МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

МОСКОВСКИЙ ЭНЕРГЕТИЧЕСКИЙ ИНСТИТУТ
(ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

Д.Г. Тигетов

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Лабораторный практикум по курсу
«Математическая статистика»
для студентов, обучающихся по направлению
«Прикладная математика и информатика»

Москва Издательство МЭИ 2014

УДК 519
Т-394

Подготовлено на кафедре прикладной математики

Рецензенты: доктор технических наук, профессор Ю.А. Горицкий
(Московский энергетический институт);
кандидат физико-математических наук В.С. Штаркман
(ФГУП НИИ «Квант»)

Тигетов Д.Г.
Т-394 Практикум по математической статистике: учебное пособие / Д.Г. Тигетов.– М. : Издательство МЭИ, 2014.– 80 с.; ил.

ISBN

Последовательно излагаются методы решения основных задач математической статистики из разделов точечного и интервального оценивания, проверки гипотез, однофакторного дисперсионного и линейного регрессионного анализов. Дополнительно изгалаются основы построения реализаций выборок различными способами.

Изложение сопровождается заданиями, предназначенными для самостоятельного решения с использованием специализированных математических пакетов или языков программирования высокого уровня.

Пособие рекомендуется студентам старших курсов, обучающихся по направлению «Прикладная математика и информатика».

УДК Т-394

ISBN

Оглавление.

Введение.

В своей деятельности человек часто сталкивается с настолько сложными процессами и явлениями, что их точное описание с помощью детерминированных формальных математических моделей оказывается практически невозможным. В силу отсутствия точных моделей исследователям приходится прибегать к моделям, в которых явление представляется не в полной мере определенным. В рамках таких моделей наблюдаемые (поддающиеся измерению) характеристики явления также приходится считать до некоторой степени неопределенными (случайными).

Случайные в рамках модели наблюдения на практике представляют собой экспериментальные числовые данные, которые позволяют сделать обоснованные выводы о ненаблюдаемых (не поддающихся непосредственному измерению) характеристиках явления. Подобные выводы основываются на специализированных методах обработки наблюдений, которые разрабатываются в разделе математики – математической статистике.

Математическая статистика включает в себя большое количество разнообразных направлений, часть которых посвящена выработке наилучших методов обработки наблюдений в различных статистических задачах. Решение каждой статистической задачи можно условно разбить на три этапа:

1) формализация наблюдений,
2) выбор метода обработки,
3) обработка экспериментальных данных.

Первый этап решения задач связан с выработкой формализованного представления наблюдаемых характеристик явления. В широком круге задач наблюдаемые характеристики представляются совокупностью случайных величин $\xi_1,\ldots, \xi_n$, для которых, например, с точностью до значения некоторого параметра θ известно их совместное распределение $F_\xi(x_1,..., x_n \mid \theta)$. Параметр θ при этом является ненаблюдаемой характеристикой (считается неизвестным) и характеризует внутреннее устройство явления, определить которое (сделать вывод о неизвестном значении параметра θ) предлагается на основе наблюдаемых величин $\xi_1, \ldots, \xi_n$.

Второй этап решения задач заключается в выборе одного из известных или разработке нового метода обработки наблюдений $\xi_1, \ldots, \xi_n$, представляемого функцией $T(\xi_1,..., \xi_n)$, называемой статистикой. Выбор статистики $T(\xi_1,..., \xi_n)$ обычно стараются осуществить наилучшим образом в соответствии с заданными требованиями или критериями качества.

Третий этап решения задачи представляет собой выполнение эксперимента и измерение наблюдаемых характеристик явления, по результатам которого формируется набор числовых данных $x_1, \ldots, x_n$, к которым применяется выбранный метод обработки – вычисляется значение $T(x_1,..., x_n)$, которое и принимается в качестве решения задачи.

В качестве примера рассмотрим бросание некоторой монеты n раз. Сформировать детерминированную модель, описывающую возможные траектории движения монеты и выпадение герба или решки в каждом бросании, весьма затруднительно. Существенно проще оказывается вероятностная модель, представляющая каждое бросание как случайный процесс, который с вероятностью p приводит к выпадению герба, и с обратной вероятностью $1-p$ – к выпадению решки.

При решении задач на первом этапе можно утверждать, что результатам n бросаний соответствует совокупность наблюдений (случайных величин) $(\xi_1,..., \xi_n)$, в которой величины ξ_i являются независимыми и принимают значение 1 с вероятностью p и значение 0 с вероятностью $1-p$.

Совокупность величин $(\xi_1,..., \xi_n)$ является наблюдаемой характеристикой явления, а вероятность выпадения герба p является ненаблюдаемой (не поддающейся измерению) характеристикой, в отношении которой могут формулироваться различные задачи, например:

1) оценить неизвестную вероятность p;

2) построить (случайный) интервал $[p_l, p_r]$, который содержит неизвестное значение p с вероятностью близкой к единице;

3) определить, какое из утверждений является более правдоподобным – «монета является симметричной» ($p=0.5$) или «монета не является симметричной» ($p \neq 0.5$);

4) зная, что монета может быть только двух типов – симметричной ($p=0.5$) и несимметричной ($p=0.7$), – определить к какому типу относится монета.

Применительно к задаче пункта 1) на втором этапе решения в качестве оценки величины p может использоваться статистика $\hat{p}(\xi_1,..., \xi_n)$:

$$\hat{p}(\xi_1,..., \xi_n) = \frac{1}{n}\sum_{i=1}^{n} \xi_i,$$

которая является несмещенной и состоятельной оценкой величины p, и среди всех несмещенных оценок имеет наименьшую дисперсию.

Для вычисления оценки $\hat{p}(\xi_1,..., \xi_n)$ на третьем этапе решения задачи пункта 1) необходимо выполнить (или смоделировать) n бросаний монеты. Предположим, что по результатам $n=5$ бросаний симметричной монеты ($p=0.5$) были получены экспериментальные данные $x_1=1$, $x_2=0$, $x_3=1$, $x_4=0$, $x_5=0$, тогда значение оценки $\hat{p}(\xi_1,..., \xi_n)$ составляет 0.4:

$$\hat{p}(x_1,\ldots,\ x_n) = \frac{1}{5}\sum_{i=1}^{n} x_i = \frac{2}{5} = 0.4\ .$$

Расхождение между значением оценки $\hat{p}(x_1,\ldots,\ x_n) = 0.4$ и точным значением $p = 0.5$ отражает свойство всех статистических выводов, которые в общем случае не являются точными и безошибочными процедурами, и для достижения приемлемой точности или уровня ошибки могут потребовать десятки тысяч наблюдений.

Обработка экспериментальных данных, в количествах исчисляемых тысячами, требует привлечения вычислительной техники и специализированного программного обеспечения, без использования которых невозможно представить себе успешное решение статистических задач.

Приводимые в пособии задачи для самостоятельного решения сформулированы таким образом, чтобы избавить читателя от первого этапа решения, связанного с формализацией наблюдаемых величин $\xi_1,\ \ldots,\ \xi_n$. Основное внимание уделяется второму этапу решения задач, при выполнении которого в центре внимания оказываются методы обработки наблюдений, и третьему этапу, на котором выполняется моделирование явлений, генерация экспериментальных данных и их обработка с использованием языков программирования высокого уровня в сочетании со специализированными математическими пакетами.

Лабораторная работа №1. **Построение реализаций случайных величин.**

1. Основные сведения.

В задачах математической статистики исходными данными являются реализации $(x_1,\ldots,\ x_n)$ совокупности случайных величин $(\xi_1,\ldots,\ \xi_n)$ с заданным распределением и количеством n.

Переменный характер количества n приводит к необходимости располагать общим методом построения реализаций $\{x_i\}_{i=1}^{\infty}$ последовательностей случайных величин $\{\xi_i\}_{i=1}^{\infty}$. Формально, последовательность $\{\xi_i(\omega)\}_{i=1}^{\infty}$ является последовательностью функций, где ω – элемент множества элементарных событий Ω вероятностного пространства, поэтому реализацию $\{x_i\}_{i=1}^{\infty}$ следует рассматривать как числовую последовательность, получаемую подстановкой некоторого фиксированного элемента $\omega^* \in \Omega$ в каждую случайную величину (функцию) последовательности $\{\xi_i(\omega)\}_{i=1}^{\infty}$:

$$x_i = \xi_i(\omega^*),$$
$$\omega^* \in \Omega\ ,\ i = 1,2,\ldots$$

Во многих практических задачах последовательность случайных величин $\{\xi_i\}_{i=1}^{\infty}$, как правило, такова, что все величины ξ_i имеют одинаковое распределение и независимы в совокупности. Для таких последовательностей методы построения реализаций $\{x_i\}_{i=1}^{\infty}$ устроены достаточно просто, поскольку реализация x_i каждой случайной величины ξ_i может быть построена отдельно (независимо) от реализаций x_j других величин ξ_j.

При построении реализаций отдельно взятых случайных величин ξ_i часто используются различные виды преобразований случайных величин [11]. Большое количество случайных величин дискретного, непрерывного и смешанного типа может быть получено в результате преобразования случайной величины с равномерным распределением $R[0,1]$ [11].

Равномерное распределение.

Среди методов построения реализаций $\{x_i\}_{i=1}^{\infty}$ последовательности $\{\xi_i\}_{i=1}^{\infty}$ случайных величин с равномерным распределением широкое распространение получили линейно-конгруэнтные методы. В простейших случаях линейно-конгруэнтного метода строится последовательность целых чисел $\{a_i\}_{i=1}^{\infty}$, в которой каждый следующий элемент a_i вычисляется через предыдущий a_{i-1}, а первый элемент a_1 выбирается произвольным образом из множества вычетов по модулю C:

$$a_i = (A \cdot a_{i-1} + B) \bmod C \tag{1.1}$$

где постоянные A, B и C – натуральные числа.

Пусть последовательность реализаций $\{x_i\}_{i=1}^{\infty}$ определяется на основе последовательности $\{a_i\}_{i=1}^{\infty}$ (1.1) следующим соотношением:

$$x_i = \frac{a_i}{C} \tag{1.2}$$

тогда последовательность $\{x_i\}_{i=1}^{\infty}$ можно считать реализацией последовательности независимых случайных величин $\{\xi_i\}_{i=1}^{\infty}$, имеющих распределение близкое к равномерному распределению $R[0,1]$.

Для получения реализации $\{y_i\}_{i=1}^{\infty}$ последовательности независимых случайных величин $\{\eta_i\}_{i=1}^{\infty}$, имеющих произвольное равномерное распределение $R[a,b]$, достаточно выполнить простое линейное преобразование:

$$y_i = a + (b - a) \cdot x_i \tag{1.3}$$

Из соотношения (1.1) следует, что количество различных числовых значений в последовательности $\{a_i\}_{i=1}^{\infty}$ не может превышать C. Более того, если при некотором номере P имеет место равенство $a_{1+P}=a_1$, то очевидно после номера P элементы последовательности начнут повторяться ($a_{k+P}=a_k$ при любом натуральном k), поэтому величину P называют *периодом*. Из (1.1) очевидным образом следует, что период $P \le C$, поэтому постоянные A, B и C следует выбирать таким образом, чтобы период P был равен C , поскольку, чем больше период P , тем больше различных числовых значений в последовательности $\{a_i\}_{i=1}^{\infty}$, и тем лучше последовательность $\{x_i\}_{i=1}^{\infty}$ соответствуют реализации последовательности $\{\xi_i\}_{i=1}^{\infty}$ независимых случайных величин, имеющих распределение $R[0,1]$.

Бинарное распределение.

Пусть требуется построить реализацию $\{y_i\}_{i=1}^{\infty}$ последовательности независимых бинарных случайных величин $\{\eta_i\}_{i=1}^{\infty}$, принимающих значения 1 с вероятностью p и 0 с вероятностью $1-p$.

Пусть $\{\xi_i\}_{i=1}^{\infty}$ является последовательностью независимых случайных величин, имеющих равномерное распределение $R[0,1]$, тогда случайные величины η_i могут быть выражены через случайные величины ξ_i следующим образом:

$$\eta_i=\begin{cases}1 & \xi_i<p\\ 0 & \xi_i\ge p\end{cases}. \tag{1.4}$$

Действительно,

$$P\{\eta_i=1\}=P\{\xi_i<p\}=p\ ,$$

$$P\{\eta_i=0\}=P\{\xi_i\ge p\}=1-P\{\xi_i<p\}=1-p\ ,$$

и независимость случайных величин η_i следует из независимости случайных величин ξ_i.

Из (1.4) следует, что реализация $\{y_i\}_{i=1}^{\infty}$ может быть построена на основе реализации $\{x_i\}_{i=1}^{\infty}$ (1.2) последовательности $\{\xi_i\}_{i=1}^{\infty}$:

$$y_i=\begin{cases}1 & x_i<p\\ 0 & x_i\ge p\end{cases}. \tag{1.5}$$

Биномиальное распределение.

Пусть требуется построить реализацию $\{z_i\}_{i=1}^{\infty}$ последовательности независимых случайных величин $\{\zeta_i\}_{i=1}^{\infty}$, имеющих биномиальное распределение $Bi\,(n,p)$.

Пусть $\{\eta_i\}_{i=1}^{\infty}$ – последовательность независимых бинарных случайных величин, принимающих значения 1 с вероятностью p и 0 с вероятностью

$1-p$, тогда случайные величины ζ_i могут быть выражены через случайные величины η_j:

$$\zeta_i = \sum_{j=(i-1)\cdot n+1}^{i\cdot n} \eta_j \tag{1.6}$$

Действительно, каждая случайная величина ζ_i по определению является количеством успехов в n независимых испытаниях, в каждом из которых успех наступает с вероятностью p, поэтому для вычисления значения ζ_i достаточно подсчитать количество единиц среди n независимых бинарных случайных величин $\eta_{(i-1)\cdot n+1}$, $\eta_{(i-1)\cdot n+2}$, …, $\eta_{(i-1)\cdot n+n}$, принимающих значение 1 с вероятностью p и 0 с вероятностью $1-p$.

Из (1.6) следует, что если $\{y_i\}_{i=1}^{\infty}$ – реализация $\{\eta_i\}_{i=1}^{\infty}$ (1.5), тогда каждый элемент последовательности $\{z_i\}_{i=1}^{\infty}$:

$$z_i = \sum_{j=(i-1)\cdot n+1}^{i\cdot n} y_j \tag{1.7}$$

Одномерное нормальное распределение.

В тех случаях, когда требуется построить реализацию $\{y_i\}_{i=1}^{\infty}$ последовательности независимых случайных величин $\{\eta_i\}_{i=1}^{\infty}$, имеющих стандартное нормальное распределение $N(0,1)$ достаточно использовать, например, следующее соотношение:

$$\eta_i = \frac{\frac{1}{n}\sum_{j=(i-1)\cdot n+1}^{i\cdot n} \xi_j - \frac{1}{2}}{\sqrt{\frac{1}{12}\cdot\frac{1}{n}}} \tag{1.8}$$

где $\{\xi_i\}_{i=1}^{\infty}$ – последовательность независимых случайных величин, имеющих равномерное распределение $R[0,1]$.

Преобразование (1.8) основано на центральной предельной теореме, согласно которой сумма независимых случайных величин ξ_j, имеющих одинаковое распределение и конечные математическое ожидание и дисперсию, имеет распределение близкое к нормальному, в этом случае нормированная сумма случайных величин ξ_j будет иметь распределение, приближено совпадающее со стандартным нормальным распределением $N(0,1)$.

В качестве n в (1.8) для большинства практических задач достаточно взять величину порядка 6 [5], в этом случае функция распределения величин η_i (1.8) совпадает с функцией распределения стандартного нормального распределения $N(0,1)$ с точностью 0.002.

Из (1.8) следует, что элементы последовательности $\{y_i\}_{i=1}^{\infty}$ определяются соотношением:

$$y_i = \frac{\frac{1}{n}\sum_{j=(i-1)\cdot n+1}^{i\cdot n} x_j - \frac{1}{2}}{\sqrt{\frac{1}{12}\cdot\frac{1}{n}}} \qquad (1.9)$$

где $\{x_i\}_{i=1}^{\infty}$ – реализация (1.2) последовательности $\{\xi_i\}_{i=1}^{\infty}$.

При построении реализации $\{z_i\}_{i=1}^{\infty}$ последовательности независимых случайных величин $\{\zeta_i\}_{i=1}^{\infty}$, имеющих произвольное нормальное распределение $N(m,\sigma^2)$, может быть использована уже построенная реализация $\{y_i\}_{i=1}^{\infty}$ (1.9) последовательности $\{\eta_i\}_{i=1}^{\infty}$.

Действительно, легко видеть, что каждая случайная величина ζ_i может быть получена из случайной величины η_i путем простого линейного преобразования:

$$\zeta_i = m + \sigma\cdot\eta_i \qquad (1.10)$$

Откуда непосредственно следует в точности такое же выражение для числовых величин z_i:

$$z_i = m + \sigma\cdot y_i \qquad (1.11)$$

Многомерное нормальное распределение.

Пусть требуется построить реализацию $\{Y_i\}_{i=1}^{\infty}$ последовательности независимых векторных случайных величин $\{\Theta_i\}_{i=1}^{\infty}$, имеющих k-мерное нормальное распределение $N(\overline{0}_k, E_k)$, где $\overline{0}_k$ – нулевой вектор размера k и E_k – единичная матрица порядка $k\times k$.

При построении реализации $\{Y_i\}_{i=1}^{\infty}$ может быть использована реализация $\{y_i\}_{i=1}^{\infty}$ последовательности величи $\{\eta_i\}_{i=1}^{\infty}$, имеющих стандартное нормальное распределение $N(0,1)$, поскольку каждую векторную величину Θ_i можно представить с помощью k случайных величин η_j:

$$\Theta_i = \begin{pmatrix} \eta_{(i-1)\cdot n+1} \\ \eta_{(i-1)\cdot n+2} \\ \dots \\ \eta_{(i-1)\cdot n+k} \end{pmatrix},$$

откуда непосредственно следует аналогичное выражение для числовых векторов Y_i:

$$Y_i = \begin{pmatrix} y_{(i-1)\cdot n+1} \\ y_{(i-1)\cdot n+2} \\ \dots \\ y_{(i-1)\cdot n+k} \end{pmatrix}, \tag{1.12}$$

где $\{y_i\}_{i=1}^{\infty}$ – реализация (1.9) последовательности $\{\eta_i\}_{i=1}^{\infty}$.

Пусть теперь требуется построить реализацию $\{Z_i\}_{i=1}^{\infty}$ последовательности независимых векторных случайных величин $\{\Psi_i\}_{i=1}^{\infty}$, имеющих произвольное k-мерное нормальное распределение $N(\overline{m}, K)$, где $\overline{m} = (m_1, ..., m_k)$ – произвольный вектор математических ожиданий и K – произвольная дисперсионная матрица.

Заметим, что каждая случайная величина Ψ_i может быть получена путем выполнения линейного преобразования величин Θ_i, имеющих k-мерное нормальное распределение $N(\overline{0}_k, E_k)$:

$$\Psi_i = \overline{m} + L\Theta_i \tag{1.13}$$

где нижнетреугольная матрица L определяется разложением:

$$K = LL^T. \tag{1.14}$$

Действительно, если Θ_i имеет нормальное распределение, то и Ψ_i имеет нормальное распределение, поскольку линейное преобразование не изменяет типа распределения. Остается лишь вычислить параметры распределения Ψ_i – математическое ожидание $M[\Psi_i]$:

$$M[\Psi_i] = M[\overline{m} + L\Theta_i] = \overline{m} + LM[\Theta_i] = \overline{m} + L\overline{0}_k = \overline{m},$$

и дисперсионную матрицу $D[\Psi_i]$:

$$D[\Psi_i] = M[(\Psi_i - M[\Psi_i])(\Psi_i - M[\Psi_i])^T] = M[(\overline{m} + L\Theta_i - \overline{m})(\overline{m} + L\Theta_i - \overline{m})^T] =$$

$$= M[(L\Theta_i)(L\Theta_i)^T] = M[L\Theta_i\Theta_i^T L^T] = LM[\Theta_i\Theta_i^T]L^T = LD[\Theta_i]L^T = LE_k L^T = LL^T.$$

Из последнего в равенства частности следует (1.14) и способ определения матрицы L: $K = D[\Psi_i] = LL^T$.

Из (1.13) следует выражение для векторов Z_i через векторы Y_i (1.12):

$$Z_i = \overline{m} + LY_i \tag{1.15}$$

Распределения общего вида.

При построении реализаций $\{y_i\}_{i=1}^{\infty}$ последовательностей независимых случайных величин $\{\eta_i\}_{i=1}^{\infty}$, имеющих заданную функцию распределения

$F_\eta(y)$, используется преобразование независимых случайных величин $\{\xi_i\}_{i=1}^{\infty}$, имеющих равномерное распределение $R[0,1]$:

$$\eta_i = \inf\{\, t : F_\eta(t) > \xi_i \}\,, \qquad (1.16)$$

Если η_i – случайные величины непрерывного типа и функция распределения $F_\eta(y)$ имеет обратную функцию $F_\eta^{-1}(x)$ на всей числовой оси (рисунок 1.1а), то соотношение (1.16) существенно упрощается:

$$\eta_i = F_\eta^{-1}(\xi_i)\,, \qquad (1.17)$$

Легко проверить, что равенство (1.17) действительно задает случайные величины η_i с нужной функцией распределения $F_\eta(y)$:

$$P\{\eta < y\} = P\{F_\eta^{-1}(\xi_i) < y\} = P\{F_\eta(F_\eta^{-1}(\xi_i)) < F_\eta(y)\} =$$
$$= P\{\xi_i < F_\eta(y)\} = F_\xi(F_\eta(y)) = F_\eta(y)\,,$$

где $F_\xi(x)$ – функция распределения случайных величин ξ_i, имеющих равномерное распределение $R[0,1]$:

$$F_\xi(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1\,, \\ 1, & 1 < x \end{cases} \qquad (1.18)$$

Из (1.17) следует выражение для числовых величин y_i:

$$y_i = F_\eta^{-1}(x_i)\,, \qquad (1.19)$$

где $\{x_i\}_{i=1}^{\infty}$ – реализация (1.2) последовательности $\{\xi_i\}_{i=1}^{\infty}$.

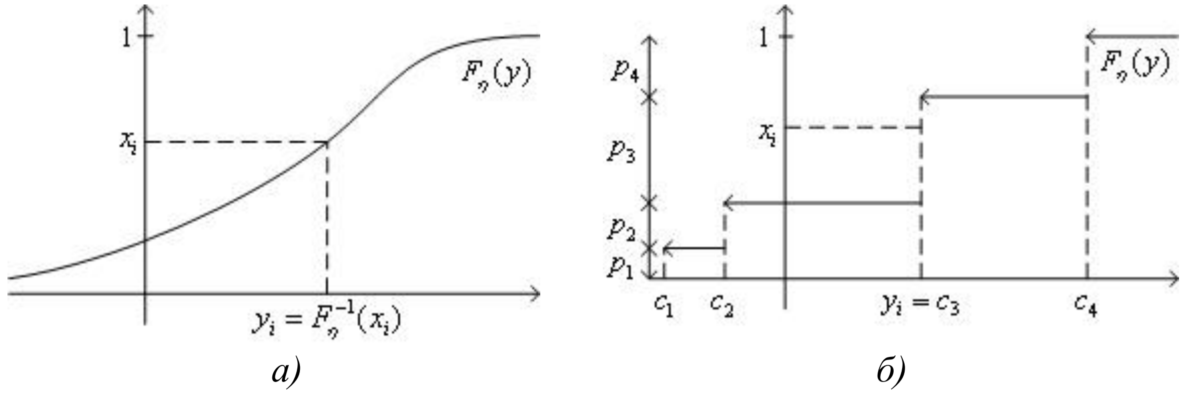


Рисунок 1.1 Построение реализаций случайных величин: а) непрерывной, б) дискретной.

Если η_i – случайные величины дискретного типа, принимающие конечное число значений c_1, c_2, …, c_k с вероятностями p_1, p_2, …, p_k (рисунок 1.1б), то соотношение (1.16) преобразуется к виду:

$$\eta_i = \begin{cases} c_1, & 0 \le \xi_i < p_1 \\ c_2, & p_1 \le \xi_i < p_2 + p_1 \\ \dots & \dots \\ c_k, & \sum_{j=1}^{k-1} p_j \le \xi_i \le \sum_{j=1}^{k} p_j \end{cases}, \tag{1.20}$$

Легко видеть, что соотношение (1.20) задает случайные величины η_i с заданным распределением $R[0,1]$:

$$P\{\eta_i = c_1\} = P\{0 \le \xi_i < p_1\} = P\{\xi_i < p_1\} - P\{\xi_i < 0\} = F_\xi(p_1) - F_\xi(0) = p_1,$$

$$P\{\eta_i = c_m\} = P\left\{\sum_{j=1}^{m-1} p_j \le \xi_i < \sum_{j=1}^{m} p_j\right\} = P\left\{\xi_i < \sum_{j=1}^{m} p_j\right\} - P\left\{\xi_i < \sum_{j=1}^{m-1} p_j\right\} =$$

$$= F_\xi\left(\sum_{j=1}^{m} p_j\right) - F_\xi\left(\sum_{j=1}^{m-1} p_j\right) = \sum_{j=1}^{m} p_j - \sum_{j=1}^{m-1} p_j = p_m, \; m = 2,\dots, k-1,$$

$$P\{\eta_i = c_k\} = P\left\{\sum_{j=1}^{k-1} p_j \le \xi_i \le \sum_{j=1}^{k} p_j\right\} = P\left\{\xi_i \le \sum_{j=1}^{k} p_j\right\} - P\left\{\xi_i < \sum_{j=1}^{k-1} p_j\right\} =$$

$$= 1 - F_\xi\left(\sum_{j=1}^{k-1} p_j\right) = 1 - \sum_{j=1}^{k-1} p_j = p_k,$$

где $F_\xi(x)$ – функция распределения (1.18) случайных величин ξ_i, имеющих равномерное распределение.

Из (1.20) следует правило вычисления числовых величин y_i:

$$y_i = \begin{cases} c_1, & 0 \le x_i < p_1 \\ c_2, & p_1 \le x_i < p_2 + p_1 \\ \dots & \dots \\ c_k, & \sum_{j=1}^{k-1} p_j \le x_i \le \sum_{j=1}^{k} p_j \end{cases}, \tag{1.21}$$

где $\{x_i\}_{i=1}^{\infty}$ – реализация (1.2) последовательности $\{\xi_i\}_{i=1}^{\infty}$.

При вычислении реализации y_i в соответствии с (1.21) может быть использован простой алгоритм, задаваемый следующей программой:

$$\begin{aligned} & m = 1 \\ & \textit{пока} \quad \left(x_i \ge \sum_{j=1}^{m} p_j\right) \quad \textit{и} \quad \left(\sum_{j=1}^{m} p_j < 1\right) \\ & \qquad m = m + 1 \\ & y_i = c_m \end{aligned} \tag{1.22}$$

Если η_i – случайные величины дискретного типа, принимающие счетное число значений c_1, c_2, … с вероятностями p_1, p_2, …, то для вычисления величин y_i аналогично (1.21) также может быть использована программа (1.22), поскольку для всякого $x_i \in [0,1)$ найдется такой номер m, для которого $x_i < \sum_{j=1}^{m} p_j$, и следовательно $y_i = c_m$.

Если η_i – случайные величины смешанного типа, то для построения реализаций y_i необходимо комбинировать соотношения (1.19), (1.21) и программу (1.22): если для значения x_i определена обратная функция $F_\eta^{-1}(x)$, то следует присвоить элементу y_i значение $F_\eta^{-1}(x_i)$, в противном случае, если значение x_i попадает в диапазон разрыва значений функции распределения в некоторой точке c_m ($F(c_m) \le x \le \lim_{y \to c_m + 0} F(y)$), то следует присвоить элементу y_i значение c_m.

Полиномиальное распределение.

Пусть требуется построить реализацию $\{Z_i\}_{i=1}^{\infty}$ последовательности независимых векторных случайных величин $\{\Psi_i\}_{i=1}^{\infty}$, имеющих полиномиальное распределение $\Pi(p_1,..., p_k; n)$.

В векторных случаных величинах $\Psi_i = (\psi_1^{(i)},..., \psi_k^{(i)})$ каждая величина $\psi_j^{(i)}$ определяется количеством появлений некоторого события A_j, имеющего вероятность p_j, в серии из n независимых испытаний. Пусть $\{\eta_i\}_{i=1}^{\infty}$ – последовательности независимых случайных величин дискретного типа, принимающих значения c_1, c_2, …, c_k с вероятностями p_1, p_2, …, p_k, тогда для определения величин Ψ_i необходимо разбить последовательность $\{\eta_i\}_{i=1}^{\infty}$ на непересекающиеся группы по n величин, и в каждой группе подсчитать количество значений c_j:

$$\psi_j^{(i)} = \sum_{l=(i-1)\cdot n+1}^{i\cdot n} I(\eta_l, c_j),$$
$$I(y,c) = \begin{cases} 1, & y = c \\ 0, & y \ne c \end{cases} \qquad (1.23)$$

Из (1.23) следует правило вычисления числовых векторов $Z_i = (z_1^{(i)},..., z_k^{(i)})$ – реализаций соответствующих случайных векторов $\Psi_i = (\psi_1^{(i)},..., \psi_k^{(i)})$:

$$z_j^{(i)} = \sum_{l=(i-1)\cdot n+1}^{i\cdot n} I(y_l, c_j),$$
$$I(y,c) = \begin{cases} 1, & y = c \\ 0, & y \neq c \end{cases} \qquad (1.24)$$

где $\{y_i\}_{i=1}^{\infty}$ – реализация (1.21) последовательности $\{\eta_i\}_{i=1}^{\infty}$.

2. Задачи.

Задача 1.

Напишите функцию, которая возвращает заданное количество элементов последовательности $\{a_i\}_{i=1}^{\infty}$, используя линейно конгруэнтный метод (1.1), и получите первые 10 элементов последовательности (задав произвольным образом элемент a_1). Определите период получаемой последовательности P. Выберите числа A, B и C таким образом, чтобы период P совпадал с числом C и по крайней мере составлял величину порядка $10^6 \div 10^7$.

Указание: в качестве чисел A, B и C рекомендуется выбирать простые числа (или взаимно простые числа).

Задача 2.

Напишите функции, которые возвращают заданное количество l элементов $x_1, \ldots, x_l$ последовательности $\{x_i\}_{i=1}^{\infty}$ – реализации последовательности независимых случайных величин $\{\xi_i\}_{i=1}^{\infty}$, имеющих:

1) равномерное распределение $R[a,b]$;

2) биномиальное распределение $Bi(n,p)$;

3) нормальное распределение $N(m,\sigma^2)$ (предусмотреть возможность задания количества суммируемых величин n в качестве параметра);

4) показательное распределение $E(a)$;

5) распределение Пуассона $Po(\lambda)$;

6) функцию распределения $F(x)$:

$$F(x) = \begin{cases} 0.1 \cdot e^{x+0.1}, & -\infty < x \le -0.1 \\ 0.5 \cdot \dfrac{x+0.1}{0.4} + 0.3, & -0.1 < x \le 0.3 \\ 1 - 0.2 \cdot e^{-(x-0.3)}, & 0.3 < x < \infty \end{cases};$$

а также величин последовательности $\{\xi_i\}_{i=1}^{\infty}$,

7) принимающих значения $c_1, \ldots, c_m$ с вероятностями $p_1, \ldots, p_m$ соответственно.

Для каждой функции пунктов 1) – 7) получите первые $l = 10$ элементов.

Задача 3.

Составьте функцию, которая возвращает матрицу порядка $k \times l$, в столбцах которой располагаются реализации z_i (1.15) независимых векторных случайных величин Ψ_i (1.13), имеющих многомерное нормальное распределение $N(\overline{m}, K)$, где $\overline{m} = (m_1,..., m_k)$ – произвольный вектор математических ожиданий и K – произвольная дисперсионная матрица. С помощью составленной функции сформируйте матрицу реализаций для $k = 5$ и $l = 10$, задав произвольным образом вектор $\overline{m}$ и матрицу K.

Задача 4.

Напишите функцию, которая возвращает матрицу порядка $k \times l$, в столбцах которой располагаются реализации z_i (1.24) независимых случайных векторов Ψ_i (1.23), имеющих полиномиальное распределение $\Pi(p_1,..., p_k; n)$, и сформируйте матрицу реализаций для произвольного вектора вероятностей $p = (p_1,..., p_k)$ с $k = 4$ и $l = 10$.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: постоянные A, B и C, текст функции получения элементов последовательности, полученные первые 10 элементов последовательности.

Задача 2: тексты функций и полученные 10 элементов последовательностей реализаций для всех 7 пунктов.

Задача 3: текст функции получения матрицы реализаций и матрицу реализаций, использованные вектор $\overline{m}$ и матрицу K.

Задача 4: текст функции получения матрицы реализаций и матрицу реализаций, использованный вектор p и n.

4. Вопросы и задания.

Опишите методы построения реализаций последовательностей случайных величин, имеющих равномерное, бинарное, биномиальное, нормальное (в том числе многомерное) и полиномиальное распределения, а также случайных величин дискретного, непрерывного и смешанного типа.

Лабораторная работа №2. **Основные характеристики выборок.**

1. Основные сведения.

Реализации выборки.

Выборкой называется вектор случайных величин $(\xi_1,..., \xi_n)$, в котором все случайные величины независимы в совокупности и имеют одинаковое распределение, число n при этом называют *объемом выборки* [10, 13].

Под *реализацией выборки* $(\xi_1,..., \xi_n)$ понимается любой числовой вектор $(x_1,..., x_n)$, полученный как значение векторной функции $(\xi_1(\omega),..., \xi_n(\omega))$ при некотором элементарном событии ω^* из множества всех элементарных событий Ω:

$$(x_1,..., x_n) = (\xi_1(\omega^*),..., \xi_n(\omega^*)),\ \omega^* \in \Omega .$$

Перебирая все возможные элементарные события множества Ω, можно получить *множество реализаций выборки* X:

$$X = \{(x_1,..., x_n) : (x_1,..., x_n) = (\xi_1(\omega),..., \xi_n(\omega)),\ \omega \in \Omega\} . \qquad (2.1)$$

При построении реализаций $(x_1^*,..., x_n^*)$ широко используются методы построения реализаций последовательностей случайных величин, рассмотренные в лабораторной работе №1. Пусть случайные величины выборки ξ_i имеют функцию распределения $F_\xi(x)$, $\{\hat{\xi}_i\}_{i=1}^{\infty}$ – последовательность независимых в совокупности случайных величин, имеющих такую же функцию распределения $F_\xi(x)$, и $\{\hat{x}_i\}_{i=1}^{\infty}$ – одна из реализаций последовательности $\{\hat{\xi}_i\}_{i=1}^{\infty}$. В этом случае, первые n элементов $(\hat{x}_1,..., \hat{x}_n)$ можно считать одной из возможных реализаций выборки $(\xi_1,..., \xi_n)$, полученной при некотором элементарном событии $\omega^* \in \Omega$:

$$(\hat{x}_1,..., \hat{x}_n) = (\xi_1(\omega^*),..., \xi_n(\omega^*)) .$$

Совершенно аналогично следующие n величин $(\hat{x}_{n+1},..., \hat{x}_{n+n})$ можно считать некоторой другой реализацией выборки $(\xi_1,..., \xi_n)$, полученной при другом элементарном событии $\tilde{\omega} \in \Omega$:

$$(\hat{x}_{n+1},..., \hat{x}_{n+n}) = (\xi_1(\tilde{\omega}),..., \xi_n(\tilde{\omega})) .$$

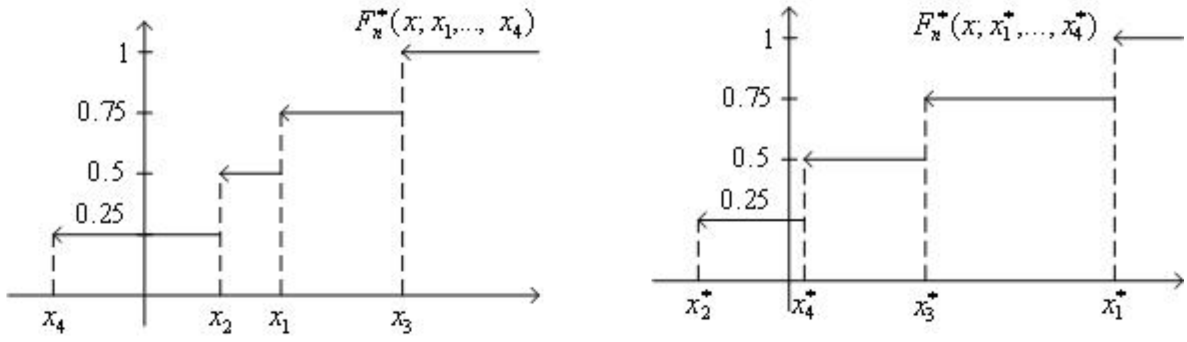
Продолжая подобным образом, из последовательности $\{\hat{x}_i\}_{i=1}^{\infty}$ можно сформировать необходимое количество реализаций выборок вида $(\hat{x}_{k\cdot n+1},..., \hat{x}_{k\cdot n+n})$, где $k = 0,1,2,...$ (при условии что период используемой последовательности оказывается достаточно большим).

Эмпирическая функция распределения.

Одной из основных характеристик выборок $(\xi_1,..., \xi_n)$ является *эмпирическая функция распределения* $F_n^*(x; \xi_1,..., \xi_n)$ [5, 6, 7], которая для каждого аргумента x определяется как отношение случайной величины $\mu_n(x; \xi_1,..., \xi_n)$, равной случайному количеству величин выборки ξ_i меньших x, к объему выборки n:

$$F_n^*(x; \xi_1,..., \xi_n) = \frac{\mu_n(x; \xi_1,..., \xi_n)}{n} . \qquad (2.2)$$

Эмпирическая функция распределения $F_n^*(x;\xi_1,..., \xi_n)$ является случайной функцией, множество реализаций которой состоит из функций $F_n^*(x;x_1,..., x_n)$, получаемых при подстановке в (2.2) вместо случайных величин ξ_i числовых значений x_i из реализаций выборки $(x_1,..., x_n) \in X$ (2.1). Каждая реализация $F_n^*(x;x_1,..., x_n)$ является кусочно-постоянной функцией (рисунки 1.2), имеющей разрывы в точках x_i величиной $\frac{1}{n}$.



а) б)

Рисунок 1.2. Примеры реализаций эмпирической функции распределения.

Основное свойство реализаций эмпирической функции распределения заключается в следующей теореме [2].

Теорема (Гливенко)

Пусть $F_n^*(x;\xi_1,..., \xi_n)$ является эмпирической функцией распределения, построенной по выборке $(\xi_1,..., \xi_n)$ из распределения $F_\xi(x)$, тогда последовательность случайных величин $\sup\limits_{-\infty<x<\infty} \left|F_n^*(x;\xi_1,..., \xi_n) - F_\xi(x)\right|$ сходится к нулю с вероятностью 1 («почти наверное») при $n \to \infty$:

$$\sup_{-\infty<x<\infty} \left|F_n^*(x;\xi_1,..., \xi_n) - F_\xi(x)\right| \xrightarrow[n\to\infty]{п.н.} 0 .$$

Приведенная теорема говорит о том, что «почти все» реализации эмпирической функции распределения $F_n^*(x;\xi_1,..., \xi_n)$ имеют «отклонение» от функции распределения $F_\xi(x)$, которое стремится к нулю с ростом объема выборки n. «Почти все» означает, что имеются реализации для которых «отклонение» не стремится к нулю с ростом n (например, такие в которых $x_1 = ... = x_n$), но вероятность получения таких реализаций равна нулю.

Вариационный ряд.

Пусть в результате статистического эксперимента была получена реализация $(x_1,..., x_n)$ выборки $(\xi_1,..., \xi_n)$. Упорядочим числовые значения x_i по возрастанию и обозначим за $x_{(1)}$ –наименьшее из значений среди x_i, за $x_{(2)}$ – второе по величине значение среди x_i, и так далее, а за $x_{(n)}$ – наибольшее значение среди x_i, другими словами, пусть $x_{(k)}$ обозначает k-ое по величине значение среди x_i, упорядоченных по возрастанию:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(k)} \le \ldots \le x_{(n)} .$$

Подобное упорядочивание и определение значений $x_{(k)}$ может быть произведено с любой реализацией $(x_1,..., x_n)$ выборки, поэтому вполне корректным будет определение такой совокупности случайных величин $\xi_{(1)}$, …, $\xi_{(n)}$, в которой каждая величина $\xi_{(k)}$ всегда принимает значение $x_{(k)}$.

Совокупность случайных величин $(\xi_{(1)},..., \xi_{(n)})$, представляющую собой упорядоченную по возрастанию совокупность величин выборки $(\xi_1,..., \xi_n)$, называют *вариационным рядом* [6, 13]:

$$\xi_{(1)} \le \xi_{(2)} \le \ldots \le \xi_{(n)} ,$$

величины $\xi_{(k)}$, взятые по отдельности, называют *порядковыми статистиками* [7].

Вариационный ряд $(\xi_{(1)},..., \xi_{(n)})$, как и выборка, является векторной случайной величиной, поэтому на практике имеется возможность получить лишь отдельные реализации вариационного ряда $(x_{(1)},..., x_{(n)})$ при упорядочивании по возрастанию величин реализации выборки $(x_1,..., x_n)$.

Частоты и гистограмма.

На практике объемы выборок n могут достигать значений порядка тысяч или десятков тысяч. Обработка числовых значений x_i в таких количествах требует наглядного графического представления, каковым, например, является гистограмма частот [1, 12].

Суть образования частот $\nu_k(\xi_1,..., \xi_n)$ заключается в том, чтобы от исходных случайных величин выборки ξ_i перейти к рассмотрению «укрупненных» величин ν_k. Указанный переход осуществляется путем разбиения числовой оси на некоторое количество m непересекающихся интервалов и полуинтервалов L_k ($k = \overline{1, m}$):

$$L_1 = (-\infty , y_1),\ L_2 = [y_1, y_2),\ \ldots,\ L_k = [y_{k-1}, y_k),\ \ldots,\ L_m = (y_{m-1}, \infty), \qquad (2.3)$$

где $y_1 < y_2 < \ldots < y_{m-1}$ – точки, выбираемые произвольным образом.

Для разбиения числовой оси на множества L_k, определяется набор *частот* $\nu_k(\xi_1,..., \xi_n)$, в котором каждая величина ν_k равна случайному количеству исходных величин выборки ξ_i попавших в множество L_k:

$$\nu_k(\xi_1,..., \xi_n) = \sum_{i=1}^{n} I(\xi_i, L_k), \qquad (2.4)$$

$$I(\xi, L) = \begin{cases} 1, & \xi \in L \\ 0, & \xi \notin L \end{cases}$$

Наряду с частотами ν_k вводятся также *относительные частоты* $\tilde{\nu}_k$, определяемые как отношение частот ν_k к объему выборки n:

$$\tilde{\nu}_k(\xi_1,..., \xi_n) = \frac{\nu_k(\xi_1,..., \xi_n)}{n}. \quad (2.5)$$

Легко видеть, что согласно определениям (2.4) и (2.5) частоты ν_k и $\tilde{\nu}_k$ являются случайными величинами. По результатам статистического эксперимента и получения реализации $(x_1,..., x_n)$ выборки $(\xi_1,..., \xi_n)$ наборы частот $(\nu_1,..., \nu_m)$ и $(\tilde{\nu}_1,..., \tilde{\nu}_m)$ приобретают соответствующие реализации $(n_1,..., n_m)$ и $(\tilde{n}_1,..., \tilde{n}_m)$, получаемые подстановкой в определения (2.4) и (2.5) числовых значений x_i вместо величин ξ_i:

$$n_k = \nu_k(x_1,..., x_n), \; \tilde{n}_k = \tilde{\nu}_k(x_1,..., x_n).$$

Наглядным графическим представлением реализаций частот $(n_1,..., n_m)$ (и реализаций относительных частот $(\tilde{n}_1,..., \tilde{n}_m)$) является *гистограмма* [10], представляющая собой столбцовую диаграмму, в которой над множествами разбиения L_k (2.3) изображаются столбцы высотой n_k (или $\tilde{n}_k$).

Выборочные моменты.

Для оценки «интегральных» характеристик неизвестной функции распределения $F_\xi(x)$, таких как моменты случайных величин ξ_i выборки, широко используются *выборочные моменты* $\hat{m}_k(\xi_1,..., \xi_n)$ и *выборочные центральные моменты* $\hat{\mu}_k(\xi_1,..., \xi_n)$ [2, 8]:

$$\hat{m}_k(\xi_1,..., \xi_n) = \frac{1}{n}\sum_{i=1}^{n} \xi_i^k,$$
$$\hat{\mu}_k(\xi_1,..., \xi_n) = \frac{1}{n}\sum_{i=1}^{n} (\xi_i - \hat{m}_1(\xi_1,..., \xi_n))^k. \quad (2.6)$$

В частности для оценки неизвестного математического ожидания величин ξ_i используется *выборочное среднее* $\hat{m}_1(\xi_1,..., \xi_n)$, а для оценки дисперсии – *выборочная дисперсия* $\hat{\mu}_2(\xi_1,..., \xi_n)$ или *исправленная выборочная дисперсия* $\tilde{\mu}_2(\xi_1,..., \xi_n)$:

$$\tilde{\mu}_2(\xi_1,..., \xi_n) = \frac{1}{n-1}\sum_{i=1}^{n} (\xi_i - \hat{m}_1(\xi_1,..., \xi_n))^2. \quad (2.7)$$

Важное свойство оценок (2.6) и (2.7) заключается в их состоятельности: с увеличением объема выборки n вероятность «большого» отклонения оценки от оцениваемой величины уменьшается и в пределе с ростом объема n стремится к нулю.

2. Задачи.

Задача 1.

Напишите функции, которые формируют заданное количество k реализаций выборки объема n с заданным распределением (на основе реализации последовательности случайных величин с таким же распределением). Используйте функции для построения реализаций из распределений:

1) равномерного $R[-1,2]$,

2) нормального $N(1,2)$,

3) показательного $E(4)$,

4) биномиального $Bi(0.3,8)$,

для $k = 3$ и $n = 10$.

Напишите функцию, которая для заданных реализаций строит соответствующие реализации вариационного ряда, и примените её к построенным реализациям выборок.

Задача 2.

Для каждого из распределений, указанных в задаче 1, постройте по 3 реализации выборки объемом 10 и для каждого распределения по отдельности по трем реализациям постройте в одной системе координат графики трех соответствующих реализаций эмпирической функции распределений (2.2).

Задача 3.

Постройте реализации выборок объемами 10, 100 и 1000 для всех распределений, указанных в задаче 1. Для каждого распределения по отдельности в одной системе координат постройте графики трех реализаций эмпирической функции распределения по заданным реализациям объемов 10, 100 и 1000 и график функции распределения.

Задача 4.

Напишите функцию, которая строит гистограмму реализаций относительных частот (2.5) по заданной реализации выборки и заданным точкам y_k, определяющих разбиение L_k (2.3). Постройте по одной реализации выборки объемом $n = 100$ для всех распределений, указанных в задаче 1, и постройте для них гистограммы реализаций относительных частот.

Задача 5.

Напишите функции вычисления реализаций выборочных моментов k-го порядка (2.6) и исправленной выборочной дисперсии (2.7). Постройте по одной реализации выборки объемом $n = 100$ для каждого из распределений, указанных в задаче 1 за исключением биномиального распределения. По реализациям выборок вычислите реализации выборочных среднего и дисперсии, исправленной выборочной дисперсии, выборочного момента 2го порядка и выборочного центрального момента 3го порядка. Сравните вычисленные величины с соответствующими математическими ожиданиями, дисперсиями, вторыми начальными и третьими центральными моментами.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: текст функции построения реализаций выборки и текст функции построения реализаций вариационного ряда, по одной реализации выборки и одной реализации вариационного ряда для каждого из указанных распределений.

Задача 2: текст программы построения графиков реализаций эмпирической функции распределения и четыре рисунка (по одному на каждое распределение), на которых изображены по три графика реализаций эмпирической функции распределения.

Задача 3: текст программы построения графиков и четыре рисунка (по одному рисунку на каждое распределение), на которых представлены четыре графика (три графика реализаций эмпирической функции распределения и график функции распределения).

Задача 4: текст функции построения гистограмм и четыре изображения гистограмм (по одному изображению для каждого из указанных распределений).

Задача 5: программы вычисления реализаций всех указанных характеристик, вычисленные реализации всех характеристик и точные значения всех указанных моментов (формулы моментов могут быть взяты из справочников).

4. Вопросы и задания.

1) Сформулируйте определения выборки, объема выборки, реализации выборки и множества реализаций выборки. В чем отличие выборки от реализации выборки?

2) Опишите метод построения реализаций выборки на основе реализации последовательности случайных величин.

3) Сформулируйте определение эмпирической функции распределения. Что представляет собой множество реализаций эмпирической функции распределения.

4) Сформулируйте определение вариационного ряда и поясните построение реализаций вариационного ряда на основе реализации выборки.

5) Приведите определения для частоты и относительной частоты. Опишите метод построения гистограмм для реализаций относительных частот.

6) Напишите статистики для выборочных моментов и выборочных центральных моментов порядка k и статистику для исправленной выборочной дисперсии.

Лабораторная работа №3. **Точечное оценивание.**

1. Основные сведения.

Точечные оценки.

Пусть наблюдения $(\xi_1,..., \xi_n)$ имеют совместную функцию распределения $F_\xi(x_1,..., x_n \mid \theta)$, где θ – параметр, значение которого неизвестно, известно лишь, что значение параметра θ взято из множества допустимых значений Θ ($\theta \in \Theta$), и требуется оценить значение выражения $\tau(\theta)$, содержащее неизвестный параметр θ и представляющее собой характеристику неизвестной функции распределения $F_\xi(x_1,..., x_n \mid \theta)$. Решением поставленной задачи может считаться любая функция $T_n(\xi_1,..., \xi_n)$, использующая только наблюдения $(\xi_1,..., \xi_n)$, с помощью которой можно оценить неизвестное значение $\tau(\theta)$.

Статистику $T_n(\xi_1,..., \xi_n)$ называют точечной оценкой, хотя в строгом формальном смысле *точечная оценка* представляет собой последовательность статистик $\{T_n(\xi_1,..., \xi_n)\}_{n=1}^{\infty}$, в которой каждая статистика $T_n(\xi_1,..., \xi_n)$ используется для оценки неизвестной величины $\tau(\theta)$ при некотором фиксированном n.

Свойства несмещенности и состоятельности точечных оценок.

Основными свойствами точечных оценок являются несмещенность и состоятельность. Несмещенность оценки означает отсутствие систематического смещения у оценки при любом допустимом значении параметра θ. Состоятельность оценки гарантирует повышение «точности» оценки с увеличением количества n наблюдаемых величин ξ_i.

Формально оценка $T_n(\xi_1,..., \xi_n)$ является *несмещенной оценкой* $\tau(\theta)$ [6, 13], если при всяком n и любом допустимом значении параметра $\theta \in \Theta$ математическое ожидание оценки $T_n(\xi_1,..., \xi_n)$ совпадает с величиной $\tau(\theta)$:

$$\forall n \forall \theta \in \Theta : M_\theta[T_n(\xi_1,..., \xi_n)] = \tau(\theta). \qquad (3.1)$$

Оценка $T_n(\xi_1,..., \xi_n)$ является *состоятельной оценкой* [6, 13] если при всяком допустимом значении параметра $\theta \in \Theta$ последовательность статистик $T_n(\xi_1,..., \xi_n)$ сходится по вероятности к величине $\tau(\theta)$:

$$\begin{gathered}\forall \theta \in \Theta : T_n(\xi_1,..., \xi_n) \xrightarrow{P} \tau(\theta), \\ n \to \infty.\end{gathered} \qquad (3.2)$$

Определить наличие свойств несмещенности и состоятельности у оценки удается только посредством теоретического анализа, тем не менее, проявление свойств несмещенности и состоятельности можно наблюдать и на практике при исследовании большого количества значений оценки.

Если оценка $T_n(\xi_1,..., \xi_n)$ является несмещенной, то в соответствии с определением (3.1) значения (реализации) оценки группируются вокруг величины $\tau(\theta)$ таким образом, что среднее значение ошибки (разности $T_n - \tau$) равно нулю. Пусть при фиксированном количестве величин, $n = n_1$, в результате m экспериментов получена серия реализаций $\{(x_1^{(j)},..., x_{n_1}^{(j)})\}_{j=1}^{m}$ наблюдаемых величин $(\xi_1,..., \xi_{n_1})$ и совокупность соответствующих реализаций t_1, …, t_m оценки $T_{n_1}(\xi_1,..., \xi_{n_1})$:

$$t_j = T_{n_1}(x_1^{(j)},..., x_{n_1}^{(j)}),$$
$$j = \overline{1, m}.$$

Реализации t_j несмещенной оценки $T_n(\xi_1,..., \xi_n)$ с большой вероятностью окажутся сосредоточенными в окрестности значения $\tau(\theta)$, в этом заключается проявление свойства несмещенности, кроме того, если вычислить среднее арифметическое всех реализаций t_j, то полученное значение при достаточно больших m окажется приближенно равным $\tau(\theta)$:

$$\frac{1}{m}\sum_{j=1}^{m} t_j \approx \tau(\theta) \tag{3.3}$$

Если оценка $T_n(\xi_1,..., \xi_n)$ является состоятельной, то в соответствии с определением (3.2) величина ошибки (разности $T_n - \tau$) стремится к нулю с ростом n, то есть распределение оценки $T_n(\xi_1,..., \xi_n)$ «сжимается» в окрестности $\tau(\theta)$. Следствием «сжатия» распределения является «более плотное сосредоточивание» реализаций оценки $T_n(\xi_1,..., \xi_n)$ в окрестности величины $\tau(\theta)$ при увеличении n.

Рассмотрим уже имеющиеся реализации t_j оценки $T_{n_1}(\xi_1,..., \xi_{n_1})$, полученные при некотором фиксированном значении $n = n_1$, выберем $n_2 >> n_1$ – и получим новые реализации $\bar{t}_1$, …, $\bar{t}_m$ оценки $T_{n_2}(\xi_1,..., \xi_{n_2})$ при увеличенном объеме n_2. Поскольку оценка $T_n(\xi_1,..., \xi_n)$ является состоятельной, то с большой вероятностью окажется, что среднее отклонение значений $\bar{t}_j$ меньше среднего отклонения значений t_j:

$$\frac{1}{n_2}\sum_{j=1}^{n_2} \left|\bar{t}_j - \tau(\theta)\right| < \frac{1}{n_1}\sum_{j=1}^{n_1} \left|t_j - \tau(\theta)\right|. \tag{3.4}$$

Повторив проделанный эксперимент несколько раз с монотонно возрастающим количеством n, легко убедиться в том, что значения состоятельной оценки с ростом n «группируются более плотно» возле оцениваемой величины $\tau(\theta)$.

Методы построения оценок.

Для решения разнообразных задач точечного оценивания существует ряд специальных методов построения оценок, наиболее простыми и часто используемыми из которых являются методы: моментов, максимального правдоподобия и порядковых статистик. Указанные методы, как правило, используются для построения оценки неизвестного значения параметра θ, которая затем может быть использована для построения оценки произвольной характеристики $\tau(\theta)$ функции распределения наблюдений.

Метод моментов.

Пусть исходные наблюдения образуют выборку $\zeta_n = (\xi_1,..., \xi_n)$ из распределения $F_\xi(x\,|\,\theta)$, где $\theta = (\theta_1,..., \theta_k)$ – неизвестный параметр из множества допустимых значений Θ и требуется по заданной построить оценки вектора θ и величины $\tau(\theta)$.

Для построения оценки по методу моментов [5, 6, 13] необходимо выбрать m моментов ν_1, ν_2, …, ν_m функции распределения $F_\xi(x\,|\,\theta)$, которые бы зависели от векторного параметра $\theta = (\theta_1, \theta_2,..., \theta_k)$:

$$\begin{cases} \nu_1 = f_1(\theta_1, \theta_2,..., \theta_k) \\ \nu_2 = f_2(\theta_1, \theta_2,..., \theta_k) \\ \qquad ... \\ \nu_m = f_m(\theta_1, \theta_2,..., \theta_k) \end{cases}. \tag{3.5}$$

Пусть систему (3.5) удается разрешить относительно переменных θ_1, θ_2, ..., θ_k, тогда эти переменные выражаются через моменты ν_1, ν_2, …, ν_m:

$$\begin{cases} \theta_1 = g_1(\nu_1, \nu_2,..., \nu_m) \\ \theta_2 = g_2(\nu_1, \nu_2,..., \nu_m) \\ \qquad ... \\ \theta_k = g_k(\nu_1, \nu_2,..., \nu_m) \end{cases}. \tag{3.6}$$

Подставляя в систему (3.6) выборочные моменты $\hat{\nu}_i(\zeta_n)$ вместо моментов ν_i, получим *оценку параметра* $\hat{\theta}(\zeta_n) = (\hat{\theta}_1(\zeta_n), \hat{\theta}_2(\zeta_n),..., \hat{\theta}_k(\zeta_n))$ *по методу моментов*:

$$\begin{cases} \hat{\theta}_1(\zeta_n) = g_1(\hat{\nu}_1(\zeta_n), \hat{\nu}_2(\zeta_n),..., \hat{\nu}_m(\zeta_n)) \\ \hat{\theta}_2(\zeta_n) = g_2(\hat{\nu}_1(\zeta_n), \hat{\nu}_2(\zeta_n),..., \hat{\nu}_m(\zeta_n)) \\ \qquad ... \\ \hat{\theta}_k(\zeta_n) = g_k(\hat{\nu}_1(\zeta_n), \hat{\nu}_2(\zeta_n),..., \hat{\nu}_m(\zeta_n)) \end{cases}. \tag{3.7}$$

Оценку $\hat{\theta}(\xi_1,..., \xi_n)$ (3.7) можно использовать при построении *оценки* $\hat{T}(\xi_1,..., \xi_n)$ *по методу моментов* для произвольной величины $\tau(\theta)$, зависящей от параметра θ, непосредственной подстановкой оценки $\hat{\theta}(\xi_1,..., \xi_n)$ вместо параметра θ:

$$\hat{T}(\xi_1,..., \xi_n) = \tau(\hat{\theta}(\xi_1,..., \xi_n)) . \qquad (3.8)$$

Оценки по методу моментов в большинстве случаев не являются несмещенными, тем не менее, если функция g_i системы (3.7) является непрерывной в окрестности точки $(\nu_1,..., \nu_m)$, образованной моментами ν_i, то оценка $\hat{\theta}_i(\zeta_n)$ является состоятельной оценкой θ_i. Если все оценки $\hat{\theta}_i(\zeta_n)$ являются состоятельными, и величина $\tau(\theta)$ как функция параметра $\theta = (\theta_1,..., \theta_k)$ является непрерывной на множестве Θ, тогда и оценка $\hat{T}(\xi_1,..., \xi_n)$ (3.8) является состоятельной. Кроме того, при некоторых широких условиях оценки $\hat{\theta}_i(\zeta_n)$ являются асимптотически нормальными (при $n \to \infty$).

Метод максимального правдоподобия.

Пусть исходные наблюдения $(\xi_1,..., \xi_n)$, не обязательно представляющие собой выборку, имеют совместную плотность вероятности $p_\xi(x_1,..., x_n | \theta)$ (или функцию вероятности в случае дискретных случайных величин ξ_i), зависящую от неизвестного параметра $\theta = (\theta_1,..., \theta_k)$ из множества допустимых значений Θ, и требуется построить оценку параметра $\theta = (\theta_1,..., \theta_k)$ и величины $\tau(\theta)$.

Определим функцию правдоподобия $L(\xi_1,..., \xi_n | \theta)$ как функцию параметра $\theta \in \Theta$, полученную подстановкой в функцию $p_\xi(x_1,..., x_n | \theta)$ величин ξ_i вместо переменных x_i:

$$L(\xi_1,..., \xi_n | \theta) = p_\xi(\xi_1,..., \xi_n | \theta) .$$

Оценка по методу максимального правдоподобия $\theta^*(\xi_1,..., \xi_n)$ [8, 10] определяется таким образом, чтобы при всех возможных реализациях наблюдений $(\xi_1,..., \xi_n)$ функция правдоподобия $L(\xi_1,..., \xi_n | \theta)$ в точке $\theta^*(\xi_1,..., \xi_n)$ принимала бы значение равное точной верхней грани на множестве допустимых значений параметра Θ:

$$L(\xi_1,..., \xi_n | \theta^*) = \sup_{\theta \in \Theta} L(\xi_1,..., \xi_n | \theta) .$$

Если точка $\theta^*(\xi_1,..., \xi_n)$ является внутренней точкой множества Θ, то согласно необходимому условию экстремума частные производные функции правдоподобия $L(\xi_1,..., \xi_n | \theta)$ в точке $\theta^*(\xi_1,..., \xi_n)$ должны быть равны нулю:

$$\left.\frac{\partial}{\partial\theta_i}L(\xi_1,\ldots,\ \xi_n\mid\theta_1,\ldots,\ \theta_k)\right|_{\theta=\theta^*}=0,\ i=\overline{1,k}. \qquad (3.9)$$

В ряде случаев бывает удобней перейти к логарифму функции правдоподобия, в этом случае справедливо условие аналогичное (3.9):

$$\left.\frac{\partial}{\partial\theta_i}\ln L(\xi_1,\ldots,\ \xi_n\mid\theta_1,\ldots,\ \theta_k)\right|_{\theta=\theta^*}=0\,,\ i=\overline{1,k}\,. \qquad (3.10)$$

Условия (3.9) и (3.10) могут быть использованы для нахождения оценки максимального правдоподобия $\theta^*(\xi_1,\ldots,\ \xi_n)$, и в случае если оценка $\theta^*(\xi_1,\ldots,\ \xi_n)$ имеет явное выражение через величины ξ_i, она может быть использована для построения оценки произвольной величины $\tau(\theta)$ путем непосредственной подстановки:

$$T^*(\xi_1,\ldots,\ \xi_n)=\tau(\theta^*(\xi_1,\ldots,\ \xi_n))\,.$$

Оценки максимального правдоподобия при выполнении некоторых условий являются состоятельными, асимптотически эффективными и асимптотически нормальными (при $n\to\infty$) оценки.

Метод порядковых статистик.

Метод порядковых статистик аналогичен методу моментов с той разницей, что необходимо выбрать m квантилей x_{p_1}, x_{p_2}, …, x_{p_m} функции распределения $F_\xi(x\mid\theta)$ уровней p_1, p_2, …, p_m, которые выражались бы через компоненты векторного параметра $\theta=(\theta_1,\theta_2,\ldots,\ \theta_k)$:

$$\begin{cases} x_{p_1}=\tilde{f}_1(\theta_1,\theta_2,\ldots,\ \theta_k) \\ x_{p_2}=\tilde{f}_2(\theta_1,\theta_2,\ldots,\ \theta_k) \\ \qquad\ldots \\ x_{p_m}=\tilde{f}_m(\theta_1,\theta_2,\ldots,\ \theta_k) \end{cases}, \qquad (3.11)$$

далее разрешая систему (3.11) относительно переменных θ_1, θ_2, ..., θ_k, можно получить систему, в которой компоненты параметра выражены через квантили:

$$\begin{cases} \theta_1=\tilde{g}_1(x_{p_1},x_{p_2},\ldots,\ x_{p_m}) \\ \theta_2=\tilde{g}_2(x_{p_1},x_{p_2},\ldots,\ x_{p_m}) \\ \qquad\ldots \\ \theta_k=\tilde{g}_k(x_{p_1},x_{p_2},\ldots,\ x_{p_m}) \end{cases}. \qquad (3.12)$$

Для получения *оценки по методу порядковых статистик* $\tilde{\theta}(\zeta_n)=(\tilde{\theta}_1(\zeta_n),\tilde{\theta}_2(\zeta_n),\ldots,\ \tilde{\theta}_k(\zeta_n))$ остается лишь подставить в систему (3.12) вместо квантилей x_{p_j} соответствующие выборочные квантили $\tilde{x}_{p_j}(\zeta_n)$:

$$\begin{cases} \tilde{\theta}_1(\zeta_n) = \tilde{g}_1(\tilde{x}_{p_1}(\zeta_n), \tilde{x}_{p_2}(\zeta_n), ..., \tilde{x}_{p_m}(\zeta_n)) \\ \tilde{\theta}_2(\zeta_n) = \tilde{g}_2(\tilde{x}_{p_1}(\zeta_n), \tilde{x}_{p_2}(\zeta_n), ..., \tilde{x}_{p_m}(\zeta_n)) \\ \quad ... \\ \tilde{\theta}_k(\zeta_n) = \tilde{g}_k(\tilde{x}_{p_1}(\zeta_n), \tilde{x}_{p_2}(\zeta_n), ..., \tilde{x}_{p_m}(\zeta_n)) \end{cases} . \qquad (3.13)$$

Построение оценки для произвольной величины $\tau(\theta)$ сводится к подстановке вместо параметра θ оценки $\tilde{\theta}(\zeta_n)$ (3.13):

$$\tilde{T}(\xi_1, ..., \xi_n) = \tau(\tilde{\theta}(\xi_1, ..., \xi_n)) . \qquad (3.14)$$

В качестве выборочных квантилей $\tilde{x}_{p_j}(\xi_1, ..., \xi_n)$ часто используются порядковые статистики (элементы вариационного ряда) $\xi_{([np_j]+1)}$ с номерами $[np_j]+1$, соответствующим уровням p_j квантилей (отсюда происходит и название метода порядковых статистик).

Согласно теореме Крамера при некоторых условиях порядковая статистика $\xi_{([np_j]+1)}$ (и выборочная квантиль $\tilde{x}_{p_j}(\xi_1, ..., \xi_n)$), является состоятельной оценкой квантиля x_{p_j}. Как и в случае метода моментов, если функция $\tilde{g}_i$ является непрерывной в некоторой окрестности точки $(x_{p_1}, ..., x_{p_m})$, то оценка $\tilde{\theta}_i(\zeta_n)$ является состоятельной оценкой компоненты θ_i параметра. Если все оценки $\tilde{\theta}_i(\zeta_n)$ являются состоятельными и функция $\tau(\theta)$ непрерывна на множестве допустимых значений параметра Θ, тогда оценка $\tilde{T}(\xi_1, ..., \xi_n)$ (3.14) также является состоятельной.

Сравнение точечных оценок.

В общем случае для оценки величины $\tau(\theta)$ могут быть предложены несколько различных оценок, и в этом случае неизбежно возникает вопрос о выборе одной из оценок на основе принятого критерии сравнения. Широко распространен критерий сравнения оценок по «точности» [5, 13], в котором в качестве «точности» оценок используется величина отклонения, усредненного по множеству всех значений оценки. В частности одним из вариантов такого усредненного отклонения является среднеквадратическое отклонение $\delta(T, \tau)$:

$$\delta(T(\xi_1, ..., \xi_n), \tau(\theta)) = M_\theta[(T(\xi_1, ..., \xi_n) - \tau(\theta))^2] .$$

При наличии двух оценок T_1 и T_2, оценка T_1 считается более точной, если её среднеквадратическое отклонение меньше:

$$\delta(T_1, \tau) < \delta(T_2, \tau) .$$

К сожалению, отклонение $\delta(T, \tau)$ может зависеть от параметра θ, так что вполне возможно при одних значениях параметра $\delta(T_1, \tau) < \delta(T_2, \tau)$, а при других – наоборот $\delta(T_1,\tau) > \delta(T_2,\tau)$. В этом случае, очевидно, оценки T_1 и T_2 несравнимы, и из них невозможно указать более точную.

На практике вычисление среднеквадратического отклонения $\delta(T,\tau)$ заменяют вычислением подходящих выборочных характеристик, например, исправленной выборочной дисперсии $\tilde{\mu}(t_1,..., t_m)$ или величиной размаха $w(t_1,..., t_m)$:

$$w(t_1,..., t_m) = \max_{1\le i\le m} t_i - \min_{1\le i\le m} t_i, \qquad (3.15)$$

где $t_1, \ldots, t_m$ – значения оценки T.

2. Задачи.

Задача 1.

По выборке $(\xi_1,..., \xi_n)$ из показательного распределения $E(\theta)$ постройте оценку параметра θ методом моментов. Возьмите в качестве значения параметра величину $\theta = 2$ и постройте $m = 20$ реализаций выборки объемом $n_1 = 10$ и столько же реализаций выборки объемом $n_2 = 50$, затем для всех реализаций вычислите построенную оценку. Вычислите величины средних отклонений (3.4) для двух множеств значений оценки, полученных при объемах n_1 и n_2 соответственно.

В одной системе координат на двух прямых, параллельных оси абсцисс и пересекающие ось ординат в точках $(0,n_1)$ и $(0,n_2)$, отметьте точками (маркерами) все полученные значения оценки для объема выборки n_1 и для объема выборки n_2 соответственно. В этой же системе координат проведите три прямые параллельные оси ординат и пересекающие ось абсцисс в точках, соответствующих двум средним арифметическим значений оценки (3.3) для объемов n_1 и n_2, а также точному значению параметра $\theta = 2$.

Подумайте, какие свойства оценки отражены на построенном рисунке, и какую форму может иметь распределение оценки при объемах выборки n_1 и n_2.

Задача 2.

По выборке $(\xi_1,..., \xi_n)$ из нормального распределения $N(\theta_1,\theta_2^2)$ постройте три оценки двумерного параметра $\theta = (\theta_1,\theta_2)$ методом порядковых статистик, используя, если это возможно, три различных набора квантилей. Задайте значение параметра равное $\theta = (2,4)$, постройте $m = 15$ реализаций выборки объемом $n = 20$ и для всех реализаций выборки вычислите значения трех построенных оценок.

Для каждой компоненты θ_i параметра в одной системе координат изобразите маркерами вычисленные значения компоненты оценки, откладывая их на трех прямых, параллельных оси абсцисс и пересекающих ось ординат, например, в точках $(0,1)$, $(0,2)$ и $(0,3)$ (на каждой прямой следует изображать значения одной их трех оценок), сами прямые при этом можно не отображать.

Используя рисунок, подумайте, какой из трех наборов квантилей приводит к наилучшей оценке.

Задача 3.

По выборке $(\xi_1,..., \xi_n)$ из равномерного распределения $R[\theta_1,\theta_2]$ постройте три оценки двумерного параметра $\theta = (\theta_1,\theta_2)$, используя методы моментов, максимального правдоподобия и порядковых статистик. Возьмите в качестве значения параметра вектор $\theta = (1,3)$, постройте по $m = 20$ реализаций выборок объемов $n_1 = 10$, $n_2 = 50$ и $n_3 = 100$ и для всех построенных реализаций выборок вычислите соответствующие значения всех трех оценок.

Для каждого объема выборки в отдельной системе координат изобразите вычисленные значения трех оценок в виде маркеров, расположенных на трех разных прямых параллельных оси абсцисс (поскольку каждая оценка является двумерной статистикой, то на одной и той же прямой располагаются и значения оценки для компоненты θ_1 и значения оценки для компоненты θ_2 параметра θ). В каждой системе координат проведите две прямые параллельные оси ординат и пересекающие ось абсцисс в точках 1 и 3, соответствующих значениям компонент параметра $\theta = (1,3)$. Все три системы координат должны иметь одинаковый масштаб.

Подумайте, о проявлении каких свойств построенных оценок свидетельствуют полученные данные и рисунки.

Для каждого объема выборки и каждой оценки вычислите размах полученных значений оценки (3.15). Постройте в одной системе координат три ломанных, каждая из которых отражает зависимость размаха значений оценки от объема выборки, откладывая объем выборки по оси абсцисс и размах по оси ординат. На основе построенных ломаных укажите, какая оценка из трех построенных является наилучшей.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: статистику, используемую для оценки параметра, указанный в задаче рисунок и текст программы расчета.

Задача 2: три оценки параметра, все 15 значений каждой оценки, указанный в задаче рисунок и текст программы расчета.

Задача 3: три оценки параметра, три рисунка со значениями оценок (по одному на каждый объем выборки), рисунок с ломаными размаха для трех оценок с указанием лучшей оценки и текст программы расчета.

4. Вопросы и задания.

1) Сформулируйте задачу точечного оценивания.

2) Какие свойства оценок считают основными? Сформулируйте определение свойств несмещенность и состоятельности и поясните смысл определений.

3) Сформулируйте определение несмещенной оценки. Поясните, каким образом располагаются значения несмещенной оценки по отношению к оцениваемой величине $\tau(\theta)$. Каким будет множество значений статистики, которая не является несмещенной оценкой величины $\tau(\theta)$.

4) Сформулируйте определение состоятельности. Используя определение, поясните, каким образом изменяются множества значений состоятельной оценки при увеличении количества наблюдений.

5) Сформулируйте метод моментов построения точечных оценок.

6) Сформулируйте метод максимального правдоподобия построения точечных оценок.

7) Сформулируйте метод порядковых статистик построения точечных оценок.

8) Поясните способ сравнения оценок на основе «точности» оценок.

Лабораторная работа №4. **Интервальное оценивание.**

1. Основные сведения.

Пусть совокупность наблюдений $(\xi_1,..., \xi_n)$ имеет распределение $F_\xi(x_1,..., x_n;\theta)$, где θ – некоторое неизвестное значение из множества допустимых значений параметра Θ. *Доверительным интервалом* для величины $\tau(\theta)$ с уровнем доверия $P_д$ называется случайный интервал $(T_1(\xi_1,..., \xi_n); T_2(\xi_1,..., \xi_n))$, в котором статистики $T_1(\xi_1,..., \xi_n)$ и $T_2(\xi_1,..., \xi_n)$ удовлетворяют условию:

$$\inf_{\theta\in\Theta} P\{T_1(\xi_1,..., \xi_n) < \tau(\theta) < T_2(\xi_1,..., \xi_n)\} = P_д. \qquad (4.1)$$

Условие (4.1) означает, что при любом допустимом значении параметра $\theta \in \Theta$ интервал $(T_1(\xi_1,..., \xi_n); T_2(\xi_1,..., \xi_n))$ содержит величину $\tau(\theta)$ с вероятностью не меньше $P_д$.

Нижней доверительной границей для величины $\tau(\theta)$ с уровнем доверия $P_д$ называется такая статистика $T_l(\xi_1,..., \xi_n)$, для которой справедливо равенство:

$$\inf_{\theta\in\Theta} P\{T_l(\xi_1,..., \xi_n) < \tau(\theta)\} = P_д. \qquad (4.2)$$

а *верхней доверительной границей* для величины $\tau(\theta)$ с уровнем доверия $P_д$ называется такая статистика $T_u(\xi_1,..., \xi_n)$, для которой выполняется следующее условие:

$$\inf_{\theta\in\Theta} P\{\tau(\theta) < T_u(\xi_1,..., \xi_n)\} = P_д . \qquad (4.3)$$

Условие (4.2) показывает, что при любом допустимом значении параметра $\theta \in \Theta$ статистика $T_l(\xi_1,..., \xi_n)$ оказывается меньше величины $\tau(\theta)$ с вероятностью не меньше $P_д$, а статистика $T_u(\xi_1,..., \xi_n)$ – больше величины $\tau(\theta)$ с вероятностью не меньше $P_д$.

Построение доверительных интервалов и границ.

Общий метод построения доверительных интервалов и границ основывается на понятии *центральной статистики* $\varphi(\xi_1,..., \xi_n;\tau(\theta))$ [5, 6], выражение которой содержит неизвестную величину $\tau(\theta)$, но распределение которой не зависит от неизвестного значения параметра θ. При наличии центральной статистики $\varphi(\xi_1,..., \xi_n;\tau(\theta))$ первый этап построения доверительного интервала заключается в нахождении такого интервала (y_1, y_2) для центральной статистики, что при любом допустимом значении параметра $\theta \in \Theta$ центральная статистика $\varphi(\xi_1,..., \xi_n;\tau(\theta))$ оказывается в интервале (y_1, y_2) с вероятностью не меньше доверительной $P_д$:

$$\forall\, \theta \in \Theta : P\{y_1 < \varphi(\xi_1,..., \xi_n;\tau(\theta)) < y_2\} \ge P_д . \qquad (4.4)$$

Значений y_1 и y_2 стараются выбрать так, чтобы интервал (y_1, y_2) оказался как можно короче, если это сделать не удается, то обычно выбирают значения y_1 и y_2 так, чтобы слева от y_1 и справа от y_2 оказалась одна и та же вероятность $\frac{1 - P_д}{2}$.

В неравенстве (4.4) существенным является отсутствие зависимости значений y_1 и y_2 от неизвестного значения параметра θ, что гарантируется независимостью распределения центральной статистики $\varphi(\xi_1,..., \xi_n;\tau(\theta))$ от параметра θ.

На втором этапе построения доверительного интервала производится разрешение двойного неравенства, стоящего под знаком вероятности в (4.4), таким образом, чтобы в центральной части оказалась неизвестная величина $\tau(\theta)$:

$$\forall\, \theta \in \Theta : P\{T_1(\xi_1,..., \xi_n) < \tau(\theta) < T_2(\xi_1,..., \xi_n)\} \ge P_д . \qquad (4.5)$$

Статистики $T_1(\xi_1,..., \xi_n)$ и $T_2(\xi_1,..., \xi_n)$, полученные в результате преобразования, образуют доверительный интервал $(T_1(\xi_1,..., \xi_n); T_2(\xi_1,..., \xi_n))$ для величины $\tau(\theta)$ с уровнем доверия не менее $P_д$.

При построении нижней доверительной границы с монотонно неубывающей по $\tau(\theta)$ центральной статистикой $\varphi(\xi_1,..., \xi_n;\tau(\theta))$ необходимо выбрать значение y таким образом, чтобы выполнялось условие:

$$\forall\,\theta \in \Theta : P\{\, y < \varphi(\xi_1,..., \xi_n;\tau(\theta))\} \geq P_\partial ,$$

и далее разрешить относительно $\tau(\theta)$ неравенство, стоящее под знаком вероятности:

$$\forall\,\theta \in \Theta : P\{T_l(\xi_1,..., \xi_n) < \tau(\theta)\} \geq P_\partial .$$

Статистика $T_l(\xi_1,..., \xi_n)$, полученная в результате преобразования, является нижней доверительной границей для $\tau(\theta)$ с уровнем доверия не менее P_∂.

Аналогичным образом с точностью до неравенств строятся нижние и верхние границы в случае монотонно невозрастающей (неубывающей) по $\tau(\theta)$ центральной статистики $\varphi(\xi_1,..., \xi_n;\tau(\theta))$.

Доверительный интервал для математического ожидания по выборке из нормального распределения с известной дисперсией.

Пусть задана выборка $(\xi_1,..., \xi_n)$ из нормального распределения $N(m,\sigma^2)$, где m – неизвестный параметр, а σ – известное числовое значение, тогда при построении доверительных интервалов и границ для m используется центральная статистика [5, 10, 13]:

$$\varphi(\xi_1,..., \xi_n;m) = \frac{\frac{1}{n}\sum_{i=1}^{n}\xi_i - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), \tag{4.6}$$

имеющая стандартное нормальное распределение $N(0,1)$, которая приводит к доверительному интервалу для m с уровнем доверия P_∂:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - y_2\frac{\sigma}{\sqrt{n}};\frac{1}{n}\sum_{i=1}^{n}\xi_i + y_2\frac{\sigma}{\sqrt{n}}\right), \tag{4.7}$$

в котором величина y_2 является квантилью уровня $\frac{1+P_\partial}{2}$ распределения $N(0,1)$.

Доверительный интервал для математического ожидания по выборке из нормального распределения с неизвестной дисперсией.

Если $(\xi_1,..., \xi_n)$ – выборка из нормального распределения $N(m,\sigma^2)$, в котором и m, и σ являются неизвестными параметрами, тогда при построении доверительных интервалов и границ для m используется следующая центральная статистика [5, 10]:

$$\varphi(\xi_1,..., \xi_n;m) = \frac{\hat{m}_1 - m}{\sqrt{\tilde{\mu}_2}}\sqrt{n} \sim T(n-1),$$

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i \,,\; \tilde{\mu}_2 = \frac{1}{n-1}\sum_{i=1}^{n} (\xi_i - \hat{m}_1)^2 \,,$$

имеющая распределение Стьюдента [7, 8] с $n-1$ степенью свободы $T(n-1)$. Доверительным интервалом для m с уровнем доверия $P_д$ будет являться следующий интервал:

$$\left(\hat{m}_1 - y_2\sqrt{\frac{\tilde{\mu}_2}{n}}; \hat{m}_1 + y_2\sqrt{\frac{\tilde{\mu}_2}{n}}\right), \tag{4.8}$$

в котором y_2 – квантиль уровня $\frac{1+P_д}{2}$ распределения Стьюдента с $n-1$ степенью свободы $T(n-1)$.

Доверительный интервал для среднеквадратичного отклонения по выборке из нормального распределения с известным математическим ожиданием.

При построении доверительных интервалов и границ для величины σ по выборке $(\xi_1,\ldots, \xi_n)$ из нормального распределения $N(m,\sigma^2)$, где m – известная величина, используется следующая центральная статистика:

$$\varphi(\xi_1,\ldots, \xi_n;\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} (\xi_i - m)^2 \sim \chi^2(n), \tag{4.9}$$

имеющая распределение «хи-квадрат» с n степенями свободы $\chi^2(n)$ [7, 8]. Доверительным интервалом для σ с уровнем доверия $P_д$ является интервал:

$$\left(\sqrt{\frac{\sum_{i=1}^{n} (\xi_i - m)^2}{y_2}}; \sqrt{\frac{\sum_{i=1}^{n} (\xi_i - m)^2}{y_1}}\right), \tag{4.10}$$

в котором y_1 и y_2 – квантили уровней $\frac{1-P_д}{2}$ и $\frac{1+P_д}{2}$ распределения $\chi^2(n)$ соответственно.

Доверительный интервал для среднеквадратичного отклонения по выборке из нормального распределения с неизвестным математическим ожиданием.

В том случае, если $(\xi_1,\ldots, \xi_n)$ – выборка из нормального распределения $N(m,\sigma^2)$ с неизвестным значением m, построение доверительных интервалов и границ для σ основывается на следующей центральной статистике:

$$\varphi(\xi_1,\ldots, \xi_n;\sigma^2) = \frac{n-1}{\sigma^2}\tilde{\mu}_2(\xi_1,\ldots, \xi_n) \sim \chi^2(n-1), \tag{4.11}$$

$$\tilde{\mu}_2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{m}_1)^2\ ,\ \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i\ ,$$

имеющей распределение «хи-квадрат» с $n-1$ степенью свободы $\chi^2(n-1)$. Доверительным интервалом для σ с уровнем доверия $P_д$ в этом случае является, например, следующий интервал:

$$\left(\sqrt{\frac{(n-1)\tilde{\mu}_2(\xi_1,...,\ \xi_n)}{y_2}};\sqrt{\frac{(n-1)\tilde{\mu}_2(\xi_1,...,\ \xi_n)}{y_1}}\right),$$
$$\tilde{\mu}_2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{m}_1)^2\ ,\ \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n}\xi_i\ , \tag{4.12}$$

в котором y_1 и y_2 являются квантилями уровней $\frac{1-P_д}{2}$ и $\frac{1+P_д}{2}$ распределения $\chi^2(n-1)$ соответственно.

Построение приближенного доверительного интервала для неизвестной вероятности случайного события.

В тех случаях, когда построение доверительных интервалов и границ сопряжено с определенными трудностями, часто прибегают к построению «приближенных» доверительных интервалов, получаемых при использовании вместо точного распределения центральной статистики некоторого приближенного (асимптотического) распределения [6, 13].

Рассмотрим серию из n независимых испытаний, в каждом из которых может произойти событие A, имеющее неизвестную вероятность p. Результатом проведения серии испытаний является выборка $(\xi_1,...,\ \xi_n)$, в которой каждая случайная величина ξ_i принимает значение 1 с неизвестной вероятностью p и значение 0 с вероятностью $1-p$. Согласно центральной предельной теореме сумма случайных величин $\sum_{i=1}^{n}\xi_i$ будет иметь асимптотически, при возрастании n, нормальное распределение с параметрами $M\left[\sum_{i=1}^{n}\xi_i\right] = np$ и $D\left[\sum_{i=1}^{n}\xi_i\right] = np(1-p)$. В таком случае, при достаточно большом n центральная статистика $\varphi(\xi_1,...,\ \xi_n;\ p)$:

$$\varphi(\xi_1,...,\ \xi_n;\ p) = \frac{\sum_{i=1}^{n}\xi_i - np}{\sqrt{np(1-p)}} \sim N(0,1) \tag{4.13}$$

имеет распределение, приближенно совпадающее со стандартным нормальным распределением $N(0,1)$.

Если применить общий метод построения доверительных интервалов к центральной статистике (4.13), то получится интервал, который после некоторого упрощения может расцениваться как приближенный доверительный интервал с уровнем доверия $P_д$:

$$\left(\hat{p} - y_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\ \hat{p} + y_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$
$$\hat{p}(\xi_1,\ldots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \tag{4.14}$$

в котором y_2 – квантиль стандартного нормального распределения $N(0,1)$ уровня $\frac{1+P_д}{2}$.

Если же в приближенном доверительном интервале (4.14) заменить корень $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ его наибольшим возможным значением:

$$\max_{0\le p\le 1}\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{n}} = \sqrt{\frac{1}{4n}},$$

тогда в качестве приближенного доверительного интервала будет получен следующий более простой (но более широкий) интервал:

$$\left(\hat{p} - y_2\sqrt{\frac{1}{4n}};\ \hat{p} + y_2\sqrt{\frac{1}{4n}}\right),$$
$$\hat{p}(\xi_1,\ldots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i, \tag{4.15}$$

в котором y_2 – квантиль стандартного нормального распределения $N(0,1)$ уровня $\frac{1+P_д}{2}$.

2. Задачи.

Задача 1.

Напишите функцию, которая для заданной матрицы реализаций выборки из нормального распределения $N(m,\sigma^2)$ и заданных величин объема выборки n, среднеквадратичного отклонения σ и уровня доверия $P_д$ возвращает матрицу реализаций доверительного интервала для m с уровнем доверия $P_д$ в соответствии с (4.7) (в каждой строке или столбце матрицы располагаются два значения: левой и правой границ доверительного интервала).

Задача 2.

Используя функцию задачи 1, исследуйте зависимость получаемых доверительных интервалов от параметров: объема выборки n, среднеквадратичного отклонения σ и уровня доверия $P_д$. Возьмите в качестве значения m величину 2, постройте по $k = 20$ реализаций выборки $(\xi_1,..., \xi_n)$ объема n из нормального распределения $N(2,\sigma^2)$ и для каждой реализации выборки вычислите границы доверительного интервала (4.7), используя наборы величин $(n,\sigma, P_д)$, представленные в следующей таблице:

№	n	σ	$P_д$
1	20	1	0.9
2	20	1	0.95
3	20	1	0.99
4	10	1	0.95
5	40	1	0.95
6	20	2	0.95
7	20	3	0.95

В системе координат, откладывая по оси абсцисс номер реализации доверительного интервала, а по оси ординат значения статистик доверительного интервала изобразите в виде двух ломанных значения левой и правой границ доверительного интервала для набора параметров 1. В этой же системе координат проведите прямую параллельную оси абсцисс и пересекающую ось ординат в точке (0, 2), соответствующей значению $m = 2$, использованному при построении реализаций выборки. Подсчитайте количество реализаций доверительного интервала, которые не содержат величину 2.

В одной системе координат изобразите в виде ломанных левые и правые границы реализаций доверительных интервалов для наборов параметров 1, 2 и 3. В двух других системах координат аналогичным образом изобразите реализации доверительных интервалов для наборов параметров 1, 4 и 5, а также 1, 6 и 7.

Задача 3.

По выборке $(\xi_1,..., \xi_n)$ из нормального распределения $N(m,\sigma^2)$ постройте нижнюю и верхнюю доверительные границы для математического ожидания m с уровнем доверия $P_д$, используя центральную статистику (4.6) и считая величину σ известной.

Напишите две функции, которые возвращают векторы реализации нижней и верхней доверительных границ для заданной матрицы реализаций выборки и заданных величин σ и $P_д$.

Постройте $k = 20$ реализаций выборки объема $n = 15$ из распределения $N(m,\sigma^2)$ для значений $m = 1$ и $\sigma = 2$, используя реализации выборки, постройте реализации нижней и верхней доверительных границ, а также реализации доверительного интервала (4.7) с одним и тем же уровнем доверия $P_д = 0.97$. В одной системе координат изобразите в виде ломанных реализации нижней и верхней доверительных границ, а также значения левой и правой границ доверительного интервала, откладывая по оси абсцисс номер реализации, а по оси ординат значения статистик.

Задача 4.

Напишите функцию, которая для заданной матрицы реализаций выборки из нормального распределения $N(m,\sigma^2)$ и заданных величин объема выборки n и уровня доверия $P_д$ возвращает матрицу реализаций доверительного интервала (4.8) для m с уровнем доверия $P_д$.

Постройте $k = 20$ реализаций выборки объема $n = 15$ из распределения $N(m,\sigma^2)$ для значений $m = 3$ и $\sigma = 1$. На основе реализаций выборки постройте реализации доверительных интервалов (4.7) и (4.8) с уровнем доверия $P_д = 0.95$. В одной системе координат изобразите в виде ломанных левые и правые границы реализаций доверительных интервалов.

Вычислите среднюю «длину» реализаций доверительного интервала (4.8) и сравните с «длиной» реализаций доверительного интервала (4.7) (вычислите отношение «длин»).

Задача 5.

Напишите функцию, которая для заданных матрицы реализаций выборки из нормального распределения $N(m,\sigma^2)$, объема выборки n, математического ожидания m и уровня доверия $P_д$ возвращает реализации доверительного интервала (4.10) для среднеквадратичного отклонения σ, и аналогичную ей функцию, возвращающую реализации доверительного интервала (4.12).

Используя составленные функции, постройте по $k = 20$ реализаций интервалов (4.10) и (4.12) с уровнем доверия $P_д = 0.95$ по реализациям выборки объема $n = 20$ из нормального распределения $N(m,\sigma^2)$ с параметром m равным 1 и σ равным 2. В одной системе координат постройте в виде ломанных левые и правые границы реализаций доверительных интервалов.

Вычислите среднюю «длину» реализаций доверительного интервала (4.10), среднюю «длину» реализаций доверительного интервала (4.12) и вычислите отношение средних «длин».

Задача 6.

Постройте нижние и верхние доверительные границы для среднеквадратичного отклонения σ по выборке объема n из нормального распределения $N(m,\sigma^2)$ в случае известного и неизвестного значения математического ожидания m , используя центральные статистики (4.9) и (4.11).

Напишите четыре функции, которые возвращают реализации построенных нижних и верхних доверительных границ на основе матрицы реализаций выборки и необходимых параметров.

Постройте $k = 20$ реализаций выборки объема $n = 25$ из нормального распределения $N(m, \sigma^2)$, взяв в качестве m значение 4 и $\sigma = 2$. Используя реализации выборки, вычислите реализации нижних и верхних доверительных границ для среднеквадратичного отклонения σ с уровнем доверия $P_\partial = 0.97$. В одной системе координат постройте в виде ломанных реализации всех доверительных границ.

Задача 7.

Напишите функции, которые возвращают реализации доверительных интервалов (4.14) и (4.15) для заданных матрицы реализаций выборки из бинарного распределения с вероятностью p и объема выборки n.

Постройте $k = 15$ реализаций выборки из бинарного распределения с вероятностью $p = 0.4$ объема $n = 20$. Используя функции, вычислите реализации доверительных интервалов (4.14) и (4.15) с уровнем доверия $P_\partial = 0.95$. Изобразите в одной системе координат в виде ломанных левые и правые границы вычисленных реализаций доверительных интервалов.

Задача 8.

По выборке $(\xi_1,..., \xi_n)$ объема n из распределения Пуассона $Po(\lambda)$ постройте нижнюю и верхнюю доверительные границы для параметра λ с заданным уровнем доверия P_∂, используя асимптотически нормальное распределении статистики $\sum_{i=1}^{n} \xi_1$.

Напишите функции, которые возвращают реализации нижней и верхней доверительных границ для параметра λ. Постройте $k = 10$ реализаций выборки объема $n = 20$ из распределения Пуассона $Po(\lambda)$ с параметром $\lambda = 2$ и вычислите реализации нижней и верхней доверительных границ с уровнем доверия $P_\partial = 0.95$. Изобразите в одной системе координат в виде ломанных вычисленные реализации доверительных границ.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: текст функции, вычисляющей реализации доверительного интервала (4.7).

Задача 2: изображение ломаных для реализаций левой и правой границ доверительного интервала для набора параметров 1 и количество реализаций интервалов, не содержащих величину 2. Три изображения, на каждом из которых представлены в виде ломанных левые и правые границы доверительных интервалов для набора параметров 1, 2 и 3, набора параметров 1, 4 и 5, а также набора параметров 1, 6 и 7.

Задача 3: статистики для нижней и верхней доверительных границ, тексты функций вычисления статистик доверительных границ и изображение, на котором в виде ломаных изображены реализации нижней и верхней доверительных границ, а также левой и правой границ доверительного интервала (4.7).

Задача 4: текст функции вычисления реализаций доверительного интервала (4.8), изображение, на котором представлены левые и правые границы доверительных интервалов (4.7) и (4.8), вычисленное отношение «длин» реализаций доверительных интервалов.

Задача 5: тексты функций вычисления реализаций доверительных интервалов (4.10) и (4.12), изображение, на котором в виде ломанных отображены левые и правые границы доверительных интервалов, и вычисленное отношение средних «длин».

Задача 6: статистики нижних и верхних доверительных границ, тексты функций вычисления статистик всех указанных доверительных границ и изображение, на котором представлены в виде ломанных реализации всех указанных доверительных границ.

Задача 7: тексты функций вычисления реализаций доверительных интервалов (4.14) и (4.15) и изображение, на котором отображены реализации доверительных интервалов (4.14) и (4.15).

Задача 8: статистики нижней и верхней доверительных границ, тексты функций вычисления статистик доверительных границ и изображение реализаций доверительных границ в виде ломанных.

4. Вопросы и задания.

1) Сформулируйте определения доверительного интервала, верхней и нижней доверительной границ.

2) Сформулируйте общий метод построения доверительных интервалов и границ с использованием центральной статистики.

3) Как изменяются левая и правая границы доверительного интервала (4.7) при увеличении и уменьшении: уровня доверия $P_д$, среднеквадратичного отклонения σ, объема выборки n?

4) Как будет изменяться отношение «длин» доверительных интервалов (4.7) и (4.8) с увеличением объема выборки n?

5) Для некоторой величины τ построены доверительный интервал (T_1, T_2), нижняя T_l и верхняя T_u доверительные границы с одинаковым уровнем доверия $P_д$. Как соотносятся между собой статистики T_1 и T_l, и статистики T_2 и T_u (больше, меньше или равны, с какой вероятностью и почему)? Можно ли из статистик T_l и T_u образовать интервал (T_l, T_u) (бу-

дет ли статистика T_l всегда меньше T_u при одной и той же реализации совокупности наблюдений)? Можно ли считать интервал (T_l, T_u) доверительным интервалом для τ, и если можно, то каков уровень доверия интервала (T_l, T_u) ?

Лабораторная работа №5. **Проверка статистических гипотез.**

1. Основные сведения.

Пусть совокупность наблюдений $(\xi_1,..., \xi_n)$ имеет некоторое распределение P^*, которое неизвестно точно, а известно лишь, что распределение P^* принадлежит некоторому классу распределений $\wp$. При заданных условиях требуется с некоторой степенью уверенности подтвердить или опровергнуть различные предположения, иначе называемые *гипотезами*, о свойствах неизвестного распределения P^*. Утверждаемым свойствам, как правило, удовлетворяет не только само распределение P^*, но, возможно, и некоторые другие распределения, которые в общем случае образуют некоторое множество $\wp_0$ ($\wp_0 \subseteq \wp$). Множество распределений $\wp_0$ соответствует проверяемой гипотезе, которую называют *основной гипотезой* и, как правило, обозначают H_0, все оставшиеся распределения из множества $\wp \setminus \wp_0$, которые называют *альтернативными распределениями*, разбивают на одно или несколько подмножеств $\wp_i$ ($i = \overline{1,k}$), каждое из которых определяет свою гипотезу H_i, которые в совокупности называют *альтернативными гипотезами*. Если множество $\wp_i$ состоит всего из одного элемента, то соответствующую гипотезу H_i называют *простой*, в противном случае (когда множество $\wp_i$ состоит из более чем одного элемента) гипотезу H_i называют *сложной*.

Основная гипотеза H_0 утверждает, что неизвестное распределение P^* принадлежит множеству $\wp_0$:

$$H_0 : P^* \in \wp_0, \tag{5.1}$$

в то время как каждая альтернативная гипотеза H_i утверждает, что неизвестное распределение P^* принадлежит множеству $\wp_i$:

$$H_i : P^* \in \wp_i.$$

Процедуру проверки основной гипотезы H_0 против альтернативных гипотез H_i ($i = \overline{1,k}$) называют критерием проверки. *Критерий проверки* является методом (способом) обработки наблюдений $(\xi_1,..., \xi_n)$, который позволяет вполне обосновано принять решение о том, можно ли считать основную гипотезу H_0 правдоподобной и принять её, или же следует признать гипотезу H_0 неправдоподобной и отклонить её.

Основной принцип проверки гипотез является практически одинаковым для всех критериев [5, 6, 8, 9]: в основе критерия проверки, как правило, располагается специальным образом сформированная *статистика критерия* $T(\xi_1,..., \xi_n)$, распределение которой должно обязательно зависеть от распределения исходных наблюдений $(\xi_1,..., \xi_n)$.

Предположим, что для некоторого малого числа α (обычно $\alpha \le 0.05$) удается найти множество Γ_α значений статистики критерия $T(\xi_1,..., \xi_n)$ такое, что вероятности события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ при всех распределениях $P_0 \in \wp_0$, определяемых основной гипотезой H_0, оказываются не больше величины α:

$$P\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha; P_0\} \le \alpha ,$$
$$P_0 \in \wp_0 . \tag{5.2}$$

В этом случае множество Γ_α может быть использовано для проверки основной гипотезы H_0. Действительно, если гипотеза H_0 верна, то согласно (5.1) неизвестное распределение P^* совокупности наблюдений $(\xi_1,..., \xi_n)$ принадлежит множеству $\wp_0$, но для всех распределений множества $\wp_0$ выполняется (5.2), следовательно (5.2) выполняется и для P^*:

$$P\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha; P^*\} \le \alpha , \tag{5.3}$$

Соотношение (5.3) показывает, что если гипотеза H_0 верна, то событие $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ должно иметь малую вероятность, не превосходящую малую величину α.

Пусть теперь в результате эксперимента получена реализация наблюдений $(x_1,..., x_n)$, для которой $T(x_1,..., x_n) \in \Gamma_\alpha$, иными словами произошло событие $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$. Появление события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ при однократном проведении эксперимента указывает на то, что вероятность события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ не является малой (появление событий с малой вероятностью при однократном проведении эксперимента на практике считается неправдоподобным). Получается противоречие: если гипотеза H_0 верна, то вероятность события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ должна быть малой (5.3), а результаты эксперимента указывают, что вероятность события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ малой не является. Полученное противоречие, хотя и лишь с некоторой долей уверенности, говорит о том, что исходное предположение о справедливости основной гипотезы H_0 или иначе предположение о том, что неизвестное распределение $P^* \in \wp_0$, было неверным, во всяком случае, результаты эксперимента не согласуются с основной гипотезой H_0 и гипотезу H_0 следуют отклонить.

Если же в результате эксперимента получена другая реализация наблюдений $(y_1,..., y_n)$, для которой $T(y_1,..., y_n) \notin \Gamma_\alpha$, иными словами событие $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ не произошло, а произошло обратное событие $\{T(\xi_1,..., \xi_n) \notin \Gamma_\alpha\}$, то в этом случае никакого противоречия не возникает. Действительно, если основная гипотеза H_0 верна, и $P^* \in \wp_0$ согласно (5.1), тогда из (5.3) следует, что вероятность обратного события:

$$P\{T(\xi_1,..., \xi_n) \notin \Gamma_\alpha \mid P^*\} = 1 - P\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha \mid P^*\} \geq 1 - \alpha ,$$

достаточно велика, если величина α мала, поэтому в этом случае полученная реализация наблюдений $(y_1,..., y_n)$ не противоречит основной гипотезе H_0, и основную гипотезу H_0 следует принять.

Таким образом, в соответствии с критерием проверки необходимо выполнить следующие действия:

1) для полученной реализации $(x_1,..., x_n)$ наблюдаемых величин $(\xi_1,..., \xi_n)$ вычислить значение статистики критерия $T(x_1,..., x_n)$;

2) определить принадлежность значения $T(x_1,..., x_n)$ области Γ_α: если вычисленное значение статистики $T(x_1,..., x_n)$ принадлежит области Γ_α, то отклонить основную гипотезу H_0, в противном случае принять основную гипотезу H_0.

Поскольку при возникновении события $\{T(\xi_1,..., \xi_n) \in \Gamma_\alpha\}$ происходит отклонение или иначе критика основной гипотезы H_0, то область Γ_α принято называть *критической областью*, а величину α – *уровнем значимости*. В общем случае уровень значимости α и критическая область Γ_α оказываются связанными условием (5.2).

В некоторых часто встречаемых случаях критическая область Γ_α имеет простой вид интервала:

$$\Gamma_\alpha = (h_\alpha ; \infty) ,$$

где h_α некоторое пороговое значение, вычисляемое в соответствии с заданным уровнем значимости α таким образом, чтобы выполнялось условие (5.2). В таких случаях проверка принадлежности $T(x_1,..., x_n) \in \Gamma_\alpha$ равносильна сравнению значения статистики $T(x_1,..., x_n)$ и порога h_α:

$$T(x_1,..., x_n) \in \Gamma_\alpha \leftrightarrow T(x_1,..., x_n) > h_\alpha .$$

Если оказывается, что значение статистики $T(x_1,..., x_n)$ больше порога h_α, то основная гипотеза H_0 отклоняется, в противном случае основная гипотеза принимается:

$$T(x_1,..., x_n) > h_\alpha \ \rightarrow \ H_0 \text{ отклоняется,} \tag{5.4}$$

$$T(x_1,..., x_n) \le h_\alpha \quad \to \quad H_0 \text{ принимается.}$$

Правило принятия решения (5.4) может быть переформулировано в терминах уровня значимости α. Пусть полученному значению статистики критерия $T(x_1,..., x_n)$ соответствует наименьший уровень значимости α^*:

$$\forall P_0 \in \wp_0 : P\{T(\xi_1,..., \xi_n) > T(x_1,..., x_n); P_0\} \le \alpha^*, \tag{5.5}$$

тогда правило, эквивалентное (5.4), имеет вид:

$$\begin{aligned} &\alpha^* < \alpha \quad \to \quad H_0 \text{ отклоняется,} \\ &\alpha^* \ge \alpha \quad \to \quad H_0 \text{ принимается.} \end{aligned} \tag{5.6}$$

Действительно, если $\alpha^* < \alpha$, тогда:

$$\forall P_0 \in \wp_0 : P\{T(\xi_1,..., \xi_n) > T(x_1,..., x_n); P_0\} < P\{T(\xi_1,..., \xi_n) > h_\alpha; P_0\},$$

откуда следует вложенность интервалов:

$$(T(x_1,..., x_n); \infty) \subset (h_\alpha; \infty),$$

или иначе:

$$T(x_1,..., x_n) > h_\alpha,$$

и в этом случае согласно правилу (5.4) гипотезу H_0 следует отклонить.

Если же $\alpha^* \ge \alpha$, то рассуждая аналогичным образом, получим в итоге, что $T(x_1,..., x_n) \le h_\alpha$ и в этом случае гипотезу H_0 следует принять. Таким образом, правила (5.4) и (5.6) являются эквивалентными, то есть для одинаковых реализаций $(x_1,..., x_n)$ совокупности наблюдений принимают одинаковые решения.

Вычисление уровня значимости α^* (5.5) может иметь самостоятельный интерес: в соответствии с критерием проверки при всех уровнях значимости $\alpha > \alpha^*$ основная гипотеза H_0 отклоняется критерием проверки, отсюда следует, что уровень значимости α^* является наименьшим уровнем значимости, при котором отклоняется основная гипотеза H_0 при полученном значении $T(x_1,..., x_n)$ статистики критерия.

Критерий хи-квадрат проверки простой гипотезы о вероятностях.

Предположим, что совокупность наблюдений $(\nu_1,..., \nu_k)$ имеет неизвестное полиномиальное распределение $\Pi(p_1,..., p_k; n)$. Основная гипотеза H_0 утверждает, что неизвестное распределение $\Pi(p_1,..., p_k; n)$ является распределением $\Pi(p_1^0,..., p_k^0; n)$, где $p^0 = (p_1^0,..., p_k^0)$ – известный (заданный) вектор вероятностей. Альтернативная гипотеза H_1 напротив утверждает, что неизвестное распределение $\Pi(p_1,..., p_k; n)$ не является распределением $\Pi(p_1^0,..., p_k^0; n)$.

Для проверки основной гипотезы H_0 против альтернативной гипотезы H_1 может быть использован критерий хи-квадрат проверки простой гипотезы [4, 8, 10, 13], статистика которого $X_n^2(\nu_1,..., \nu_k \mid p_0)$ имеет следующий вид:

$$X_n^2(\nu_1,..., \nu_k \mid p_0) = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k} \frac{\nu_i^2}{np_i^0} - n \qquad (5.7)$$

Если основная гипотеза H_0 верна, то с ростом n распределение статистики $X_n^2(\nu_1,..., \nu_k \mid p_0)$ стремится к распределению $\chi^2(k-1)$ (согласно теореме Пирсона [8]). Если же верна альтернативная гипотеза H_1, то с «большой» вероятностью статистика критерия $X_n^2(\nu_1,..., \nu_k \mid p_0)$ принимает «достаточно большие» значения, поэтому в качестве критических областей в критерии хи-квадрат используются области вида $\Gamma_\alpha = (h_\alpha; \infty)$.

Если при постановке задачи задан уровень значимости α, то величина порогового значения h_α приближенно определяется как квантиль распределения $\chi^2(k-1)$ уровня $1-\alpha$:

$$h_\alpha \approx F^{-1}_{\chi^2(k-1)}(1-\alpha) \qquad (5.8)$$

где $F_{\chi^2(k-1)}(x)$ – функция распределения $\chi^2(k-1)$.

Если же уровень значимости α не задан, а в результате эксперимента получена реализация наблюдений $(y_1,..., y_k)$ и соответствующее значение статистики $X^* = X_n^2(y_1,..., y_k \mid p_0)$, то наименьший уровень значимости α^*, при котором отклоняется основная гипотеза, может быть приближенно вычислен из следующего соотношения:

$$\begin{aligned} &\alpha^* \approx 1 - F_{\chi^2(k-1)}(X^*), \\ &X^* = X_n^2(y_1,..., y_k \mid p_0). \end{aligned} \qquad (5.9)$$

Соотношения (5.8) и (5.9) являются приближенными, поэтому этими соотношениями на практике можно с уверенностью пользоваться только в случаях, когда:

$$\begin{aligned} &np_i^0 \geq 10, \\ &i = \overline{1, k}. \end{aligned} \qquad (5.10)$$

Критерий хи-квадрат проверки сложной гипотезы о вероятностях.

Пусть, как и ранее, совокупность наблюдений $(\nu_1,..., \nu_k)$ имеет неизвестное полиномиальное распределение $\Pi(p_1,..., p_k; n)$. Основная сложная гипотеза H_0 утверждает, что неизвестное распределение $\Pi(p_1,..., p_k; n)$ совпадает с одним из распределений вида $\Pi(p_1^0(\theta),..., p_k^0(\theta); n)$, где $p_i^0(\theta)$ – известные (заданные) функции, и θ – d-мерный параметр ($d < k$) из некоторого множества допустимых значений Θ. Альтернативная гипотеза H_1

заключается в том, что неизвестное распределение $\Pi(p_1,..., p_k; n)$ не совпадает ни с одним из распределений вида совпадает с одним из распределений вида $\Pi(p_1^0(\theta),..., p_k^0(\theta); n)$.

Для проверки основной гипотезы H_0 против альтернативной гипотезы H_1 может быть использован критерий хи-квадрат проверки сложной гипотезы [6, 8] со статистикой:

$$X_n^2(\nu_1,..., \nu_k \mid \theta^*) = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0(\theta^*))^2}{np_i^0(\theta^*)} = \sum_{i=1}^{k} \frac{\nu_i^2}{np_i^0(\theta^*)} - n$$

где $\theta^* = \theta^*(\nu_1,..., \nu_k)$ – оценка максимального правдоподобия, построенная по наблюдениям $(\nu_1,..., \nu_k)$ при условии, что совокупность наблюдений имеет распределение $\Pi(p_1^0(\theta),..., p_k^0(\theta); n)$.

Если верна основная гипотеза H_0, то с увеличением n распределение статистики $X_n^2(\nu_1,..., \nu_k \mid \theta^*)$ стремится к распределению $\chi^2(k-d-1)$ (согласно теореме Фишера [8]) если же верна альтернативная гипотеза H_1, то статистика $X_n^2(\nu_1,..., \nu_k \mid \theta^*)$ с «большой» вероятностью принимает «достаточно большие» значения.

Для заданного уровня значимости α следует использовать критическую область $\Gamma_\alpha = (h_\alpha; \infty)$, в которой величина порогового значения h_α приближенно равна квантилю распределения $\chi^2(k-d-1)$ уровня $1-\alpha$. Если по результатам эксперимента получено значение статистики $X^* = X_n^2(y_1,..., y_k \mid \theta^*)$, то наименьший уровень значимости α^*, при котором отклоняется основная гипотеза, может быть вычислен приближенно из следующего равенства:

$$\alpha^* \approx 1 - F_{\chi^2(k-d-1)}(X^*),$$

$$X^* = X_n^2(y_1,..., y_k \mid \theta^*),$$

где $F_{\chi^2(k-d-1)}(x)$ – функция распределения $\chi^2(k-d-1)$.

Проверка гипотезы о независимости признаков.

Предположим, что дискретная случайная величина ξ может принимать значения из множества натуральных чисел от 1 до k, и аналогично дискретная случайная величина η может принимать значения среди натуральных чисел от 1 до m. Пусть в результате проведения серии из n независимых экспериментов получена совокупность наблюдений $\|\nu_{ij}\|$ в виде матрицы, в которой каждая случайная величина ν_{ij} равна количеству произошедших событий $\{\xi = i, \eta = j\}$ ($1 \le i \le k$, $1 \le j \le m$).

В условиях задачи необходимо проверить основную гипотезу H_0 о независимости величин ξ и η, против альтернативной гипотезы H_1, утверждающей, что величины ξ и η являются зависимыми.

Сформулированная задача проверки основной гипотезы H_0 о независимости может быть сведена к задаче проверки сложной гипотезы о вероятностях [8, 13] со статистикой

$$X_n^2(\nu_{11},..., \nu_{km};\theta^*) = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(\nu_{ij} - n\theta^*_{\xi,i}\theta^*_{\eta,j})^2}{n\theta^*_{\xi,i}\theta^*_{\eta,j}} =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\nu_{ij}^2}{n\theta^*_{\xi,i}\theta^*_{\eta,j}} - n, \qquad (5.11)$$

$$\theta^*_{\xi,i} = \frac{\sum_{j=1}^{m}\nu_{ij}}{n}, \quad \theta^*_{\eta,j} = \frac{\sum_{i=1}^{k}\nu_{ij}}{n}.$$

Для заданного уровня значимости α в качестве критической области Γ_α следует выбрать интервал $(h_\alpha;\infty)$, где h_α – квантиль уровня $1-\alpha$ распределения $\chi^2((k-1)(m-1))$. В случае, если в результате эксперимента получена реализация наблюдений $(y_{11},..., y_{km})$ и значение статистики $X^* = X_n^2(y_{11},..., y_{km};\theta^*)$, то наименьшее значение уровня значимости α^*, при котором отклоняется гипотеза о независимости, определяется следующим приближенно равенством:

$$\alpha^* \approx 1 - F_{\chi^2((k-1)(m-1))}(X^*),$$
$$X^* = X_n^2(y_{11},..., y_{km};\theta^*), \qquad (5.12)$$

где $F_{\chi^2((k-1)(m-1))}(x)$ – функция распределения $\chi^2((k-1)(m-1))$.

Критерий согласия Колмогорова.

Пусть задана выборка $(\xi_1,..., \xi_n)$ из неизвестного распределения с непрерывной и возрастающей функцией распределения $F_\xi(x)$, и рассматривается простая основная гипотеза H_0, утверждающая что неизвестная функция распределения $F_\xi(x)$ совпадает с известной (заданной) функцией распределения $F_0(x)$, против альтернативной гипотезы H_1, которая утверждает, что неизвестная функция распределения $F_\xi(x)$ не совпадает с $F_0(x)$.

Для проверки основной гипотезы H_0 против альтернативной гипотезы H_1 может использоваться критерий согласия Колмогорова [6, 13] со статистикой $D_n(\xi_1,..., \xi_n \mid F_0)$:

$$D_n(\xi_1,\ldots,\ \xi_n \mid F_0) = \sqrt{n}\sup_{-\infty<x<\infty}\left|F_n^*(x;\xi_1,\ldots,\ \xi_n) - F_0(x)\right|, \qquad (5.13)$$

где $F_n^*(x;\xi_1,\ldots,\ \xi_n)$ – эмпирическая функция распределения, построенная по выборке $(\xi_1,\ldots,\ \xi_n)$.

В случае, когда основная гипотеза H_0 не верна статистика $D_n(\xi_1,\ldots,\ \xi_n \mid F_0)$ с «большой» вероятностью принимает «большие» значения. Если же основная гипотеза H_0 верна, то распределение статистики $D_n(\xi_1,\ldots,\ \xi_n \mid F_0)$ при больших n приближенно совпадает с распределением Колмогорова с функцией распределения $K(t)$:

$$K(t) = \begin{cases} \sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2t^2} & ,t>0 \\ 0 & ,t\le 0 \end{cases} \qquad (5.14)$$

Для заданного уровня значимости α в качестве критической области Γ_α следует выбрать множество $(h_\alpha;\infty)$, где h_α – квантиль функции распределения $K(t)$ уровня $1-\alpha$. Если в результате эксперимента получено значение статистики $D^* = D_n(x_1,\ldots,\ x_n \mid F_0)$, то наименьший уровень значимости α^*, при котором отклоняется основная гипотеза, приближенно определяется из следующего соотношения:

$$\begin{gathered}\alpha^* \approx 1 - K(D^*), \\ D^* = D_n(x_1,\ldots,\ x_n \mid F_0)\end{gathered} \qquad (5.15)$$

Критерий Колмогорова-Смирнова.

Пусть имеются две выборки: $\xi^{(n)} = (\xi_1,\ldots,\ \xi_n)$ из неизвестного распределения $F_\xi(x)$ и $\eta^{(m)} = (\eta_1,\ldots,\ \eta_m)$ из неизвестного распределения $F_\eta(x)$ такие, что выборки $\xi^{(n)}$ и $\eta^{(m)}$ независимы, а функции распределения $F_\xi(x)$ и $F_\eta(x)$ непрерывны, и требуется проверить основную сложную гипотезу H_0 о том, что функции распределения $F_\xi(x)$ и $F_\eta(x)$ совпадают. Альтернативная сложная гипотеза H_1 заключается в том, что функции распределения $F_\xi(x)$ и $F_\eta(x)$ различны.

Для проверки основной гипотезы H_0 против альтернативной гипотезы H_1 может использоваться критерий Колмогорова-Смирнова [13] со статистикой $D_{n,m}(\xi^{(n)},\eta^{(m)})$:

$$D_{n,m}(\xi^{(n)},\eta^{(m)}) = \sqrt{\frac{nm}{n+m}}\sup_{-\infty<x<\infty}\left|F_{\xi,n}^*(x;\xi^{(n)}) - F_{\eta,m}^*(x;\eta^{(m)})\right|, \qquad (5.16)$$

где $F^*_{\xi,n}(x;\xi^{(n)})$ – эмпирическая функция распределения, построенная для выборки $\xi^{(n)}$, и $F^*_{\eta,m}(x;\eta^{(m)})$ – эмпирическая функция распределения, построенная для выборки $\eta^{(m)}$.

Если основная гипотеза H_0 верна, то распределение статистики $D_{n,m}(\xi^{(n)},\eta^{(m)})$ стремится к распределению Колмогорова с функцией распределения (5.17) с ростом n и m. Если же основная гипотеза H_0 не верна, то статистика $D_{n,m}(\xi^{(n)},\eta^{(m)})$ с «большой» вероятностью принимает «большие» значения.

Для заданного уровня значимости α в качестве критической области Γ_α следует взять интервал $(h_\alpha;\infty)$, где h_α – квантиль функции распределения $K(t)$ уровня $1-\alpha$, а наименьший уровень значимости α^*, при котором отклоняется основная гипотеза, вычисляется приближенно из следующего соотношения:

$$\begin{gathered}\alpha^* \approx 1 - K(D^{**}),\\ D^{**} = D_{n,m}(x^{(n)}, y^{(m)}),\end{gathered} \tag{5.17}$$

где $x^{(n)}$ – реализация выборки $\xi^{(n)}$, и $y^{(m)}$ – реализация выборки $\eta^{(m)}$, полученные в результате проведения эксперимента.

Критерий Фишера.

Пусть совокупность наблюдений состоит из двух выборок: $\xi^{(n)} = (\xi_1,..., \xi_n)$ из нормального распределения $N(m_\xi,\sigma^2_\xi)$ и $\eta^{(m)} = (\eta_1,..., \eta_m)$ из нормального распределения $N(m_\eta,\sigma^2_\eta)$, причем выборки $\xi^{(n)}$ и $\eta^{(m)}$ независимы, а параметры m_ξ, σ^2_ξ, m_η и σ^2_η неизвестны. В условиях задачи необходимо проверить основную гипотезу H_0, которая утверждает равенство дисперсий, $\sigma^2_\xi = \sigma^2_\eta$, против альтернативной гипотезы H_1, которая утверждает, что дисперсии не равны, $\sigma^2_\xi \neq \sigma^2_\eta$.

Для проверки основной гипотезы может использоваться критерий Фишера [13] со статистикой $f_{n,m}(\xi^{(n)},\eta^{(m)})$:

$$f_{n,m}(\xi^{(n)},\eta^{(m)}) = \frac{\tilde{\mu}_2(\xi^{(n)})}{\tilde{\mu}_2(\eta^{(m)})}, \tag{5.18}$$

где $\tilde{\mu}_2(\xi^{(n)})$ и $\tilde{\mu}_2(\eta^{(m)})$ – исправленные выборочные дисперсии:

$$\tilde{\mu}_2(\xi^{(n)}) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \hat{m}_1(\xi^{(n)}))^2, \ \hat{m}_1(\xi^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i,$$

$$\tilde{\mu}_2(\eta^{(m)}) = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j - \hat{m}_1(\eta^{(m)}))^2, \ \hat{m}_1(\eta^{(m)}) = \frac{1}{m}\sum_{j=1}^{m}\eta_j.$$

Если основная гипотеза H_0 верна, то статистика $f_{n,m}(\xi^{(n)},\eta^{(m)})$ имеет распределение Фишера $F(n-1,m-1)$, и при одновременном увеличении n и m значения статистики $f_{n,m}(\xi^{(n)},\eta^{(m)})$ «концентрируются» в окрестности единицы. Если же основная гипотеза H_0 не верна, то с большой вероятностью наблюдаются отклонения значений статистики $f_{n,m}(\xi^{(n)},\eta^{(m)})$ от единицы.

В качестве критической области Γ_α в критерии Фишера может быть выбрана двусторонняя область следующего вида:

$$\Gamma_\alpha = (0;h_{\alpha/2}) \cup (h_{1-\alpha/2};\infty), \tag{5.19}$$

где $h_{\alpha/2}$ и $h_{1-\alpha/2}$ – квантили распределения Фишера $F(n-1,m-1)$ уровней $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ соответственно.

Критерий Стьюдента.

Пусть заданы две выборки: $\xi^{(n)}=(\xi_1,\ldots,\xi_n)$ из нормального распределения $N(m_\xi,\sigma^2)$ и $\eta^{(m)}=(\eta_1,\ldots,\eta_m)$ из нормального распределения $N(m_\eta,\sigma^2)$ такие, что выборки $\xi^{(n)}$ и $\eta^{(m)}$ независимы и параметры m_ξ, m_η и σ^2 неизвестны, и необходимо проверить основную гипотезу H_0 о равенстве математических ожиданий, $m_\xi = m_\eta$, против альтернативной гипотезы H_1, утверждающей отсутствие равенства, $m_\xi \neq m_\eta$.

Для проверки основной гипотезы H_0 против альтернативной H_1 может использоваться критерий Стьюдента [4, 13] со статистикой $t_{n,m}(\xi^{(n)},\eta^{(m)})$:

$$t_{n,m}(\xi^{(n)},\eta^{(m)}) = \sqrt{\frac{nm}{n+m}}\frac{\hat{m}_1(\xi^{(n)})-\hat{m}_1(\eta^{(m)})}{\sqrt{\dfrac{(n-1)\tilde{\mu}_2(\xi^{(n)})+(m-1)\tilde{\mu}_2(\eta^{(m)})}{n+m-2}}}, \tag{5.20}$$

в которой $\hat{m}_1(\xi^{(n)})$ и $\hat{m}_1(\eta^{(m)})$ – выборочные средние, а $\tilde{\mu}_2(\xi^{(n)})$ и $\tilde{\mu}_2(\eta^{(m)})$ – исправленные выборочные дисперсии:

$$\tilde{\mu}_2(\xi^{(n)}) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\hat{m}_1(\xi^{(n)}))^2,\ \hat{m}_1(\xi^{(n)}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i,$$

$$\tilde{\mu}_2(\eta^{(m)}) = \frac{1}{m-1}\sum_{j=1}^{m}(\eta_j-\hat{m}_1(\eta^{(m)}))^2,\ \hat{m}_1(\eta^{(m)}) = \frac{1}{m}\sum_{j=1}^{m}\eta_j.$$

Если основная гипотеза H_0 верна, то статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ имеет распределение Стьюдента $T(n+m-2)$. Если же напротив основная гипотеза H_0 не верна, то статистика $t_{n,m}(\xi^{(n)},\eta^{(m)})$ с «большой» вероятностью принимает «достаточно большие» по абсолютной величине значения и в качестве критической области необходимо брать двустороннюю область Γ_α:

$$\Gamma_\alpha = (-\infty; h_{\alpha/2}) \cup (h_{1-\alpha/2}; \infty), \tag{5.21}$$

где $h_{\alpha/2}$ и $h_{1-\alpha/2}$ – квантили распределения Стьюдента $T(n+m-2)$ уровней $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ соответственно.

2. Задачи.

Задача 1.

Напишите функцию проверки гипотезы по критерию хи-квадрат проверки простой гипотезы о вероятностях, которая принимает в качестве аргументов реализацию наблюдений $(\nu_1,..., \nu_k)$, вектор гипотетических вероятностей $p^0 = (p_1^0,..., p_k^0)$ и уровень значимости α, и возвращает:

1) номер принятой гипотезы (0, если принимается основная гипотеза H_0, и 1, если принимается альтернативная гипотеза H_1);

2) вычисленное значение статистики $X_n^2(\nu_1,..., \nu_k \mid p_0)$ (5.7);

3) величину порога h_α (5.8).

С помощью составленной функции выполните проверку указанных гипотез в следующих вариантах.

Сформируйте $l = 10$ реализаций совокупности наблюдений $(\nu_1,..., \nu_k)$, имеющих полиномиальное распределение $\Pi(0.2, 0.45, 0.15, 0.2; 100)$. Для основной гипотезы H_0 с вектором вероятностей $p^0 = (0.2, 0.45, 0.15, 0.2)$ выполните проверку условия практической применимости (5.10) и для каждой реализации выполните проверку гипотезы H_0 с уровнем значимости $\alpha = 0.05$. Подсчитайте, сколько раз из общего количества l критерием была принята верная гипотеза H_0.

Сформируйте $l = 10$ реализаций совокупности наблюдений $(\nu_1,..., \nu_k)$, имеющих полиномиальное распределение $\Pi(0.4, 0.25, 0.15, 0.2; 100)$. Для основной гипотезы H_0 с вектором вероятностей $p^0 = (0.2, 0.45, 0.15, 0.2)$ выполните проверку условия практической применимости (5.10) и для каждой реализации выполните проверку основной гипотезы H_0 с уровнем значимости $\alpha = 0.1$. Подсчитайте, сколько раз из общего количества l критерием была отклонена неверная гипотеза H_0.

Задача 2.

Для критерия хи-квадрат проверки простой гипотезы о вероятностях напишите функцию, которая для заданных реализации наблюдений $(\nu_1,..., \nu_k)$ и вектора гипотетических вероятностей $p^0 = (p_1^0,..., p_k^0)$ возвращает:

1) наименьший уровень значимости α^* (5.9);

2) значение статистики $X_n^2(\nu_1,..., \nu_k \mid p_0)$ (5.7).

Сформируйте $l = 10$ реализаций совокупности наблюдений $(\nu_1,..., \nu_k)$, имеющей полиномиальное распределение $\Pi(0.3, 0.2, 0.35, 0.15; 70)$, и для каждой реализации вычислите наименьший уровень значимости α^*, при котором отклоняется основная гипотеза H_0 с вектором вероятностей $p^0 = (0.3, 0.2, 0.35, 0.15)$.

Задача 3.

Для критерия проверки гипотезы о независимости признаков напишите функцию, которая принимает в качестве аргумента матрицу реализаций величин $\|\nu_{ij}\|$ и возвращает:

1) наименьшее значение уровня значимости α^* (5.12), при котором отклоняется гипотеза о независимости признаков;

2) значение статистики $X_n^2(\nu_{11},..., \nu_{km}; \theta^*)$ (5.11).

Постройте реализации случайных величин ξ и η для каждого из двух вариантов.

В первом варианте сформируйте $l = 30$ реализаций случайной величины ξ, которая с равными вероятностями принимает одно значение из множества $\{1, 2, 3\}$, и столько же реализаций случайной величины η, которая с вероятностью 0.2 принимает значение 1 и с вероятностью 0.8 значение 2.

Во втором варианте сформируйте $l = 30$ реализаций случайной величины ξ, которая с равными вероятностями принимает одно значение из множества $\{1, 2, 3\}$, и столько же реализаций случайной величины η, которая принимает значение 1 с вероятностью 0.8 и 2 с вероятностью 0.2, если реализация ξ оказалась равной 1 или 2, и принимает значение 1 с вероятностью 0.1 и 2 с вероятностью 0.9, если соответствующая реализация ξ оказалась равной 3.

В каждом из двух вариантов, считая, что реализации случайных величин с одинаковым номером образуют пару, сформируйте матрицу реализаций случайных величин ν_{ij} и вычислите с помощью составленной функции наименьший уровень значимости (5.12), при котором отклоняется основная гипотеза, и значение статистики (5.11).

Задача 4.

В соответствии с критерием согласия Колмогорова напишите функцию, которая принимает в качестве аргументов реализацию выборки $(x_1,..., x_n)$ и вектор значений функции распределения $(F_0(x_1),..., F_0(x_n))$ и возвращает:

1) наименьшее значение уровня значимости α^* (5.15), при котором отклоняется основная гипотеза;

2) значение статистики $D_n(\xi_1,..., \xi_n \mid F_0)$ (5.13).

Постройте реализацию выборки объема $n = 30$ из нормального распределения $N(2,4)$ и проверьте с помощью составленной функции две гипотезы:

1) гипотезу о том, что распределение случайных величин выборки является нормальным распределением $N(2,4)$;

2) гипотезу о том, что распределение случайных величин выборки является показательным $E(0.5)$;

считая, что уровень значимости $\alpha = 0.03$.

Постройте реализацию выборки объема $n = 30$ из показательного распределения $E(2)$ и проверьте с помощью составленной функции две гипотезы:

1) гипотезу о том, что распределение случайных величин выборки является нормальным распределением $N(0.5,1)$;

2) гипотезу о том, что распределение случайных величин выборки является показательным $E(2)$;

считая, что уровень значимости $\alpha = 0.03$.

Задача 5.

Для критерия Колмогорова-Смирнова напишите функцию, которая принимает в качестве аргументов реализации выборок $\xi^{(n)}$ и $\eta^{(m)}$ и возвращает:

1) наименьший уровень значимости α^* (5.17), при котором принимается основная гипотеза;

2) значение статистики $D_{n,m}(\xi^{(n)},\eta^{(m)})$ (5.16).

Постройте две реализации выборки объемами $n = 30$ и $m = 20$ из нормального распределения $N(3,1)$ и проверьте гипотезу о совпадении функций распределения выборок по критерию Колмогорова-Смирнова для уровня значимости $\alpha = 0.03$.

Постройте реализацию выборки объемом $n = 40$ из нормального распределения $N(2,2)$ и реализацию выборки объемом $m = 35$ из нормального распределения $N(3,1)$ и проверьте гипотезу о совпадении функций распределения по критерию Колмогорова-Смирнова для уровня значимости $\alpha = 0.05$.

Постройте реализацию выборки объемом $n = 25$ из нормального распределения $N(1,1)$ и реализацию выборки объемом $m = 30$ из показательного распределения $E(1)$ и проверьте гипотезу о совпадении функций распределения по критерию Колмогорова-Смирнова для уровня значимости $\alpha = 0.02$.

Задача 6.

Напишите функцию проверки гипотезы по критерию Фишера, которая принимает в качестве аргументов две реализации выборок $\xi^{(n)}$ и $\eta^{(m)}$ и уровень значимости α и возвращает:

1) результат проверки гипотезы: значение 0, если принимается основная гипотеза H_0, и значение 1, в противном случае (принимается альтернативная гипотеза H_1);

2) значение статистики $f(\xi^{(n)},\eta^{(m)})$ (5.18);

3) значения величин $h_{\alpha/2}$ и $h_{1-\alpha/2}$, участвующие в образовании критической области (5.19).

Постройте реализацию выборки объема $n = 30$ из нормального распределения $N(1,2)$ и реализацию выборки объема $m = 25$ из нормального распределения $N(5,2)$ и проверьте гипотезу о равенстве дисперсий по критерию Фишера для уровня значимости $\alpha = 0.04$.

Постройте реализацию выборки объема $n = 20$ из нормального распределения $N(2,2)$ и реализацию выборки объема $m = 30$ из нормального распределения $N(2,4)$ и проверьте гипотезу о равенстве дисперсий по критерию Фишера для уровня значимости $\alpha = 0.05$.

Задача 7.

Для критерия Стьюдента напишите функцию, которая принимает в качестве аргументов реализации выборок $\xi^{(n)}$ и $\eta^{(m)}$ и возвращает:

1) результат проверки гипотезы: значение 0, если принимается основная гипотеза H_0, и значение 1, в противном случае (принимается альтернативная гипотеза H_1);

2) значение статистики $t(\xi^{(n)},\eta^{(m)})$ (5.20);

3) значения величин $h_{\alpha/2}$ и $h_{1-\alpha/2}$, участвующие в образовании критической области (5.21).

Сформируйте две реализации выборки объемами $n = 30$ и $m = 25$ из нормального распределения $N(1,2)$ и с помощью составленной функции проверьте гипотезу о равенстве математических ожиданий по критерию Стьюдента для уровня значимости $\alpha = 0.02$.

Сформируйте реализацию выборки объема $n = 25$ из нормального распределения $N(3,2)$ и реализацию выборки объема $m = 30$ из нормального распределения $N(0,2)$ и с помощью составленной функции проверьте гипотезу о равенстве математических ожиданий по критерию Стьюдента для уровня значимости $\alpha = 0.03$.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: текст функции проверки гипотезы, и для каждого набора из $l = 10$ реализаций наблюдений пороговое значение h_α (5.8), соответствующее заданному уровню значимости α, а для каждой реализации: результат проверки гипотезы и значение статистики (5.7), дополнительно количества случаев, в которых принимается и отклоняется основная гипотеза.

Задача 2: текст функции, вычисляющей наименьший уровень значимости α^* (5.9), значение статистики (5.7), и результаты вычислений для каждой реализации.

Задача 3: текст функции вычисления наименьшего уровня значимости (5.12) и значения статистики (5.11), для каждого из двух вариантов матрицу реализаций наблюдений $\|v_{ij}\|$ и вычисленные наименьший уровень значимости (5.12) и значение статистики (5.11).

Задача 4: текст функции вычисления наименьшего уровня значимости (5.15) и значения статистики (5.13) и для всех четырех гипотез: вычисленные значения наименьшего уровня значимости (5.15) и статистики (5.13), а также результаты проверки гипотез.

Задача 5: текст функции вычисления наименьшего уровня значимости (5.17) и значения статистики (5.16), для всех трех гипотез: результат проверки гипотезы, вычисленные значения наименьшего уровня значимости (5.17) и статистики (5.16).

Задача 6: текст функции проверки гипотезы по критерию Фишера, и для каждой из двух гипотез: результат проверки гипотезы, вычисленное значение статистики (5.18) и пороговые значения $h_{\alpha/2}$ и $h_{1-\alpha/2}$ критической области (5.19).

Задача 7: текст функции проверки гипотезы по критерию Стьюдента, и для каждой из двух гипотез: результат проверки гипотезы, вычисленные значения статистики (5.20) и пороговые значения $h_{\alpha/2}$ и $h_{1-\alpha/2}$ критической области (5.21).

4. Вопросы и задания.

1) Что такое основная и альтернативная гипотезы?

2) Что представляет собой критерий проверки, статистика критерия, критическая область и уровень значимости?

3) В чем заключается основной принцип проверки гипотез?

4) Объясните принцип проверки гипотез, основанный на сравнении с пороговым значением и уровнем значимости в критериях с односторонней критической областью.

По каждому из рассмотренных в работе критериев необходимо ответить на следующие вопросы:

1) Что представляет собой совокупность наблюдений?

2) В чем заключается основная гипотеза?

3) В чем заключается альтернативная гипотеза?

4) Какова статистика критерия?

5) Какое распределение имеет статистика критерия, если верна основная гипотеза?

6) Какие значения принимает статистика критерия, если верна альтернативная гипотеза?

7) Какой вид имеет критическая область?

8) Как по заданному значению статистики критерия вычисляется наименьший уровень значимости, при котором отклоняется основная гипотеза?

Лабораторная работа №6. **Однофакторный дисперсионный анализ.**

1. Основные сведения.

Основная задача дисперсионного анализа [4, 8, 12, 13] заключается в ответе на вопрос: влияет ли определенный фактор (либо факторы) на наблюдаемый результат. Если в задаче оценивается влияние только одного фактора, то анализ является однофакторным.

Предположим, имеется один единственный фактор, значения (уровни) которого можно условно пронумеровать натуральными числами от 1 до k. При каждом значении фактора с номером j производится серия экспериментов, по результатам которой записываются значения совокупности наблюдаемых и представляющих интерес величин $\xi^{(j)} = (\xi_1^{(j)},..., \xi_{n_j}^{(j)})$.

Формально будем считать, что объединение всех наблюдений $\xi^{(j)}$ по всем значениям фактора $j = \overline{1,k}$ представляет собой совокупность выборок $\xi = (\xi^{(1)},..., \xi^{(k)})$:

$$\xi^{(1)} = (\xi_1^{(1)},..., \xi_{n_1}^{(1)}),$$

$$\ldots,$$

$$\xi^{(k)} = (\xi_1^{(k)},..., \xi_{n_k}^{(k)}),$$

таких, что:

1) $\xi^{(j)}$ – выборка из нормального распределения $N(a_j, \sigma^2)$, $j = \overline{1,k}$;

2) $\xi^{(1)}, \ldots, \xi^{(k)}$ совместно независимы;

3) числовые значения a_j и σ неизвестны;

4) объемы выборок n_j известны (заданы).

При этом предполагается, что каждое значение фактора с номером j может влиять только на величину соответствующего среднего значения a_j, и поэтому степень влияния фактора оценивается мерой различия величин a_j. Если средние значения a_j различаются существенно при различных значениях фактора, то отсюда следует, что фактор имеет сильное влияние на наблюдаемый результат, если же средние значения a_j различаются несущественно, то, следовательно, и влияние фактора невелико.

Таким образом, суть задачи дисперсионного анализа заключается в том, чтобы на основе выборок $\xi^{(1)}$, …, $\xi^{(k)}$ определить различаются ли существенно неизвестные математические ожидания (средние значения) a_j.

Формально решение задачи сводится к проверке основной гипотезы H_0 о том, что все неизвестные величины a_j равны между собой:

$$H_0 : a_1 = a_2 = ... = a_k ,$$

против альтернативной гипотезы H_1, утверждающей, что некоторые величины a_j являются различными.

Критерий проверки основной гипотезы H_0 заключается в сравнении двух оценок неизвестной величины дисперсии σ^2: внутригрупповой дисперсии $\hat{s}^2$ и межгрупповой дисперсии $\tilde{s}^2$ (отсюда и происходит название анализа – дисперсионный).

Внутригрупповая дисперсия $\hat{s}^2$ представляет собой сумму выборочных дисперсий $\hat{\mu}_2(\xi^{(j)})$ с весами, пропорциональными объемам выборок n_j:

$$\hat{s}^2 = \sum_{j=1}^{k} \frac{n_j}{n} \hat{\mu}_2(\xi^{(j)}) , \qquad (6.1)$$

$$\hat{\mu}_2(\xi^{(j)}) = \frac{1}{n_j} \sum_{i=1}^{n_j} (\xi_i^{(j)} - \hat{m}_1(\xi^{(j)}))^2 , \; n = \sum_{j=1}^{k} n_j .$$

Величины выборочных дисперсий $\hat{\mu}_2(\xi^{(j)})$ являются оценками дисперсии σ^2 и не зависят от величин a_j, поэтому значения внутригрупповой дисперсии $\hat{s}^2$ сопоставимы с величиной дисперсии σ^2 и не зависят от того являются ли величины a_j одинаковыми (верна основная гипотеза H_0) или различными (верна альтернативная гипотеза H_1). Формально можно показать [8, 13], что случайная величина $\frac{n\hat{s}^2}{\sigma^2}$ всегда (независимо от того верна гипотеза H_0 или нет) имеет распределение $\chi^2(n-k)$:

$$\frac{n\hat{s}^2}{\sigma^2} \sim \chi^2(n-k), \tag{6.2}$$

Межгрупповая дисперсия $\tilde{s}^2$ представляет собой меру «разброса» выборочных средних $\hat{m}_1(\xi^{(j)})$ по каждой выборке относительно общего среднего $\hat{m}_1(\xi)$ по всем выборкам $\xi^{(1)}$, …, $\xi^{(k)}$:

$$\tilde{s}^2 = \sum_{j=1}^{k} \frac{n_j}{n} (\hat{m}_1(\xi^{(j)}) - \hat{m}_1(\xi))^2 , \tag{6.3}$$

$$\hat{m}_1(\xi^{(j)}) = \frac{1}{n_j} \sum_{i=1}^{n_j} \xi_i^{(j)} ,\ \hat{m}_1(\xi) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \xi_i^{(j)} = \sum_{j=1}^{k} \frac{n_j}{n} \hat{m}_1(\xi^{(j)}) ,\ n = \sum_{j=1}^{k} n_j ,$$

Каждое выборочное среднее $\hat{m}_1(\xi^{(j)})$ зависит от соответствующей величины a_j (поскольку $\hat{m}_1(\xi^{(j)})$ является оценкой математического ожидания a_j), поэтому межгрупповая дисперсия $\tilde{s}^2$, представляющая «разброс» выборочных средних $\hat{m}_1(\xi^{(j)})$, характеризует и меру «разброса» (различия) величин a_j. Можно показать [8, 13], что если основная гипотеза H_0 верна, то случайная величина $\frac{n\tilde{s}^2}{\sigma^2}$ имеет распределение $\chi^2(k-1)$:

$$H_0 \text{ верна} \rightarrow \frac{n\tilde{s}^2}{\sigma^2} \sim \chi^2(n-k), \tag{6.4}$$

Если величины a_j одинаковы (верна основная гипотеза H_0), то значения межгрупповой дисперсии $\tilde{s}^2$ (6.3) по величине сопоставимыми с дисперсией σ^2 и значениями внутригрупповой дисперсии $\hat{s}^2$ (6.1). Если же величины a_j существенно различаются (верна альтернативная гипотеза H_1), тогда значения межгрупповой дисперсии $\tilde{s}^2$ становятся существенно больше величины дисперсии σ^2 и значений внутригрупповой дисперсии $\hat{s}^2$. Таким образом, сравнивая величины межгрупповой $\tilde{s}^2$ и внутригрупповой дисперсий $\hat{s}^2$, можно сделать вывод о том, являются ли величины a_j одинаковыми (почти одинаковыми) или существенно различными.

Формально в качестве статистики критерия $T(\xi)$ используется «отношение» межгрупповой (6.3) и внутригрупповой (6.1) дисперсий:

$$T(\xi) = \frac{\dfrac{\tilde{s}^2}{k-1}}{\dfrac{\hat{s}^2}{n-k}} . \tag{6.5}$$

Если верна основная гипотеза H_0 (величины a_j одинаковы), то значения статистики критерия $T(\xi)$ с «большой» вероятностью располагаются в окрестности единицы, если же верна альтернативная гипотеза H_1 (величины a_j различны), то значения статистики критерия $T(\xi)$ с «большой» вероятностью оказываются намного больше единицы. Отсюда следует, что «большие» значения статистики $T(\xi)$ свидетельствуют против основной гипотезы H_0 и в качестве критических областей Γ_α следует выбирать интервалы вида $(h_\alpha; \infty)$, где h_α – пороговое значение, определяемое в соответствии с заданным уровнем значимости α.

Из (6.2) и (6.4) следует, что если основная гипотеза H_0 верна, то статистика критерия $T(\xi)$ (6.5) имеет распределение Фишера $F(k-1, n-k)$:

$$T(\xi) = \frac{\tilde{s}^2}{k-1}\frac{n-k}{\hat{s}^2} \sim F(k-1, n-k), \tag{6.6}$$

поэтому при заданном уровне значимости α в качестве h_α следует использовать квантиль распределения Фишера $F(k-1, n-k)$ уровня $1-\alpha$:

$$h_\alpha = F^{-1}_{F(k-1,n-k)}(1-\alpha) \tag{6.7}$$

где $F_{F(k-1,n-k)}(x)$ – функция распределения $F(k-1, n-k)$.

Из (6.6) также следует, что если по результатам эксперимента получено значение статистики критерия (6.5) T^*, то наименьший уровень значимости α^*, при котором отклоняется основная гипотеза H_0, может быть найден из следующего равенства:

$$\alpha^* = 1 - F_{F(k-1,n-k)}(T^*) \tag{6.8}$$

2. Задачи.

Перед выполнением задач рекомендуется составить функции вычисления внутригрупповой $\hat{s}^2$ (6.1) и межгрупповой $\tilde{s}^2$ (6.3) дисперсий.

Задача 1.

Пусть совокупность наблюдений $\xi = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ образована тремя выборками $\xi^{(1)}$, $\xi^{(2)}$ и $\xi^{(3)}$:

1) $\xi^{(1)} = (\xi_1^{(1)}, ..., \xi_{n_1}^{(1)})$ – выборка из нормального распределения $N(a_1, 0.25)$ объемом $n_1 = 35$;

2) $\xi^{(2)} = (\xi_1^{(2)}, ..., \xi_{n_2}^{(2)})$ – выборка из нормального распределения $N(a_2, 0.25)$ объемом $n_2 = 30$;

3) $\xi^{(3)} = (\xi_1^{(3)}, ..., \xi_{n_3}^{(3)})$ – выборка из нормального распределения $N(a_3, 0.25)$ объемом $n_3 = 45$;

где $a_1 = a_2 = a_3 = 1$.

Постройте реализации $(x_1^{(j)},..., x_{n_j}^{(j)})$ трех выборок $\xi^{(j)}$ $(j = \overline{1,3})$ и проверьте основную гипотезу H_0 о равенстве всех параметров, $a_1 = a_2 = a_3$ для уровня значимости $\alpha = 0.05$: вычислите внутригрупповую $\hat{s}^2$ (6.1) и межгрупповую $\tilde{s}^2$ (6.3) дисперсии, значение статистики критерия $T(\xi)$ (6.5) и сравните с пороговым значением h_α (6.7).

В одной системе координат изобразите точками компоненты $x_i^{(j)}$ реализаций всех трех выборок, откладывая по оси абсцисс номер компоненты i и по оси ординат числовые значения $x_i^{(j)}$.

Задача 2.

Пусть совокупность наблюдений $\xi = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ образована тремя выборками $\xi^{(1)}$, $\xi^{(2)}$ и $\xi^{(3)}$:

1) $\xi^{(1)} = (\xi_1^{(1)},..., \xi_{n_1}^{(1)})$ – выборка из нормального распределения $N(a_1, 0.25)$ объемом $n_1 = 40$, где $a_1 = 1$;

2) $\xi^{(2)} = (\xi_1^{(2)},..., \xi_{n_2}^{(2)})$ – выборка из нормального распределения $N(a_2, 0.25)$ объемом $n_2 = 55$, где $a_2 = 2.5$;

3) $\xi^{(3)} = (\xi_1^{(3)},..., \xi_{n_3}^{(3)})$ – выборка из нормального распределения $N(a_3, 0.25)$ объемом $n_3 = 50$, где $a_3 = -1$.

Постройте реализации $(x_1^{(j)},..., x_{n_i}^{(j)})$ трех выборок $\xi^{(j)}$ $(j = \overline{1,3})$ и проверьте основную гипотезу H_0 о равенстве всех параметров $a_1 = a_2 = a_3$: вычислите внутригрупповую $\hat{s}^2$ (6.1) и межгрупповую $\tilde{s}^2$ (6.3) дисперсии, значение статистики критерия $T(\xi)$ (6.5) и наименьший уровень значимости α^* (6.8) при котором отклоняется основная гипотеза H_0.

В одной системе координат изобразите точками компоненты $x_i^{(j)}$ реализаций всех трех выборок, откладывая по оси абсцисс номер компоненты i и по оси ординат числовые значения $x_i^{(j)}$.

Задача 3.

Пусть совокупность наблюдений $\xi = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)}, \xi^{(4)})$ образована четырьмя выборками $\xi^{(1)}$, $\xi^{(2)}$, $\xi^{(3)}$ и $\xi^{(4)}$:

1) $\xi^{(1)} = (\xi_1^{(1)},..., \xi_{n_1}^{(1)})$ – выборка из нормального распределения $N(a_1, 2)$ объемом $n_1 = 30$, где $a_1 = 2$;

2) $\xi^{(2)} = (\xi_1^{(2)},..., \xi_{n_2}^{(2)})$ – выборка из нормального распределения $N(a_2, 2)$ объемом $n_2 = 25$, где $a_2 = 2.3$;

3) $\xi^{(3)} = (\xi_1^{(3)},..., \xi_{n_3}^{(3)})$ – выборка из нормального распределения $N(a_3, 2)$ объемом $n_3 = 40$, где $a_3 = 1.8$;

4) $\xi^{(4)} = (\xi_1^{(4)},..., \xi_{n_4}^{(4)})$ – выборка из нормального распределения $N(a_4, 2)$ объемом $n_4 = 35$, где $a_4 = 1.95$;

Постройте реализации $(x_1^{(j)},..., x_{n_j}^{(j)})$ четырех выборок $\xi^{(j)}$ $(j = \overline{1,4})$ и проверьте основную гипотезу H_0 о равенстве всех параметров $a_1 = a_2 = a_3 = a_4$: вычислите внутригрупповую $\hat{s}^2$ (6.1) и межгрупповую $\tilde{s}^2$ (6.3) дисперсии, значение статистики критерия $T(\xi)$ (6.5) и наименьший уровень значимости α^* (6.8) при котором принимается основная гипотеза H_0.

В одной системе координат изобразите точками компоненты $x_i^{(j)}$ реализаций всех четырех выборок, откладывая по оси абсцисс номер компоненты i и по оси ординат числовые значения $x_i^{(j)}$.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: тексты программ, использованные при вычислениях, вычисленные значения внутригрупповой дисперсии (6.1), межгрупповой дисперсии (6.3), статистики критерия (6.5) и порогового значения (6.7), результат проверки основной гипотезы, изображение с реализациями выборок.

Задача 2: тексты программ, использованные при вычислениях, вычисленные значения внутригрупповой дисперсии (6.1), межгрупповой дисперсии (6.3), статистики критерия (6.5) и наименьшего уровня значимости (6.8), результат проверки основной гипотезы, изображение с реализациями выборок.

Задача 3: тексты программ, использованные при вычислениях, вычисленные значения внутригрупповой дисперсии (6.1), межгрупповой дисперсии (6.3), статистики критерия (6.5) и наименьшего уровня значимости (6.8), результат проверки основной гипотезы, изображение с реализациями выборок.

4. Вопросы и задания.

1) Сформулируйте задачу однофакторного дисперсионного анализа.

2) В чем заключаются основная и альтернативная гипотезы дисперсионного анализа.

3) Как вычисляются внутригрупповая и межгрупповая дисперсии? Что отражают указанные дисперсии?

4) Какое распределение имеет внутригрупповая дисперсия?

5) Какое распределение имеет межгрупповая дисперсия в том случае, когда основная гипотеза верна?

6) В чем заключается критерий проверки основной гипотезы: какова статистика критерия, как определяется критическая область и наименьший уровень значимости, при котором принимается основная гипотеза?

Лабораторная работа №7. **Различение двух простых гипотез.**

1. Основные сведения.

Пусть совокупность наблюдений $(\xi_1,..., \xi_n)$ имеет распределение с функцией плотности вероятности $f_\xi(x_1,..., x_n \mid \theta)$, зависящей от параметра θ, который может принимать всего два различных значения θ_0 и θ_1. Значения θ_0 и θ_1 известны, но какое из этих значений принимает параметр θ неизвестно, поэтому в отношении параметра θ выдвигаются две гипотезы: основная гипотеза H_0, утверждающая, что параметр θ принимает значение θ_0, и альтернативная гипотеза H_1, утверждающая, что параметр θ принимает значение θ_1. Задача проверки основной гипотезы H_0 против альтернативной гипотезы H_1 в указанной постановке называется задачей различения двух простых гипотез [6, 13].

Пусть X – множество всех возможных реализаций наблюдений $(\xi_1,..., \xi_n)$, и K – некоторый критерий проверки основной гипотезы H_0 против альтернативной гипотезы H_1. Критерий K для каждой реализации выборки либо принимает основную гипотезу H_0, либо принимает альтернативную гипотезу H_1. Пусть X_0 – множество всех реализаций выборки, для которых критерий K принимает основную гипотезу H_0, а множество X_1 – множество реализаций выборки, для которых критерий K принимает альтернативную гипотезу H_1. Таким образом, если в результате эксперимента происходит событие $\{(\xi_1,..., \xi_n) \in X_0\}$, то критерий K принимает гипотезу H_0, а если происходит событие $\{(\xi_1,..., \xi_n) \in X_1\}$, то критерий K принимает гипотезу H_1:

$$\{(\xi_1,..., \xi_n) \in X_0\} \to K \text{ принимает } H_0,$$
$$\{(\xi_1,..., \xi_n) \in X_1\} \to K \text{ принимает } H_1.$$

Легко видеть, что пара множеств (X_0, X_1) образует разбиение множества X, и каждый критерий K однозначным образом определяет соответствующую пару множеств (X_0, X_1), образующих разбиение множества X. Верно и обратное, каждая пара множеств (X_0, X_1), образующих разбиение множества X, однозначно определяет некоторый критерий K.

В общем случае, критерий $K=(X_0, X_1)$ не является безошибочной процедурой принятия решения, поскольку, например, событие $\{(\xi_1,..., \xi_n)\in X_1\}$, которое приводит к отклонению гипотезы H_0, может происходить, и том случае, когда гипотеза H_0 является верной. Вероятность такого ошибочного решения $\alpha(K)$:

$$\alpha(K)=P\{(\xi_1,..., \xi_n)\in X_1 \mid \theta_0)\},$$

вычисляемая при условии, что совокупность наблюдений $(\xi_1,..., \xi_n)$ имеет плотность вероятности $f_\xi(x_1,..., x_n \mid \theta_0)$, называется *вероятностью ошибки первого рода*.

Ошибка другого рода происходит тогда, когда возникает событие $\{(\xi_1,..., \xi_n)\in X_0\}$, приводящее к отклонению гипотезы H_1, в том случае, когда гипотеза H_1 является верной. Вероятность такой ошибки $\beta(K)$:

$$\beta(K)=P\{(\xi_1,..., \xi_n)\in X_0 \mid \theta_1)\},$$

вычисляемая при условии, что совокупность наблюдений $(\xi_1,..., \xi_n)$ имеет плотность вероятности $f_\xi(x_1,..., x_n \mid \theta_1)$, называется *вероятностью ошибки второго рода*.

Вероятности ошибок первого и второго родов $\alpha(K)$ и $\beta(K)$ являются числовыми характеристиками критериев, которые позволяют в некоторых случаях производить сравнение критериев. Конечно, желательно было бы иметь такой критерий, для которого вероятности $\alpha(K)$ и $\beta(K)$ принимают как можно меньшие значения, но, к сожалению, в общем случае одновременно минимизировать обе вероятности не представляется возможным, поэтому широко распространены различные способы сравнения критериев, приводящие к оптимальным в том или ином смысле критериям.

Минимаксный, байесовский и наиболее мощный критерии.

Если критерии сравнивать, используя наибольшую из вероятностей, то есть величину $\delta(K)=\max\{\alpha(K), \beta(K)\}$, то среди всех критериев необходимо найти такой критерий K_m, для которого величина $\delta(K_m)$ не больше аналогичной величины $\delta(K)$ других критериев:

$$\delta(K_m)=\min_K \max\{\alpha(K), \beta(K)\},$$

такой критерий K_m называется *минимаксным*.

Если критерии сравнивать с использованием линейной функции потерь $L(K)$:

$$L(K)=a_0\alpha(K)+b_0\beta(K),$$

где $a_0 > 0$ и $b_0 > 0$ заданные числа, то среди всех критериев необходимо найти такой критерий K_b, для которого функция потерь $L(K)$ принимает наименьшее значение по сравнению с другими критериями:

$$L(K_b) = \min_K L(K),$$

такой критерий K_b называется *байесовским*.

Если рассматривать класс критериев с ограниченной вероятностью ошибки первого рода $K_\alpha = \{K : \alpha(K) \le \alpha\}$, то в таком классе необходимо найти критерий $K_p \in K_\alpha$, для которого вероятность ошибки второго рода $\beta(K_p)$ минимальна в классе K_α:

$$\forall K \in K_\alpha : \beta(K_p) \le \beta(K),$$

такой критерий K_p называется *наиболее мощным критерием в классе* K_α.

Критерии отношения вероятностей (КОВ).

Оказывается минимаксный, байесовский и наиболее мощный критерии могут быть построены в явном виде, поскольку принадлежат множеству критериев отношения вероятностей [9]. Каждый *критерий отношения вероятностей* (КОВ) $K_r(h)$ определяется пороговым значением $h \ge 0$ и выполняет проверку гипотезы по следующему правилу:

$$\begin{aligned} Z(\xi_1,..., \xi_n) < h &\ \rightarrow\ K_r(h) \text{ принимает } H_0, \\ Z(\xi_1,..., \xi_n) \ge h &\ \rightarrow\ K_r(h) \text{ принимает } H_1, \end{aligned} \tag{7.1}$$

где $Z(\xi_1,..., \xi_n)$ – *отношение правдоподобия*:

$$Z(\xi_1,..., \xi_n) = \frac{f_\xi(\xi_1,..., \xi_n \mid \theta_1)}{f_\xi(\xi_1,..., \xi_n \mid \theta_0)}, \tag{7.2}$$

и $f_\xi(x_1,..., x_n \mid \theta)$ – плотность вероятности наблюдаемых величин $(\xi_1,..., \xi_n)$.

Каждый КОВ $K_r(h)$ полностью определяется значением h, поэтому вероятности ошибок первого и второго родов для КОВ являются функциями порога h:

$$\alpha(h) = \alpha(K_r(h)),\ \beta(h) = \beta(K_r(h)).$$

КОВ $K_r(h_m)$, в котором величина h_m определяется равенством $\alpha(h_m) = \beta(h_m)$, является минимаксным:

$$\begin{aligned} K_m &= K_r(h_m), \\ h_m &: \alpha(h_m) = \beta(h_m). \end{aligned} \tag{7.3}$$

КОВ $K_r(h_b)$, в котором величина $h_b = \dfrac{a_0}{b_0}$, является байесовским:

$$K_b = K_r(h_b), \tag{7.4}$$

$$h_b = \frac{a_0}{b_0}.$$

КОВ $K_r(h_\alpha)$, в котором величина h_α удовлетворяет уравнению $\alpha(h_\alpha) = \alpha$ при заданной величине α, является наиболее мощным в классе критериев $K_\alpha = \{K : \alpha(K) \le \alpha\}$:

$$\begin{gathered} K_p = K_r(h_\alpha), \\ h_\alpha : \alpha(h_\alpha) = \alpha . \end{gathered} \qquad (7.5)$$

Последовательные критерии отношения вероятностей (ПКОВ).

В рассмотренной выше постановке задачи различения двух простых гипотез количество величин n было фиксировано и требовалось найти критерий, который минимизирует либо обе вероятности ошибок (минимаксный критерий), либо их линейную комбинацию (байесовский критерий), либо минимизирует вероятность ошибки второго рода, при фиксированной верхней границе для вероятности первого рода (наиболее мощный критерий).

Задача различения двух простых гипотез может рассматриваться и в другой постановке, в которой заданы (фиксированы) верхние границы для вероятностей первого и второго родов, а количество наблюдаемых величин n фиксированным не является. В приведенной постановке требуется найти такой критерий, который в среднем требует минимальное количество величин n для принятия решения, при условии что вероятности ошибок не превышают установленные верхние границы. Решение принимается критерием при таком n, при котором количество информации оказывается достаточным для принятия «надежного» решения, то есть такого решения, которое может оказаться ошибочным с вероятностями ошибок не более заданных.

Формально постановка последней задачи имеет следующий вид: пусть наблюдения образуют счетную последовательность случайных величин $(\xi_1, \xi_2, \ldots)$, для которой известна, например, последовательность функций плотности вероятности $f_{\xi,k}(x_1, \ldots, x_k \mid \theta)$ первых k величин последовательности (параметр θ, как и ранее, может принимать одно из двух значений θ_0 и θ_1). В такой постановке для проверки основной гипотезы H_0 против альтернативной гипотезы H_1 могут применяться последовательные критерии [3].

На каждом шаге $k = 0,1,2,\ldots$ *последовательный критерий* располагает совокупностью величин $\xi^{(k)} = (\xi_1, \ldots, \xi_k)$ (при $k = 0$ величины отсутствуют), на основе которого может принимать одно из следующих решений:

1) остановиться и принять гипотезу H_0;

2) продолжить и перейти к шагу $k+1$, получив совокупность величин $\xi^{(k+1)} = (\xi_1,..., \xi_k, \xi_{k+1})$;

3) остановиться и принять гипотезу H_1.

Легко видеть, что последовательный критерий может выполнять различное количество шагов до остановки, поэтому для последовательных критериев дополнительно рассматривается случайная величина $\nu(K)$ равная количеству шагов до остановки и принятия одной из гипотез. Кроме того как и всякий критерий каждый последовательный критерий K характеризуется вероятностями ошибок первого и второго родов $\alpha(K)$ и $\beta(K)$.

Оказывается, что среди всех последовательных критериев с ограниченными вероятностями ошибок $\alpha(K) < \alpha$ и $\beta(K) < \beta$ существует критерий с наименьшими средними количествами шагов до остановки $M_{\theta_0}[\nu(K)]$ и $M_{\theta_1}[\nu(K)]$ (где математические ожидания $M_{\theta_i}[\nu(K)]$ вычисляются при условии, что параметр θ принимает значение θ_i), и таким критерием является последовательный критерий отношения вероятностей $K_s(A_0, A_1)$, в котором постоянные A_0 и A_1 выбираются в соответствии с величинами α и β.

Последовательный критерий отношения вероятностей (ПКОВ) [3, 5, 6] $K_s(A_0, A_1)$ на нулевом шаге $k = 0$ всегда выполняет переход к шагу $k = 1$, а на каждом последующем шаге $k = 1,2$ выполняет одно из следующих действий:

1) останавливается и принимает гипотезу H_0, если:

$$Z(\xi_1,..., \xi_n) \le A_0$$

2) переходит к следующему шагу $k+1$, если:

$$A_0 < Z(\xi_1,..., \xi_n) < A_1$$

3) останавливается и принимает гипотезу H_1, если:

$$Z(\xi_1,..., \xi_n) \ge A_1 .$$

где $Z(\xi_1,..., \xi_n)$ – отношение правдоподобия (7.2).

К сожалению, для заданных величин α и β в общем случае весьма затруднительно построить ПКОВ $K_s^* = K_s(A_0, A_1)$, для которого $\alpha(K_s^*) = \alpha$ и $\beta(K_s^*) = \beta$, потому на практике ограничиваются построением критерия $\tilde{K}_s$:

$$\tilde{K}_s = K_s\left(\frac{\beta}{1-\alpha}, \frac{1-\beta}{\alpha}\right), \tag{7.6}$$

для которого с некоторой точностью выполняются приближенные равенства $\alpha(\tilde{K}_s) \approx \alpha$ и $\beta(\tilde{K}_s) \approx \beta$.

Если в последовательности величин $(\xi_1, \xi_2, ...)$, образованной исходными наблюдениями, при любом n величины $(\xi_1, ..., \xi_n)$ являются совместно независимыми, то достаточно просто может быть получено приближенное выражение для среднего количества шагов до остановки $\nu(\tilde{K}_s)$ критерия $\tilde{K}_s$.

Пусть $f_\xi(x \mid \theta)$ – общая плотность вероятности величин ξ_i, тогда если верна основная гипотеза H_0, то приближенно среднее значение величины $\nu(\tilde{K}_s)$ определяется выражением [5, 6]:

$$M_{\theta_0}[\nu(\tilde{K}_s)] \approx \frac{\ln\left(\frac{\beta}{1-\alpha}\right)(1-\alpha) + \ln\left(\frac{1-\beta}{\alpha}\right)\alpha}{M_{\theta_0}\left[\ln\left(\frac{f_\xi(\xi \mid \theta_1)}{f_\xi(\xi \mid \theta_0)}\right)\right]}, \qquad (7.7)$$

где ξ – случайная величина, имеющая плотность вероятности $f_\xi(x \mid \theta_0)$. Если же верна альтернативная гипотеза H_1, то приближенно среднее значение величины $\nu(\tilde{K}_s)$ определяется выражением [5, 6]:

$$M_{\theta_1}[\nu(\tilde{K}_s)] \approx \frac{\ln\left(\frac{\beta}{1-\alpha}\right)\beta + \ln\left(\frac{1-\beta}{\alpha}\right)(1-\beta)}{M_{\theta_1}\left[\ln\left(\frac{f_\xi(\xi \mid \theta_1)}{f_\xi(\xi \mid \theta_0)}\right)\right]}, \qquad (7.8)$$

где ξ – случайная величина, имеющая плотность вероятности $f_\xi(x \mid \theta_1)$.

2. Задачи.

Задача 1.

Задана выборка $(\xi_1, ..., \xi_n)$ из нормального распределения $N(m, \sigma^2)$, в котором параметр m может принимать два различных известных значения m_0 и m_1 ($m_0 < m_1$), и σ – известное числовое значение.

При фиксированном объеме выборки n постройте минимаксный K_m (7.3), байесовский K_b (7.4) (считая, что a_0 и b_0 заданы) и наиболее мощный критерии K_p (считая величину α заданной), выведите соотношения для приближенного расчета вероятностей ошибок первого и второго родов.

Выполните проверку основной гипотезы H_0, утверждающей, что $m = m_0 = 0$, против альтернативной гипотезы H_1, утверждающей, что $m = m_1 = 2$, используя построенные минимаксный, байесовский (с $a_0 = 5$ и

$b_0 = 1$) и наиболее мощный (с $\alpha = 0.05$) критерии, для следующих двух вариантов.

В первом варианте постройте $l = 10$ реализаций выборки объемом $n = 100$ из нормального распределения $N(0,4)$, и для каждой реализации выполните проверку основной гипотезы H_0 каждым из трех построенных критериев. Для каждого критерия подсчитайте количество случаев отклонения основной гипотезы и вычислите приближенно вероятность ошибки первого рода.

Во втором варианте постройте $l = 10$ реализаций выборки объемом $n = 100$ из нормального распределения $N(2,4)$, и для каждой реализации выполните проверку основной гипотезы H_0 каждым из трех построенных критериев. Для каждого критерия подсчитайте количество случаев отклонения альтернативной гипотезы и вычислите приближенно вероятность ошибки второго рода.

Задача 2.

Задана последовательность $(\xi_1, \xi_2, ...)$ случайных величин, имеющих показательное распределение $E(\theta)$, где параметр $\theta \in \{\theta_0, \theta_1\}$ ($\theta_0 < \theta_1$). В отношении параметра θ выдвигается основная гипотеза H_0, утверждающая что $\theta = \theta_0$, и альтернативная гипотеза H_1, утверждающая что $\theta = \theta_1$.

Постройте ПКОВ $\tilde{K}_s$ (7.6) проверки основной гипотезы H_0 против альтернативной H_1, считая, что величины α и β заданы.

Считая, что $\alpha = 0.05$ и $\beta = 0.03$, $\theta_0 = 1$ и $\theta_1 = 2$, выполните проверку основной гипотезы с помощью ПКОВ $\tilde{K}_s$ для $l = 20$ различных реализаций последовательности $(\xi_1, \xi_2, ...)$ в двух следующих вариантах.

В первом варианте формируйте реализации последовательности $(\xi_1, \xi_2, ...)$, считая, что все величины ξ_i имеют показательное распределение $E(\theta_0)$. Для каждой реализации выполните проверку основной гипотезы H_0 по критерию $\tilde{K}_s$, определите количество шагов, выполненных $\tilde{K}_s$ до остановки, и сравните со значением $M_{\theta_0}[\nu(\tilde{K}_s)]$ (7.7). Для любых трех реализаций в одной системе координат постройте соответствующие реализации отношения правдоподобия (7.2) в виде ломанных и прямые, соответствующие границам $\frac{\beta}{1-\alpha}$ и $\frac{1-\beta}{\alpha}$. Подсчитайте количество случаев отклонения основной гипотезы.

Во втором варианте формируйте реализации последовательности $(\xi_1, \xi_2, ...)$, считая, что все величины ξ_i имеют показательное распределение $E(\theta_1)$. Для каждой реализации выполните проверку основной гипотезы H_0

по критерию $\tilde{K}_s$, определите количество шагов, выполненных $\tilde{K}_s$ до остановки, и сравните со значением $M_{\theta_1}[\nu(\tilde{K}_s)]$ (7.8). Для любых трех реализаций в одной системе координат постройте соответствующие реализации отношения правдоподобия (7.2) в виде ломанных и прямые, соответствующие границам $\frac{\beta}{1-\alpha}$ и $\frac{1-\beta}{\alpha}$. Подсчитайте количество случаев отклонения альтернативной гипотезы.

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: вывод минимаксного, байесовского и наиболее мощного критериев и вероятностей ошибок первого и второго родов для каждого критерия, тексты программ проверки основной гипотезы по каждому критерию, для всех реализаций выборки результаты расчетов и проверки основной гипотезы, и для каждого варианта количество случаев отклонения основной гипотезы.

Задача 2: вывод ПКОВ $\tilde{K}_s$ (7.6), текст программы проверки гипотезы по ПКОВ $\tilde{K}_s$, для каждой реализации результат проверки основной гипотезы, количество шагов до остановки и вычисленные приближенные значения среднего количества шагов (7.7) и (7.8), для каждого варианта количество случаев отклонения основной и альтернативной гипотез, и изображения реализаций отношения правдоподобия и прямых, соответствующих границам ПКОВ $\tilde{K}_s$.

4. Вопросы и задания.

1) Сформулируйте задачу различения двух простых гипотез с фиксированным количеством n наблюдаемых величин.

2) Что представляет собой критерий проверки и как определяются вероятности ошибок первого и второго родов?

3) Сформулируйте определения минимаксного, байесовского и наиболее мощного критериев в задаче различения двух простых гипотез.

4) Сформулируйте определение КОВ. Каким образом представляются минимаксный, байесовский и наиболее мощный критерии в виде соответствующих КОВ?

5) Сформулируйте задачу различения двух простых гипотез с нефиксированным количеством n наблюдаемых величин.

6) Каков общий принцип проверки гипотезы последовательными критериями?

7) Сформулируйте определение ПКОВ и поясните правило принятия решения ПКОВ в задаче различения двух простых гипотез.

Лабораторная работа №8. **Линейный регрессионный анализ.**

1. Основные сведения.

Регрессионный анализ – раздел математической статистики, в котором разрабатываются методы определения зависимости некоторой выбранной случайной $\tilde{\eta}$ от набора случайных величин $(\xi_1,..., \xi_k)$ [6, 7, 8, 12, 13].

В частности, в некоторых задачах регрессионного анализа предполагается, что зависимость случайной величины $\tilde{\eta}$ от величин $(\xi_1,..., \xi_k)$ имеет некоторый предопределенный параметрический вид, выражаемый соотношением:

$$\tilde{\eta} = f(\xi_1,..., \xi_k; \tilde{\theta}_1,..., \tilde{\theta}_m) + \varepsilon , \qquad (8.1)$$

в котором $f(x_1,..., x_k; \theta_1,..., \theta_m)$ – заданная, известная функция, $\tilde{\theta} = (\tilde{\theta}_1,..., \tilde{\theta}_m)$ – фиксированный, неизвестный числовой вектор и ε – некоторая неизвестная остаточная случайная величина.

Задача линейной регрессии.

В задачах линейной регрессии [6, 7, 13] дополнительно предполагается, что функция $f(x_1,..., x_k; \theta_1,..., \theta_m)$, отражающая зависимость случайной величины $\tilde{\eta}$ от величин $(\xi_1,..., \xi_k)$ в (8.1), является линейной по переменным $\theta_1, \ldots, \theta_m$:

$$f(x_1,.., x_k; \theta_1,..., \theta_m) = \sum_{j=1}^{m} \varphi_j(x_1,..., x_k)\theta_j \qquad (8.2)$$

где $\varphi_j(x_1,..., x_k)$ – известные функции ($j = \overline{1, m}$). Таким образом, с учетом (8.2) соотношение (8.1) преобразуется к следующему виду:

$$\tilde{\eta} = \sum_{j=1}^{m} \varphi_j(\xi_1,..., \xi_k)\tilde{\theta}_j + \varepsilon , \qquad (8.3)$$

Пусть проведена серия из n экспериментов, и в каждом i-ом эксперименте случайный вектор $(\xi_1,..., \xi_k)$ принимал значение реализации $(x_{i1},..., x_{ik})$, тогда результатам серии из n экспериментов соответствует совокупность наблюдений $\eta = (\eta_1,.., \eta_n)$, которые согласно (8.3) имеют следующий вид:

$$\begin{gathered}
\eta_1 = \sum_{j=1}^{m} \varphi_j(x_{11},..., x_{1k})\tilde{\theta}_j + \varepsilon_1 , \\
\ldots, \\
\eta_n = \sum_{j=1}^{m} \varphi_j(x_{n1},..., x_{nk})\tilde{\theta}_j + \varepsilon_n ,
\end{gathered} \qquad (8.4)$$

или в матричной форме:

$$\eta = Z\tilde{\theta} + \varepsilon,$$

$$Z = \begin{pmatrix} \varphi_1(x_{11},\dots, x_{1k}) & \dots & \varphi_m(x_{11},\dots, x_{1k}) \\ \dots & \dots & \dots \\ \varphi_1(x_{n1},\dots, x_{nk}) & \dots & \varphi_m(x_{n1},\dots, x_{nk}) \end{pmatrix}, \quad (8.5)$$

$$\varepsilon = (\varepsilon_1,\dots, \varepsilon_n).$$

Задача линейной регрессии заключается в нахождении наилучшей (по некоторому критерию сравнения) оценки неизвестного вектора $\tilde{\theta} = (\tilde{\theta}_1,\dots, \tilde{\theta}_m)$ на основе наблюдений $\eta = (\eta_1,\dots, \eta_n)$, заданных соотношениями (8.4), известных функций $\varphi_j(x_1,\dots, x_k)$ $(j = \overline{1,m})$ и числовых векторов $(x_{i1},\dots, x_{ik})$ $(i = \overline{1,n})$.

Оценка по методу наименьших квадратов.

Если в качестве критерия сравнения оценок используется *среднеквадратичное отклонение* $\delta(\theta)$:

$$\delta(\theta) = \|\eta - Z\theta\|_2^2, \quad (8.6)$$

то оценка $\hat{\theta} = (\hat{\theta}_1,\dots, \hat{\theta}_m)$, для которой среднеквадратичное отклонение $\delta(\theta)$ принимает наименьшее возможное значение, может быть получена по методу наименьших квадратов, при этом сама оценка $\hat{\theta} = (\hat{\theta}_1,\dots, \hat{\theta}_m)$ называется *оценкой по методу наименьших квадратов*. Можно показать, что оценка $\hat{\theta} = (\hat{\theta}_1,\dots, \hat{\theta}_m)$ удовлетворяет *нормальному уравнению*:

$$Z^T Z\hat{\theta} = Z^T\eta \quad (8.7)$$

и, кроме того, если ранг матрицы Z равен m, то нормальное уравнение имеет единственное решение $\hat{\theta}$. Из (8.7) легко видеть, что оценка $\hat{\theta} = (\hat{\theta}_1,\dots, \hat{\theta}_m)$ является случайной величиной, поскольку зависит от совокупности случайных величин $\eta = (\eta_1,\dots, \eta_n)$.

Свойства оценки по методу наименьших квадратов.

Если вектор случайных величин $\varepsilon = (\varepsilon_1,\dots, \varepsilon_n)$ в (8.4) обладает свойствами:

1) $M[\varepsilon_i] = 0$, $i = \overline{1,n}$;

2) $D[\varepsilon_i] = \sigma^2 > 0$, $i = \overline{1,n}$;

3) cov($\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, $i, j = \overline{1,n}$,

тогда оценка по методу наименьших квадратов $\hat{\theta}$ является несмещенной оценкой неизвестного вектора $\tilde{\theta}$, компоненты оценки $\hat{\theta}$ имеют наименьшую дисперсии из всех линейных по совокупности наблюдений η оценок, а дисперсионная матрица $D[\hat{\theta}]$ выражается в явном виде [6]:

$$D[\hat{\theta}] = \sigma^2 (Z^T Z)^{-1} \tag{8.8}$$

Оценка остаточной дисперсии.

Если величина *остаточной дисперсии* σ^2 неизвестна, то для приближенного вычисления σ^2 может быть использована оценка $\hat{\sigma}^2(\eta_1,..., \eta_n)$:

$$\hat{\sigma}^2(\eta_1,..., \eta_n) = \frac{\delta(\hat{\theta})}{n - m}, \tag{8.9}$$

которая является несмещенной оценкой остаточной дисперсии σ^2.

Коэффициенты детерминации.

Оценка по методу наименьших квадратов $\hat{\theta}$ может использоваться для построения регрессионных значений $\hat{\eta} = Z\hat{\theta}$, которые во многих приложениях используются как аппроксимация величин η. С целью определения качества получаемой аппроксимации и регрессии в целом используются числовая характеристика R^2, называемая *коэффициентом детерминации*:

$$R^2 = 1 - \frac{\|\eta - \hat{\eta}\|_2^2}{\|\eta - \overline{\eta}\|_2^2},$$
$$\overline{\eta} = \left(\frac{1}{n}\sum_{i=1}^{n} \eta_i,..., \frac{1}{n}\sum_{i=1}^{n} \eta_i \right), \tag{8.10}$$

где $\overline{\eta}$ – вектор, составленный из средних значений величин, образующих совокупность наблюдений η. Значения коэффициента детерминации близкие к единице являются признаком регрессии хорошего качества: чем ближе значение коэффициента детерминации к единице, тем «более точно» регрессионные значения $\hat{\eta}$ воспроизводят (аппроксимируют) реализацию наблюдений η, и тем «лучше» случайная величина $\tilde{\eta}$ описывается линейной комбинацией $\sum_{j=1}^{m} \varphi_j(\xi_1,..., \xi_k)\hat{\theta}_j$ из соотношения (8.3). Значения коэффициента детерминации близкие к нулю указывают на регрессию плохого качества, что возможно является следствием неудачного выбора функций φ_j или «слабой» зависимости величины $\tilde{\eta}$ от величин $\xi_1, \ldots, \xi_k$.

При сравнении вариантов регрессии с различным количеством случайных величин ξ_i используется *скорректированный коэффициент детерминации* R^2_{adj}:

$$R^2_{adj} = 1 - \frac{\frac{1}{n-k}\|\eta - \hat{\eta}\|_2^2}{\frac{1}{n}\|\eta - \overline{\eta}\|_2^2}, \tag{8.11}$$

Значения R_{adj}^2 близкие к единице являются признаками регрессии хорошего качества, а близкие к нулю – плохого, поэтому из двух вариантов регрессии лучшим вариантом следует считать тот вариант, для которого скорректированный коэффициент детерминации больше.

Задача нормальной линейной регрессии.

Если вектор случайных величин $\varepsilon = (\varepsilon_1,..., \varepsilon_n)$ в (8.4) имеет нормальное распределение $N(\overline{0}_n, \sigma^2 E_n)$, где $\overline{0}_n$ – нулевой вектор порядка n и E_n – единичная матрица порядка $n \times n$, то задача нахождения оценки вектора $\tilde{\theta} = (\tilde{\theta}_1,..., \tilde{\theta}_m)$ на основе наблюдений $\eta = (\eta_1,.., \eta_n)$ (8.4) называется *задачей нормальной линейной регрессии* [6, 13]. В условиях задачи нормальной линейной регрессии дополнительно могут быть построены доверительные интервалы для неизвестных компонент вектора $\tilde{\theta}$ и величины σ^2, а также критерий проверки гипотезы о равенстве нулю всех компонент вектора $\tilde{\theta}$, то есть о независимости величины $\tilde{\eta}$ от величин $\varphi_j(\xi_1,..., \xi_k)$, $j = \overline{1, m}$.

Доверительные интервалы для компонент вектора $\tilde{\theta}$ и остаточной дисперсии σ^2.

Можно показать [6], что доверительными интервалами для компонент $\tilde{\theta}_i$ неизвестного вектора $\tilde{\theta}$ с уровнем доверия $P_д$ являются интервалы следующего вида:

$$\left(\hat{\theta}_i - y\sqrt{\left\|(Z^T Z)^{-1}\right\|_{i,i} \frac{\delta(\hat{\theta})}{n-m}}; \hat{\theta}_i + y\sqrt{\left\|(Z^T Z)^{-1}\right\|_{i,i} \frac{\delta(\hat{\theta})}{n-m}} \right), \quad (8.12)$$

$$i = \overline{1, m},$$

в которых $\hat{\theta}_i$ – компоненты оценки по методу наименьших квадратов $\hat{\theta}$, $\left\|(Z^T Z)^{-1}\right\|_{i,i}$ – диагональный элемент матрицы $\left(Z^T Z\right)^{-1}$, расположенный в i-ой строке и i-ом столбце, $\delta(\hat{\theta}) = \left\|\eta - Z\hat{\theta}\right\|_2^2$ – значение среднеквадратичного отклонения для оценки $\hat{\theta}$ и y – квантиль распределения Стьюдента $T(n-m)$ уровня $\frac{1+P_д}{2}$.

Доверительным интервалом для величины остаточной дисперсии σ^2 с уровнем доверия $P_д$ будет следующий интервал:

$$\left(\frac{\delta(\hat{\theta})}{y_2}; \frac{\delta(\hat{\theta})}{y_1} \right), \quad (8.13)$$

в котором y_1 и y_2 – квантили распределения хи-квадрат $\chi^2(n-m)$ уровней $\frac{1-P_д}{2}$ и $\frac{1+P_д}{2}$ соответственно.

Критерий проверки гипотезы об отсутствии зависимости.

С целью проверки основной гипотезы H_0 о том, что неизвестный вектор $\tilde{\theta}$ является нулевым вектором (то есть в соотношении (8.4) нет зависимости величины $\tilde{\eta}$ от величин $\varphi_j(\xi_1,..., \xi_k)$, $j=\overline{1,m}$), против альтернативной гипотезы H_1 о том, вектор $\tilde{\theta}$ не является нулевым, используется критерий со статистикой $T(\eta_1,..., \eta_n)$:

$$T(\eta_1,.., \eta_n) = \frac{n-m}{m}\frac{\hat{\theta}^T Z^T Z\hat{\theta}}{\delta(\hat{\theta})}, \tag{8.14}$$

которая имеет распределение Фишера $F(m,n-m)$, если основная гипотеза H_0 верна, и с «большой» вероятностью принимает «большие» значения, если верна альтернативная гипотеза H_1. Наименьший уровень значимости α^*, при котором отклоняется основная гипотеза H_0, определяется из следующего соотношения:

$$\alpha^* = 1 - F_{F(m,n-m)}(T(\eta_1,..., \eta_n)) \tag{8.15}$$

где $F_{F(m,n-m)}(t)$ – функция распределения $F(m,n-m)$.

2. Задачи.

Перед выполнением задач рекомендуется составить отдельные функции для вычисления: оценки по методу наименьших квадратов из уравнения (8.7), среднеквадратичного отклонения (8.6), дисперсионной матрицы (8.8), оценки остаточной дисперсии (8.9), коэффициента детерминации (8.10), скорректированного коэффициента детерминации (8.11), реализаций доверительных интервалов (8.12) и (8.13), а также наименьшего уровня значимости (8.15).

Задача 1.

Пусть случайные величины $\tilde{\eta}$ и ξ_1 связаны соотношением:

$$\tilde{\eta} = 1 + 2\xi_1 + \varepsilon,$$

причем случайная величина ξ_1 имеет равномерное распределение $R[-1,2]$, и согласно (8.3) вектор $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) = (1,2)$, а функции $\varphi_1(\xi_1) = 1$ и $\varphi_2(\xi_1) = \xi_1$.

Постройте $n = 50$ реализаций $x_{i,1}$ случайной величины ξ_1 и сформируйте реализацию наблюдений $\eta = (\eta_1,.., \eta_n)$:

$$\eta_i = 1 + 2x_{i,1} + \varepsilon_i,$$

$$(i = \overline{1,n}),$$

считая, что случайные величины ε_i имеют равномерное распределение $R[-0.5,0.5]$ (при вычислении реализаций η_i вместо ε_i необходимо подставлять реализации случайной величины, имеющей равномерное распределение $R[-0.5,0.5]$).

Сформируйте матрицу Z, найдите оценку по методу наименьших квадратов $\hat{\theta}$ из нормального уравнения (8.7), вычислите дисперсионную матрицу $D[\hat{\theta}]$ (8.8) и среднеквадратичное отклонение $\delta(\hat{\theta})$ согласно (8.6). Сравните вектор $\tilde{\theta}$ с оценкой $\hat{\theta}$. Вычислите оценку остаточной дисперсии $\hat{\sigma}^2$ (8.9) и сравните с величиной остаточной дисперсии σ^2. Вычислите коэффициент детерминации R^2 в соответствии с (8.10).

В одной системе координат постройте n точек $(x_{i,1},\eta_i)$ и отрезки прямых $y = \tilde{\theta}_1 + \tilde{\theta}_2 x$ и $y = \hat{\theta}_1 + \hat{\theta}_2 x$ для $x \in [-1,2]$.

Задача 2.

Пусть случайные величины $\tilde{\eta}$ и ξ_1 связаны соотношением:

$$\tilde{\eta} = 3 + 2\xi_1 \sin \xi_1 + \varepsilon,$$

где случайная величина ξ_1 имеет равномерное распределение $R[0,\pi]$ и вектор $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) = (3,2)$.

Постройте $n = 100$ реализаций $x_{i,1}$ случайной величины ξ_1 и сформируйте реализацию наблюдений $\eta = (\eta_1,..., \eta_n)$:

$$\eta_i = 3 + 2x_{i,1} \sin x_{i,1} + \varepsilon_i,$$

$$(i = \overline{1,n}),$$

для двух различных вариантов:

1) случайные величины ε_i имеют равномерное распределение $R[-0.2,0.2]$;

2) случайные величины ε_i имеют равномерное распределение $R[-1,1]$.

Для каждого из двух вариантов вычислите: оценку по методу наименьших квадратов $\hat{\theta}$ (8.7), дисперсионную матрицу $D[\hat{\theta}]$ (8.8), среднеквадратичное отклонение $\delta(\hat{\theta})$ (8.6), оценку остаточной дисперсии $\hat{\sigma}^2$ (8.9) (сравните с величиной остаточной дисперсии σ^2), коэффициент детерминации R^2 (8.10).

Для каждого из двух вариантов по отдельности в одной системе координат постройте n точек $(x_{i,1},\eta_i)$ и графики кривых $y = \tilde{\theta}_1 + \tilde{\theta}_2 x \sin x$ и $y = \hat{\theta}_1 + \hat{\theta}_2 x \sin x$ для $x \in [0,\pi]$.

Задача 3.

Пусть случайные величины $\tilde{\eta}$ и (ξ_1,ξ_2) связаны соотношением:

$$\tilde{\eta} = 2 - 3\xi_1^2 + \xi_1\xi_2 + \varepsilon,$$

где случайная величина ξ_1 имеет нормальное распределение $N(0,1)$, а случайная величина ξ_2 – показательное $E(0.5)$, вектор $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3) = (2,-3,1)$.

Постройте $n = 80$ реализаций $x_{1,i}$ случайной величины ξ_1 и столько же реализаций $x_{2,i}$ случайной величины ξ_2. Используя построенные реализации $x_{1,i}$ и $x_{2,i}$, сформируйте реализацию наблюдений $\eta = (\eta_1,.., \eta_n)$:

$$\eta_i = 2 - 3x_{i,1}^2 + x_{i,1}x_{i,2} + \varepsilon_i,$$

$$(i = \overline{1,n}),$$

считая, что случайные величины ε_i имеют равномерное распределение $R[-0.3,0.3]$.

Рассмотрите следующие варианты построения регрессии:

1) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2\xi_1 + \varepsilon$;

2) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2\xi_2 + \tilde{\theta}_3\xi_2^2$;

3) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2\xi_1^2 + \tilde{\theta}_3\xi_1\xi_2 + \varepsilon$;

4) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2\xi_1^2 + \tilde{\theta}_3\xi_1\xi_2 + \tilde{\theta}_4\xi_2^2 + \varepsilon$.

Для каждого варианта регрессии сформируйте свою матрицу Z, вычислите оценку по методу наименьших квадратов $\hat{\theta}$ из нормального уравнения (8.7) и скорректированный коэффициент детерминации R_{adj}^2 (8.11).

Определите наилучший вариант регрессии, основываясь на вычисленных значениях R_{adj}^2. Для наилучшего варианта регрессии определите дисперсионную матрицу $D[\hat{\theta}]$ (8.8) и оценку остаточной дисперсии $\hat{\sigma}^2$ (8.9).

Задача 4.

Пусть случайная величина $\tilde{\eta}$ и случайные величины (ξ_1, ξ_2, ξ_3) связаны соотношением:

$$\tilde{\eta} = 1 + 2\xi_1^2 \cos\xi_2 - \xi_2 \ln\xi_3 + \varepsilon,$$

где случайная величина ξ_1 имеет нормальное распределение $N(2,4)$, ξ_2 – равномерное $R\left[-\frac{\pi}{2}; \frac{\pi}{2}\right]$ и ξ_3 – показательное $E(0.25)$, а вектор $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3) = (1,2,-1)$.

Постройте $n = 20$ реализаций $x_{i,1}$ случайной величины ξ_1, столько же реализаций $x_{i,2}$ случайной величины ξ_2 и столько же реализаций $x_{i,3}$ случайной величины ξ_3. Используя построенные реализации $x_{i,1}$, $x_{i,2}$ и $x_{i,3}$ сформируйте реализацию наблюдений $\eta = (\eta_1,.., \eta_n)$:

$$\eta_i = 1 + 2x_{i,1}^2 \cos x_{i,2} - x_{i,2} \ln x_{i,3} + \varepsilon_i,$$
$$(i = \overline{1,n}),$$
считая, что случайные величины ε_i имеют нормальное распределение $N(0, \sigma^2)$, где $\sigma^2 = 0.25$.

Для каждого из двух представленных вариантов регрессии:

1) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2 \xi_1 + \tilde{\theta}_3 \xi_2 + \tilde{\theta}_4 \xi_3 + \varepsilon$;

2) $\tilde{\eta} = \tilde{\theta}_1 + \tilde{\theta}_2 \xi_1^2 \cos \xi_2 + \tilde{\theta}_3 \xi_2 \ln \xi_3 + \varepsilon$;

вычислите оценку по методу наименьших квадратов $\hat{\theta}$ (8.7) и скорректированный коэффициент детерминации R^2_{adj} (8.11).

Для варианта регрессии с большей величиной R^2_{adj} :

1) вычислите дисперсионную матрицу $D[\hat{\theta}]$ (8.8) и оценку остаточной дисперсии $\hat{\sigma}^2$ (8.9);

2) определите для каждой компоненты θ_i оценки $\hat{\theta}$ реализацию доверительного интервала (8.12) с уровнем доверия $P_д = 0.95$;

3) вычислите реализацию доверительного интервала для остаточной дисперсии (8.13) с уровнем доверия $P_д = 0.97$;

4) найдите наименьший уровень значимости α^* (8.15) при котором отклоняется гипотеза о том, что вектор $\tilde{\theta}$ является нулевым, используя статистику (8.14).

3. Отчет по работе.

В отчет по работе необходимо включить следующие данные.

Задача 1: оценку по методу наименьших квадратов из уравнения (8.7), дисперсионную матрицу (8.8), среднеквадратичное отклонение (8.6), значение оценки остаточной дисперсии (8.9), значение коэффициента детерминации (8.10), изображение с точками и отрезками двух прямых.

Задача 2: для каждого варианта оценку по методу наименьших квадратов из уравнения (8.7), дисперсионную матрицу (8.8), среднеквадратичное отклонение (8.6), оценку остаточной дисперсии (8.9), коэффициент детерминации (8.10), изображения с точками и кривыми.

Задача 3: для каждого из четырех вариантов вычисленную из уравнения (8.7) оценку по методу наименьших квадратов, и значение скорректированного коэффициента детерминации (8.11). Для наилучшего варианта дисперсионную матрицу (8.8) и оценку остаточной дисперсии (8.9).

Задача 4: для каждого из двух вариантов вычисленную оценку по методу наименьших квадратов и значение скорректированного коэффициента детерминации (8.11). Для лучшего из двух вариантов дисперсионную матрицу (8.8), значение оценки остаточной дисперсии (8.9), реализации доверительных интервалов для всех компонент оценки (8.12) и остаточной дисперсии (8.13), наименьший уровень значимости (8.15), при котором принимается гипотеза о нулевом векторе $\tilde{\theta}$.

4. Вопросы и задания.

1) Сформулируйте задачу линейной регрессии.

2) Как вычисляется среднеквадратичное отклонение? Каким свойством по отношению к другим оценкам обладает оценка по методу наименьших квадратов?

3) Выведите нормальное уравнение (8.7).

4) Какое выражение имеет дисперсионная матрица оценки по методу наименьших квадратов? Как вычисляется оценка остаточной дисперсии? При каких условиях, накладываемых на случайные величины ε_i, получены выражения для дисперсионной матрицы и оценки остаточной дисперсии?

5) Как вычисляется коэффициент детерминации и скорректированный коэффициент детерминации? Каким образом определяется качество построенной регрессии по этим двум коэффициентам?

6) Сформулируйте задачу нормальной линейной регрессии.

7) При каком условии, распространяемом на случайные величины ε_i, могут быть получены доверительные интервалы для компонент вектора $\tilde{\theta}$ и остаточной дисперсии σ^2 .

8) Выведите доверительные интервалы для компонент вектора $\tilde{\theta}$, используя известное распределение оценки по методу наименьших квадратов $\hat{\theta}$ и среднеквадратичного отклонения $\delta(\hat{\theta})$.

9) Выведите доверительный интервал остаточной дисперсии σ^2, используя известное распределение среднеквадратичного отклонения $\delta(\hat{\theta})$.

10) Как проверяется основная гипотеза о том, что вектор $\tilde{\theta}$ является нулевым вектором, против альтернативной гипотезы утверждающей, что вектор $\tilde{\theta}$ не нулевой.

Список литературы.

1. **Айвазян С.А.** Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С.А. Айвазян, В.С. Мхитарян.– изд. 2-е, испр. – М.: ЮНИТИ-ДАНА, 2001.

Т.1 : Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика.– 656 с.– ISBN 5-238-00304-8.

Т.2 : Айвазян С.А. Основы эконометрики.– 432 с.– ISBN 5-238-00305-8.

2. **Боровков А.А.** Математическая статистика: учебник / А.А. Боровков. М.:Наука, 1984.– 472 с.
3. **Вальд А.** Последовательный анализ: монография / А. Вальд. М.: Гос. издат. физ.-мат. лит., 1960.– 328 с.
4. **Ван дер Варден Б.Л.** Математическая статистика: монография / Б.Л. Ван дер Варден. М.: Изд. иностр. лит., 1960.– 435 с.
5. **Горицкий Ю.А.** Основы математической статистики: учебное пособие / Ю.А. Горицкий. М.: Изд-во МЭИ, 1977.– 84 с.
6. **Ивченко Г.И.** Математическая статистика: учебное пособие для втузов / Г.И. Ивченко, Ю.И. Медведев. М.: Высш. шк., 1984.– 248 с.
7. **Козлов М.В.** Введение в математическую статистику: учебник / М.В. Козлов, А.В. Прохоров.– М.: Изд-во МГУ, 1987.– 264 с.
8. **Крамер Г.** Математические методы статистики: монография / Г. Крамер. М.: Мир, 1975.– 648 с.
9. **Леман Э.** Проверка статистических гипотез: монография / Э. Леман. 2-е изд., испр. М.: Наука, 1979.– 408 с.
10. **Мазманишвили А.С.** Математическая статистика: учебное пособие к практическим занятиям / А.С. Мазманишвили.– Харьков: НТУ ХПИ, 2003.– 217 с.– ISBN 966-593-270-5.
11. **Соболь И.М.** Численные методы Монте-Карло: учебное пособие / И.М. Соболь.– М.: Наука, 1973.– 312 с.
12. **Тюрин Ю.Н.** Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров; под. ред. В.Э. Фигурнова.– изд. 3-е, перераб. и доп.– М.: ИНФРА-М, 2002.– 528 с.– ISBN 5-86225-662-8.
13. **Чернова Н.И.** Математическая статистика: учебное пособие / Н.И. Чернова.– Новосибирск: Изд-во НГУ, 2007.– 148 с.– ISBN 978-5-94356-523-6.

Учебное пособие

Тигетов Давид Георгиевич

ПРАКТИКУМ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Лабораторный практикум по курсу
«Математическая статистика»
для студентов, обучающихся по направлению
«Прикладная математика»

Редактор издательства Г.Ф. Раджабова

Темплан издания МЭИ 2014 (II), учеб.		Печать офсетная	
Подписано в печать		Формат 60x84/16	Физ.печ.л
Тираж	Изд.№	Заказ	Цена

Издательство МЭИ, 111250, Москва, ул. Кразноказарменная, д. 14
Отпечатано в ООО «Типография-Н», 141292, Московская обл., г. Красноармейск,
просп. Испытателей, д. 25/2