**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**
**«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОМЫШЛЕННЫХ ТЕХНОЛОГИЙ И ДИЗАЙНА»**

---

**ВЫСШАЯ ШКОЛА ТЕХНОЛОГИИ И ЭНЕРГЕТИКИ**
**Кафедра высшей математики**

# Элементы математической статистики

**Методические указания для студентов всех форм обучения**
**(III семестр)**

**Санкт-Петербург**
**2019**

УДК 519.22/.25(07)

Элементы математической статистики: методические указания для студентов всех форм обучения (III семестр) / сост.: Н.Л. Белая, Е.Г. Иванова, М.Э. Юдовин, Е.В.Федорова - ВШТЭ СПбГУПТД. СПб., 2019.–15с.

В методических указаниях содержится краткое изложение основных разделов курса математической статистики: случайные события и случайные величины, выборочный метод, статистические оценки параметров распределения, проверка статистических гипотез. Предназначены для студентов заочной формы обучения (III семестр)

Подготовлены и рекомендованы к изданию кафедрой высшей математики ВШТЭ СПбГУПТД (протокол № 1 от 03.09.2019 г.).

Утверждены к изданию институтом энергетики и автоматики ВШТЭ СПбГУПТД (протокол № 3 от 26.11.2019г.).



Редактор и корректор Т.А. Смирнова
Техн. редактор Л.Я.Титова Темплан 2019 г., поз.130

---

Подп. к печати 19.12.2019. Формат 60х84/16. Бумага тип №1.
Печать офсетная. Объем 1,0 печ.л.; 1,0 уч.-изд.л.
Тираж 70 экз. Изд. № 130. Цена «С». Заказ

---

Ризограф Высшей школы технологии и энергетики ВШТЭ СПбГУПТД, 198095, Санкт-Петербург, ул. Ивана Черных, 4.

## *ПРЕДИСЛОВИЕ*

Статистические исследования проводились ещё в глубокой древности. Однако в науку статистика превратилась только после аксиоматического построения теории вероятностей.

Математической статистикой называется наука, разрабатывающая методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений.

Методы математической статистики носят абстрактный характер, приспособлены для обработки экспериментального материала любой природы, поэтому применимы в любых конкретных науках, технике, экономике и т.д.

## *1. СЛУЧАЙНАЯ ВЕЛИЧИНА*

***Определение .*** *Случайной величиной* называется такая величина которая в результате опыта может принимать различные значения, причем заранее неизвестно, какое именно. Можно говорить только о вероятности, с которой случайная величина принимает каждое конкретное значение.

Случайные величины принято обозначать большими латинскими буквами X, Y, Z, а их возможные значения – маленькими латинскими буквами с индексами $x_i, y_i, z_i$ .

Случайная величина бывает дискретной и непрерывной.

***Определение.*** *Дискретная случайная величина* – это величина, значения которой можно перенумеровать (пересчитать).

***Определение.*** *Непрерывная случайная величина* – это такая величина, значения которой заполняют целиком некоторый промежуток числовой оси или всю числовую ось.

***Определение.*** Любое соотношение, устанавливающее связь между всеми возможными значениями случайной величины и их вероятностями, называется *законом распределения* случайной величины.

Закон распределения полностью описывает случайную величину.

***Определение.*** *Законом распределения дискретной случайной* величины называется таблица, в первой строке которой стоят возможные значения случайной величины $x_i$ , а во второй – их вероятности $p_i$ :

| $x_i$ | $x_1$ | $x_2$ | … | $x_n$ |
|---|---|---|---|---|
| $p_i$ | $p_1$ | $p_2$ | … | $p_n$ |

Для того, чтобы задать закон распределения непрерывной случайной величины, сначала введем функцию распределения.

***Определение.*** *Функцией распределения случайной величины* называют функцию F(x), определяющую вероятность того, что случайная величина X в результате испытания примет значение, меньше x, т:е.:

$$F\left(x\right)=P\left(X<x\right).$$

Непрерывную случайную величину задают, используя функцию, которую называют плотностью распределения или дифференциальной функцией.

***Определение.*** *Плотностью распределения вероятностей* непрерывной случайной величины X называют функцию f(x)- первую производную от функции распределения F(x).

$$F'\left(x\right)=f\left(x\right)$$

Законом распределения непрерывной случайной величины называют плотность распределения этой величины. Закон распределения полностью описывает случайную величину. Но зачастую достаточно указать только отдельные числовые параметры. Такие характеристики, назначение которых – выразить в сжатой форме наиболее существенные особенности распределения, называют *числовыми характеристиками случайной величины*. В настоящем курсе мы введем только некоторые, наиболее часто применяемые.

***Определение.*** *Математическое ожидание дискретной случайной величины* X - это величина

$$M(X) = \sum_{i=1}^{n} x_i p_i ,$$

где $x_i$ – значения случайной величины, $p_i$ – их вероятности, n– число возможных значений случайной величины.

***Определение.*** *Математическое ожидание непрерывной случайной величины* X – это величина

$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx,$$

где f(x) – плотность распределения вероятностей.

Математическое ожидание – это некоторое среднее значение, вокруг которого группируются значения случайной величины.

Отклонением случайной величины называют разность между случайной величиной и ее математическим ожиданием.

***Определение***. *Дисперсией случайной величины* X называется математическое ожидание квадрата отклонения X от ее математического ожидания.

$$D(X) = M\left(\left(X - M(X)\right)^2\right)$$

Дисперсия показывает, как рассеяны возможные значения случайной величины около ее математического ожидания. Для вычисления дисперсии пользуются формулой:

$$D(X) = M\left(X^2\right) - \left(M(X)\right)^2.$$

Числовой характеристики случайной величины, кроме математического ожидания и дисперсии, является и *среднее квадратичное отклонение* $\sigma = \sqrt{D(X)}$ . В дальнейшем числовые значения математического ожидания и дисперсии будут обозначаться символами $\mu$ *и* $\sigma^2$*соответственно*:

$$\mu = \sum_{i=1}^{+\infty} x_i p_i \; ; \quad \mu = \int_{-\infty}^{+\infty} x f(x) dx \; ;$$

$$\sigma^2 = \sum_{i=1}^{+\infty} (x_i - \mu)^2 p_i \; ; \quad \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx.$$

Вспомним еще несколько понятий.

***Определение***. *Квантиль распределения* – функция, обратная к функции распределения. Число $x_\beta$, определяемое уравнением $F(x_\beta) = \beta$, называется $\beta$*-квантилью* распределения.

***Определение***. *Случайной выборкой* объема $n$ называется набор значений $(x_1,...,x_n)$ случайной величины, полученных в результате $n$ независимых опытов. Эти значения называют в статистике *наблюдениями*.

Выборка имеет числовые характеристики, аналогичные характеристикам случайной величины:

*выборочное среднее* $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i;$

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

*выборочное стандартное отклонение* $s = \sqrt{s^2}.$

## *2. ЗАВИСИМОСТЬ МЕЖДУ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ*

Пусть имеем переменные $x$ и $y$ .

Предполагается, что значения независимой переменной $x$ можно задавать, а значения $y$ определяются из эксперимента и, следовательно, содержат случайные ошибки.

Предположим, что между переменными величинами $x$ и $y$ существует линейная зависимость

$$y = a_0 + a_1 x, \tag{1}$$

причем коэффициенты в уравнении (1) неизвестны. Зависимость (1) является простейшей из возможных, однако при изучении реальных зависимостей ее можно использовать в качестве первого приближения и при необходимости перейти к более сложной модели. Формула (1) определяет строгую линейную зависимость, в ней не учитывается, что значения $y$ доступны лишь в результате эксперимента и, значит, содержат случайные ошибки. Поэтому более реалистичной является модель

$$Y = a_0 + a_1 x + E \tag{2}$$

где $Y, E -$ случайные величины.

Проведем $n$ опытов и получим $n$ пар наблюдений $(x_i, y_i)$, $i = 1, \dots, n$. В соответствии с принятой линейной моделью (2) эти наблюдения можно представить в виде

$$y_i = a_0 + a_1 x_i + \varepsilon_i\ ,\ i = 1, \dots\ , n. \qquad (3)$$

Здесь $\varepsilon_i$ – случайные ошибки. Система (3) имеет $n$ уравнений с $(n+2)$ неизвестными $a_0, a_1, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$.

Очевидно, такие системы имеют бесконечно много решений, поэтому из них нельзя найти точные значения коэффициентов $a_0$ и $a_1$. Однако существуют методы, позволяющие найти их приближенные значения .

Один из таких методов – ***метод наименьших квадратов*** (МНК).

### *3. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ*

Далее мы предполагаем, что случайные ошибки $\varepsilon_i$ независимы и имеют нормальное распределение с нулевым математическим ожиданием и неизвестной, но одинаковой дисперсией.

Обозначим $e_i = y_i - \alpha_0 - \alpha_1 x_i$ где $\alpha_0, \alpha_1$– произвольные числа, и составим сумму

$$S(a_0, a_1) = \sum_{i=1}^{n} e_i^2 \qquad (4)$$

Числа $e_i$ называются остатками. Они характеризуют отклонение по вертикали точек $(x_i, y_i)$ от произвольной прямой.

*Метод наименьших квадратов* заключается в том, что коэффициенты $\alpha_0, \alpha_1$ выбираются таким образом, чтобы сумма квадратов остатков была *наименьшей*.

Предположим, что сумма $S(\alpha_0, \alpha_1)$ принимает наименьшее значение при $\alpha_0 = \hat{a}_0,\ \alpha_1 = \hat{a}_1$. В математической статистике доказывается, что величины $\hat{a}_0, \hat{a}_1$ являются *статистическими оценками* неизвестных параметров $a_0, a_1$.

Геометрический смысл описанного подхода заключается в следующем. На рис.1 изображены экспериментальные точки и произвольная прямая

$$y = \alpha_0 + \alpha_1 x,$$

знаком ★ отмечены точки, полученные в результате эксперимента; вертикальные отрезки изображают остатки $e_i$ ; модуль остатка – это расстояние по вертикали от экспериментальной точки $(x_i, y_i)$ до прямой. Очевидно, что невозможно выбрать прямую так, чтобы она проходила через все точки $(x_i, y_i)$. Поэтому выбирается на основе определенного критерия некоторое среднее положение прямой, наилучшее согласно выбранному нами критерию. Наиболее часто используемым критерием является минимум суммы квадратов остатков.

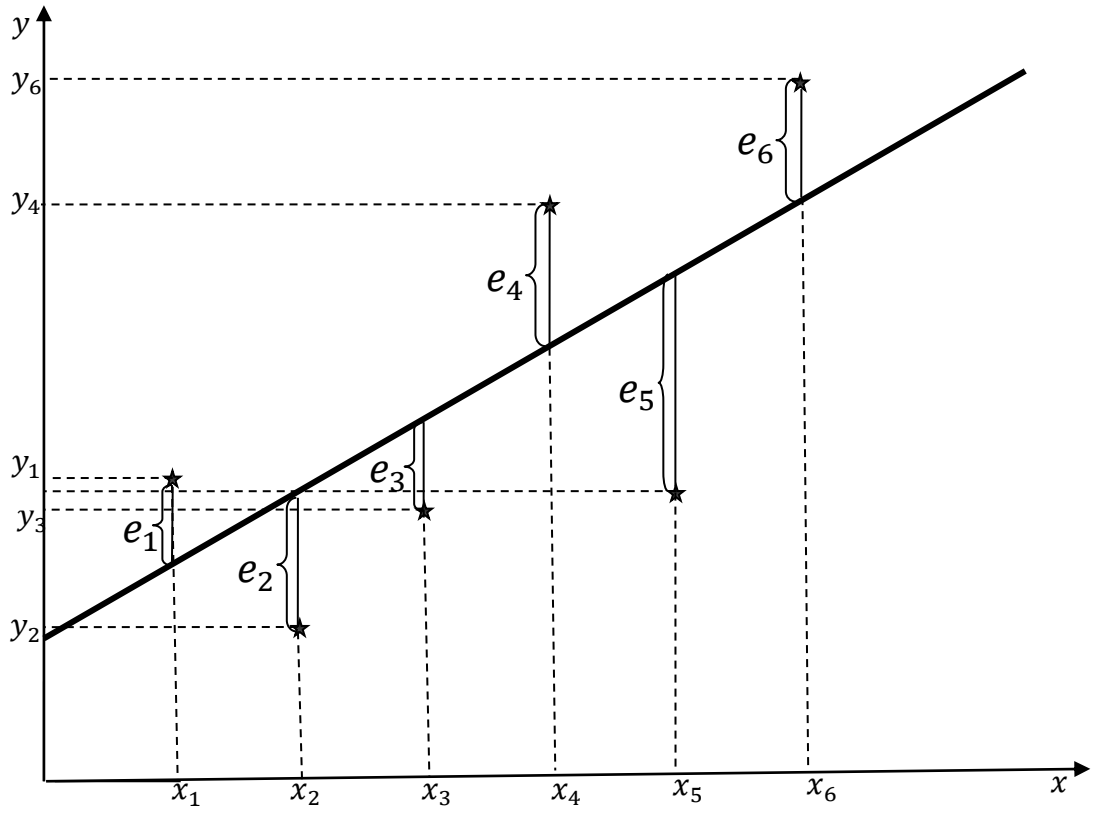


Рис.1

Таким образом, задача сводится к отысканию значений параметров $(\alpha_0, \alpha_1)$, при которых функция $S(\alpha_0, \alpha_1)$ имеет наименьшее значение.

## *4. ФОРМУЛЫ ДЛЯ ВЫЧИСЛЕНИЯ КОЭФФИЦИЕНТОВ* $\hat{a}_0, \hat{a}_1$

Для отыскания значений переменных $\alpha_0, \alpha_1$, минимизирующих $S(\alpha_0, \alpha_1)$, используем необходимое условие экстремума:

$$\begin{cases} \frac{\partial S}{\partial \alpha_0} = 0 \\ \frac{\partial S}{\partial \alpha_1} = 0 \end{cases}. \tag{5}$$

Решение системы (5) дается формулами :

$$\hat{a}_1 = \frac{S_{XY}}{S_X}, \quad \hat{a}_0 = \bar{y} - \hat{\alpha}_1 \bar{x},$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad S_X = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{6}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \; S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{7}$$

Коэффициенты $\hat{a}_0$, $\hat{a}_1$ являются случайными величинами, так как они зависят от значений случайной величины $Y$. Как правило, они не всегда совпадают с "истинными" коэффициентами. Однако можно доказать, что их математические ожидания равны соответственно $a_0$, $a_1$. На языке математической статистики это означает, что $\hat{a}_0$, $\hat{a}_1$являются *несмещенными оценками* коэффициентов $a_0$, $a_1$.

Таким образом, вместо точного уравнения (1) с неизвестными коэффициентами получено приближенное уравнение

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x \tag{8}$$

Остатки для этого уравнения равны

$$\hat{e}_i = y_i - \hat{y}_i,$$

где $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i$.

Для любого заданного $x$ величина $\hat{y}$, определяемая уравнением (11), является несмещенной оценкой для $y$. Обозначим $S_R = S(\hat{a}_0, \hat{a}_1)$. В курсе математической статистики доказывается, что величина

$$s^2 = S_R/(n-2) \qquad (9)$$

является несмещенной оценкой для дисперсии случайных ошибок .

## *5. ПРОВЕРКА ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ*

Если некоторый из вычисленных коэффициентов $\hat{a}_0$, $\hat{a}_1$ оказался малым, то естественно предположить, что соответствующий истинный коэффициент равен нулю, а его оценка не равна нулю из-за случайных ошибок или, как говорят в этом случае, вычисленный коэффициент незначимо отличается от нуля.

Обозначим через $\boldsymbol{H_0}$ гипотезу: $a_1 = 0$, иначе говоря, гипотеза состоит в том, что переменная $y$ не зависит от $x$. Гипотеза проверяется следующим образом.

Вычислим полную сумму квадратов $S_Y$, остаточную сумму квадратов $S_R$ и $F$- отношение по формулам:

$$S_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2 \; ; \; S_R = S(\hat{a}_0, \hat{a}_1); \; F = \frac{(n-2)(S_Y - S_R)}{S_R}. \qquad (10)$$

Если гипотеза $\boldsymbol{H_0}$ верна, то случайная величина $F$ имеет распределение Фишера со степенями свободы 1 и $(n-2)$. Задаем уровень значимости $\boldsymbol{\alpha}$. Из таблицы находим $(1-\boldsymbol{\alpha})$-квантиль этого распределения. Обозначим его через $F_{\text{крит}}$. Для проверки гипотезы применяется следующее правило:

*Если* $F \le F_{\text{крит}}$, *то гипотеза* $\boldsymbol{H_0}$ *принимается, если же* $F > F_{\text{крит}}$, *то* $\boldsymbol{H_0}$ *отвергается.*

***Замечание.*** Принятие этой гипотезы не означает, что она обязательно верна. Утверждается лишь, что при имеющихся экспериментальных данных и на выбранном уровне значимости нет оснований отвергать гипотезу. Увеличив $\boldsymbol{\alpha}$, мы, возможно, должны будем отвергнуть гипотезу. Аналогично, отвергая гипотезу, мы не можем гарантировать, что она неверна.

## *6. ОЦЕНКА ТОЧНОСТИ ПРИБЛИЖЕННОЙ МОДЕЛИ*

Предположим, что коэффициенты в уравнении (8) оказались значимыми. Если мы хотим использовать $\widehat{\boldsymbol{y}}$ в качестве приближенного значения для $\boldsymbol{y}$ при произвольном $\boldsymbol{x}$, то возникает вопрос о точности этого приближения. Определим выражение

$$\Delta y(x) = t_{\gamma,m} \cdot s\sqrt{\frac{1}{n} + (x - \bar{x})^2/S_X}\ , \qquad (11)$$

где $s$ и $S_X$ определяются формулами (6) и (9), а коэффициент $t$ определяется из таблицы квантилей распределения Стьюдента по заданному уровню значимости $\boldsymbol{\alpha}$ и числу степеней свободы $m = n - 2$. Например, при $\alpha = 0{,}05$ и $n = 12$ имеем $\gamma = 1 - \alpha/2 = 0{,}975$ и $m = 12 - 2 = 10$. Тогда из таблицы распределения Стьюдента находим $t_{\gamma,m} = 2{,}23$.

Доверительный интервал для $\boldsymbol{y}$ зависит от $\boldsymbol{x}$ и имеет вид

$$\hat{y}(x) - \Delta y(x) < y(x) < \hat{y}(x) + \Delta y(x)\ , \qquad (12)$$

где $\hat{y}(x)$ вычисляется по формуле (8), а $\Delta y(x)$ – по формуле (11).

Из формул (9) и (11) видно, что длина и центр этого интервала зависят от $\boldsymbol{x}$, причем наименьшая длина достигается при $x = \bar{x}$, а по мере удаления

от $\bar{x}$ длина интервала увеличивается, а значит, уменьшается точность оценки величины $y$ .

На рис. 2 показана полоса, определенная неравенствами (12). Наименьшее вертикальное сечение этой полосы будет при $x = \bar{x}$, а при удалении в обе стороны от этого сечения полоса расширяется.

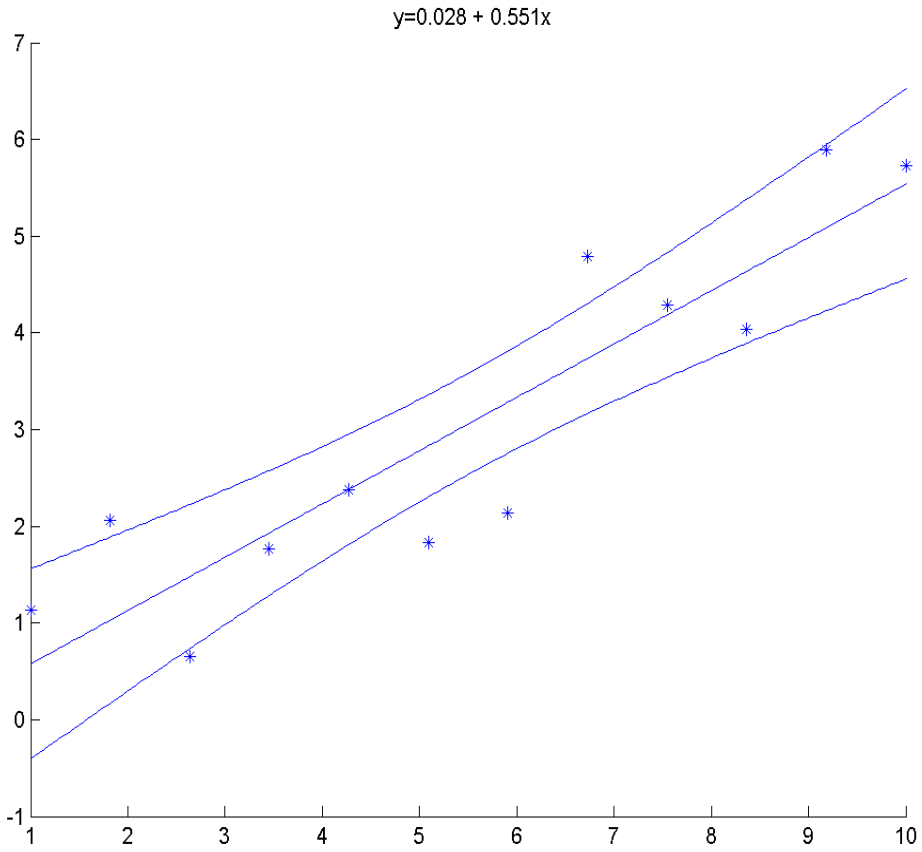


Рис.2

## *7. ПРОВЕРКА АДЕКВАТНОСТИ МОДЕЛИ*

Предыдущие рассуждения были основаны на предположении, что модель (1) адекватна, т.е. верна. Эта гипотеза также может быть проверена методами математической статистики. Для ее проверки необходимо иметь некоторое число дополнительных наблюдений, на основе которых строится оценка дисперсии случайной ошибки, независимая от оценки по

формуле (9). Например, если для каждого $x_i$ получено $k$ повторных наблюдений $y_{i,j},\ i = 1,2,\dots,n;\ j = 1,2,\dots,k$, то независимая от $s^2$ оценка имеет вид:

$$s_1^2 = \frac{S_1}{kn - n} \quad ,$$

где $S_1 = \sum_{i=1}^{n} \sum_{j=1}^{k} \left(y_{i,j} - y_{i\cdot}\right)^2,\ \ y_{i\cdot} = \frac{1}{k} \sum_{j=1}^{k} y_{i,j}$ .

В математической статистике доказано, что величина

$$F = \frac{(S_R - S_1)/(n - k)}{S_1/(kn - n)}$$

имеет распределение Фишера со степенями свободы

$$(n - k),(kn - n).$$

Проверка гипотезы об адекватности модели (1) производится по той же схеме, что и проверка значимости коэффициента. Находим $(1 - \alpha)$-квантиль распределения Фишера $F_{\text{крит}}$ . Если $F \le F_{\text{крит}}$ , то гипотеза принимается, если же $F > F_{\text{крит}}$, то гипотеза отвергается.

***Библиографический список***

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2011.

2. Письменный Д.Т. Конспект лекций по теории вероятностей, математической статистике и случайным процессам: учебн. пособие для вузов. – М.: Айрис-Пресс, 2010.

3. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики (для технических приложений). – М.: Наука, 1969.

***Содержание***

Предисловие……..……..……3

1. Случайная величина…….……...3

2. Зависимость между случайными величинами…..…….7

3. Метод наименьших квадратов…….……..…..8

4. Формулы для вычисления коэффициентов $\hat{a}_0, \hat{a}_1$ ……..……10

5. Проверка значимости коэффициентов……..…...…11

6. Оценка точности приближенной модели……..12

7. Проверка адекватности модели……..……..……..13

Библиографический список……..……..……...….14