№ 2770

В.А. Карасев
Г.Д. Лёвшина

Теория вероятностей и математическая статистика

Математическая статистика

Практикум

№ 2770

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «МИСиС»

Кафедра математики

В.А. Карасев
Г.Д. Лёвшина

Теория вероятностей и математическая статистика

Математическая статистика

Практикум

Рекомендовано редакционно-издательским советом университета

МИСиС

Москва 2016

УДК 519.2
К21

Р е ц е н з е н т
канд. экон. наук, проф. *В.Ф. Михин*

Карасев В.А.

К21 Теория вероятностей и математическая статистика : математическая статистика : практикум / В.А. Карасев, Г.Д. Лёвшина. – М. : Изд. Дом МИСиС, 2016. – 120 с.
ISBN 978-5-906846-01-3

Практикум предназначен для подготовки к занятиям, контрольным работам и к выполнению лабораторных работ (типовых расчетов) по курсу «Теория вероятностей и математическая статистика» разделу «Математическая статистика». Изложение теоретического материала сопровождается большим количеством примеров решения типовых задач, дается подробное описание выполнения лабораторных работ (типовых расчетов), разбирается решение типовых вариантов контрольных работ. В заключение приводятся варианты контрольных работ с ответами.

Предназначен для студентов всех направлений.

УДК 519.2

ISBN 978-5-906846-01-3

СОДЕРЖАНИЕ

Предисловие 5

1. Практические занятия по математической статистике 6

1.1. Первичная обработка результатов эксперимента и оценка основных параметров генеральной совокупности 6

1.2. Оценка математического ожидания по неравноточным измерениям 12

1.3. Оценка дисперсии по нескольким сериям экспериментов 14

1.4. Построение гистограммы распределения 16

1.5. Некоторые используемые в статистике распределения 17

1.6. Построение доверительных интервалов 20

1.6.1. Доверительный интервал для математического ожидания 21

1.6.2. Доверительный интервал для среднего квадратического отклонения 21

1.6.3. Доверительный интервал для дисперсии 22

1.7. Проверка статистических гипотез 25

1.7.1. Основные понятия 25

1.7.2. Проверка гипотез о дисперсии нормального распределения 29

1.7.3. Проверка гипотез о математических ожиданиях нормального распределения 35

1.7.4. Проверка гипотезы о виде распределения генеральной совокупности 43

1.8. Регрессионный анализ. Построение линейной и квадратичной регрессионных моделей 46

1.8.1. Оценка коэффициентов регрессии 46

1.8.2. Проверка гипотезы об адекватности регрессионной модели 49

1.8.3. Построение доверительных интервалов для коэффициентов регрессии 50

1.8.4. Задача регрессии для линейной функции 51

1.8.5. Задача регрессии для квадратичной функции 53

1.9. Линейный корреляционный анализ 62
1.9.1. Двумерный случайный вектор, его выборочные характеристики 62
1.9.2. Построение доверительного интервала для коэффициента корреляции. Проверка гипотезы о существовании линейной зависимости 68
2. Указания по выполнению типовых расчетов (лабораторных работ) «Обработка основных типов экспериментальных данных» 71
2.1. Типовой расчет 1. Сравнение двух случайных выборок (первичная обработка данных, проверка статистических гипотез) 71
2.1.1. Первичная обработка результатов измерений 72
2.1.2. Построение доверительных интервалов 73
2.1.3. Проверка гипотез о равенстве дисперсий, о равенстве математических ожиданий 73
2.1.4. Проверка гипотезы о нормальном распределении объединения двух случайных выборок 77
2.1.5. Построение гистограмм распределения объединения двух случайных выборок 81
2.2. Типовой расчет 2. Обработка данных методами регрессионного анализа 84
2.2.1. Задача 1 85
2.2.2. Задача 2 89
2.3. Типовой расчет 3. Обработка данных методами линейного корреляционного анализа 96
2.3.1. Нахождение оценок числовых характеристик двумерного случайного вектора. Расчет оценки коэффициента корреляции 96
2.3.2. Нахождение уравнений линейной регрессии 97
2.3.3. Построение доверительного интервала для коэффициента корреляции ρ. Проверка гипотезы о существовании линейной зависимости между величинами X и Y 97
2.4. Примерные варианты контрольной работы 97
Библиографический список 106
Приложение 107

ПРЕДИСЛОВИЕ

Пособие является продолжением практикума «Теория вероятностей и математическая статистика» раздел «Теория вероятностей» (библ. № 2454), вышедшего в 2015 г. В нем приведены необходимые сведения об основных современных методах математической статистики и обработки экспериментальных данных на примерах, заимствованных из области металлургии, металловедения и экономики.

Математическая статистика – это наука, которая методами теории вероятностей на основании результатов наблюдений изучает закономерности в массовых случайных явлениях. Математическая статистика (не путать со статистикой – разделом экономической теории) указывает способы сбора и группировки статических данных (результатов наблюдений или экспериментов), разрабатывает методы их обработки для оценки характеристик распределения, для установления зависимости случайной величины от других, для проверки статистических гипотез о виде распределения или значениях его параметров. Математическая статистика возникла и развивалась параллельно с теорией вероятностей.

Данный практикум адресуется студентам всех специальностей НИТУ «МИСиС». Практикум подготовлен в соответствии с требованиями государственных образовательных стандартов. Рассматриваются следующие разделы курса: основные выборочные характеристики, точечные и интервальные оценки, проверка статистических гипотез, регрессионный и корреляционный анализ.

Согласно учебному плану предусмотрено выполнение трех типовых расчетов (лабораторных работ). Исходя из этого сформирована структура практикума, состоящая из двух разделов. Первый содержит теоретический материал и методические указания по решению задач для подготовки к практическим занятиям. Приведены необходимые для решения задач понятия и формулы, рассмотрено большое количество примеров решения типовых задач, охватывающих основные темы курса математической статистики.

Во второй части практикума разбираются все три типовые расчеты (лабораторные работы), приводятся подробные примеры их выполнения. Затем даны типовые варианты контрольной работы с ответами. В приложении представлены статистические таблицы, необходимые для решения данных задач.

Студентам, желающим ознакомиться с более полным и строгим изложением основ математической статистики, рекомендуем литературу, приведенную в библиографическом списке.

1. ПРАКТИЧЕСКИЕ ЗАНЯТИЯ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

1.1. Первичная обработка результатов эксперимента и оценка основных параметров генеральной совокупности

Математическая статистика позволяет с помощью математических методов обрабатывать, систематизировать и использовать численные результаты эксперимента для получения практических выводов.

Под *генеральной совокупностью* в математической статистике понимается множество (гипотетическое) всех возможных результатов измерения некоторой величины, которые могут быть получены в данных условиях. Тем же самым понятием в теории вероятностей является случайная величина X.

Реальная серия повторных измерений этой величины x_1, x_2,..., x_n трактуется как случайная выборка из генеральной совокупности, или просто *случайная выборка*. Число n называется *объемом* случайной выборки. Приведем примеры.

1. Проведена серия повторных измерений одной и той же физической величины в одних и тех же условиях. Разброс результатов обусловлен погрешностью измерительной аппаратуры.

2. Измеряется некоторая характеристика одинаковых изделий, изготовленных при поточном производстве. Разброс результатов обусловлен особенностями технологии производства.

3. Измеряется некоторая характеристика людей определенного пола и интервала возрастов, например рост. Разброс результатов обусловлен природными факторами.

В статистике принята следующая математическая модель подобных экспериментов. *Каждый элемент случайной выборки рассматривается как отдельная случайная величина.* Относительно этих случайных величин, которые в дальнейшем будем обозначать прописными буквами, известна некоторая априорная информация.

Случайная выборка называется *повторной*, если все входящие в нее случайные величины $X_1, X_2, ..., X_n$ взаимно независимы и имеют одинаковую функцию распределения $F(x)$, причем ту же, что и наблюдаемая случайная величина X. На практике это, в частности, означает, что измерения производятся *независимо* друг от друга (полученные результаты одних измерений не влияют на возможные результаты других). Величины имеют одинаковые математические ожидания $M(X_i) = a$, т.е., результаты измерений свободны от *систематических ошибок* (результаты в среднем не смещены относительно истинного значения $M(X) = a$), и одинаковые дисперсии $D(X_i) = D(X) = \sigma^2$, что называется *равноточностью* измерений (например, измерения физической величины проведены на одном и том же приборе при одинаковых условиях).

Если случайные величины $X_1, X_2, ..., X_n$, образующие повторную случайную выборку, имеют нормальное распределение с одинаковыми параметрами a, σ , т.е., $X_i \sim N(a; \sigma)$, то такая выборка называется *нормальной*, а соответствующая генеральная совокупность – *нормальной генеральной совокупностью.*

В математической статистике рассматривают и *неповторные* выборки, в которых нарушается хотя бы одно из указанных условий: взаимная независимость, одинаковость функции распределения. Слово «повторная» обычно опускается, и пишут просто «выборка». Для неповторной выборки обязательно пишут «неповторная выборка».

Распределение случайной величины X характеризуется рядом *параметров* (математическое ожидание, дисперсия и т.д.). Эти параметры называют *параметрами генеральной совокупности.* Важной задачей математической статистики является нахождение по случайной выборке приближенных значений каждого из параметров, называемых *точечными оценками параметров*, или просто *оценками*. Таким образом, *оценкой параметра* β называется функция $f(X_1, X_2, ..., X_n)$ от случайной выборки, значение которой принимается в качестве приближенного для данного параметра и обозначается $\tilde{\beta}$:

$$\beta \approx \tilde{\beta} = f(X_1, X_2, ..., X_n)\,. \tag{1.1}$$

Так как оценка зависит от случайной выборки, то она, в свою очередь, является случайной величиной. Для одного и того же параметра β по одной и той же выборке можно построить много различных оценок. Для сравнения оценок между собой введены специальные характеристики.

Оценка называется *несмещенной*, если ее математическое ожидание равно истинному значению параметра, т.е. $M(\tilde{\beta}) = \beta$. Несмещенная оценка обеспечивает близость в среднем значений оценки к значению оцениваемого параметра, т.е. не дает систематической ошибки.

Оценка называется *состоятельной*, если при $n \to \infty$ она сходится по вероятности к истинному значению оцениваемого параметра, т.е.

$$\forall \varepsilon > 0 \ \lim_{n\to\infty} P(\left|\tilde{\beta}_n - \beta\right| > \varepsilon) = 0, \tag{1.2}$$

где $\tilde{\beta}_n$ – оценка параметра β, найденная по выборке объема n.

Смысл понятия состоятельности заключается в том, что с увеличением объема выборки оценка стремится к истинному значению параметра.

Точностью оценки $\tilde{\beta}$ называется средний квадрат отклонения оценки от β: $q^2(\tilde{\beta}) = M[(\tilde{\beta} - \beta)^2]$. Для несмещенных оценок точность определяется величиной дисперсии оценки: $q^2(\tilde{\beta}) = D(\tilde{\beta})$. Чем меньше $q = \sqrt{q^2(\tilde{\beta})}$, тем оценка лучше (точнее). *Наилучшей линейной оценкой* параметра β называется такая его линейная несмещенная оценка, которая имеет наименьшую дисперсию в классе всех линейных несмещенных оценок.

Пусть задана повторная случайная выборка $X_1, X_2, \ldots, X_n$. За *оценку математического ожидания* a принимается среднее арифметическое элементов выборки:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \tag{1.3}$$

Оценкой дисперсии σ^2 при *известном математическом ожидании* a является величина S_0^2:

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - a)^2 . \tag{1.4}$$

Оценкой дисперсии σ^2 при *неизвестном математическом ожидании* является величина S^2, которую называют *эмпирической дисперсией*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 . \tag{1.5}$$

Оценкой среднего квадратического отклонения σ при этом являются, соответственно, величины

$$S_0 = \sqrt{S_0^2}\,; \quad S = \sqrt{S^2}\,. \tag{1.6}$$

Оценки математического ожидания и дисперсии, найденные по формулам (1.3) – (1.5), являются несмещенными и состоятельными. Среднее арифметическое (1.3) является наилучшей линейной оценкой математического ожидания для повторной случайной выборки.

Оценка параметра $\sigma = \sqrt{\sigma^2}$ с помощью значений (1.6) является состоятельной, но смещенной (ее смещение убывает с увеличением n). Число $k = n - 1$ в формуле (1.5) называется *числом степеней свободы* оценки S^2.

Для практических расчетов формулу (1.5) удобно преобразовать к виду

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \tag{1.7}$$

Вычисление среднего значения $\overline{X}$ и оценки дисперсии S^2 упрощается, если отсчет значений X_i вести от подходящим образом выбранных начала отсчета C и масштабного коэффициента h, т.е. если сделать линейную замену:

$$X_i = C + hU_i \;\; (i = 1, 2, ..., n)\,. \tag{1.8}$$

При такой замене формулы (1.3) и (1.5) – (1.6) принимают вид

$$\overline{X} = C + h\overline{U}, \qquad \overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i\,; \tag{1.9}$$

$$S^2 = \frac{h^2}{n-1}\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2\right). \tag{1.10}$$

Для контроля вычислений весь расчет повторяют с другим началом отсчета *C,* результаты должны совпадать с точностью до возможных ошибок округления.

Задача 1.1. В табл. 1.1 в первом столбце записаны результаты $n = 18$ независимых равноточных измерений заряда электрона $q = x \cdot 10^{-10}$, полученных Милликеном. Вычислить оценки математического ожидания, дисперсии и среднего квадратического отклонения величины X, провести контроль расчетов.

Решение

Выбираем $C = 4{,}780$ и, полагая $h = 10^{-3}$, подсчитываем значения $u_i = (x_i - C)/h = (x_i - 4{,}780)/10^{-3}$ и u_i^2. Суммы чисел второго и третьего столбцов дают возможность рассчитать $\overline{X}$ и S^2:

$$\overline{U} = 13/18 = 0{,}72, \quad \overline{X} = 4{,}780 + 0{,}72 \cdot 10^{-3} = 4{,}7807;$$
$$S^2 = 10^{-6}\left(6239 - 13^2/18\right)/17 = 3{,}66 \cdot 10^{-4},$$

откуда $S = \sqrt{S^2} = \sqrt{3{,}66 \cdot 10^{-4}} = 1{,}91 \cdot 10^{-2}$.

В последних двух столбцах приведены расчеты при другом начале отсчета $C_1 = 4{,}790$, т.е. при замене $V_i = (X_i - 4{,}790)/10^{-3}$. Эти расчеты дают те же значения $\overline{X}$ и S:

$$\overline{V} = 167/18 = -9{,}2, \quad \overline{X} = 4{,}790 - 9{,}28 \cdot 10^{-3} = 4{,}7809;$$
$$S^2 = 10^{-6}\left(7779 - 167^2/18\right)/17 = 3{,}66 \cdot 10^{-4}.$$

Таблица 1.1

Исходные данные и результаты расчета к задаче 1.1

Исходные данные X	Расчет		Контроль	
	U	U^2	V	V^2
4,761	–19	361	–29	841
4,792	12	144	2	4
4,758	–22	484	–32	1024
4,764	–16	256	–26	676
4,810	30	900	20	400
4,799	19	361	9	81
4,797	17	289	7	49
4,790	10	100	0	0
4,747	–33	1089	–43	1849
4,769	–11	121	–21	441
4,806	26	676	16	256
4,779	–1	1	–11	121
4,785	5	25	–5	25
4,790	10	100	0	0
4,777	–3	9	–13	169
4,749	–31	961	–41	1686
4,781	1	1	–9	81
4,799	19	361	9	81
Σ	13	6239	–167	7779

При большом числе исходных данных их предварительно группируют, т.е. весь диапазон значений X разбивают на l равных *интервалов*, подсчитывают число исходных данных, попавших в каждый j-й интервал, и относят это число (частоту m_j) к середине интервала X_j ($j = 1, 2, ..., l$). Затем середины этих интервалов кодируют по формуле (1.8), выбирая за новое начало отсчета C середину одного из интервалов, а за масштабный коэффициент h – длину интервала. При таком кодировании все значения U_j будут целыми числами, которые для соседних интервалов отличаются на единицу. Расчет по сгруппированным данным дает лишь приближенные значения оценки математического ожидания $\overline{X} \approx C + h\overline{U}$ и оценки дисперсии $S^2 \approx h^2 S_u^2$, где

$$\overline{U} = \frac{1}{n}\sum_{j=1}^{l} U_j m_j; \qquad (1.11)$$

$$S_u^2 = \frac{1}{n-1}\sum_{j=1}^{l}(U_j - \overline{U})^2 m_j = \left(\sum_{j=1}^{l} U_j^2 m_j - n\overline{U}^2\right)\Big/(n-1). \qquad (1.12)$$

Оценка дисперсия S^2, вычисленная по сгруппированным данным, оказывается меньше оценки дисперсии, найденной по несгруппированным результатам эксперимента на величину, приблизительно равную $h^2/12$. Это следует учитывать при округлении значения S, сохраняя лишь один сомнительный знак; значение среднего $\overline{X}$ округляют при этом до единиц того разряда, который сохранен в значении S. Подобными соображениями можно руководствоваться и в тех случаях, когда результаты измерений округлены с учетом цены деления шкалы измерительного прибора (при этом обычными методами через цену деления шкалы оценивают погрешность вычисления S^2).

Задача 1.2. Проведено 52 эксперимента, их результаты оказались в диапазоне 22,75–26,75. Этот диапазон разбит на 8 интервалов длины $h = 0{,}5$, и сгруппированные данные эксперимента приведены в первых столбцах табл. 1.2. Вычислить оценки математического ожидания, дисперсии и среднего квадратического отклонения, провести контроль расчетов.

Решение

Для расчета выбрано начало отсчета $C = 24{,}5$, а для контроля $C_1 = 25{,}0$. Кодировка имеет соответственно вид $U = (X - 24{,}5)/0{,}5$; $V = (X - 25{,}0)/0{,}5$.