котором с надежностью $\gamma = 0,95$ точность оценки математического ожидания a равна $\delta = 0,5$, если $\sigma = 5$.

Так как $\delta = \delta_\gamma \frac{\sigma}{\sqrt{n}}$, то $n = \left(\delta_\gamma \frac{\sigma}{\delta}\right)^2$. В данном случае $\delta_\gamma = 1,96$ и

$$n = \left(1,96 \cdot \frac{5}{0,5}\right)^2 = (39,2)^2 \approx 1536.$$

II. Нахождение доверительного интервала для параметра a , если дисперсия σ^2 не известна.

Несмещенной точечной оценкой для параметра a является выборочное среднее

$$\bar{x}_e = \frac{1}{n} \sum_{i=1}^n \xi_i. \text{ При этом } M\bar{x}_e = a \text{ и } D\bar{x}_e = \frac{\sigma^2}{n}.$$

Так как дисперсия σ^2 нам не известна, то приблизим ее точечной несмещенной оценкой

$$s^2 = \frac{n}{n-1} D_e, \text{ то есть возьмем } D\bar{x}_e \approx \frac{s^2}{n}. \text{ Введем нормированную случайную величину}$$

$$t = \frac{\bar{x}_e - M\bar{x}_e}{\sqrt{D\bar{x}_e}} = \frac{\bar{x}_e - a}{s} \sqrt{n} = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n \xi_i - a}{\sqrt{\frac{n}{n-1} D_e}} = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n (\xi_i - a)}{\sqrt{\frac{n}{n-1} D_e}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\xi_i - a}{\sigma}}{\sqrt{\frac{n}{n-1} \cdot \frac{D_e}{\sigma^2}}}.$$

Перейдем к случайным величинам $\xi'_i = \frac{\xi_i - a}{\sigma}$, которые распределены нормально с

параметрами $a' = 0$ и $\sigma' = 1$. При этом $D'_e = \frac{D_e}{\sigma^2}$ и тогда

$$t = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{n}{n-1} D'_e}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n \xi'^2_i - \left(\frac{1}{n} \sum_{i=1}^n \xi'_i \right)^2 \right]}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n \xi'^2_i - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i \right)^2 \right]}}.$$

Введем в n -мерном пространстве выборок $(\xi'_1, \xi'_2, \dots, \xi'_n)$ новую ортогональную систему

координат $(\eta_1, \eta_2, \dots, \eta_n)$ так, что $\eta_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i$, $\eta_j = \sum_{i=1}^n \alpha_{ji} \xi'_i$ и $\sum_{j=1}^n \alpha_{ji} \alpha_{jm} = \delta_{im} = \begin{cases} 1, i = m \\ 0, i \neq m \end{cases}$.

Тогда

$$\sum_{j=1}^n \eta_j^2 = \sum_{j=1}^n \left(\sum_{k=1}^n \alpha_{jk} \xi'_k \right)^2 = \sum_{j=1}^n \left(\sum_{i,m=1}^n \alpha_{ji} \alpha_{jm} \xi'_i \xi'_m \right) = \sum_{i,m=1}^n \xi'_i \xi'_m \sum_{j=1}^n \alpha_{ji} \alpha_{jm} = \sum_{i=1}^n \xi'^2_i \text{ и}$$

$$t = \frac{\eta_1}{\sqrt{\frac{1}{n-1} \left[\sum_{j=1}^n \eta_j^2 - \eta_1^2 \right]}} = \frac{\eta_1}{\sqrt{\frac{1}{n-1} \sum_{j=2}^n \eta_j^2}}.$$

Поскольку случайные величины η_j , так же как и ξ'_i , распределены нормально с параметрами 0 и 1, то случайная величина t имеет распределение Стьюдента с $n-1$ степенью свободы.

Из уравнения $P\{|t| < t_\gamma\} = \gamma$ по таблице 3 находим соответствующее значение $t_\gamma = t(\gamma, n)$.

Тогда $|t| = \frac{\sqrt{n}}{s} |\bar{x}_e - a| < t_\gamma$, $|\bar{x}_e - a| < \frac{s}{\sqrt{n}} t_\gamma$ и искомый доверительный интервал

$$a \in \left(\bar{x}_e - \frac{s}{\sqrt{n}} t_\gamma; \bar{x}_e + \frac{s}{\sqrt{n}} t_\gamma \right).$$

Полученное выражение для доверительного интервала совпадает с предыдущим случаем (когда дисперсия σ^2 известна) с заменой σ на s и δ_γ на t_γ .

Так как с ростом n распределение Стьюдента стремится к нормальному распределению, то при достаточно больших n $t_\gamma = \delta_\gamma$ и доверительный интервал находят по формуле

$$a \in \left(\bar{x}_e - \frac{s}{\sqrt{n}} \delta_\gamma; \bar{x}_e + \frac{s}{\sqrt{n}} \delta_\gamma \right).$$

Пример 3. Из генеральной совокупности, распределенной нормально, извлечена выборка объема $n = 25$:

x_i	-1,0	-0,5	0,0	0,5	1,0	1,5	2,0
n_i	2	3	5	8	4	2	1

Найти с надежностью $\gamma = 0,95$ доверительный интервал для математического ожидания a .

По заданной выборке вычислим выборочное среднее и исправленную выборочную дисперсию

$$\bar{x}_e = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{1}{25} (-2 - 1,5 + 4 + 4 + 3 + 2) = 0,38$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_e)^2 n_i = \frac{1}{24} (1,38^2 \cdot 2 + 0,88^2 \cdot 3 + 0,38^2 \cdot 5 + 0,12^2 \cdot 8 + 0,72^2 \cdot 4 + 1,12^2 \cdot 2 + 1,62^2) = \frac{14,176}{24} \approx 0,5907 \text{ и } s = \sqrt{0,5907} \approx 0,7686.$$

По таблице 3 находим $t_\gamma = t(0,95; 25) = 2,064$. Тогда

$$a \in \left(0,38 - \frac{0,7686}{5} \cdot 2,064; 0,38 + \frac{0,7686}{5} \cdot 2,064 \right), \text{ то есть } a \in (0,063; 0,697).$$

III. Нахождение доверительного интервала для параметра σ , если математическое ожидание a не известно.

Несмещенной точечной оценкой для дисперсии σ^2 является исправленная выборочная дисперсия $s^2 = \frac{n}{n-1} D_e$. Будем искать доверительный интервал $(s - \delta; s + \delta)$ так, чтобы

$$P\{s - \delta < \sigma < s + \delta\} = \gamma.$$

Положим $s - \delta = \frac{\sqrt{nD_e}}{t_1}$ и $s + \delta = \frac{\sqrt{nD_e}}{t_2}$, где t_1 и t_2 неизвестные пока величины.

Тогда неравенство $s - \delta < \sigma < s + \delta$ перейдет в неравенство $\frac{\sqrt{nD_e}}{t_1} < \sigma < \frac{\sqrt{nD_e}}{t_2}$ или

$$\begin{cases} \sqrt{nD_e} < \sigma t_1 \\ \sqrt{nD_e} > \sigma t_2 \end{cases}, \text{ откуда } \sigma t_2 < \sqrt{nD_e} < \sigma t_1 \text{ или } t_2 < \frac{\sqrt{nD_e}}{\sigma} < t_1.$$

Таким образом получаем условие $P\{s - \delta < \sigma < s + \delta\} = P\left\{t_2 < \frac{\sqrt{nD_\epsilon}}{\sigma} < t_1\right\} = \gamma$.

Так как математическое ожидание a неизвестно, то приблизим его выборочным средним $\bar{x}_\epsilon$ и введем нормированные случайные величины $\xi'_i = \frac{\xi_i - \bar{x}_\epsilon}{\sigma}$, распределенные

нормально. Тогда $D'_\epsilon = \frac{D_\epsilon}{\sigma^2}$ и $\frac{\sqrt{nD_\epsilon}}{\sigma} = \sqrt{nD'_\epsilon} = \sqrt{\sum_{i=1}^n \xi_i'^2 - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i'\right)^2}$.

Как и в предыдущем пункте осуществим линейное ортогональное преобразование

координат $\eta_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i'$, $\eta_j = \sum_{i=1}^n \alpha_{ji} \xi_i'$ и $\sum_{j=1}^n \alpha_{ji} \alpha_{jm} = \delta_{im} = \begin{cases} 1, i = m \\ 0, i \neq m \end{cases}$. При этом случайные величины η_j также нормированы и распределены нормально.

Там же мы показали, что $\sum_{j=1}^n \eta_j^2 = \sum_{i=1}^n \xi_i'^2$. Тогда

$$\sqrt{nD'_\epsilon} = \sqrt{\sum_{i=1}^n \xi_i'^2 - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i'\right)^2} = \sqrt{\sum_{j=1}^n \eta_j^2 - \eta_1^2} = \sqrt{\sum_{j=2}^n \eta_j^2}$$

и мы приходим к соотношению $P\left\{t_2 < \sqrt{\sum_{j=2}^n \eta_j^2} < t_1\right\} = \gamma$.

Случайная величина $\chi = \sqrt{\sum_{j=2}^n \eta_j^2}$ имеет χ распределение с $n-1$ степенью свободы и нам известна плотность $p_\chi(x)$ ее функции распределения (см. § 6).

Для нахождения величин t_1 и t_2 получаем уравнение $P\{t_2 < \chi < t_1\} = \int_{t_2}^{t_1} p_\chi(x) dx = \gamma$.

Преобразуем выражения для t_1 и t_2 , чтобы получить в этом уравнении одно неизвестное

$$t_1 = \frac{\sqrt{nD_\epsilon}}{s - \delta} = \frac{\sqrt{n \cdot \frac{n-1}{n} s^2}}{s - \delta} = \frac{s\sqrt{n-1}}{s - \delta} = \frac{\sqrt{n-1}}{1 - q}, \text{ где } q = \frac{\delta}{s}.$$

$$\text{Аналогично } t_2 = \frac{\sqrt{nD_\epsilon}}{s + \delta} = \frac{s\sqrt{n-1}}{s + \delta} = \frac{\sqrt{n-1}}{1 + q}.$$

Получили уравнение относительно q : $\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} p_\chi(x) dx = \gamma$.

По таблице 4 находим его решение $q = q(\gamma, n)$. Тогда $\delta = sq$ и доверительный интервал имеет вид $\sigma \in (s - \delta; s + \delta) = (s(1 - q); s(1 + q))$. При этом, если $1 - q < 0$, то так как $\sigma > 0$, берется интервал $\sigma \in (0; s(1 + q))$.

Пример 4. Проведено $n = 20$ измерений одним прибором без систематической ошибки некоторой величины. Найдена исправленная выборочная дисперсия $s = 0,8$. Найти точность прибора с надежностью $\gamma = 0,99$.

Ошибки измерений подчинены нормальному закону, а точность прибора характеризуется среднеквадратичным отклонением случайных ошибок измерений. Таким образом, задача сводится к отысканию доверительного интервала для σ с заданной надежностью $\gamma = 0,99$

По таблице 4 находим $q = q(0,99; 20) = 0,58$. Тогда

$$\sigma \in (0,8(1-0,58); 0,8(1+0,58)) = (0,336; 1,264).$$

§ 10.2. Построение доверительного интервала для параметра λ распределения Пуассона.

Пусть случайная величина ξ распределена по закону Пуассона

$$P\{\xi = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0, k = 0, 1, 2, \dots$$

По выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ найдем с надежностью γ доверительный интервал для

параметра λ . Несмещенной точечной оценкой параметра λ является $\lambda^* = \bar{x}_\epsilon = \frac{1}{n} \sum_{i=1}^n \xi_i$.

При этом $M\lambda^* = M\bar{x}_\epsilon = \lambda$ и $D\lambda^* = D\bar{x}_\epsilon = D\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D\xi_i = \frac{1}{n^2} n D\xi = \frac{\lambda}{n}$, так как для распределения Пуассона $M\xi = \lambda$ и $D\xi = \lambda$.

Введем нормированную случайную величину

$$\eta = \frac{\lambda^* - M\lambda^*}{\sqrt{D\lambda^*}} = \frac{\bar{x}_\epsilon - \lambda}{\sqrt{\frac{\lambda}{n}}} = \sqrt{\frac{n}{\lambda}} (\bar{x}_\epsilon - \lambda) = \sqrt{\frac{n}{\lambda}} \left(\frac{1}{n} \sum_{i=1}^n \xi_i - \lambda \right),$$

которая по центральной предельной теореме асимптотически нормальна с параметрами $a = 0$ и $\sigma = 1$. Тогда $P\{|\eta| < \delta\} = 2\Phi(\delta) = \gamma$. Теперь по таблице 2 значений функции $\Phi(x)$

находим решение δ_γ уравнения $\Phi(\delta) = \frac{\gamma}{2}$.

Далее из неравенства $|\eta| = \sqrt{\frac{n}{\lambda}} \cdot |\bar{x}_\epsilon - \lambda| < \delta_\gamma$ следует найти интервал изменения λ .

Запишем это неравенство в виде системы

$$\begin{cases} \sqrt{\frac{n}{\lambda}} \cdot (\bar{x}_\epsilon - \lambda) < \delta_\gamma \\ \sqrt{\frac{n}{\lambda}} \cdot (\bar{x}_\epsilon - \lambda) > -\delta_\gamma \end{cases} \quad \text{или} \quad \begin{cases} \sqrt{n} \cdot \bar{x}_\epsilon - \lambda \sqrt{n} < \delta_\gamma \sqrt{\lambda} \\ \sqrt{n} \cdot \bar{x}_\epsilon - \lambda \sqrt{n} > -\delta_\gamma \sqrt{\lambda} \end{cases}, \text{ то есть } \begin{cases} \lambda \cdot \sqrt{n} + \delta_\gamma \sqrt{\lambda} - \bar{x}_\epsilon \sqrt{n} > 0 \\ \lambda \cdot \sqrt{n} - \delta_\gamma \sqrt{\lambda} - \bar{x}_\epsilon \sqrt{n} < 0 \end{cases}.$$

Решим эти квадратичные неравенства относительно $\sqrt{\lambda}$.

Нулями первого квадратного трехчлена будут $(\sqrt{\lambda})_{1,2} = -\frac{\delta_\gamma}{2\sqrt{n}} \pm \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}$ и решением

первого неравенства с учетом того, что $\sqrt{\lambda} > 0$ будет $\sqrt{\lambda} > -\frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}$.

Нулями второго квадратного трехчлена будут $(\sqrt{\lambda})_{1,2} = \frac{\delta_\gamma}{2\sqrt{n}} \pm \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}$ и его решение

есть $\sqrt{\lambda} < \frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}$. Таким образом, решением системы неравенств будет

$$-\frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}} < \sqrt{\lambda} < \frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}.$$

Отсюда искомый доверительный интервал есть

$$\left(-\frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}\right)^2 < \lambda < \left(\frac{\delta_\gamma}{2\sqrt{n}} + \sqrt{\bar{x}_e + \frac{\delta_\gamma^2}{4n}}\right)^2.$$

Пример 5. Случайная величина ξ распределена по закону Пуассона. По выборке объема $n=100$

x_i	0	1	2	3	4
n_i	58	28	11	2	1

Найти с надежностью $\gamma = 0,95$ доверительный интервал для параметра λ .

Вычислим сначала $\bar{x}_e = \frac{1}{100}(28 + 22 + 11 + 2 + 1) = 0,6$. Затем по таблице 2 из уравнения

$\Phi(\delta) = \frac{\gamma}{2} = \frac{0,95}{2} = 0,475$ находим $\delta_\gamma = 1,96$. Согласно полученной формуле

$$\left(-\frac{1,96}{20} + \sqrt{0,6 + \frac{1,96^2}{400}}\right)^2 < \lambda < \left(\frac{1,96}{20} + \sqrt{0,6 + \frac{1,96^2}{400}}\right)^2, \text{ то есть } 0,4662 < \lambda < 0,8788.$$

§ 10.3. Построение доверительного интервала для вероятности p биномиального распределения.

Пусть случайная величина ξ имеет биномиальное распределение

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k}.$$

Ее математическое ожидание и дисперсия соответственно равны $M\xi = np$ и $D\xi = npq$.

Будем считать, что известна относительная частота $\omega = \frac{m}{n}$ появления события в n

независимых испытаниях. Найдем доверительный интервал для вероятности p наступления события в одном испытании.

В качестве точечной оценки для вероятности p возьмем относительную частоту

$p^* = \omega = \frac{m}{n}$. Эта оценка будет несмещенной, так как $\xi = m$ и

$$M(p^*) = M(\omega) = M\left(\frac{m}{n}\right) = \frac{1}{n} M(m) = \frac{1}{n} \cdot np = p$$

и ее дисперсия

$$D(p^*) = D(\omega) = D\left(\frac{m}{n}\right) = \frac{1}{n^2} D(m) = \frac{1}{n^2} \cdot npq = \frac{pq}{n} = \frac{p(1-p)}{n}.$$

Введем нормированную случайную величину

$$\eta = \frac{p^* - M(p^*)}{\sqrt{D(p^*)}} = \frac{\omega - p}{\sqrt{p(1-p)}} \sqrt{n}.$$

По теореме Муавра-Лапласа она асимптотически нормальна с параметрами $a = 0$ и $\sigma = 1$.

Тогда $P\{|\eta| < \delta\} = 2\Phi(\delta) = \gamma$. Теперь по таблице 2 значений функции $\Phi(x)$ находим

решение δ_γ уравнения $\Phi(\delta) = \frac{\gamma}{2}$.

Доверительный интервал для вероятности p найдем, решив неравенство

$\sqrt{\frac{n}{p(1-p)}} \cdot |\omega - p| < \delta_\gamma$. Для этого возведем его в квадрат и приведем подобные члены

$$(\delta_\gamma^2 + n)p^2 - (\delta_\gamma^2 + 2n\omega)p + n\omega^2 < 0.$$

Найдем дискриминант $D = (\delta_\gamma^2 + 2n\omega)^2 - 4n\omega^2(\delta_\gamma^2 + n) = 4n^2\delta_\gamma^2\left(\frac{\delta_\gamma^2}{4n^2} + \frac{\omega(1-\omega)}{n}\right)$ и корни

квадратного трехчлена $p_{1,2} = \frac{n}{\delta_\gamma^2 + n} \left[\omega + \frac{\delta_\gamma^2}{2n} \pm \delta_\gamma \sqrt{\frac{\delta_\gamma^2}{4n^2} + \frac{\omega(1-\omega)}{n}} \right]$. Тогда решение

квадратичного неравенства, то есть доверительный интервал имеет вид

$$\frac{n}{\delta_\gamma^2 + n} \left[\omega + \frac{\delta_\gamma^2}{2n} - \delta_\gamma \sqrt{\frac{\delta_\gamma^2}{4n^2} + \frac{\omega(1-\omega)}{n}} \right] < p < \frac{n}{\delta_\gamma^2 + n} \left[\omega + \frac{\delta_\gamma^2}{2n} + \delta_\gamma \sqrt{\frac{\delta_\gamma^2}{4n^2} + \frac{\omega(1-\omega)}{n}} \right].$$

Если n достаточно велико, то величины $\frac{\delta_\gamma^2}{2n}$ и $\frac{\delta_\gamma^2}{4n^2}$ малы и ими можно пренебречь, а

величина $\frac{n}{\delta_\gamma^2 + n} \approx 1$. В этом случае получаем упрощенную формулу

$$\omega - \delta_\gamma \sqrt{\frac{\omega(1-\omega)}{n}} < p < \omega + \delta_\gamma \sqrt{\frac{\omega(1-\omega)}{n}}.$$

Пример 6. Проведено $n = 160$ независимых испытаний, в которых событие A появилось $m = 16$ раз. Найти с надежностью $\gamma = 0,95$ доверительный интервал для вероятности p появления события A в одном испытании.

В данном случае $\omega = \frac{m}{n} = \frac{16}{160} = 0,1$. По таблице 2 находим решение $\delta_\gamma = 1,96$ уравнения

$\Phi(\delta) = \frac{\gamma}{2} = \frac{0,95}{2} = 0,475$. По упрощенной формуле находим доверительный интервал

$$0,1 - \sqrt{\frac{0,1 \cdot 0,9}{160}} < p < 0,1 + \sqrt{\frac{0,1 \cdot 0,9}{160}}, \text{ то есть } 0,0954 < p < 0,1046.$$

Задания.

1. Выборка из большой партии электроламп содержит 100 ламп. Средняя продолжительность горения лампы выборки оказалась равной 1000 часов. Найти с надежностью 0,95 доверительный интервал для средней продолжительности a горения лампы всей партии, если известно, что среднеквадратичное отклонение продолжительности горения лампы $\sigma = 40$ часов. Предполагается, что продолжительность горения ламп распределена нормально.
2. Глубина моря измеряется прибором, систематическая ошибка которого равна нулю, а случайные ошибки распределены нормально со среднеквадратичным отклонением $\sigma = 20$ метров. Сколько надо сделать независимых измерений, чтобы определить глубину с ошибкой не более 15 метров при надежности 90%?
3. Из генеральной совокупности, распределенной нормально, извлечена выборка

x_i	1	2	3	4	5	6	7
n_i	2	3	7	12	9	2	1

Найти доверительный интервал для оценки с надежностью 0,95 неизвестного математического ожидания a .

4. Из генеральной совокупности, распределенной нормально, извлечена выборка

x_i	-1,5	-1,0	-0,5	0	0,5	1,0	1,5
n_i	2	4	8	12	7	5	2

Найти доверительный интервал, покрывающий среднеквадратичное отклонение σ с надежностью 0,99.

5. Случайная величина ξ (число поврежденных при перевозке изделий в одном контейнере) распределена по закону Пуассона. В выборке приведено распределение числа поврежденных изделий в 500 контейнерах (в первой строке указано количество x_i поврежденных изделий в одном контейнере, во второй строке приведена частота n_i - число контейнеров, содержащих x_i поврежденных изделий):

x_i	0	1	2	3	4	5	6	7
n_i	199	169	87	31	9	3	1	1

Найти с надежностью 0,99 доверительный интервал для параметра λ распределения Пуассона.

6. Игровой автомат должен обеспечивать появление выигрыша в одном случае из 100 бросаний монеты в автомат. При проверке игрового автомата произведено 500 испытаний, причем выигрыш появился 4 раза. Найти доверительный интервал, покрывающий неизвестную вероятность появления выигрыша с надежностью 0,999. Обеспечивает ли этот игровой автомат заявленную вероятность выигрыша?

§ 11. Статистическая проверка гипотез о значениях параметров распределений.

В данном разделе пойдет речь о проверке каких-либо предположений (гипотез) относительно неизвестных параметров известного закона распределения случайной величины. Решение о том, справедлива ли рассматриваемая гипотеза или нет, должно основываться лишь на знании выборки $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$, то есть на основании некоторой статистики. Таким образом, чтобы определить процедуру принятия решения по выборке ζ , мы должны задать в той или иной форме отображение выборочного пространства на множество рассматриваемых гипотез. Такое отображение называют статистическим критерием.

Пусть задан закон распределения случайной величины ξ , но не известен его параметр θ . Относительно значений этого параметра мы и будем выдвигать гипотезы.

Статистическая гипотеза H называется простой, если она состоит в том, что $\theta = \theta_0$, где θ_0 некоторое фиксированное значение, и сложной, если $\theta \in (\theta_1; \theta_2)$.

Например, для биномиального распределения $P_n \{ \xi = k \} = C_n^k p^k (1-p)^{n-k}$

гипотеза $H : p = \frac{1}{2}$ является простой, а гипотеза $H : p > \frac{1}{2}$ является сложной.

Задачу проверки статистических гипотез будем ставить следующим образом. Дана выборка $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$. Относительно значения параметра θ имеется:

основная гипотеза $H_0 : \theta = \theta_0$

и конкурирующая или альтернативная гипотеза $H_1 : \theta = \theta_1$

(для наглядности изложения гипотезы возьмем простыми).

Мы должны построить такой статистический критерий, который позволял бы нам заключить согласуется ли выборка ζ с гипотезой H_0 или нет, то есть принимаем ли мы эту гипотезу или нет. Будем строить статистические критерии при помощи так называемого критического множества. Выборку $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ можно рассматривать как точку n -мерного пространства выборок. В этом пространстве выборок выделим такое множество S , называемое критическим, что при попадании выборки ζ в S гипотеза H_0 отвергается, а в остальных случаях она принимается. Получаемый при помощи критического множества S статистический критерий принято называть S -критерием.

Пусть $p(x, \theta)$ плотность функции распределения, тогда вероятность попадания выборки ζ в критическое множество S равна $W(S, \theta) = \int_S p(x, \theta) dx$. Эту вероятность называют еще функцией мощности S -критерия.

С каждым S -критерием связаны ошибки двух родов:

ошибка I-го рода $\alpha = W(S, \theta_0) = \int_S p(x, \theta_0) dx$ - отвергаем гипотезу H_0 , когда она верна, то

есть, принимаем альтернативную гипотезу, хотя верна основная;

ошибка II-го рода $\beta = W(\bar{S}, \theta_1) = \int_{\bar{S}} p(x, \theta_1) dx = 1 - \int_S p(x, \theta_1) dx$ - отвергаем гипотезу H_1 ,

когда она верна, то есть, принимаем основную гипотезу, хотя верна альтернативная.

Вероятность α ошибки I-го рода называют уровнем значимости критерия, а величину

$W(S, \theta_1) = 1 - \beta$ называют мощностью критерия.

Очевидно, что S -критерий будет тем лучше, чем меньше будут ошибки α и β , то есть чем меньше уровень значимости α и чем больше мощность $W(S, \theta_1)$ этого критерия.

Однако при заданном числе испытаний n невозможно ни при каком выборе критической области S одновременно сделать как угодно малыми обе ошибки α и β . Поэтому, сначала задается уровень значимости α и рассматривается множество $\{S_\alpha\}$ всех S -критериев с уровнем значимости α . Далее среди них выбирается критерий S^* , для которого мощность максимальна

$$W(S^*, \theta_1) = \max_{S \in \{S_\alpha\}} W(S, \theta_1).$$

Критерий S^* , удовлетворяющий этим условиям, называют оптимальным или наиболее мощным критерием.

К сожалению, оптимальный критерий, удовлетворяющий этим условиям, существует не для всех типов распределений. Поэтому видоизменим схему построения наиболее мощного статистического критерия. Введем функцию $\varphi(x)$ такую, что $0 \leq \varphi(x) \leq 1$ и рассмотрим φ -критерий с функцией мощности $W(\varphi, \theta) = \int_S \varphi(x) p(x, \theta) dx$.

Для этого φ -критерия ошибки I-го и II-го рода равны соответственно

$$\alpha = W(\varphi, \theta_0) = \int_S \varphi(x) p(x, \theta_0) dx \text{ и } \beta = 1 - W(\varphi, \theta_1) = 1 - \int_S \varphi(x) p(x, \theta_1) dx,$$

а мощность критерия равна $W(\varphi, \theta_1) = 1 - \beta$.

Зададим уровень значимости α и рассмотрим множество $\{\varphi_\alpha\}$ всех φ -критериев с уровнем значимости α . Выберем среди них критерий φ^* , для которого мощность максимальна

$$W(\varphi^*, \theta_0) = \alpha \text{ и } W(\varphi^*, \theta_1) = \max_{\varphi \in \{\varphi_\alpha\}} W(\varphi, \theta_1) \quad (1)$$

Такой φ -критерий будем называть наиболее мощным или оптимальным. Покажем, что в такой постановке задача уже будет всегда разрешимой.

Для сокращения записей введем обозначения

$$p_0(x) = p(x, \theta_0); p_1(x) = p(x, \theta_1); M_0 \varphi = \int \varphi(x) p_0(x) dx; M_1 \varphi = \int \varphi(x) p_1(x) dx$$

(здесь M — символ математического ожидания).

Оптимальный критерий будем искать среди критериев, которые определяются

отношением правдоподобия $\frac{p_1(x)}{p_0(x)}$.

Теорема Неймана-Пирсона. Для любого $\alpha \in [0; 1]$ существуют такие числа $c \geq 0$ и

$\varepsilon \in [0; 1]$, что φ^* -критерий с функцией

$$\varphi^*(x) = \begin{cases} 1, & \text{если } \frac{p_1(x)}{p_0(x)} > c, \text{ то есть } p_1(x) > c \cdot p_0(x) \\ \varepsilon, & \text{если } \frac{p_1(x)}{p_0(x)} = c, \text{ то есть } p_1(x) = c \cdot p_0(x) \\ 0, & \text{если } \frac{p_1(x)}{p_0(x)} < c, \text{ то есть } p_1(x) < c \cdot p_0(x) \end{cases}$$

является наиболее мощным критерием согласно (1) с уровнем значимости α .

Доказательство. Рассмотрим следующую функцию от c в предположении, что верна

$$\text{гипотеза } H_0: g(c) = P \left\{ \frac{p_1(x)}{p_0(x)} > c \mid H_0 \right\}.$$

Функция $1 - g(c) = P \left\{ \frac{p_1(x)}{p_0(x)} \leq c \mid H_0 \right\}$ — есть функция распределения случайной величины

$\frac{p_1(x)}{p_0(x)}$, а значит, она не убывает и непрерывна слева. Тогда функция

$g(c) = 1 - P \left\{ \frac{p_1(x)}{p_0(x)} \leq c \mid H_0 \right\}$ не возрастает и также непрерывна слева. Кроме того

$$g(+0) = 1 \text{ и } g(+\infty) = 0.$$

Найдем c_α из условия $g(c_\alpha) \leq \alpha < g(c_\alpha - 0)$.

А именно, если $g(c_\alpha) < g(c_\alpha - 0)$ (функция $g(c)$ терпит скачок в точке c_α), то возьмем

$$\varepsilon_\alpha = \frac{\alpha - g(c_\alpha)}{g(c_\alpha - 0) - g(c_\alpha)}.$$

Если же $g(c_\alpha) = g(c_\alpha - 0)$ (функция $g(c)$ непрерывна в точке c_α),

то возьмем $\varepsilon_\alpha = 0$. При этом если $g(c_\alpha) = \alpha$ на некотором промежутке (функция $g(c)$

постоянна на этом промежутке), то примем за c_α любую точку этого промежутка,

например левый конец.

Положив в выражении для $\varphi^*(x)$ $c = c_\alpha$ и $\varepsilon = \varepsilon_\alpha$, получим требуемую функцию.

Действительно, покажем, что построенный φ^* -критерий имеет уровень значимости α и обладает свойством оптимальности, то есть его мощность максимальна.

Найдем сначала

$$\begin{aligned}
M_0 \varphi^* &= \int \varphi^*(x) p_0(x) dx = \int_{p_1(x) > c_\alpha p_0(x)} 1 \cdot p_0(x) dx + \int_{p_1(x) = c_\alpha p_0(x)} \varepsilon_\alpha \cdot p_0(x) dx + \int_{p_1(x) < c_\alpha p_0(x)} 0 \cdot p_0(x) dx = \\
&= \int_{p_1(x) > c_\alpha p_0(x)} p_0(x) dx + \frac{\alpha - g(c_\alpha)}{g(c_\alpha - 0) - g(c_\alpha)} \int_{p_1(x) = c_\alpha p_0(x)} p_0(x) dx = g(c_\alpha) + \\
&+ \frac{\alpha - g(c_\alpha)}{g(c_\alpha - 0) - g(c_\alpha)} (g(c_\alpha - 0) - g(c_\alpha)) = g(c_\alpha) + \alpha - g(c_\alpha) = \alpha.
\end{aligned}$$

Таким образом, уровень значимости этого критерия равен α .

Пусть φ любой другой критерий с уровнем значимости α : $M_0 \varphi = \alpha$. Покажем, что

$M_1 \varphi^* \geq M_1 \varphi$. Рассмотрим интеграл

$$\begin{aligned}
\int (\varphi^*(x) - \varphi(x)) (p_1(x) - c_\alpha p_0(x)) dx &= \int_{\substack{\varphi^*(x) > \varphi(x) \\ (p_1(x) \geq c_\alpha p_0(x))}} (\varphi^*(x) - \varphi(x)) (p_1(x) - c_\alpha p_0(x)) dx + \\
&+ \int_{\substack{\varphi^*(x) < \varphi(x) \\ (p_1(x) < c_\alpha p_0(x))}} (\varphi^*(x) - \varphi(x)) (p_1(x) - c_\alpha p_0(x)) dx \geq 0, \text{ откуда}
\end{aligned}$$

$$\int (\varphi^*(x) - \varphi(x)) p_1(x) dx \geq c_\alpha \int (\varphi^*(x) - \varphi(x)) p_0(x) dx, \text{ то есть}$$

$$M_1 \varphi^* - M_1 \varphi \geq c_\alpha (M_0 \varphi^* - M_0 \varphi) = c_\alpha (\alpha - \alpha) \geq 0 \text{ и } M_1 \varphi^* \geq M_1 \varphi.$$

Таким образом, мы построили наиболее мощный критерий.

Мы провели все рассуждения, используя плотность $p(x, \theta)$ случайной величины, то есть считая, что она абсолютно-непрерывна. Но все остается в силе и для дискретных случайных величин с заменой плотности на вероятности $p(\xi_i, \theta)$, а интегралы на суммы.

Рассмотрим теперь применение построенного критерия Неймана-Пирсона для проверки гипотез о значении параметров распределений.

§ 11.1. Проверка гипотез о значении параметров нормального распределения.

Пусть случайная величина ξ распределена по нормальному закону и $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ выборка объема n . Поскольку нормальное распределение имеет два параметра a и σ , то рассмотрим следующие случаи.

I. Проверка гипотез о значении математического ожидания a , если дисперсия σ^2 известна.

Пусть требуется проверить основную гипотезу

$$H_0: a = a_0$$

при альтернативной гипотезе (правая односторонняя)

$$H_1: a = a_1 > a_0$$

Совместная плотность распределения вектора ζ есть

$$p_j(\zeta) = p_j(\xi_1) p_j(\xi_2) \dots p_j(\xi_n) = \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i - a_j)^2}, \quad j = 0, 1$$

Тогда отношение правдоподобия примет вид

$$\begin{aligned}
\frac{p_1(\zeta)}{p_0(\zeta)} &= \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i - a_1)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i - a_0)^2 \right] = \\
&= \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i^2 - 2a_1\xi_i + a_1^2 - \xi_i^2 + 2a_0\xi_i - a_0^2) \right] = \exp \left[-\frac{1}{2\sigma^2} \left(n(a_1^2 - a_0^2) - 2(a_1 - a_0) \sum_{i=1}^n \xi_i \right) \right] = \\
&= \exp \left[-\frac{1}{2\sigma^2} \left(n(a_1^2 - a_0^2) - 2n(a_1 - a_0) \bar{x}_s \right) \right] = \exp \left[-\frac{n(a_1 - a_0)}{2\sigma^2} (a_1 + a_0 - 2\bar{x}_s) \right].
\end{aligned}$$

Критическая область S определяется при некотором c из равенства

$$P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = \alpha.$$

рассмотрим неравенство

$$\frac{p_1(\zeta)}{p_0(\zeta)} = \exp \left[-\frac{n(a_1 - a_0)}{2\sigma^2} (a_1 + a_0 - 2\bar{x}_s) \right] > c \quad (2)$$

и разрешим его относительно $\bar{x}_s$:

$$-\frac{n(a_1 - a_0)}{2\sigma^2} (a_1 + a_0 - 2\bar{x}_s) > \ln c \text{ и с учетом того, что } a_1 > a_0 \text{ получим}$$

$$\bar{x}_s > \frac{a_1 + a_0}{2} + \frac{\sigma^2}{n(a_1 - a_0)} \ln c.$$

Обозначим через $c_1 = \frac{a_1 + a_0}{2} + \frac{\sigma^2}{n(a_1 - a_0)} \ln c$ - это тоже еще неизвестная величина, так как

c неизвестно. Таким образом, $\bar{x}_s > c_1$ и

$$P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = P \{ \bar{x}_s > c_1 \} = \alpha.$$

Задача нахождения критической области S в n -мерном пространстве выборок сведена к задаче нахождения критической области для одномерного случая (относительно выборочного среднего).

Выборочное среднее $\bar{x}_s$ имеет математическое ожидание $M\bar{x}_s = a$ и дисперсию $D\bar{x}_s = \frac{\sigma^2}{n}$. Введем нормированную случайную величину $\eta = \frac{\bar{x}_s - M\bar{x}}{\sqrt{D\bar{x}_s}} = \frac{\bar{x}_s - a}{\sigma} \sqrt{n}$,

которая по центральной предельной теореме распределена асимптотически нормально с параметрами $a = 0$ и $\sigma = 1$. Тогда

$$P \{ \bar{x}_s > c_1 \mid H_0 \} = P \left\{ \frac{\bar{x}_s - a_0}{\sigma} \sqrt{n} > \frac{c_1 - a_0}{\sigma} \sqrt{n} \right\} = P \{ \eta > c_\alpha \},$$

$$\text{где } c_\alpha = \frac{c_1 - a_0}{\sigma} \sqrt{n} = \frac{\sqrt{n}}{\sigma} \left(\frac{a_1 - a_0}{2} + \frac{\sigma^2}{n(a_1 - a_0)} \ln c \right) \text{ и } \eta = \frac{\bar{x}_s - a_0}{\sigma} \sqrt{n}.$$

Теперь из уравнения $P \{ \eta > c_\alpha \} = \alpha$ найдем c_α . Для этого рассмотрим

$$P\{\eta > c_\alpha\} = \frac{1}{\sqrt{2\pi}} \int_{c_\alpha}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_0^{c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{2} - \Phi(c_\alpha).$$

Тогда для нахождения c_α получаем уравнение $\frac{1}{2} - \Phi(c_\alpha) = \alpha$, то есть $\Phi(c_\alpha) = \frac{1}{2} - \alpha$.

По таблице 2 значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ находим критическую точку c_α .

Теперь, если $\eta = \frac{\bar{x}_s - a_0}{\sigma} \sqrt{n} < c_\alpha$, то у нас нет оснований отвергать гипотезу H_0 , если же $\eta > c_\alpha$, то гипотезу H_0 отвергаем.

На рисунке 4 изображена критическая область S , вероятность попадания в которую равна α .

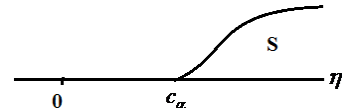


Рис. 4

Предположим, что нам известно гипотетическое значение a_1 . Тогда мы можем найти мощность $W = 1 - \beta$ построенного правостороннего критерия. Для этого вычислим

$$\beta = P\left\{\frac{p_1(\zeta)}{p_0(\zeta)} < c \mid H_1\right\} = P\{\bar{x}_s < c_1 \mid H_1\} = P\left\{\frac{\bar{x}_s - a_1}{\sigma} \sqrt{n} < \frac{c_1 - a_1}{\sigma} \sqrt{n}\right\} = P\{\eta < c_\beta\},$$

где $c_\beta = \frac{c_1 - a_1}{\sigma} \sqrt{n} = \frac{c_1 - a_0}{\sigma} \sqrt{n} - \frac{a_1 - a_0}{\sigma} \sqrt{n} = c_\alpha - \lambda$ и $\lambda = \frac{a_1 - a_0}{\sigma} \sqrt{n}$.

Тогда

$$\beta = P\{\eta < c_\beta\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c_\beta} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_0^{c_\beta} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(c_\beta) = \frac{1}{2} + \Phi(c_\alpha - \lambda).$$

Таким образом, зная c_α и λ , можно найти β , а значит и мощность

$$W = 1 - \beta = \frac{1}{2} - \Phi(c_\beta) = \frac{1}{2} - \Phi(c_\alpha - \lambda)$$

построенного правостороннего критерия.

Наконец, если заданы ошибки α и β (или мощность $W = 1 - \beta$), то можно найти объем выборки, обеспечивающий эти ошибки.

Действительно, из уравнения $\Phi(c_\alpha) = \frac{1}{2} - \alpha$ находим c_α , а из уравнения $\Phi(c_\beta) = \beta - \frac{1}{2}$ находим c_β . Далее, из соотношения

$$c_\beta = c_\alpha - \lambda, \text{ то есть } \frac{a_1 - a_0}{\sigma} \sqrt{n} = c_\alpha - c_\beta \text{ находим } n = \sigma^2 \left(\frac{c_\alpha - c_\beta}{a_1 - a_0} \right)^2.$$

Пример 1. По выборке объема $n = 16$, извлеченной из генеральной совокупности, распределенной нормально со среднеквадратичным отклонением $\sigma = 4$, найдено $\bar{x}_s = 2,3$. Требуется при уровне значимости $\alpha = 0,05$:

- 1) Проверить основную гипотезу $H_0: a = a_0 = 2$ при альтернативной гипотезе $H_1: a = a_1 > 2$;

- 2) Найти мощность построенного правостороннего критерия при значении $a_1 = 3$;
- 3) Найти объем выборки n , при котором мощность критерия будет равна $W = 0,6$.

Из уравнения $\Phi(c_\alpha) = \frac{1}{2} - \alpha = 0,45$ по таблице 2 значений функции $\Phi(x)$ находим

$$c_\alpha = 1,65. \text{ Вычислим } \eta = \frac{\bar{x}_6 - a_0}{\sigma} \sqrt{n} = \frac{2,3 - 2}{4} \cdot 4 = 0,3.$$

Так как $\eta = 0,3 < c_\alpha = 1,65$, то у нас нет оснований отвергать гипотезу H_0 .

Найдем далее $c_\beta = c_\alpha - \lambda = c_\alpha - \frac{a_1 - a_0}{\sigma} \sqrt{n} = 1,65 - \frac{3 - 2}{4} \cdot 4 = 0,65$ и мощность построенного

правостороннего критерия $W = 1 - \beta = \frac{1}{2} - \Phi(c_\beta) = 0,5 - 0,24 = 0,26$.

Найдем теперь объем выборки n , при котором мощность критерия будет равна $W = 0,6$.

В этом случае $\beta = 1 - W = 1 - 0,6 = 0,4$ и из уравнения $\Phi(c_\beta) = \beta - \frac{1}{2} = 0,4 - 0,5 = -0,1$ по

таблице 2 с учетом того, что функция $\Phi(x)$ нечетна, находим $c_\beta = -0,25$. Тогда

$$n = \sigma^2 \left(\frac{c_\alpha - c_\beta}{a_1 - a_0} \right)^2 = 16 \left(\frac{1,65 + 0,25}{3 - 2} \right)^2 \approx 58.$$

Рассмотрим теперь случай, когда альтернативная гипотеза левая односторонняя

$$H_0 : a = a_0$$

$$H_1 : a = a_1 < a_0$$

Критическая область S и в этом случае определяется при некотором c из равенства

$$P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = \alpha. \text{ Но в этом случае } a_1 < a_0 \text{ и, разрешая неравенство (2) относительно}$$

$\bar{x}_6$, получим $\bar{x}_6 < -c'_1$, где $c'_1 = -\frac{a_1 + a_0}{2} - \frac{\sigma^2}{n(a_1 - a_0)} \ln c$. Таким образом,

$$\alpha = P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = P \left\{ \bar{x}_6 < -c'_1 \mid H_0 \right\} = P \left\{ \frac{\bar{x}_6 - a_0}{\sigma} \sqrt{n} < -\frac{c'_1 + a_0}{\sigma} \sqrt{n} \right\} = P \{ \eta < -c_\alpha \},$$

где $\eta = \frac{\bar{x}_6 - a_0}{\sigma} \sqrt{n}$ и $c_\alpha = \frac{c'_1 + a_0}{\sigma} \sqrt{n}$.

Теперь из уравнения $P \{ \eta < -c_\alpha \} = \alpha$ следует найти c_α . Для этого рассмотрим

$$P \{ \eta < -c_\alpha \} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_0^{-c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(-c_\alpha) = \frac{1}{2} - \Phi(c_\alpha).$$

Таким образом, для нахождения c_α получаем тоже уравнение, что и в предыдущем случае

$$\Phi(c_\alpha) = \frac{1}{2} - \alpha.$$

По таблице 2 находим значение c_α . Если $\eta = \frac{\bar{x}_e - a_0}{\sigma} \sqrt{n} < -c_\alpha$, то гипотезу H_0 отвергаем, иначе принимаем.

На рисунке 5 изображена критическая область S , вероятность попадания в которую равна α .

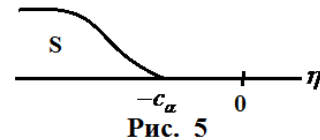


Рис. 5

Если нам известно гипотетическое значение a_1 , то мы можем найти мощность $W = 1 - \beta$ построенного левостороннего критерия. Для этого вычислим

$$\beta = P\left\{\frac{p_1(\zeta)}{p_0(\zeta)} < c \mid H_1\right\} = P\left\{\bar{x}_e > -c'_1 \mid H_1\right\} = P\left\{\frac{\bar{x}_e - a_1}{\sigma} \sqrt{n} > -\frac{c'_1 + a_1}{\sigma} \sqrt{n}\right\} = P\{\eta > -c_\beta\},$$

где $c_\beta = \frac{c'_1 + a_1}{\sigma} \sqrt{n} = \frac{c'_1 + a_0}{\sigma} \sqrt{n} + \frac{a_1 - a_0}{\sigma} \sqrt{n} = c_\alpha + \lambda$, где $\lambda = \frac{a_1 - a_0}{\sigma} \sqrt{n}$ тоже, что и в предыдущем случае. Тогда

$$\beta = P\{\eta > -c_\beta\} = \frac{1}{\sqrt{2\pi}} \int_{-c_\beta}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{-c_\beta}^0 e^{-\frac{t^2}{2}} dt = \frac{1}{2} - \Phi(-c_\beta) = \frac{1}{2} + \Phi(c_\beta).$$

Таким образом, $\beta = \frac{1}{2} + \Phi(c_\beta) = \frac{1}{2} + \Phi(c_\alpha + \lambda)$ и мощность критерия

$$W = 1 - \beta = \frac{1}{2} - \Phi(c_\alpha + \lambda).$$

Наконец, если заданы уровень значимости α и мощность W , то найдя c_α и c_β соответственно из уравнений $\Phi(c_\alpha) = \frac{1}{2} - \alpha$ и $\Phi(c_\beta) = \beta - \frac{1}{2} = \frac{1}{2} - W$, из соотношения

$$c_\beta = c_\alpha + \lambda = c_\alpha + \frac{a_1 - a_0}{\sigma} \sqrt{n} \text{ найдем объем выборки } n = \sigma^2 \left(\frac{c_\beta - c_\alpha}{a_1 - a_0} \right)^2, \text{ обеспечивающий эти}$$

заданные величины.

Рассмотрим еще случай, когда альтернативная гипотеза двусторонняя

$$H_0: a = a_0 \quad (3)$$

$$H_1: a = a_1 \neq a_0$$

Так как условие $a_1 \neq a_0$ эквивалентно тому, что $a_1 > a_0$ или $a_1 < a_0$, то мы можем воспользоваться результатами двух предыдущих случаев

$$\alpha = P\{|\eta| > c_\alpha\} = P\{\eta > c_\alpha\} + P\{\eta < -c_\alpha\} = \frac{1}{2} - \Phi(c_\alpha) + \frac{1}{2} - \Phi(-c_\alpha) = 1 - 2\Phi(c_\alpha)$$

и для нахождения c_α получаем уравнение $\Phi(c_\alpha) = \frac{1 - \alpha}{2}$.

Теперь, если $|\eta| = \left| \frac{\bar{x}_e - a_0}{\sigma} \sqrt{n} \right| > c_\alpha$, то гипотезу H_0 отвергаем, иначе принимаем.

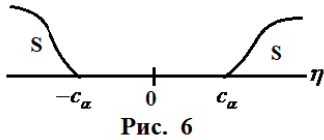


Рис. 6

На рисунке 6 изображена критическая область S , состоящая из двух интервалов, вероятность попадания в каждый из них равна $\frac{\alpha}{2}$. Поскольку η есть нормированная нормально

распределенная случайная величина, а распределение такой величины симметрично относительно нуля, то и критические точки симметричны относительно нуля. В этом случае достигается наибольшая мощность критерия.

Для нахождения мощности построенного двустороннего критерия найдем

$$\begin{aligned}\beta &= P\{\bar{x}_e < c_1 | H_1\} + P\{\bar{x}_e > -c_1' | H_1\} = P\left\{\frac{\bar{x}_e - a_1}{\sigma} \sqrt{n} < \frac{c_1 - a_1}{\sigma} \sqrt{n}\right\} + \\ &+ P\left\{\frac{\bar{x}_e - a_1}{\sigma} \sqrt{n} > -\frac{c_1' + a_1}{\sigma} \sqrt{n}\right\} = P\{\eta < c_\alpha - \lambda\} + P\{\eta > -c_\alpha - \lambda\} = \\ &= P\{-c_\alpha - \lambda < \eta < c_\alpha - \lambda\} = \Phi(c_\alpha - \lambda) - \Phi(-c_\alpha - \lambda) = \Phi(c_\alpha - \lambda) + \Phi(c_\alpha + \lambda).\end{aligned}$$

Тогда $W = 1 - [\Phi(c_\alpha - \lambda) + \Phi(c_\alpha + \lambda)]$.

Замечание. Находя двустороннюю критическую область при уровне значимости α , мы одновременно находим и доверительный интервал для математического ожидания a с надежностью $\gamma = 1 - \alpha$. Действительно, при проверке двусторонней гипотезы (3), мы

требовали, чтобы вероятность попадания случайной величины $\eta = \frac{\bar{x}_e - a_0}{\sigma} \sqrt{n}$ в

двустороннюю критическую область S была равна α . Следовательно, вероятность попадания величины η в область принятия гипотезы $(-c_\alpha; c_\alpha)$ равна $\gamma = 1 - \alpha$. Другими

словами с надежностью γ выполняется неравенство $-c_\alpha < \frac{\bar{x}_e - a_0}{\sigma} \sqrt{n} < c_\alpha$, откуда

$$\bar{x}_e - c_\alpha \frac{\sigma}{\sqrt{n}} < a < \bar{x}_e + c_\alpha \frac{\sigma}{\sqrt{n}}, \text{ где } \Phi(c_\alpha) = \frac{1 - \alpha}{2} = \frac{\gamma}{2}, \text{ а это и есть доверительный интервал}$$

для математического ожидания a .

Хотя отыскание двусторонней критической области и доверительного интервала приводит к одинаковым результатам, их истолкование различно: двусторонняя критическая область определяет границы (критические точки), между которыми заключено $(1 - \alpha)\%$ числа наблюдаемых значений величины η , найденных при повторении опытов; доверительный же интервал определяет границы (концы интервала), между которыми в $\gamma = (1 - \alpha)\%$ опытов заключено истинное значение параметра a .

II. Проверка гипотез о значении математического ожидания a , если дисперсия σ^2 не известна.

Пусть необходимо проверить гипотезы

$$H_0: a = a_0$$

$$H_1: a = a_1 > a_0 \text{ (правая односторонняя).}$$

Как и в предыдущем пункте, критическая область S определяется при некотором c из соотношения

$$\alpha = P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = P \left\{ \bar{x}_\epsilon > c_1 \mid H_0 \right\}.$$

Дальнейшие рассуждения аналогичны тем, которые мы делали в § 10 при нахождении доверительного интервала для математического ожидания для случая, когда дисперсия не известна.

Выборочное среднее $\bar{x}_\epsilon$ имеет математическое ожидание $M\bar{x}_\epsilon = a$ и дисперсию

$D\bar{x}_\epsilon = \frac{\sigma^2}{n}$. Поскольку дисперсия σ^2 не известна, то по выборке $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ найдем

исправленную выборочную дисперсию $s^2 = \frac{n}{n-1} D_\epsilon$ и примем $D\bar{x}_\epsilon \approx \frac{s^2}{n}$. Введем

нормированную случайную величину

$$t = \frac{\bar{x}_\epsilon - M\bar{x}_\epsilon}{\sqrt{D\bar{x}_\epsilon}} = \frac{\bar{x}_\epsilon - a}{s} \sqrt{n} = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n \xi_i - a}{\sqrt{\frac{n}{n-1} D_\epsilon}} = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n (\xi_i - a)}{\sqrt{\frac{n}{n-1} D_\epsilon}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\xi_i - a}{\sigma}}{\sqrt{\frac{n}{n-1} \cdot \frac{D_\epsilon}{\sigma^2}}}.$$

Перейдем к случайным величинам $\xi'_i = \frac{\xi_i - a}{\sigma}$, которые распределены нормально с

параметрами $a' = 0$ и $\sigma' = 1$. Для нее $D'_\epsilon = \frac{D_\epsilon}{\sigma^2}$ и тогда

$$t = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{n}{n-1} D'_\epsilon}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n \xi_i'^2 - \left(\frac{1}{n} \sum_{i=1}^n \xi'_i \right)^2 \right]}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i}{\sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n \xi_i'^2 - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i \right)^2 \right]}}.$$

Введем в n -мерном пространстве выборок $(\xi'_1, \xi'_2, \dots, \xi'_n)$ новую ортогональную систему

координат $(\eta_1, \eta_2, \dots, \eta_n)$ так, что $\eta_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi'_i$, $\eta_j = \sum_{i=1}^n \alpha_{ji} \xi'_i$ и $\sum_{j=1}^n \alpha_{ji} \alpha_{jm} = \delta_{im} = \begin{cases} 1, i = m \\ 0, i \neq m \end{cases}$.

Мы ранее показали, что $\sum_{j=1}^n \eta_j^2 = \sum_{i=1}^n \xi_i'^2$. Тогда

$$t = \frac{\eta_1}{\sqrt{\frac{1}{n-1} \left[\sum_{j=1}^n \eta_j^2 - \eta_1^2 \right]}} = \frac{\eta_1}{\sqrt{\frac{1}{n-1} \sum_{j=2}^n \eta_j^2}}.$$

Поскольку случайные величины η_j , так же как и ξ'_i , распределены нормально с

параметрами 0 и 1, то случайная величина t имеет распределение Стьюдента с $n-1$ степенью свободы. Таким образом

$$\alpha = P\{\bar{x}_s > c_1 | H_0\} = P\left\{\frac{\bar{x}_s - a_0}{s} \sqrt{n} > \frac{c_1 - a_0}{s} \sqrt{n}\right\} = P\{t > c_\alpha\}, \text{ где } c_\alpha = \frac{c_1 - a_0}{s} \sqrt{n}.$$

По таблице 5 критических точек распределения Стьюдента для случая односторонней критической области находим $c_\alpha = c_\alpha(\alpha; n-1)$. Теперь, если

$$t = \frac{\bar{x}_s - a_0}{s} \sqrt{n} > c_\alpha, \text{ то гипотезу } H_0 \text{ отвергаем, иначе принимаем.}$$

Совершенно аналогично рассматривается случай левой односторонней гипотезы

$$H_0 : a = a_0$$

$$H_1 : a = a_1 < a_0$$

$$\text{Имеем } \alpha = P\left\{\frac{p_1(\zeta)}{p_0(\zeta)} > c | H_0\right\} = P\{\bar{x}_s < -c_1' | H_0\} = P\{t < -c_\alpha\}.$$

По таблице 5 критических точек распределения Стьюдента для случая односторонней критической области находим $c_\alpha = c_\alpha(\alpha; n-1)$. Теперь, если

$$t = \frac{\bar{x}_s - a_0}{s} \sqrt{n} < -c_\alpha, \text{ то гипотезу } H_0 \text{ отвергаем, иначе принимаем.}$$

Для случая двусторонней гипотезы

$$H_0 : a = a_0$$

$$H_1 : a = a_1 \neq a_0,$$

объединяя два предыдущих случая, получим уравнение $\alpha = P\{|t| > c_\alpha\}$.

По таблице 5 критических точек распределения Стьюдента для случая двусторонней

критической области находим $c_\alpha = c_\alpha(\alpha; n-1)$. Если $|t| = \left|\frac{\bar{x}_s - a_0}{s} \sqrt{n}\right| > c_\alpha$, то гипотезу H_0

отвергаем, иначе принимаем.

Как мы уже отмечали ранее, с ростом n распределение Стьюдента стремится к нормальному распределению и при достаточно больших n значение c_α следует находить из уравнений $\Phi(c_\alpha) = \frac{1}{2} - \alpha$ для односторонних гипотез и $\Phi(c_\alpha) = \frac{1-\alpha}{2}$ для двусторонней гипотезы, которые мы получили в пункте I при рассмотрении случая с известной дисперсией.

Пример 2. Из генеральной совокупности, распределенной по нормальному закону,

извлечена выборка объема $n = 25$ и по ней найдены выборочная средняя $\bar{x}_s = 102,5$ и исправленное среднеквадратичное отклонение $s = 7,5$. Требуется при уровне значимости $\alpha = 0,02$ проверить основную гипотезу $H_0 : a = a_0 = 100$ при конкурирующей гипотезе $H_1 : a = a_1 \neq a_0$.

По уровню значимости $\alpha = 0,02$ и числу степеней свободы $n-1 = 24$ по таблице 5 критических точек распределения Стьюдента для двусторонней критической области находим $c_\alpha = 2,49$. Далее вычислим $t = \frac{\bar{x}_s - a_0}{s} \sqrt{n} = \frac{102,5 - 100}{7,5} \cdot 5 = 1,67$.

Так как $|t| = 1,67 < c_\alpha = 2,49$, то гипотезу H_0 принимаем, то есть отличие $\bar{x}_s = 102,5$ от гипотетического значения $a = 100$ не существенно для данного уровня значимости.

III. Сравнение математических ожиданий двух случайных величин, дисперсии которых известны.

Пусть даны две случайные величины ξ и η , распределенные нормально с параметрами a_1, σ_1 и a_2, σ_2 соответственно. Будем считать, что их дисперсии σ_1^2 и σ_2^2 известны, а относительно математических ожиданий имеются гипотезы

$$H_0 : a_1 = a_2$$

$$H_1 : a_1 \neq a_2$$

По выборкам $\zeta_1 = (\xi_1, \xi_2, \dots, \xi_n)$ и $\zeta_2 = (\eta_1, \eta_2, \dots, \eta_m)$ объемов n и m соответственно, находим выборочные средние

$$\bar{x}_s = \frac{1}{n} \sum_{i=1}^n \xi_i \text{ и } \bar{y}_s = \frac{1}{m} \sum_{i=1}^m \eta_i.$$

Как и в предыдущих случаях рассмотрим нормированную случайную величину

$$z = \frac{(\bar{x}_s - \bar{y}_s) - M(\bar{x}_s - \bar{y}_s)}{\sqrt{D(\bar{x}_s - \bar{y}_s)}}.$$

При гипотезе H_0 имеем $M(\bar{x}_s - \bar{y}_s) = M\bar{x}_s - M\bar{y}_s = a_1 - a_2 = 0$,

$$D(\bar{x}_s - \bar{y}_s) = D\bar{x}_s + D\bar{y}_s = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}, \text{ то есть } z = \frac{\bar{x}_s - \bar{y}_s}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}.$$

По центральной предельной теореме случайная величина z распределена асимптотически нормально с параметрами $a = 0$ и $\sigma = 1$. Так как альтернативная гипотеза двусторонняя,

$$\text{то } \alpha = P\{|z| > c_\alpha \mid H_0\} = P\{z > c_\alpha\} + P\{z < -c_\alpha\} = \frac{1}{2} - \Phi(c_\alpha) + \frac{1}{2} - \Phi(-c_\alpha) = 1 - 2\Phi(c_\alpha)$$

и для нахождения критической точки c_α приходим к уже известному нам уравнению

$$\Phi(c_\alpha) = \frac{1 - \alpha}{2}.$$

Зная уровень значимости α , находим по таблице 2 значений функции $\Phi(x)$ значение c_α .

Теперь, если $|z| < c_\alpha$, то гипотезу H_0 принимаем, если же $|z| > c_\alpha$, то гипотезу H_0 отвергаем.

В случае, когда конкурирующая гипотеза односторонняя (правая или левая)

$$H_0 : a_1 = a_2 \text{ или } H_0 : a_1 = a_2$$

$$H_1 : a_1 > a_2 \quad H_1 : a_1 < a_2,$$

то значение c_α находится из уравнения $\Phi(c_\alpha) = \frac{1}{2} - \alpha$ и гипотезу H_0 принимаем при $z < c_\alpha$ для правой односторонней гипотезы и принимаем при $z > -c_\alpha$ для левой односторонней гипотезы.

Пример 3. Из двух генеральных совокупностей, распределенных нормально с дисперсиями $\sigma_1^2 = 80$ и $\sigma_2^2 = 100$ соответственно, по двум независимым выборкам объема $n = 40$ и $m = 50$ найдены выборочные средние $\bar{x}_s = 130$ и $\bar{y}_s = 140$. При уровне значимости $\alpha = 0,01$ требуется проверить гипотезы

$$H_0 : a_1 = a_2$$

$$H_1 : a_1 \neq a_2.$$

$$\text{Вычислим } z = \frac{\bar{x}_s - \bar{y}_s}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{130 - 140}{\sqrt{\frac{80}{40} + \frac{100}{50}}} = -\frac{10}{\sqrt{4}} = -5.$$

Из уравнения $\Phi(c_\alpha) = \frac{1-\alpha}{2} = \frac{0,99}{2} = 0,495$ по таблице 2 значений функции $\Phi(x)$ находим $c_\alpha = 2,58$. Так как $|z| = 5 > c_\alpha = 2,58$, то гипотезу H_0 отвергаем, то есть математические ожидания разнятся значительно.

IV. Проверка гипотез о значении дисперсии σ^2 .

Пусть при уровне значимости α требуется проверить гипотезы

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 = \sigma_1^2 > \sigma_0^2 \quad (\text{правая односторонняя}).$$

Согласно теореме Неймана-Пирсона составим отношение правдоподобия

$$\frac{p_1(\zeta)}{p_0(\zeta)}, \text{ где } p_j(\zeta) = \left(\frac{1}{\sigma_j \sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma_j^2} \sum_{i=1}^n (\xi_i - a)^2}, \quad j = 0, 1.$$

$$\text{В данном случае } \frac{p_1(\zeta)}{p_0(\zeta)} = \left(\frac{\sigma_0}{\sigma_1} \right)^n e^{-\frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (\xi_i - a)^2}.$$

Критическая область S определяется при некотором c из равенства

$$P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = \alpha.$$

$$\text{Рассмотрим неравенство } \frac{p_1(\zeta)}{p_0(\zeta)} = \left(\frac{\sigma_0}{\sigma_1} \right)^n e^{-\frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (\xi_i - a)^2} > c$$

и разрешим его относительно $\sum_{i=1}^n (\xi_i - a)^2$:

$$\exp \left[-\frac{1}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum_{i=1}^n (\xi_i - a)^2 \right] > \left(\frac{\sigma_1}{\sigma_0} \right)^n \cdot c, \quad \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2 \sigma_1^2} \sum_{i=1}^n (\xi_i - a)^2 \ln \left[\left(\frac{\sigma_1}{\sigma_0} \right)^n \cdot c \right]$$

и так как $\sigma_1 > \sigma_0$, то $\sum_{i=1}^n (\xi_i - a)^2 > \frac{2\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \left[\left(\frac{\sigma_1}{\sigma_0} \right)^n \cdot c \right]$.

Обозначим через $c_1 = \frac{2\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln \left[\left(\frac{\sigma_1}{\sigma_0} \right)^n \cdot c \right]$. Таким образом

$$\alpha = P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = P \left\{ \sum_{i=1}^n (\xi_i - a)^2 > c_1 \mid H_0 \right\}.$$

Поскольку математическое ожидание a неизвестно, то заменим его несмещенной точечной оценкой $\bar{x}_e$. Тогда $\sum_{i=1}^n (\xi_i - a)^2 \approx \sum_{i=1}^n (\xi_i - \bar{x}_e)^2 = nD_e$. Перейдем к нормированным случайным величинам $\xi'_i = \frac{\xi_i - \bar{x}_e}{\sigma}$, распределенным нормально. Тогда $D'_e = \frac{D_e}{\sigma^2}$ и

$$\alpha = P \{ nD_e > c_1 \mid H_0 \} = P \left\{ \frac{nD_e}{\sigma_0^2} > \frac{c_1}{\sigma_0^2} \right\} = P \left\{ nD'_e > \frac{c_1}{\sigma_0^2} \right\}.$$

Введем случайную величину (при гипотезе H_0)

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{nD_e}{\sigma_0^2} = nD'_e = \sum_{i=1}^n \xi_i'^2 - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i' \right)^2.$$

В n -мерном пространстве выборок $(\xi'_1, \xi'_2, \dots, \xi'_n)$ введем новую ортогональную систему

координат $(\eta_1, \eta_2, \dots, \eta_n)$ так, что $\eta_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i'$, $\eta_j = \sum_{i=1}^n \alpha_{ji} \xi_i'$ и $\sum_{j=1}^n \alpha_{ji} \alpha_{jm} = \delta_{im} = \begin{cases} 1, i = m \\ 0, i \neq m \end{cases}$.

При этом случайные величины η_j также нормированы и распределены нормально.

Ранее при построении доверительных интервалов мы показали, что $\sum_{j=1}^n \eta_j^2 = \sum_{i=1}^n \xi_i'^2$.

Тогда $\chi^2 = \sum_{j=1}^n \eta_j^2 - \eta_1^2 = \sum_{j=2}^n \eta_j^2$ имеет распределение Пирсона или χ^2 с $n-1$ степенью свободы. Таким образом

$$\alpha = P \{ \chi^2 > c_\alpha \}, \text{ где } c_\alpha = \frac{c_1}{\sigma_0^2}.$$

По таблице 6 критических точек распределения χ^2 находим $c_\alpha = c_\alpha(\alpha; n-1)$. Теперь, если для найденного значения $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} > c_\alpha$, то гипотезу H_0 отвергаем, иначе принимаем.

Если альтернативная гипотеза левая односторонняя

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 = \sigma_1^2 < \sigma_0^2$$

то для нахождения критической области получаем равенство

$$\alpha = P \left\{ \frac{p_1(\zeta)}{p_0(\zeta)} > c \mid H_0 \right\} = P \left\{ \sum_{i=1}^n (\xi_i - a)^2 < c_1 \mid H_0 \right\} = P \left\{ nD'_e < \frac{c_1}{\sigma_0^2} \right\} = P \left\{ \chi^2 < c_\alpha \right\}.$$

По той же таблице 6 критических точек распределения χ^2 находим $c_\alpha = c_\alpha(1-\alpha; n-1)$ и тогда, если для найденного значения $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} < c_\alpha$, то основную гипотезу H_0

отвергаем, иначе принимаем.

В случае двусторонней гипотезы

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 = \sigma_1^2 \neq \sigma_0^2,$$

объединяя два предыдущих случая, по таблице 6 критических точек распределения χ^2 находим две критические точки $c'_\alpha = c_\alpha\left(\frac{\alpha}{2}; n-1\right)$ и $c''_\alpha = c_\alpha\left(1-\frac{\alpha}{2}; n-1\right)$ и тогда, если $c''_\alpha < \chi^2 < c'_\alpha$, то гипотезу H_0 принимаем, иначе отвергаем.

Пример 4. Из генеральной совокупности, распределенной нормально, извлечена выборка объема $n = 30$ и по ней найдена исправленная выборочная дисперсия $s^2 = 12,5$.

Требуется при уровне значимости $\alpha = 0,01$ проверить основную гипотезу

$$H_0: \sigma^2 = \sigma_0^2 = 14 \text{ при конкурирующей гипотезе } H_1: \sigma^2 = \sigma_1^2 < 14.$$

Вычислим сначала $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29 \cdot 12,5}{14} = 25,9$. Учитывая, что гипотеза левая

односторонняя по таблице 6 критических точек распределения χ^2 находим

$c_\alpha = c_\alpha(1-\alpha; n-1) = c_\alpha(0,99; 29) = 14,3$. Так как $\chi^2 = 25,9 > c_\alpha = 14,3$, то гипотезу H_0

принимаем, то есть различие между найденной исправленной выборочной дисперсией s^2 и гипотетическим значением σ_0^2 не значительно.

§ 11.2. Проверка гипотез о значении вероятности p биномиального распределения.

Пусть случайная величина ξ имеет биномиальное распределение

$$P\{\xi = k\} = C_n^k p^k q^{n-k}, q = 1-p, \text{ при этом } M\xi = np \text{ и } D\xi = npq.$$

I. Проверка гипотез о значении вероятности p появления события в одном испытании.

Рассмотрим сначала случай, когда альтернативная гипотеза правая односторонняя

$$H_0: p = p_0$$

$$H_1: p = p_1 > p_0$$

Отношение правдоподобия для биномиального распределения имеет вид

$$\frac{p_1(x)}{p_0(x)} = \frac{C_n^x p_1^x (1-p_1)^{n-x}}{C_n^x p_0^x (1-p_0)^{n-x}} = \left(\frac{1-p_1}{1-p_0} \right)^n \cdot \left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right)^x.$$

Критическая область S определяется при некотором c из условия $P\left\{\frac{p_1(x)}{p_0(x)} > c \mid H_0\right\} = \alpha$.

Рассмотрим неравенство $\frac{p_1(x)}{p_0(x)} = \left(\frac{1-p_1}{1-p_0}\right)^n \cdot \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x > c$ и разрешим его

относительно x :

$$\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x > c \left(\frac{1-p_0}{1-p_1}\right)^n ; x \ln \frac{p_1(1-p_0)}{p_0(1-p_1)} > \ln \left[c \left(\frac{1-p_0}{1-p_1}\right)^n \right]$$

Так как $p_1 > p_0$, то $\frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1} > 1$ и $\ln \frac{p_1(1-p_0)}{p_0(1-p_1)} > 0$, откуда

$$x > \ln \left[c \left(\frac{1-p_0}{1-p_1}\right)^n \right] \cdot \left[\ln \frac{p_1(1-p_0)}{p_0(1-p_1)} \right]^{-1}.$$

Таким образом $P\left\{\frac{p_1(x)}{p_0(x)} > c \mid H_0\right\} = P\{x > c_1 \mid H_0\} = \alpha$,

$$\text{где } c_1 = \ln \left[c \left(\frac{1-p_0}{1-p_1}\right)^n \right] \cdot \left[\ln \frac{p_1(1-p_0)}{p_0(1-p_1)} \right]^{-1}.$$

Введем случайную величину $\eta = \frac{x - Mx}{\sqrt{Dx}} = \frac{x - np}{\sqrt{npq}}$, которая по теореме Муавра-Лапласа

асимптотически нормальна с параметрами $a = 0, \sigma = 1$. Если задается частота $\omega = \frac{x}{n}$

наступления события в n независимых испытаниях, то $\eta = \frac{x - np}{\sqrt{npq}} = \frac{\omega - p}{\sqrt{p(1-p)}} \cdot \sqrt{n}$.

Тогда $\alpha = P\{x > c_1 \mid H_0\} = P\left\{\frac{x - np_0}{\sqrt{np_0(1-p_0)}} > \frac{c_1 - np_0}{\sqrt{np_0(1-p_0)}}\right\} = P\{\eta > c_\alpha\}$,

где $c_\alpha = \frac{c_1 - np_0}{\sqrt{np_0(1-p_0)}}$. Так как

$$P\{\eta > c_\alpha\} = \frac{1}{\sqrt{2\pi}} \int_{c_\alpha}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_0^{c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{2} - \Phi(c_\alpha),$$

то для нахождения c_α получаем уже известное нам уравнение $\Phi(c_\alpha) = \frac{1}{2} - \alpha$.

Теперь, если $\eta = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\omega - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} > c_\alpha$, то гипотезу H_0 отвергаем, иначе

принимаем.

Если задано гипотетическое значение вероятности p_1 , то можно найти мощность построенного правостороннего критерия. Для этого найдем сначала

$$\beta = P\left\{\frac{p_1(x)}{p_0(x)} < c \mid H_1\right\} = P\{x < c_1 \mid H_1\} = P\left\{\frac{x - np_1}{\sqrt{np_1(1-p_1)}} < \frac{c_1 - np_1}{\sqrt{np_1(1-p_1)}}\right\} = P\{\eta < c_\beta\}, \text{ где}$$

$$c_\beta = \frac{c_1 - np_1}{\sqrt{np_1(1-p_1)}} = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot \frac{c_1 - np_0}{\sqrt{np_0(1-p_0)}} + \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot \frac{n(p_0 - p_1)}{\sqrt{np_0(1-p_0)}} =$$

$$= \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda \text{ и } \lambda = \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}} \sqrt{n}.$$

Учитывая, что $P\{\eta < c_\beta\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c_\beta} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_0^{c_\beta} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(c_\beta)$ получим

$$\beta = \frac{1}{2} + \Phi(c_\beta) = \frac{1}{2} + \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right) \text{ и мощность построенного критерия равна}$$

$$W = 1 - \beta = \frac{1}{2} - \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right).$$

Если же заданы уровень значимости α и мощность W (или ошибка II-го рода β), то можно найти минимальный объем выборки n , обеспечивающий эти величины.

Действительно, из уравнения $\Phi(c_\alpha) = \frac{1}{2} - \alpha$ находим c_α , а из уравнения $\Phi(c_\beta) = \beta - \frac{1}{2}$

находим c_β . Далее, учитывая связь между ними $c_\beta = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}} \sqrt{n}$,

получим $(p_0 - p_1) \sqrt{n} = \sqrt{p_1(1-p_1)} \cdot c_\beta - \sqrt{p_0(1-p_0)} \cdot c_\alpha$, откуда

$$n = \left(\frac{\sqrt{p_1(1-p_1)} \cdot c_\beta - \sqrt{p_0(1-p_0)} \cdot c_\alpha}{p_0 - p_1} \right)^2.$$

Проведенные рассуждения и полученные формулы аналогичны тем, которые были получены для нормального распределения при проверке гипотез о значении математического ожидания a при известной дисперсии σ^2 .

Аналогично рассматривается случай левосторонней альтернативной гипотезы

$$H_0: p = p_0$$

$$H_1: p = p_1 < p_0$$

С учетом того, что $\frac{p_1(1-p_0)}{p_0(1-p_1)} < 1$ и $\ln \frac{p_1(1-p_0)}{p_0(1-p_1)} < 0$ получим

$$\alpha = P\left\{\frac{p_1(x)}{p_0(x)} > c \mid H_0\right\} = P\{x < -c'_1 \mid H_0\}, \text{ где } c'_1 = -\ln \left[c \left(\frac{1-p_0}{1-p_1} \right)^n \right] \cdot \left[\ln \frac{p_1(1-p_0)}{p_0(1-p_1)} \right]^{-1}.$$

Далее имеем $\alpha = P\{x < -c'_1 \mid H_0\} = P\left\{\frac{x - np_0}{\sqrt{np_0(1-p_0)}} < -\frac{c'_1 + np_0}{\sqrt{np_0(1-p_0)}}\right\} = P\{\eta < -c_\alpha\}$, где

$$c_\alpha = \frac{c'_1 + np_0}{\sqrt{np_0(1-p_0)}}. \text{ Учитывая, что}$$

$$P\{\eta < -c_\alpha\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_0^{-c_\alpha} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(-c_\alpha) = \frac{1}{2} - \Phi(c_\alpha),$$

для определения c_α получаем тоже уравнение $\Phi(c_\alpha) = \frac{1}{2} - \alpha$.

Если теперь $\eta = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\omega - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} < -c_\alpha$, то гипотезу H_0 отвергаем, иначе принимаем.

По заданной гипотетической вероятности p_1 найдем мощность построенного левостороннего критерия

$$\beta = P\left\{\frac{p_1(x)}{p_0(x)} < c \mid H_1\right\} = P\{x > -c'_1 \mid H_1\} = P\{\eta > -c_\beta\}, \text{ где}$$

$$\begin{aligned} c_\beta &= \frac{c'_1 + np_1}{\sqrt{np_1(1-p_1)}} = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot \frac{c'_1 + np_0}{\sqrt{np_0(1-p_0)}} - \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot \frac{n(p_0 - p_1)}{\sqrt{np_0(1-p_0)}} = \\ &= \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \lambda \text{ и } \lambda = \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}} \sqrt{n}. \end{aligned}$$

$$\text{Учитывая, что } P\{\eta > -c_\beta\} = \frac{1}{\sqrt{2\pi}} \int_{-c_\beta}^{\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{-c_\beta}^0 e^{-\frac{t^2}{2}} dt =$$

$$= \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \int_0^{-c_\beta} e^{-\frac{t^2}{2}} dt = \frac{1}{2} - \Phi(-c_\beta) = \frac{1}{2} + \Phi(c_\beta) \text{ получим}$$

$$\beta = \frac{1}{2} + \Phi(c_\beta) = \frac{1}{2} + \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \lambda\right) \text{ и мощность построенного критерия равна}$$

$$W = 1 - \beta = \frac{1}{2} - \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \lambda\right).$$

Если же заданы ошибки I-го и II-го рода α и β , то найдя c_α и c_β соответственно из уравнений

$$\Phi(c_\alpha) = \frac{1}{2} - \alpha \text{ и } \Phi(c_\beta) = \beta - \frac{1}{2} \text{ и учитывая связь между ними}$$

$$c_\beta = \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}} \sqrt{n}, \text{ найдем минимальный объем выборки, обеспечивающий}$$

$$\text{эти ошибки } n = \left(\frac{\sqrt{p_0(1-p_0)} \cdot c_\alpha - \sqrt{p_1(1-p_1)} \cdot c_\beta}{p_0 - p_1} \right)^2.$$

Если же альтернативная гипотеза двусторонняя

$$H_0 : p = p_0$$

$$H_1 : p = p_1 \neq p_0$$

то, объединяя два рассмотренных выше случая, получим

$$\alpha = P\{|\eta| > c_\alpha\} = P\{\eta > c_\alpha\} + P\{\eta < -c_\alpha\} = \frac{1}{2} - \Phi(c_\alpha) + \frac{1}{2} - \Phi(c_\alpha) = 1 - 2\Phi(c_\alpha)$$

и для нахождения c_α получаем уравнение $\Phi(c_\alpha) = \frac{1-\alpha}{2}$.

Если теперь $|\eta| = \frac{|x - np_0|}{\sqrt{np_0(1-p_0)}} = \frac{|\omega - p_0|}{\sqrt{p_0(1-p_0)}} \sqrt{n} > c_\alpha$, то гипотезу H_0 отвергаем, иначе

принимаем.

Найдем мощность построенного двустороннего критерия

$$\begin{aligned} \beta &= P\{-c_\beta < \eta < c_\beta\} = P\left\{-\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda < \eta < \sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right\} = \\ &= \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right) - \Phi\left(-\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right) = \\ &= \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right) + \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \lambda\right) \text{ и} \\ W &= 1 - \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha + \lambda\right) - \Phi\left(\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha - \lambda\right). \end{aligned}$$

Пример 5. По 100 независимым испытаниям найдена относительная частота $\omega = 0,14$.

При уровне значимости $\alpha = 0,05$ требуется проверить гипотезы

$$H_0 : p = p_0 = 0,2$$

$$H_1 : p = p_1 \neq 0,2.$$

Найти мощность построенного двустороннего критерия для гипотетического значения $p_1 = 0,1$.

Из уравнения $\Phi(c_\alpha) = \frac{1-\alpha}{2} = \frac{0,95}{2} = 0,475$ по таблице 2 значений функции $\Phi(x)$ находим

$$c_\alpha = 1,96. \text{ Вычислим } \eta = \frac{\omega - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} = \frac{0,14 - 0,2}{\sqrt{0,2 \cdot 0,8}} \sqrt{100} = \frac{-0,06}{0,4} \cdot 10 = -1,5.$$

Так как $|\eta| = 1,5 < c_\alpha = 1,96$, то у нас нет оснований отвергать гипотезу H_0 . Таким образом наблюдаемая относительная частота $\omega = 0,14$ незначимо отличается от гипотетической вероятности $p_0 = 0,2$.

Для нахождения мощности построенного критерия вычислим сначала

$$\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} \cdot c_\alpha = \sqrt{\frac{0,2 \cdot 0,8}{0,1 \cdot 0,9}} \cdot 1,96 = \frac{4}{3} \cdot 1,96 = 2,61 \text{ и}$$

$$\lambda = \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}} \sqrt{n} = \frac{0,2 - 0,1}{\sqrt{0,1 \cdot 0,9}} \sqrt{100} = \frac{0,1}{0,3} \cdot 10 = 3,33.$$

Тогда мощность

$$W = 1 - \Phi(2,61 + 3,33) - \Phi(2,61 - 3,33) = 1 - \Phi(5,94) - \Phi(-0,72) = 1 - 0,5 + 0,264 = 0,764.$$

II. Сравнение двух вероятностей биномиального распределения.

Пусть имеется две генеральные совокупности, распределенные по биномиальному закону, с неизвестными вероятностями p_1 и p_2 наступления события A в этих совокупностях. Из этих генеральных совокупностей извлечены выборки объемов n_1 и n_2 соответственно и пусть число появлений события A в этих выборках равны m_1 и m_2

соответственно, то есть известны относительные частоты $\omega_1 = \frac{m_1}{n_1}$ и $\omega_2 = \frac{m_2}{n_2}$.

Предположим, что требуется проверить гипотезу о равенстве вероятностей наступления события A в этих двух генеральных совокупностях

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Как было показано ранее в §10.3, точечной несмещенной оценкой для вероятности p

является относительная частота $\omega = \frac{m}{n} : M(\omega) = p$, при этом ее дисперсия $D(\omega) = \frac{pq}{n}$.

Введем нормированную случайную величину

$$\eta = \frac{\omega_1 - \omega_2 - M(\omega_1 - \omega_2)}{\sqrt{D(\omega_1 - \omega_2)}},$$

которая будет асимптотически нормальной с параметрами $a = 0$ и $\sigma = 1$.

При гипотезе H_0 вероятности $p_1 = p_2 = p$, где p их общее значение. Тогда

$$M(\omega_1 - \omega_2) = M(\omega_1) - M(\omega_2) = p - p = 0$$

$$D(\omega_1 - \omega_2) = D(\omega_1) + D(\omega_2) = \frac{pq}{n_1} + \frac{pq}{n_2} = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Кроме того, объединяя обе выборки в одну, получим следующую оценку для их общей вероятности $p \approx \frac{m_1 + m_2}{n_1 + n_2}$.

$$\text{Таким образом, } \eta = \frac{\omega_1 - \omega_2}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2} \left(1 - \frac{m_1 + m_2}{n_1 + n_2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

Так как гипотеза двусторонняя, то, как мы уже показывали ранее, критическая область S определяется при некотором c_α из соотношения $P\{|\eta| > c_\alpha \mid H_0\} = \alpha$, что приводит нас к

уже известному уравнению $\Phi(c_\alpha) = \frac{1-\alpha}{2}$.

Теперь, если $|\eta| < c_\alpha$, то гипотезу H_0 принимаем, то есть вероятности p_1 и p_2 разнятся незначительно, если же $|\eta| > c_\alpha$, то гипотезу H_0 отвергаем.

В случае односторонних гипотез

$$H_0 : p_1 = p_2 \text{ или } H_0 : p_1 \neq p_2$$

$$H_1 : p_1 > p_2 \quad H_1 : p_1 < p_2$$

то значение c_α находится из уравнения $\Phi(c_\alpha) = \frac{1}{2} - \alpha$.

Гипотезу H_0 принимаем, если $\eta < c_\alpha$ при правосторонней гипотезе и $\eta > -c_\alpha$ при левосторонней гипотезе. В противных случаях гипотезу H_0 отвергаем.

Пример 6. Первое предприятие изготовило партию из 800 изделий и из них 15 оказалось бракованными. Второе предприятие изготовило партию в 1000 изделий и из них бракованными оказалось 25 изделий. Требуется при уровне значимости $\alpha = 0,05$ проверить гипотезу о равенстве вероятностей изготовления бракованного изделия обоими предприятиями:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Из уравнения $\Phi(c_\alpha) = \frac{1-\alpha}{2} = \frac{0,95}{2} = 0,475$ по таблице 2 находим $c_\alpha = 1,96$.

$$\text{Вычислим } \eta = \frac{\frac{15}{800} - \frac{25}{1000}}{\sqrt{\frac{40}{1800} \left(1 - \frac{40}{1800}\right) \left(\frac{1}{800} + \frac{1}{1000}\right)}} = \frac{0,01875 - 0,025}{\sqrt{\frac{1}{45} \cdot \frac{44}{45} \cdot 0,00225}} = \frac{-0,28125}{0,3146} = -0,89.$$

Так как $|\eta| = 0,89 < c_\alpha = 1,96$, то у нас нет оснований отвергать нулевую гипотезу. Таким образом, вероятности изготовления бракованного изделия у обоих предприятий различаются незначительно.

Задания.

1. Из генеральной совокупности, распределенной по нормальному закону с известным среднеквадратичным отклонением $\sigma = 5$, извлечена выборка объема $n = 36$ и по ней найдена выборочная средняя $\bar{x}_s = 13,5$. Требуется при уровне значимости $\alpha = 0,01$: 1) проверить основную гипотезу $H_0 : a = a_0 = 15$ при конкурирующей гипотезе $H_1 : a = a_1 < 15$; 2) найти мощность построенного левостороннего критерия проверки для гипотетического значения $a_1 = 14$; 3) найти минимальный объем выборки n , при котором мощность критерия равна 0,8.
2. Из генеральной совокупности, распределенной по нормальному закону с известным среднеквадратичным отклонением $\sigma = 5$, извлечена выборка объема $n = 49$ и по ней найдена выборочная средняя $\bar{x}_s = 27,5$. Требуется при уровне значимости $\alpha = 0,02$: 1) проверить основную гипотезу $H_0 : a = a_0 = 26$ при конкурирующей гипотезе $H_1 : a = a_1 \neq 26$; 2) найти мощность построенного левостороннего критерия проверки для гипотетического значения $a_1 = 28$.

3. Из генеральной совокупности, распределенной по нормальному закону, извлечена выборка

x_i	2	4	6	8	10	12
n_i	2	6	10	12	8	3

Требуется при уровне значимости $\alpha = 0,01$ проверить основную гипотезу $H_0 : a = a_0 = 7$ при конкурирующей гипотезе $H_1 : a = a_1 > 7$.

4. Из генеральной совокупности, распределенной по нормальному закону, извлечена выборка

x_i	2	2,5	3	3,5	4	4,5
n_i	1	3	8	6	2	1

Требуется при уровне значимости $\alpha = 0,02$ проверить основную гипотезу $H_0 : a = a_0 = 3$ при конкурирующей гипотезе $H_1 : a = a_1 \neq 3$.

5. Из двух генеральных совокупностей, распределенных нормально с дисперсиями $\sigma_1^2 = 60$ и $\sigma_2^2 = 80$ соответственно, по двум независимым выборкам объема $n = 30$ и $m = 40$ найдены выборочные средние $\bar{x}_e = 130$ и $\bar{y}_e = 125$. При уровне значимости $\alpha = 0,05$ требуется проверить гипотезы

$$H_0 : a_1 = a_2$$

$$H_1 : a_1 > a_2.$$

6. Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать $\sigma_0^2 = 0,1$. В результате измерений 25 случайно отобранных изделий получена следующая выборка

x_i	2,8	3,0	3,2	3,4	3,6
n_i	2	6	9	7	1

Предполагается, что размер изделия x_i есть нормально распределенная случайная величина. Требуется при уровне значимости $\alpha = 0,05$ проверить, обеспечивает ли станок требуемую точность. Принять нулевую гипотезу $H_0 : \sigma^2 = 0,1$ при конкурирующей гипотезе $H_1 : \sigma^2 > 0,1$.

7. Из генеральной совокупности, распределенной по нормальному закону, извлечена выборка

x_i	56	58	60	62	64
n_i	1	4	10	3	2

Требуется при уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0 : \sigma^2 = \sigma_0^2 = 2$ при конкурирующей гипотезе $H_1 : \sigma^2 \neq 2$.

8. Предприятие рассылает рекламные каталоги возможным заказчикам. Как показал опыт, вероятность того, что организация, получившая каталог, закажет рекламируемое изделие, равна 0,08. Предприятие разослало 1000 каталогов новой улучшенной формы и получило 100 заказов. Можно ли при уровне значимости $\alpha = 0,05$ считать, что новая форма каталога оказалась эффективнее старой? Принять нулевую гипотезу $H_0 : p = p_0 = 0,08$ при конкурирующей гипотезе

$H_1: p = p_1 > 0,08$. Найти мощность построенного правостороннего критерия для гипотетического значения $p_1 = 0,1$.

9. При $n = 100$ выстрелов по цели каждым из двух орудий зарегистрировано соответственно $m_1 = 10$ и $m_2 = 12$ промахов. При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу $H_0: p_1 = p_2$ о равенстве вероятностей промаха обоих орудий при конкурирующей гипотезе $H_1: p_1 \neq p_2$.

§ 12. Статистическая проверка гипотез о виде закона распределения.

В предыдущем разделе были рассмотрены так называемые параметрические критерии, предназначенные для проверки гипотез о значении параметров распределений, если закон распределения случайной величины известен. В данном разделе речь пойдет о непараметрических критериях, предназначенных для проверки гипотез о виде закона распределения.

Пусть $\zeta = (\xi_1, \xi_2, \dots, \xi_n)$ выборка из генеральной совокупности, закон распределения которой неизвестен. Основная гипотеза H_0 утверждает, что закон распределения имеет вполне конкретный вид (нормальный, показательный или какой-либо другой) и задается функцией распределения $F_\xi(x)$. Альтернативная гипотеза здесь никак не конкретизируется, то есть гипотеза H_1 утверждает «не H_0 » и речь идет просто о согласии данных выборки и гипотезы H_0 . Наиболее известными критериями согласия для данной задачи являются критерий Колмогорова и критерий Пирсона или χ^2 (хи-квадрат).

§ 12.1. Критерий Колмогорова.

Этот критерий применяется в том случае, когда гипотетическая функция распределения $F_\xi(x)$ абсолютно непрерывна. Критерий Колмогорова базируется на теореме Гливленко, которая говорит о сходимости по вероятности эмпирической функции распределения $F_n(x)$ к теоретической функции распределения $F_\xi(x)$

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{-\infty < x < \infty} |F_n(x) - F_\xi(x)| \leq \varepsilon \right\} = 1.$$

В качестве критерия выбирается величина

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F_\xi(x)|.$$

Она замечательна тем, что распределение величины D_n при гипотезе H_0 не зависит от вида функции $F_\xi(x)$, то есть в качестве гипотетической функции $F_\xi(x)$ может выступать произвольная абсолютно непрерывная функция распределения. Кроме того она имеет предельное распределение, а именно

$$\lim_{n \rightarrow \infty} P \left\{ \sqrt{n} D_n \leq x \right\} = K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 x^2} = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 x^2}, & x > 0 \\ 0, & x \leq 0 \end{cases}.$$

Критическая область S при заданном уровне значимости α и достаточно больших n (уже при $n \geq 20$) определяется соотношением

$$\alpha = P\{\sqrt{n}D_n > c_\alpha \mid H_0\} \approx 1 - K(c_\alpha).$$

Критическое значение c_α определяется из уравнения $K(c_\alpha) = 1 - \alpha$ при помощи таблицы 9 значений функции $K(x)$. Если $\sqrt{n}D_n < c_\alpha$, то гипотезу H_0 принимаем, иначе отвергаем.

Пример 1. Из генеральной совокупности извлечена выборка объема $n = 200$

x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9	2,1	2,3
n_i	6	9	24	27	30	26	23	22	20	8	3

Используя критерий Колмогорова, при уровне значимости $\alpha = 0,01$, проверить гипотезу о нормальном распределении генеральной совокупности.

Сначала найдем точечные несмещенные оценки параметров нормального распределения

$$a \approx \bar{x}_s = \frac{1}{n} \sum_{i=1}^{11} x_i n_i = \frac{1}{200} (0,3 \cdot 6 + 0,5 \cdot 9 + \dots + 2,3 \cdot 3) = 1,2620$$

$$\sigma^2 \approx s^2 = \frac{1}{n-1} \sum_{i=1}^{11} (x_i - \bar{x}_s)^2 n_i = \frac{1}{199} [(0,3 - 1,262)^2 \cdot 6 + \dots + (2,3 - 1,262)^2 \cdot 3] = 0,2381,$$

откуда $\sigma \approx s = \sqrt{0,2381} = 0,4880$.

Далее вычислим значения теоретической функции распределения $F_\xi(x)$ в точках x_i

$$F_\xi(x_i) = \frac{1}{2} + \Phi\left(\frac{x_i - a}{\sigma}\right) = \frac{1}{2} + \Phi\left(\frac{x_i - 1,262}{0,488}\right).$$

При помощи таблицы 2 вычислим

$$F_\xi(0,3) = 0,5 + \Phi\left(\frac{0,3 - 1,262}{0,488}\right) = 0,5 + \Phi(-1,97) = 0,5 - 0,4756 = 0,0244 \text{ и так далее.}$$

Затем вычисляем значения эмпирической функции распределения $F_n(x)$ в точках x_i . Так как эта функция терпит скачки в точках x_i , то будем брать полу сумму левого и правого

$$\text{предельных значений } F_n(0,3) = \frac{1}{2} \left(0 + \frac{6}{200}\right) = 0,015; F_n(0,5) = \frac{1}{2} \left(\frac{6}{200} + \frac{15}{200}\right) = 0,0525 \text{ и}$$

так далее. Результаты вычислений занесем в следующую таблицу

x_i	n_i	$F_{\xi}(x_i)$	$F_n(x_i)$	$ F_{\xi}(x_i) - F_n(x_i) $
0,3	6	0,0244	0,0150	0,0093
0,5	9	0,0592	0,0525	0,0067
0,7	24	0,1247	0,1350	0,0103
0,9	27	0,2291	0,2625	0,0334
1,1	30	0,3700	0,4050	0,0350
1,3	26	0,5310	0,5450	0,0140
1,5	23	0,6871	0,6675	0,0196
1,7	22	0,8153	0,7800	0,0353
1,9	20	0,9045	0,8850	0,0195
2,1	8	0,9570	0,9550	0,0020
2,3	3	0,9833	0,9875	0,0042

Из таблицы находим $D_n = \sup_x |F_{\xi}(x) - F_n(x)| = 0,0353$ и $\sqrt{n}D_n = \sqrt{200} \cdot 0,0353 = 0,499$.

По таблице 7 находим решение уравнения $K(c_{\alpha}) = 1 - \alpha = 0,99$: $c_{\alpha} = 1,63$.

Так как $\sqrt{n}D_n = 0,499 < c_{\alpha} = 1,63$, то у нас нет оснований отвергать гипотезу о нормальном распределении генеральной совокупности.

§ 12.2. Критерий Пирсона или χ^2 .

Одним из наиболее универсальных методов статистической проверки гипотез о виде закона распределения является метод, предложенный К. Пирсоном в 1900 году и носящий его имя. Пусть из генеральной совокупности извлечена выборка объема n , которую запишем в виде вариационного ряда

x_1	x_2	$\dots$	x_k
n_1	n_2	$\dots$	n_k

входящие в нее частоты n_i ($n_1 + n_2 + \dots + n_k = n$) называют эмпирическими частотами.

Пусть основная гипотеза H_0 утверждает, что закон распределения генеральной совокупности описывается функцией распределения $F_{\xi}(x)$. Найдем по выборке точечные оценки параметров, входящих в функцию распределения, если они не заданы самой гипотезой. Например, основная гипотеза – имеет место биномиальное распределение с неизвестной вероятностью p наступления события в одном испытании, или – имеет место биномиальное распределение с вероятностью $p = 0,3$ наступления события в одном испытании. Далее находятся так называемые теоретические частоты n_i' . Теоретическая частота показывает сколько раз случайная величина примет значение x_i , если бы был справедлив гипотетический закон распределения случайной величины. Для этого согласно

гипотетическому закону вычисляются вероятности $p_i = P\{\xi = x_i\}$, а затем, учитывая приближенное равенство $p_i \approx \frac{n'_i}{n}$, находят $n'_i = np_i$. Если же выборка задается в виде интервалов и соответствующим им эмпирических частот, то вероятности p_i есть вероятности попадания случайной величины в соответствующий интервал. После этого вводится случайная величина

$$\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}.$$

Эта случайная величина, во-первых, не зависит от вида функции $F_\xi(x)$, то есть в роли гипотетической функции может выступать любая функция распределения, а во-вторых, при достаточно больших объемах выборки она имеет предельное распределение. Этим предельным распределением является распределение Пирсона или χ^2 с $m = k - r - 1$ степенью свободы. В последней формуле k есть число различных вариантов (или интервалов) в выборке, а r есть число определяемых по выборке параметров гипотетического распределения. На практике предельное распределение можно использовать для расчетов с хорошим приближением уже при $n \geq 50$ и $n_i \geq 5$ ($i = 1, 2, \dots, k$).

Малочисленные частоты ($n_i < 5$) следует объединить; в этом случае и соответствующие им теоретические частоты также суммируются. Если производилось объединение частот, то при определении числа степеней свободы $m = k - r - 1$ в качестве значения k следует принять число групп выборки, оставшихся после объединения частот.

При заданном уровне значимости α критическая область S определяется из соотношения $P\{\chi^2 > c_\alpha \mid H_0\} = \alpha$. Критическое значение c_α находится по таблице 6 критических точек распределения χ^2 с учетом числа степеней свободы. Если $\chi_n^2 < c_\alpha$, то гипотезу H_0 о законе распределения принимаем, иначе отвергаем.

Рассмотрим вычисление теоретических частот для различных видов распределений.

I. Нахождение теоретических частот при проверке гипотез о нормальном распределении для случая, когда выборка задана в виде последовательности равноотстоящих вариант и соответствующих им частот.

Пусть вариационный ряд состоит из равноотстоящих вариант $x_{i+1} - x_i = h$. По выборке находим точечные несмещенные оценки параметров нормального распределения $a \approx \bar{x}_s$ и $\sigma^2 \approx s^2$. Перейдем к нормированным случайным величинам $u_i = \frac{x_i - a}{\sigma} \approx \frac{x_i - \bar{x}_s}{s}$.

Можно считать, что $P\{x = x_i\} = P\{u = u_i\} \approx \frac{1}{s\sqrt{2\pi}} e^{-\frac{u_i^2}{2}} = \frac{1}{s} \varphi(u_i)$ и

$$p_i = P\{x_i \leq x < x_{i+1}\} = P\{u_i \leq u < u_{i+1}\} \approx \frac{h}{s} \varphi(u_i), \text{ откуда } n'_i = np_i = \frac{nh}{s} \varphi(u_i).$$

Теперь, зная n_i и n'_i , находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$ и сравниваем с c_α , которое находится по

таблице 6 критических точек распределения χ^2 с $m = k - 2 - 1 = k - 3$ степенями свободы ($r = 2$, так как оба параметра нормального распределения мы находили по выборке).

Если $\chi_n^2 < c_\alpha$, то гипотезу о нормальном распределении генеральной совокупности принимаем, иначе отвергаем.

Пример 2. Используя критерий Пирсона, при уровне значимости $\alpha = 0,05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки объема $n = 200$

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

Находим точечные оценки параметров нормального распределения

$$a \approx \bar{x}_s = \frac{1}{n} \sum_{i=1}^9 x_i n_i = 12,63 \text{ и } \sigma^2 \approx s^2 = \frac{1}{n-1} \sum_{i=1}^9 (x_i - \bar{x}_s)^2 n_i = 22,1539, \text{ откуда } \sigma \approx s = 4,7068.$$

Далее вычисляем теоретические частоты $n'_i = \frac{nh}{s} \varphi(u_i) = \frac{200 \cdot 2}{4,7068} \varphi(u_i) = 84,9834 \cdot \varphi(u_i)$, где

$$u_i = \frac{x_i - \bar{x}_s}{s} = \frac{x_i - 12,63}{4,7068}, \text{ а значения функции } \varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \text{ можно найти, используя}$$

таблицу 1. Результаты вычислений представим в виде таблицы

x_i	n_i	u_i	$\varphi(u_i)$	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
5	15	-1,6211	0,1072	9,1121	3,8046
7	26	-1,1961	0,1951	16,5790	5,3534
9	25	-0,7712	0,2963	25,1819	0,0013
11	30	-0,3463	0,3757	31,9303	0,1167
13	26	0,0786	0,3977	33,7990	1,7996
15	21	0,5035	0,3514	29,8669	2,6324
17	24	0,9284	0,2593	22,0324	0,1757
19	20	1,3534	0,1597	13,5681	3,0490
21	13	1,7783	0,0821	6,9753	5,2036

Теперь находим $\chi_n^2 = \sum_{i=1}^9 \frac{(n_i - n_i')^2}{n_i'} = 22,1364$, а по таблице 6 по $\alpha = 0,05$ и числу степеней свободы $m = 9 - 2 - 1 = 6$ находим $c_\alpha = 12,6$. Так как $\chi_n^2 = 22,14 > c_\alpha = 12,6$, то гипотезу о нормальном распределении генеральной совокупности отвергаем.

II. Нахождение теоретических частот при проверке гипотез о нормальном распределении для случая, когда выборка задана в виде последовательности интервалов одинаковой длины и соответствующих им частот.

Пусть выборка задана в виде интервалов и соответствующих им частот

$(x_0; x_1)$	$(x_1; x_2)$	$\dots$	$(x_{k-1}; x_k)$
n_1	n_2	$\dots$	n_k

По серединам интервалов $x_i^* = \frac{x_i + x_{i-1}}{2}$ находим точечные несмещенные оценки

параметров нормального распределения $a \approx \bar{x}_s$ и $\sigma^2 \approx s^2$. Перейдем к нормированной случайной величине $u = \frac{x - a}{\sigma} \approx \frac{x - \bar{x}_s}{s}$ и рассмотрим интервалы $(u_0; u_1), (u_1; u_2), \dots, (u_{k-1}; u_k)$

где $u_0 = -\infty, u_i = \frac{x_i - \bar{x}_s}{s} (i = 1, 2, \dots, k-1), u_k = +\infty$.

Учитывая, что $p_i = P\{x_{i-1} \leq x < x_i\} = P\{u_{i-1} \leq u < u_i\} = \Phi(u_i) - \Phi(u_{i-1})$, находим теоретические частоты $n_i' = np_i = n[\Phi(u_i) - \Phi(u_{i-1})]$.

Теперь, зная n_i и n_i' , находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i'}$ и сравниваем с c_α , которое находится по

таблице 6 критических точек распределения χ^2 с $m = k - 2 - 1 = k - 3$ степенями свободы

($r = 2$, так как оба параметра нормального распределения мы находили по выборке).

Если $\chi_n^2 < c_\alpha$, то гипотезу о нормальном распределении генеральной совокупности принимаем, иначе отвергаем.

Пример 3. Используя критерий Пирсона, при уровне значимости $\alpha = 0,05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки объема $n = 300$

$(x_{i-1}; x_i)$	$(-20; -10)$	$(-10; 0)$	$(0; 10)$	$(10; 20)$	$(20; 30)$	$(30; 40)$	$(40; 50)$
n_i	20	47	80	89	40	16	8

По серединам интервалов находим точечные оценки параметров нормального распределения

$$a \approx \bar{x}_e = \frac{1}{n} \sum_{i=1}^7 x_i^* n_i = 10,4 \text{ и } \sigma^2 \approx s^2 = \frac{1}{n-1} \sum_{i=1}^7 (x_i^* - \bar{x}_e)^2 n_i = 187,4649, \text{ откуда } \sigma \approx s = 13,6918.$$

Далее находим концы интервалов $u_i = \frac{x_i - \bar{x}_e}{s} = \frac{x_i - 10,4}{13,6918}$ и теоретические частоты

$$n'_i = n [\Phi(u_i) - \Phi(u_{i-1})] \text{ с использованием таблицы 2 значений функции } \Phi(x).$$

Результаты вычислений представим в виде таблицы

$(x_{i-1}; x_i)$	n_i	$(u_{i-1}; u_i)$	$n'_i = n [\Phi(u_i) - \Phi(u_{i-1})]$	$\frac{(n_i - n'_i)^2}{n'_i}$
$(-\infty; -10)$	20	$(-\infty; -1,4899)$	20,4358	0,0093
$(-10; 0)$	47	$(-1,4899; -0,7596)$	46,6901	0,0021
$(0; 10)$	80	$(-0,7596; -0,0292)$	79,3781	0,0049
$(10; 20)$	89	$(-0,0292; 0,7012)$	81,0146	0,7871
$(20; 30)$	40	$(0,7012; 1,4315)$	49,6390	1,8717
$(30; 40)$	16	$(1,4315; 2,1619)$	18,2433	0,2770
$(40; 50)$	8	$(2,1619; +\infty)$	4,5941	2,5250

Теперь находим $\chi_n^2 = \sum_{i=1}^7 \frac{(n_i - n'_i)^2}{n'_i} = 5,477$, а по таблице 6 по $\alpha = 0,05$ и числу степеней свободы $m = 7 - 2 - 1 = 4$ находим $c_\alpha = 9,5$. Так как $\chi_n^2 = 5,477 < c_\alpha = 9,5$, то гипотезу о нормальном распределении генеральной совокупности принимаем.

III. Нахождение теоретических частот при проверке гипотез о показательном распределении.

Предположим, что выдвинута гипотеза о том, что генеральная совокупность имеет показательное распределение с плотностью

$$p_\xi(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \text{ где параметр } \lambda > 0.$$

Пусть выборка задана в виде интервалов и соответствующих им частот

$(x_0; x_1)$	$(x_1; x_2)$	$\dots$	$(x_{k-1}; x_k)$
n_1	n_2	$\dots$	n_k

где $n_1 + n_2 + \dots + n_k = n$.

По серединам интервалов $x_i^* = \frac{x_i + x_{i-1}}{2}$ найдем точечную несмещенную оценку параметра

$\lambda \approx \frac{1}{\bar{x}_e}$ и вероятности попадания случайной величины в каждый из интервалов

$$p_i = P\{x_{i-1} \leq x < x_i\} = \int_{x_{i-1}}^{x_i} \lambda e^{-\lambda x} dx \approx \frac{1}{\bar{x}_e} \int_{x_{i-1}}^{x_i} e^{-\frac{x}{\bar{x}_e}} dx = \frac{1}{\bar{x}_e} \left(-\bar{x}_e \right) e^{-\frac{x}{\bar{x}_e}} \Big|_{x_{i-1}}^{x_i} = e^{-\frac{x_{i-1}}{\bar{x}_e}} - e^{-\frac{x_i}{\bar{x}_e}},$$

откуда теоретические частоты равны $n_i' = n \left(e^{-\frac{x_{i-1}}{\bar{x}_e}} - e^{-\frac{x_i}{\bar{x}_e}} \right)$.

Теперь, зная n_i и n_i' , находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i'}$ и сравниваем с c_α , которое находится по

таблице 6 критических точек распределения χ^2 с $m = k - 1 - 1 = k - 2$ степенями свободы ($r = 1$, так как параметр показательного распределения мы находили по выборке).

Если $\chi_n^2 < c_\alpha$, то гипотезу о нормальном распределении генеральной совокупности принимаем, иначе отвергаем.

Пример 4. В результате испытания $n = 200$ элементов на продолжительность безотказной работы получено следующее эмпирическое распределение

$(x_{i-1}; x_i)$	(0;5)	(5;10)	(10;15)	(15;20)	(20;25)	(25;30)
n_i	133	45	15	4	2	1

При уровне значимости $\alpha = 0,01$ используя критерий Пирсона требуется проверить гипотезу о том, что продолжительность безотказной работы элементов распределено по показательному закону.

По серединам интервалов найдем $\bar{x}_e = \frac{1}{n} \sum_{i=1}^6 x_i^* n_i = 5$, тогда $\lambda \approx 0,2$. Затем находим

теоретические частоты $n_i' = n(e^{-0,2x_i} - e^{-0,2x_{i-1}})$. При этом объединим малочисленные частоты ($4+2+1=7$), а соответствующие им интервалы объединим в один интервал (15;30).

Результаты вычислений занесем в таблицу

$(x_{i-1}; x_i)$	n_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
(0;5)	133	126,4241	0,3420
(5;10)	45	46,5088	0,0489
(10;15)	15	17,1096	0,2601
(15;30)	7	9,4616	0,6404

Теперь находим $\chi_n^2 = \sum_{i=1}^4 \frac{(n_i - n'_i)^2}{n'_i} = 1,2914$, а по таблице 6 по $\alpha = 0,01$ и числу степеней свободы $m = 4 - 1 - 1 = 2$ находим $c_\alpha = 9,2$. Так как $\chi_n^2 = 1,2914 < c_\alpha = 9,2$, то гипотезу о показательном распределении принимаем.

IV. Нахождение теоретических частот при проверке гипотез о равномерном распределении.

Предположим, что выдвинута гипотеза о том, что генеральная совокупность распределена равномерно на некотором промежутке $(a; b)$ с плотностью

$$p_\xi(x) = \begin{cases} \frac{1}{b-a}, & x \in (a; b) \\ 0, & x \notin (a; b) \end{cases}.$$

Пусть выборка задана в виде интервалов и соответствующих им частот

$(x_0; x_1)$	$(x_1; x_2)$	$\dots$	$(x_{k-1}; x_k)$
n_1	n_2	$\dots$	n_k

где $n_1 + n_2 + \dots + n_k = n$.

Во-первых, следует найти точечные оценки a^* и b^* для концов интервала $(a; b)$. Для этого следует воспользоваться методом моментов. Для равномерного распределения из соотношений $M_\xi = \frac{a+b}{2} \approx \bar{x}_e$ и $D\xi = \frac{(b-a)^2}{12} \approx s^2$ получаем систему уравнений

$$\begin{cases} a^* + b^* = 2\bar{x}_e \\ b^* - a^* = 2s\sqrt{3} \end{cases}, \text{ из которой находим } \begin{cases} a^* = \bar{x}_e - s\sqrt{3} \\ b^* = \bar{x}_e + s\sqrt{3} \end{cases}.$$

По серединам интервалов $x_i^* = \frac{x_i + x_{i-1}}{2}$ находим $\bar{x}_e = \frac{1}{n} \sum_{i=1}^k x_i^* n_i$ и $s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i^* - \bar{x}_e)^2 n_i$, а затем a^* и b^* .

Учитывая, что плотность функции распределения $p_\xi(x) \approx \frac{1}{b^* - a^*}$, находим теоретические

$$\text{частоты } n'_1 = n \cdot \frac{x_1 - a^*}{b^* - a^*}, n'_i = n \cdot \frac{x_i - x_{i-1}}{b^* - a^*} (i = 2, 3, \dots, k-1), n'_k = n \cdot \frac{b^* - x_{k-1}}{b^* - a^*}.$$

Далее находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$ и сравниваем с c_α , которое находится по таблице 6

критических точек распределения χ^2 с $m = k - 2 - 1 = k - 3$ степенями свободы ($r = 2$, так как два параметра равномерного распределения мы находили по выборке). Если $\chi_n^2 < c_\alpha$, то гипотезу о равномерном распределении генеральной совокупности принимаем, иначе отвергаем.

Если в формулировке задачи указано конкретное значение интервала $(a; b)$, то их точечные оценки находить не надо, теоретические частоты $n'_i = n \cdot \frac{x_i - x_{i-1}}{b - a}$ ($i = 1, 3, \dots, k$) и при нахождении c_α число степеней свободы $m = k - 1$.

Пример 5. Произведено $n = 200$ испытаний, в результате каждого из которых событие А появлялось в различные моменты времени. В итоге было получено эмпирическое распределение (в первой строке указаны интервалы времени в минутах, а во второй – число появлений события А в данном интервале)

$(x_{i-1}; x_i)$	(2;4)	(4;6)	(6;8)	(8;10)	(10;12)	(12;14)	(14;16)	(16;18)	(18;20)
n_i	23	19	18	29	24	18	23	25	21

Требуется при уровне значимости $\alpha = 0,05$, используя критерий Пирсона проверить гипотезу о том, что время появления событий распределено равномерно.

По серединам интервалов $x_i^* = \frac{x_i + x_{i-1}}{2}$ находим $\bar{x}_e = \frac{1}{n} \sum_{i=1}^9 x_i^* n_i = 11,0900$ и

$s^2 = \frac{1}{n-1} \sum_{i=1}^9 (x_i^* - \bar{x}_e)^2 n_i = 26,3436$, откуда $s = 5,1326$. Далее находим $a^* = 2,2001$,

$b^* = 19,9799$ и теоретические частоты, которые занесем в таблицу

$(x_{i-1}; x_i)$	n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
(2;4)	23	20,2468	0,3744
(4;6)	19	22,4974	0,5437
(6;8)	18	22,4974	0,8991
(8;10)	29	22,4974	1,8725
(10;12)	24	22,4974	0,1004
(12;14)	18	22,4974	0,8991
(14;16)	23	22,4974	0,0112
(16;18)	25	22,4974	0,2784
(18;20)	21	22,2716	0,0726

Теперь находим $\chi_n^2 = \sum_{i=1}^9 \frac{(n_i - n'_i)^2}{n'_i} = 5,0583$, а по таблице 6 по $\alpha = 0,05$ и числу степеней

свободы $m = 9 - 2 - 1 = 6$ находим $c_\alpha = 12,6$. Так как $\chi_n^2 = 5,0583 < c_\alpha = 12,6$, то гипотезу о равномерном распределении принимаем.

V. Нахождение теоретических частот при проверке гипотез о биномиальном распределении.

Предположим, что произведено n серий испытаний. Каждая серия испытаний состоит из m опытов. В каждом из опытов событие A может появиться с вероятностью p . В итоге получено следующее эмпирическое распределение

x_i	0	1	...	m
n_i	n_0	n_1	...	n_m

В первой строке указано число x_i появлений события A в одном опыте, во второй строке – число опытов, в которых зарегистрировано x_i появлений события A .

Требуется проверить гипотезу о том, что случайная величина – число появлений события A распределена по биномиальному закону.

Если вероятность p не задана, то по выборке находим ее точечную несмещенную оценку $p \approx \frac{1}{m} \bar{x}_e$. Далее по формуле Бернулли находим вероятности

$$p_i = P_m(i) = C_m^i p^i (1-p)^{m-i} \text{ и теоретические частоты } n_i' = n \cdot p_i.$$

Затем находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n_i')^2}{n_i'}$ и сравниваем с c_α , которое находится по таблице 6

критических точек распределения χ^2 с $m = k - 1 - 1 = k - 2$ степенями свободы, если вероятность p определялась по выборке, или с $m = k - 1$ степенями свободы, если вероятность p задана.

Если $\chi_n^2 < c_\alpha$, то гипотезу о биномиальном распределении генеральной совокупности принимаем, иначе отвергаем.

Пример 6. Произведено $n = 100$ серий испытаний. Каждая серия испытаний состояла из $m = 10$ опытов. В каждом из опытов вероятность появления события A равна $p = 0,3$. В итоге получено следующее эмпирическое распределение

x_i	0	1	2	3	4	5
n_i	2	10	27	32	23	6

Требуется при уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что число появлений события A распределено по биномиальному закону.

Так как вероятность $p = 0,3$ задана, то по формуле Бернулли находим вероятности

$$p_0 = C_{10}^0 \cdot 0,3^0 \cdot 0,7^{10} = 0,0282, \quad p_1 = C_{10}^1 \cdot 0,3 \cdot 0,7^9 = 0,1211, \quad p_2 = 0,2335, \quad p_3 = 0,2668,$$

$p_4 = 0,2001, \quad p_5 = 0,1029$, затем теоретические частоты и заносим их в таблицу. Поскольку частота $n_0 = 2$ малочисленная, объединим ее с частотой $n_1 = 10$; соответствующие

теоретические частоты также объединим $n_0' + n_1' = 2,82 + 12,11 = 14,93$.

x_i	n_i	n'_i	$\frac{(n_i - n'_i)^2}{n'_i}$
0,1	12	14,93	0,5750
2	27	23,35	0,5706
3	32	26,68	1,0608
4	23	20,01	0,4468
5	6	10,29	1,7886

Теперь находим $\chi_n^2 = \sum_{i=1}^5 \frac{(n_i - n'_i)^2}{n'_i} = 4,4418$, а по таблице 6 по $\alpha = 0,05$ и числу степеней свободы $m = 5 - 1 = 4$ находим $c_\alpha = 9,5$. Так как $\chi_n^2 = 4,4418 < c_\alpha = 9,5$, то гипотезу о биномиальном распределении принимаем.

VI. Нахождение теоретических частот при проверке гипотез о распределении Пуассона.

Пусть задано эмпирическое распределение дискретной случайной величины

x_i	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k

Требуется проверить гипотезу о том, что данная случайная величина распределена по закону Пуассона $P\{\xi = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$, где параметр $\lambda > 0$.

По выборке находим точечную несмещенную оценку параметра $\lambda \approx \bar{x}_s$, а затем

теоретические частоты $n'_i = n \cdot \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$. Затем находим $\chi_n^2 = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$ и сравниваем с c_α ,

которое находится по таблице 6 критических точек распределения χ^2 с $m = k - 1 - 1 = k - 2$ степенями свободы.

Если $\chi_n^2 < c_\alpha$, то гипотезу о том, что генеральная совокупность распределена по закону Пуассона, принимаем, иначе отвергаем.

Пример 7. Отдел технического контроля проверил $n = 200$ партий одинаковых изделий и получил следующее эмпирическое распределение (в первой строке указано количество x_i нестандартных изделий в одной партии; во второй строке – количество партий, содержащих x_i нестандартных изделий)

x_i	0	1	2	3	4
n_i	116	56	22	4	2

Требуется при уровне значимости $\alpha = 0,025$ проверить гипотезу о том, что число нестандартных изделий распределено по закону Пуассона.

Сначала найдем выборочную среднюю

$$\bar{x}_e = \frac{1}{n} \sum_{i=1}^5 x_i n_i = \frac{1}{200} (116 \cdot 0 + 56 \cdot 1 + 22 \cdot 2 + 4 \cdot 3 + 2 \cdot 4) = 0,6.$$

Тогда точечная оценка параметра $\lambda \approx 0,6$. Теперь найдем теоретические частоты

$$n_0' = 200 \cdot \frac{0,6^0}{0!} e^{-0,6} = 200 \cdot 0,5488 = 109,76; n_1' = 200 \cdot \frac{0,6}{1!} e^{-0,6} = 200 \cdot 0,3293 = 65,86;$$

$$n_2' = 19,76; n_3' = 3,96; n_4' = 0,60.$$

Объединим малочисленные частоты ($4+2=6$) и соответствующие им теоретические частоты ($3,96+0,60=4,56$). Результат представим в виде таблицы

x_i	n_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
0	116	109,76	0,3548
1	56	65,86	1,4762
2	22	19,76	0,2539
3;4	6	4,56	0,4547

Теперь находим $\chi_n^2 = \sum_{i=1}^4 \frac{(n_i - n_i')^2}{n_i'} = 2,5396$, а по таблице 6 по $\alpha = 0,025$ и числу степеней

свободы $m = 4 - 1 - 1 = 2$ находим $c_\alpha = 7,4$. Так как $\chi_n^2 = 2,5396 < c_\alpha = 7,4$, то гипотезу о распределении Пуассона принимаем.

Задания.

1. Из генеральной совокупности извлечена выборка

$(x_i; x_{i+1})$	1	2	3	4	5	6	7	8
n_i	259	167	109	74	70	47	40	34

Используя критерий Колмогорова, при уровне значимости $\alpha = 0,01$, проверить гипотезу о показательном распределении генеральной совокупности.

2. Используя критерий Пирсона, при уровне значимости $\alpha = 0,05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки

x_i	-5	-3	-1	1	3	5	7
n_i	8	22	44	57	40	23	6

3. Используя критерий Пирсона, при уровне значимости $\alpha = 0,05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности с эмпирическим распределением выборки

$(x_i; x_{i+1})$	(-45;-30)	(-30;-15)	(-15;0)	(0;15)	(15;30)	(30;45)	(45;60)
n_i	22	45	60	76	42	34	21

4. В итоге испытаний $n = 800$ элементов на время безотказной работы, получено следующее эмпирическое распределение выборки

$(x_i; x_{i+1})$	(0; 10)	(10; 20)	(20; 30)	(30; 40)	(40; 50)	(50; 60)	(60; 70)
n_i	305	215	150	80	30	15	5

Здесь в первой строке указаны интервалы времени в часах, а во второй – количество отказавших элементов в соответствующем интервале. Используя критерий Пирсона при уровне значимости $\alpha = 0,01$ проверить гипотезу о том, что время безотказной работы элементов распределено по показательному закону.

5. Часы, выставленные в витринах часовых магазинов, показывают случайное время. Выдвигается естественная гипотеза, что показания этих часов распределены равномерно в интервале (0; 12). Наблюдения над 500 часами в различных магазинах дали следующую выборку, где весь интервал разбит на 12 часовых интервалов

$(x_{i-1}; x_i)$	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12
n_i	36	47	41	47	49	45	32	37	40	41	37	48

Используя критерий Пирсона проверить гипотезу о равномерном распределении при уровне значимости $\alpha = 0,05$.

6. В библиотеке случайно отобрано 200 выборок по 5 книг. Регистрировалось число поврежденных книг. В итоге получено следующее эмпирическое распределение выборки

x_i	0	1	2	3	4
n_i	72	77	34	14	3

Здесь в первой строке указано число поврежденных книг в одной выборке, а во второй количество выборок, содержащих x_i поврежденных книг. Требуется, используя критерий Пирсона, при уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что случайная величина – число поврежденных книг распределена по биномиальному закону.

7. В итоге проверки $n = 250$ партий одинаковых изделий на стандартность получено следующее эмпирическое распределение выборки

x_i	0	1	2	3	4
n_i	136	86	22	4	2

Здесь в первой строке указано количество x_i нестандартных изделий в одной партии, а во второй – количество партий, содержащих x_i нестандартных изделий. Используя критерий Пирсона при уровне значимости $\alpha = 0,01$ проверить гипотезу о том, что случайная величина – число нестандартных изделий распределено по закону Пуассона.

§ 13. Элементы теории корреляции.

В курсе «Теория вероятности» был введен коэффициент корреляции двух случайных величин ξ и η

$$\rho(\xi, \eta) = M \left(\frac{\xi - M\xi}{\sqrt{D\xi}} \cdot \frac{\eta - M\eta}{\sqrt{D\eta}} \right)$$

и получены его свойства:

- 1) если случайные величины ξ и η независимы, то коэффициент корреляции $\rho(\xi, \eta) = 0$;
- 2) $|\rho(\xi, \eta)| \leq 1$;
- 3) если случайные величины ξ и η зависят линейно $\eta = \alpha\xi + \beta$, то коэффициент корреляции максимален по модулю $|\rho(\xi, \eta)| = 1$.

В случае, когда коэффициент корреляции $\rho(\xi, \eta) \neq 0$, случайные величины ξ и η называются коррелированными. В этом случае между ними существует определенная зависимость $\eta = g(\xi)$. Функцию $g(\xi)$ называют функцией регрессии η на ξ . Там же был рассмотрен случай линейной регрессии $g(\xi) \approx \alpha\xi + \beta$ и при помощи метода наименьших квадратов получено уравнение прямой линии среднеквадратичной регрессии η на ξ

$$\eta = \rho(\xi, \eta) \frac{\sqrt{D\eta}}{\sqrt{D\xi}} (\xi - M\xi) + M\eta \quad (1)$$

Рассмотрим здесь этот вопрос с точки зрения статистики. Во многих задачах математической статистики требуется установить и оценить зависимость изучаемой случайной величины ξ от одной или нескольких других случайных величин.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. В этом случае статистическую зависимость называют корреляционной.

Если мы рассмотрим условное математическое ожидание $M(\eta | \xi = x)$, то эту функцию от x называют функцией регрессии η на ξ . Аналогично, условное математическое ожидание $M(\xi | \eta = y)$ называют функцией регрессии ξ на η .

Например, пусть ξ урожай зерна, а η количество внесенных удобрений. С одинаковых по площади участков земли при равном количестве внесенных удобрений снимают различные урожаи, то есть ξ не является функцией от η . Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, средний урожай является функцией от внесенных удобрений, то есть ξ связано с η корреляционной зависимостью.

Рассмотрим для определенности выборку для двух случайных величин ξ и η :

$$(\xi_1; \eta_1), (\xi_2; \eta_2), \dots, (\xi_n; \eta_n).$$

Здесь значения ξ_i и η_k могут повторяться. Поэтому обозначим через $x_1, x_2, \dots, x_t$ и через $y_1, y_2, \dots, y_s$ различные значения этих величин. Пусть значение x_1 встречается n_{x_1} раз, значение x_2 встречается n_{x_2} раз и так далее; значение y_1 встречается n_{y_1} раз, значение y_2 встречается n_{y_2} раз и так далее. При этом $\sum_{i=1}^t n_{x_i} = \sum_{k=1}^s n_{y_k} = n$. Пара $(x_i; y_k)$ встречается $n_{x_i y_k}$ раз. Тогда выборку удобно представить в виде так называемой корреляционной таблицы

$\eta \setminus \xi$	x_1	x_2	$\dots$	x_t	n_{y_i}
y_1	$n_{x_1 y_1}$	$n_{x_2 y_1}$	$\dots$	$n_{x_t y_1}$	n_{y_1}
y_2	$n_{x_1 y_2}$	$n_{x_2 y_2}$	$\dots$	$n_{x_t y_2}$	n_{y_2}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
y_s	$n_{x_1 y_s}$	$n_{x_2 y_s}$	$\dots$	$n_{x_t y_s}$	n_{y_s}
n_{x_i}	n_{x_1}	n_{x_2}	$\dots$	n_{x_t}	n

По этой таблице мы можем найти:

выборочные средние $\bar{x}_\epsilon = \frac{1}{n} \sum_{i=1}^t x_i n_{x_i}$ и $\bar{y}_\epsilon = \frac{1}{n} \sum_{k=1}^s y_k n_{y_k}$, которые являются несмещенными точечными оценками для математических ожиданий $M\xi$ и $M\eta$;

выборочные дисперсии $D_x = \frac{1}{n} \sum_{i=1}^t (x_i - \bar{x}_\epsilon)^2 n_{x_i}$ и $D_y = \frac{1}{n} \sum_{k=1}^s (y_k - \bar{y}_\epsilon)^2 n_{y_k}$, которые являются точечными оценками для дисперсий $D\xi$ и $D\eta$.

Кроме того, мы можем найти условные выборочные средние $\bar{y}_{x_i} = \frac{1}{n_{x_i}} \sum_{k=1}^s y_k n_{x_i y_k}$,

являющиеся несмещенными точечными оценками для условных математических ожиданий $M(\eta | \xi = x_i)$. Тогда выборочная функция регрессии η на ξ определяется таблицей

x_i	x_1	x_2	$\dots$	x_t
$\bar{y}_{x_i}$	$\bar{y}_{x_1}$	$\bar{y}_{x_2}$	$\dots$	$\bar{y}_{x_t}$

Аналогично, условные выборочные средние $\bar{x}_{y_k} = \frac{1}{n_{y_k}} \sum_{i=1}^t x_i n_{x_i y_k}$ есть несмещенные

точечные оценки для условных математических ожиданий $M(\xi | \eta = y_k)$ и их значения определяют выборочную функцию регрессии ξ на η .

Для коэффициента корреляции

$$\rho(\xi, \eta) = M\left(\frac{\xi - M\xi}{\sqrt{D\xi}} \cdot \frac{\eta - M\eta}{\sqrt{D\eta}}\right) = \frac{M(\xi \cdot \eta) - M\xi \cdot M\eta}{\sqrt{D\xi} \cdot \sqrt{D\eta}}$$

точной оценкой является выборочный коэффициент корреляции

$$r_{\epsilon} = \frac{\frac{1}{n} \sum_{i=1}^t \sum_{k=1}^s x_i y_k n_{x_i y_k} - \bar{x}_{\epsilon} \cdot \bar{y}_{\epsilon}}{\sqrt{D_x} \sqrt{D_y}}.$$

Поставим теперь задачу найти приближенное значение выборочной функции регрессии η на ξ в виде линейной, квадратичной или какой-либо другой функции. Для этого воспользуемся методом наименьших квадратов.

Найдем уравнение прямой линии выборочной среднеквадратичной регрессии η на ξ

$$\bar{y}_x \approx \alpha x + \beta.$$

Коэффициенты α и β находятся из условия, чтобы сумма квадратов отклонений

$F(\alpha, \beta) = \sum_{i=1}^t [\bar{y}_{x_i} - (\alpha x_i + \beta)]^2$ была наименьшей. Для этого следует найти от функции

$F(\alpha, \beta)$ частные производные по переменным α и β , приравнять их нулю и так далее.

Но мы практически все это проделали ранее в курсе «Теория вероятности» при получении уравнения (1), только надо в нем заменить в нем ξ и η на x и $\bar{y}_x$, а математические ожидания, дисперсии и коэффициент корреляции на их точечные оценки. В результате получим выборочное уравнение прямой линии среднеквадратичной регрессии η на ξ

$$\bar{y}_x = r_{\epsilon} \frac{\sqrt{D_y}}{\sqrt{D_x}} (x - \bar{x}_{\epsilon}) + \bar{y}_{\epsilon}. \quad (2)$$

Совершенно аналогично находится выборочное уравнение прямой линии среднеквадратичной регрессии ξ на η

$$\bar{x}_y = r_{\epsilon} \frac{\sqrt{D_x}}{\sqrt{D_y}} (y - \bar{y}_{\epsilon}) + \bar{x}_{\epsilon}.$$

При построении квадратичной линии выборочной среднеквадратичной регрессии η на ξ $\bar{y}_x \approx \alpha x^2 + \beta x + \gamma$ коэффициенты находятся из условия минимума функции

$F(\alpha, \beta, \gamma) = \sum_{i=1}^t [\bar{y}_{x_i} - (\alpha x_i^2 + \beta x_i + \gamma)]^2$. Так же поступают и при нахождении линий

выборочной среднеквадратичной более высоких порядков.

Пример. Найти выборочное уравнение прямой линии среднеквадратичной регрессии по корреляционной таблице

$\eta \backslash \xi$	20	25	30	35	40	n_y
16	4	6	-	-	-	10
26	-	8	10	-	-	18
36	-	-	32	3	9	44
46	-	-	4	12	6	22
56	-	-	-	1	5	6
n_x	4	14	46	16	20	$n = 100$

Для удобства вычислений перейдем к условным вариантам $u_i = \frac{x_i - 30}{5}$, $v_k = \frac{y_k - 36}{10}$. Тогда корреляционная таблица примет вид

$v \backslash u$	-2	-1	0	1	2	n_v
-2	4	6	-	-	-	10
-1	-	8	10	-	-	18
0	-	-	32	3	9	44
1	-	-	4	12	6	22
2	-	-	-	1	5	6
n_u	4	14	46	16	20	$n = 100$

Найдем сначала выборочные средние

$$\bar{u}_e = \frac{1}{100}(-2 \cdot 4 - 1 \cdot 14 + 0 \cdot 46 + 1 \cdot 16 + 2 \cdot 20) = 0,34, \text{ откуда } \bar{x}_e = 5 \cdot \bar{u}_e + 30 = 31,70$$

$$\bar{v}_e = \frac{1}{100}(-2 \cdot 10 - 1 \cdot 18 + 0 \cdot 44 + 1 \cdot 22 + 2 \cdot 6) = -0,04, \text{ откуда } \bar{y}_e = 10 \cdot \bar{v}_e + 36 = 35,60$$

выборочные дисперсии

$$D_u = \frac{1}{100} (4 \cdot 4 + 1 \cdot 14 + 0 \cdot 46 + 1 \cdot 16 + 4 \cdot 20) - 0,34^2 = 1,14, \text{ откуда } D_x = 25 \cdot D_u = 25,8$$

$$D_v = \frac{1}{100} (4 \cdot 10 + 1 \cdot 18 + 0 \cdot 44 + 1 \cdot 22 + 4 \cdot 6) - 0,04^2 = 1,04, \text{ откуда } D_y = 100 \cdot D_v = 104$$

и выборочный коэффициент корреляции, для которого в условных вариантах справедлива

формула
$$r_{\epsilon} = \frac{\frac{1}{n} \sum_{i=1}^t \sum_{k=1}^s u_i v_k n_{u_i v_k} - \bar{u}_{\epsilon} \cdot \bar{v}_{\epsilon}}{\sqrt{D_u} \sqrt{D_v}}.$$
 Вычислим сначала

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^t \sum_{k=1}^s u_i v_k n_{u_i v_k} &= \frac{1}{100} ((-2) \cdot (-2) \cdot 4 + (-2) \cdot (-1) \cdot 6 + (-1) \cdot (-1) \cdot 8 + (-1) \cdot 0 \cdot 10 + 0 \cdot 0 \cdot 32 + \\ &+ 1 \cdot 1 \cdot 12 + 1 \cdot 2 \cdot 6 + 2 \cdot 1 \cdot 1 + 2 \cdot 2 \cdot 5) = 0,82, \end{aligned}$$

тогда $r_{\epsilon} = \frac{0,82 - 0,34 \cdot (-0,04)}{\sqrt{1,14} \sqrt{1,04}} = 0,76.$ Подставим найденные значения в выборочное

уравнение прямой линии среднеквадратичной регрессии η на ξ

$$\bar{y}_x = 0,76 \frac{\sqrt{104}}{\sqrt{25,8}} (x - 31,7) + 35,6 \text{ или } \bar{y}_x = 1,45x - 10,36.$$

Найдем из корреляционной таблицы условные выборочные средние

$$\bar{v}_{u_1} = \frac{1}{4} (-2 \cdot 4) = -2; \bar{v}_{u_2} = \frac{1}{16} (-2 \cdot 6 - 1 \cdot 8) = -1,25; \bar{v}_{u_3} = \frac{1}{46} (-1 \cdot 10 + 0 \cdot 32 + 1 \cdot 4) = -0,13;$$

$$\bar{v}_{u_4} = \frac{1}{16} (0 \cdot 3 + 1 \cdot 12 + 2 \cdot 1) = 0,875; \bar{v}_{u_5} = \frac{1}{20} (0 \cdot 9 + 1 \cdot 6 + 2 \cdot 5) = 0,8, \text{ откуда}$$

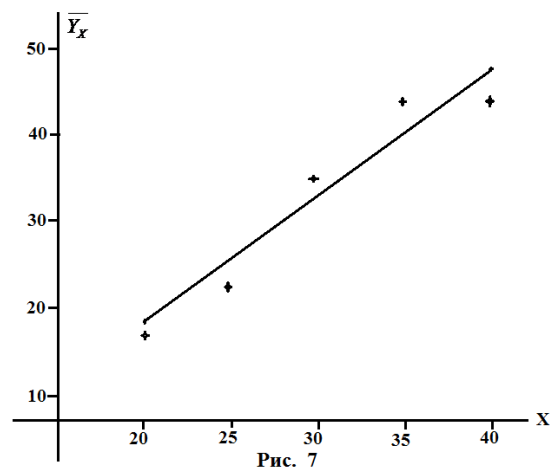
$$\bar{y}_{x_1} = 10 \cdot \bar{v}_{u_1} + 36 = 16; \bar{y}_{x_2} = 10 \cdot \bar{v}_{u_2} + 36 = 23,5; \bar{y}_{x_3} = 10 \cdot \bar{v}_{u_3} + 36 = 34,7;$$

$$\bar{y}_{x_4} = 10 \cdot \bar{v}_{u_4} + 36 = 44,75; \bar{y}_{x_5} = 10 \cdot \bar{v}_{u_5} + 36 = 44.$$

Таким образом, мы нашли значения выборочной функции регрессии η на ξ

x	20	25	30	35	40
$\bar{y}_x$	16	23,5	34,7	44,75	44

Для иллюстрации полученного решения на рисунке 7 изображен график найденной прямой линии выборочной среднеквадратичной регрессии, а также значения выборочной функции регрессии (отмечены крестиками). Из рисунка видно, что прямая линия выборочной среднеквадратичной регрессии достаточно хорошо аппроксимирует функцию регрессии.



Задание.

Найти выборочные уравнения прямых линий среднеквадратичной регрессии η на ξ и ξ на η по корреляционной таблице

$\eta \backslash \xi$	5	10	15	20	25	30	35	n_y
100	-	-	-	-	-	6	1	7
120	-	-	-	-	-	4	2	6
140	-	-	8	10	5	-	-	23
160	3	4	3	-	-	-	-	10
180	2	1	-	1	-	-	-	4
n_x	5	5	11	11	5	10	3	$n = 50$

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

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
165.										
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица 2 значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$.

x	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0909	0948	0987	1026	1103	1064	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2356	2389	2421	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2703	2734	2764	2793	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3437	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3728	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3906	3925	3943	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4648	4656	4664	4671	4678	4686	4692	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4874	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4924	4927	4929	4930	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4958	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3,0	0,4986		3,1	4990	3,2	49931	3,3	49952	3,4	49966
3,5	4998		3,6	4998	3,7	49989	3,8	49993	3,9	49995
4,0	499968									
4,5	499997									
5,0	49999997									

Таблица 3 значений $t_\gamma = t(\gamma, n)$

n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,001	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Таблица 4 значений $q = q(\gamma, n)$

n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Таблица 5 критических точек распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Таблица 6 критических точек распределения χ^2

Число степеней свободы k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,89
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Таблица 7 значений функции распределения Колмогорова $K(x) = 1 + 2 \sum_{m=1}^{\infty} (-1)^m e^{-2m^2 x^2}$

x	0	1	2	3	4	5	6	7	8	9
0,5	0,0361	0,0428	0,0503	0,0585	0,0675	0,0772	0,0876	0,0987	0,1104	0,1228
0,6	0,1357	0,1492	0,1633	0,1778	0,1927	0,2080	0,2236	0,2396	0,2556	0,2722
0,7	0,2888	0,3055	0,3223	0,3391	0,3560	0,3728	0,3896	0,4064	0,4230	0,4395
0,8	0,4559	0,4720	0,4880	0,5038	0,5194	0,5347	0,5497	0,5645	0,5791	0,5933
0,9	0,6073	0,6209	0,6343	0,6473	0,6601	0,6725	0,6846	0,6964	0,7079	0,7191
1,0	0,7300	0,7406	0,7508	0,7608	0,7704	0,7798	0,7889	0,7976	0,8061	0,8143
1,1	0,8223	0,8300	0,8374	0,8445	0,8514	0,8580	0,8644	0,8706	0,8765	0,8823
1,2	0,8878	0,8930	0,8981	0,9030	0,9076	0,9121	0,9164	0,9206	0,9245	0,9283
1,3	0,9319	0,9354	0,9387	0,9418	0,9449	0,9478	0,9505	0,9531	0,9557	0,9580
1,4	0,9603	0,9625	0,9646	0,9665	0,9684	0,9702	0,9718	0,9734	0,9750	0,9764
1,5	0,9778	0,9791	0,9803	0,9815	0,9826	0,9836	0,9846	0,9855	0,9864	0,9873
1,6	0,9880	0,9888	0,9895	0,9901	0,9908	0,9914	0,9919	0,9924	0,9929	0,9934
1,7	0,9938	0,9942	0,9946	0,9950	0,9953	0,9956	0,9959	0,9962	0,9965	0,9967
1,8	0,9969	0,9971	0,9973	0,9975	0,9977	0,9979	0,9980	0,9982	0,9983	0,9984
1,9	0,9985	0,9986	0,9987	0,9988	0,9989	0,9990	0,9991	0,9991	0,9992	0,9993
2,0	0,9993	0,9994	0,9994	0,9995	0,9995	0,9996	0,9996	0,9996	0,9997	0,9997

Ответы.

§ 7: 1) $\lambda \approx \lambda^* = \bar{x}_g = 0,43$. 2) $p \approx p^* = \frac{1}{x_g} = \frac{1}{3,4} = 0,29$.

3) $a \approx a^* = \bar{x}_g = 1,16$; $\sigma \approx \sigma^* = \sqrt{D_g} = \sqrt{0,2372} = 0,487$.

4) $\bar{x}_g = 12,31$; $D_g = 33,7829$; $a \approx a^* = \bar{x}_g - \sqrt{3D_g} = 2,24$; $b \approx b^* = \bar{x}_g + \sqrt{3D_g} = 22,38$.

5) $\bar{x}_g = 27,3333$; $D_g = 122,8889$; $\alpha \approx \alpha^* = \frac{\bar{x}_g}{D_g} = 6,08$; $\beta \approx \beta^* = \frac{\bar{x}_g}{D_g} = 0,22$.

§ 8: 1) $\ln L(\zeta, \lambda) = n\bar{x}_g \ln \lambda - n\lambda - \ln C$; $C = (x_1)^{n_1} \dots (x_k)^{n_k} - \text{const}$; $\lambda \approx \lambda^* = \bar{x}_g = 0,585$.

2) $\ln L(\zeta, p) = n \ln p + n(\bar{x}_g - 1) \ln(1 - p)$; $p \approx p^* = \frac{1}{x_g} = \frac{1}{4,25} = 0,2353$.

3) $\ln L(\zeta, \lambda) = n \ln \lambda - \lambda n \bar{x}_g$; $\lambda \approx \lambda^* = \frac{1}{x_g} = \frac{1}{9,3667} = 0,1068$.

4) $\ln L(\zeta, a) = -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^k (x_i - a)^2 n_i$; $a \approx a^* = \bar{x}_g = 10,366$.

5) $\ln L(\zeta, \sigma) = -n \ln \sigma - n \ln \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^k (x_i - a)^2 n_i$; $\sigma^2 \approx \sigma^{*2} = \frac{1}{n} \sum_{i=1}^k (x_i - a)^2 n_i = 14,25$;

$\sigma^* = 3,775$.

6) $\ln L(\zeta, \beta) = \alpha n \ln \beta - \beta n \bar{x}_g + \ln C$; $C = x_1^{(\alpha-1)n_1} \dots x_k^{(\alpha-1)n_k} [\Gamma(\alpha)]^{-n} - \text{const}$;

$\beta \approx \beta^* = \frac{\alpha}{x_g} = \frac{1,5}{6,16} = 0,2435$.

§ 9: $\ln L(\zeta, a) = -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\xi_i - a)^2$; $\frac{\partial}{\partial a} \ln L(\zeta, a) = \frac{n}{\sigma^2} (\bar{x}_g - a)$ - выполнено

условие существования наиболее эффективной оценки; $D(\bar{x}_g) = \frac{\sigma^2}{n}$ и $I(a) = \frac{n}{\sigma^2}$.

§ 10: 1) $\bar{x}_g = 1000$; $n = 100$; $\delta_\gamma = 1,96$; $992,16 < a < 1007,87$.

2) $\sigma = 20$; $\delta = 15$; $\delta_\gamma = 1,65$; $n = \left(\delta_\gamma \frac{\sigma}{\delta} \right)^2 \approx 5$.

3) $\bar{x}_g = 3,9167$; $s = 1,339$; $t_\gamma = 2,03$; $3,4636 < a < 4,3697$.

4) $s = 0,7467$; $q = 0,5$; $0,37335 < \sigma < 1,1205$.

5) $\bar{x}_g = 1$; $\delta_\gamma = 2,58$; $0,8911 < \lambda < 1,1222$.

6) $\omega = 0,008$; $\delta_\gamma = 3,32$; $0 < p < 0,132$. Обеспечивает.

§ 11: 1) 1) $\eta = -1,8 > c_\alpha = -2,33$. Гипотезу H_0 принимаем. 2) $w = 0,1292$; 3) $n = 150$.

2) 1) $|\eta| = 2,1 < c_\alpha = 2,33$. Гипотезу H_0 принимаем. 2) $w = 0,6808$.

3) $\bar{x}_g = 7,3171$; $s = 2,5927$; $t = 0,7831 < c_\alpha = 2,42$. Гипотезу H_0 принимаем.

4) $\bar{x}_g = 3,1905$; $s = 0,5804$; $|t| = 1,504 < c_\alpha = 2,53$. Гипотезу H_0 принимаем.

- 5) $z = 2,5 > c_\alpha = 1,65$. Гипотезу H_0 отвергаем, математические ожидания разнятся значительно.
- 6) $s^2 = 0,0416$; $\chi^2 = 9,984 < c_\alpha = 36,4$. Гипотезу H_0 принимаем.
- 7) $s^2 = 3,99$; $\chi^2 = 37,9 \notin (8,91; 32,9)$. Гипотезу H_0 отвергаем.
- 8) $\eta = 2,33 > c_\alpha = 1,65$. Гипотезу H_0 отвергаем, новая форма рекламы эффективней старой. Мощность построенного критерия $w = 0,317$.
- 9) $\eta = -0,452$; $|\eta| < c_\alpha = 2,58$. Гипотезу H_0 принимаем.

§ 12: 1) $\lambda \approx \frac{1}{\bar{x}_s} = \frac{1}{3}$; $\sqrt{n}D_n = 3,44 < c_\alpha = 1,63$. Гипотезу о показательном распределении отвергаем.

2) $a \approx \bar{x}_s = 0,92$; $\sigma \approx s = 2,799$; $\chi_n^2 = 1,51 < c_\alpha = 9,5$. Гипотезу о нормальном распределении принимаем.

3) $a \approx \bar{x}_s = 5,35$; $\sigma \approx s = 24,518$; $\chi_n^2 = 7,245 < c_\alpha = 9,5$. Гипотезу о нормальном распределении принимаем.

4) $\bar{x}_s = 17,25$; $\lambda \approx \frac{1}{\bar{x}_s} = 0,058$; $\chi_n^2 = 32,212 > c_\alpha = 15,1$. Гипотезу о показательном распределении отвергаем.

5) $\bar{x}_s = 0,995$; $p \approx \frac{1}{m} \bar{x}_s = 0,199$; $n'_0 = 65,9466$; $n'_1 = 81,9187$; $n'_2 = 40,7037$; $n'_3 = 10,1124$; $n'_4 = 1,2562$; $\chi_n^2 = 2,5641 < c_\alpha = 6$. Гипотезу о биномиальном распределении принимаем.

6) $\lambda \approx \bar{x}_s = 0,6$; $n'_0 = 137,2029$; $n'_1 = 82,3217$; $n'_2 = 24,6965$; $n'_3 = 4,9393$; $n'_4 = 0,7409$; $\chi_n^2 = 0,9663 < c_\alpha = 9,2$. Гипотезу о распределении по закону Пуассона принимаем.

§ 13: $\bar{y}_x = -2,15x + 181,8$; $\bar{x}_y = -0,33y + 65,7$.

Литература.

1. Гнеденко Б.В. Курс теории вероятностей. М: Наука, 1988. – 448 с.
2. Гихман И.И., Скороход А.В., Ядренко М.И. Теория вероятностей и математическая статистика. К: Вища школа, 1978. – 408 с.
3. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. М: Издательство ЛКИ, 2010. – 600 с.
4. Крамер Г. Математические методы статистики. М: Мир, 1975. – 648 с.
5. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М: Высшая школа, 1979. – 400 с.
6. Процеров Ю.С. Теория вероятностей. Учебно-методическое пособие. Одесса, 2015. – 122 с. (Размещено на сайте Научной библиотеки ОНУ: www.lib.onu.edu.ua).