Федеральное агентство по образованию

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

СОЦИАЛЬНО-ЭКОНОМИЧЕСКАЯ СТАТИСТИКА

Методические указания
по выполнению лабораторных работ

Санкт-Петербург
Издательство СПбГЭТУ «ЛЭТИ»
2017

УДК 581.587.8 Я75

Социально-экономическая статистика: Методические указания по выполнению лабораторных работ / Сост.: А. Э. Сулейманкадиева. СПб.: Изд-во СПбГЭТУ «ЛЭТИ», 2017. 20 с.

Содержат описания лабораторных работ по основным темам общей теории статистики, целью которых является изучение методов и приемов сбора и обработки статистической информации. Они включают задания по выполнению группировки, сводки статистических данных по совокупности, анализу взаимосвязей между явлениями и признаками с использованием компьютерных программ для обработки данных.

Предназначены для студентов специальностей 220501, 080503, 080507 факультета экономики и менеджмента «ЛЭТИ» направления 080500 дневной и очно-заочной форм обучения.

Утверждено
редакционно-издательским советом университета
в качестве методических указаний

СОДЕРЖАНИЕ

Редактор Н. В. Лукина

Подписано в печать . Формат 60×84 1/16.
Бумага офсетная. Печать офсетная. Печ. л. 1,25.
Гарнитура «Times». Тираж 140 экз. Заказ

Издательство СПбГЭТУ «ЛЭТИ»
197376, С.-Петербург, ул. Проф. Попова, 5

ОБЩИЕ ПОЛОЖЕНИЯ

Целью лабораторных работ является получение навыков и изучение приемов многомерного статистического анализа с помощью Microsoft Excel и других компьютерных программ.

При выполнении лабораторных работ каждый вариант задания может выполняться как индивидуальным исполнителем, так и группой из двух-трех человек. В том случае, если вариант задания является общим для нескольких студентов, перед каждым из них должен быть поставлен ряд индивидуальных вопросов.

Методические указания включают лабораторные работы по следующим темам: 1) «Группировка, сводка и статистические таблицы»; 2) «Дисперсионный анализ, эмпирическая и аналитическая регрессия» и 3) «Корреляционный анализ, множественная линейная регрессия». Особое внимание уделено вопросам построения типологических структурных, аналитических и комбинационных группировок, выявлению закономерностей явлений и обобщению полученных результатов анализа статистической информации; исследованию взаимосвязи между различными явлениями и признаками совокупности. После выполнения аудиторной части лабораторной работы каждый студент составляет отчет о работе, производит анализ и расчеты с использованием ПК. Отчет оформляется в соответствии с требованиями, предъявляемыми к выполнению и оформлению лабораторной работы и сдается преподавателю на проверку. Если работа оценивается положительно, то преподавателем назначается время для ее защиты студентом.

Основные требования к выполнению лабораторной работы

1. Каждый студент самостоятельно подбирает исходные данные по совокупности, которую характеризуют количественные и качественные признаки (например, в качестве совокупности могут быть использованы: продукция, предприятия, недвижимость, услуги, образовательная сфера и др.). Необходимая статистическая информация по выбранной совокупности подбирается из ежегодных статистических, экономических, финансовых журналов, газет и других специализированных изданий, интернета. Например, могут быть использованы такие издания, как «Россия в цифрах», «Эксперт», «Financial Times», «Top-Manager», «Российский экономический

журнал», «Вопросы экономики», «Финансовые известия», «СтройБизнесМаркет» и др.

2. Число наблюдений (единиц совокупности) должно быть не меньше 35, но не больше 50. Количество признаков – 5–6, из которых 4 являются количественными.

3. В начале каждой работы ставится цель, определяются задачи и подробно излагается алгоритм выполнения работы. Все результаты анализа (как промежуточные, так и конечные) должны быть прокомментированы в виде обобщений и выводов (например, выявление закономерностей развития явления во времени, тесноты взаимосвязи между анализируемыми признаками явления и др.).

Основные требования к оформлению отчета по лабораторной работе

Отчет должен содержать: 1) титульный лист, на котором указывается наименование высшего учебного заведения, факультет, кафедра, дисциплина, номер и название лабораторной работы; фамилия, имя и отчество студента, курс, номер его учебной группы; фамилия и инициалы, ученая степень и должность преподавателя, проверяющего работу; 2) исходные данные для выполнения лабораторной работы и их источники; 3) подробное описание хода выполнения лабораторной работы; 4) построение определенных в методических указаниях графиков, схем, таблиц; 5) пояснительную записку с изложением результатов статистической обработки исходных данных, выводов и заключений, которые должны подтверждаться приведением цифровых данных из расчетной части.

Все страницы работы и приводимые формулы должны быть пронумерованы, а на формулы в тексте должны быть даны ссылки.

Лабораторная работа № 1. ГРУППИРОВКА, СВОДКА И СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Целью объединения, сведения данных в статистических таблицах является выявление обобщающих закономерностей, характерных для изучаемой совокупности объектов наблюдения как целостной системы. Статистическая группировка представляет собой последующее разбиение совокупности на однородные группы по существенным для конкретной задачи анализа признакам. В таких однородных группах наблюдения,

образованных из объектов, сходных по одним признакам, изучается близость или различие тех же объектов по другим признакам. С помощью группировок решаются следующие *задачи*:

1. Разделение всей совокупности на качественно однородные группы – выделение социально-экономических типов. Эти группировки называются *типологическими* (например, группировки хозяйственных объектов по формам собственности, населения – по общественным группам и др.).

2. Характеристика структуры явления и структурных сдвигов. Такие группировки называются *структурными* (например, определение значения каждого вида транспорта в транспортном балансе страны, изучение состава населения по полу, возрасту и другим признакам и т. д.).

3. Изучение взаимосвязей между отдельными признаками изучаемого явления. Эти группировки называются *аналитическими* (например, группировка предприятий определенной отрасли экономики по уровню производительности труда с целью определения ее влияния на себестоимость продукции).

Для выполнения группировки сначала выбирается группировочный признак и определяется число выделяемых групп. Деление группировок на типологические, структурные и аналитические условно. Типологические и структурные группировки в качестве основания группировки используют содержательные признаки объектов. При этом типологические группы (если возможно) упорядочены по качественному их взаимному соотношению (например, последовательно указаны с определенной точки зрения плохие, средние, хорошие объекты). Структурные группировки основаны на выделении групп в порядке их возрастания или убывания с указанием их относительного объема. При аналитической группировке основанием для деления служит величина некоторого факторного признака, различие в уровне которого влияет на величину определенного результативного признака. Разновидностью аналитических группировок являются *комбинационные* группировки, которые характеризуются включением в качестве группировочных двух и более признаков объектов и позволяют выявить взаимосвязь между признаками-факторами и признаком-результатом.

Признак, на основе которого производится подразделение единиц совокупности на группы, называется *группировочным признаком,* или *основанием группировки*. Группировка может выполняться по одному

признаку (простая группировка) или по нескольким (комбинационная группировка). Выбор группировочного признака основывается на анализе качественной природы изучаемого явления.

Группировочные признаки могут быть *количественными* и *качественными* (*атрибутивными*). Количественные признаки имеют цифровое выражение (стаж работы, размер дохода и др.). Качественные признаки регистрируются в виде текстовой записи (например, социальная группа населения, профессия рабочих, отрасли экономики и др.). При группировке по количественному признаку число групп определяется в зависимости от характера изменения признака и задач исследования. При группировке по качественному признаку число групп определяется количеством соответствующих наименований, если число этих наименований не очень велико. Если признак имеет большое количество разновидностей, то при группировке ряд наименований объединяют в одну группу. Для обоснованного объединения их в группы разрабатываются классификации.

Отличающиеся друг от друга значения варьирующего признака в конкретной совокупности наблюдений называются *вариантами.* Ряды распределения показывают, сколько раз встречается в совокупности каждая из вариант (частота вариант) или какая доля всех наблюдений соответствует отдельным вариантам (частость). При небольшом количестве вариант ряд распределения состоит из перечисления всех вариант и указания частоты каждой из них (то есть принимает определенные значения из конечного набора значений, выражаемых, как правило, целыми числами). Такой ряд называют *дискретным*. Если число вариант значительно, образуют *интервальный* ряд распределения.

Интервальные ряды, позволяя повысить доступность восприятия информации о форме распределения данных по оси признака, приводят к укрупнению, объединению данных в группы. Внутри каждой из них все изначально разные наблюдения принимаются равными некоторому характерному групповому значению. Это позволяет качественно просто характеризовать различия групп, существенно упрощает расчеты, но может снизить точность описания характера распределения при расчетах, общих для всей совокупности параметров.

Интервалы группировки (разница между максимальным и минимальным значениями признака в группе) могут быть а) *равными,* б) *неравными* (*постепенно увеличивающимися*) и в) *специализированными*.

При построении равноинтервальной группировки ширина интервала группировки (*i*) равна размаху вариации (*R*), деленной на количество интервалов (1.1):

$$i = x_{\max} - x_{\min} / n \quad , \qquad (1.1)$$

где *n* – число групп.

Число групп (*n*) определяется по формуле Стерджесса (1.2):

$$n = 1{,}000 + 3{,}322 \lg N, \qquad (1.2)$$

где *N* – объем совокупности количество единиц совокупности.

При значительной неравномерности распределения данных по оси признака структура этого распределения может быть выявлена путем применения группировки на неравные интервалы с примерно одинаковым числом наблюдений в каждом интервале. Неравные интервалы (постепенно увеличивающиеся) применяются в структурных, аналитических и комбинационных группировках. Логическим требованием при составлении разнообъемных рядов является следующее: нельзя допускать попадания одних и тех же вариант в разные интервалы.

Специализированные интервалы используются в типологических группировках. Границы устанавливаются там, где намечается переход от одного качества к другому. Наметить точки перехода можно только на основе теоретического анализа, используя для выделения типов не отдельные, изолированные признаки, а совокупность признаков, характеризующих различные стороны изучаемого явления.

Интервалы группировки могут быть *закрытыми* и *открытыми*. Закрытые интервалы – это обычные интервалы, имеющие как нижние (то есть «от»), так и верхние (то есть «до») границы. Открытые интервалы – это интервалы, имеющие какую-либо одну границу – верхнюю или нижнюю. Они применяются тогда, когда признак изменяется неравномерно в широких пределах, причем большие (или малые) значения признака встречаются нечасто.

Следующей за группировкой ступенью систематизации и обобщения материалов статистического материала является *статистическая сводка*. Под статистической сводкой понимается подсчет числа единиц в подгруппах

и группах, выделенных при группировке, и подведение итогов по количественным признакам.

Результаты группировки и сводки материалов оформляются в виде *статистических таблиц*. В статистической таблице выделяются два элемента: *подлежащее* (обычно помещается в первой вертикальной или в горизонтальной графе) – перечень единиц или групп, на которые подразделена вся масса единиц совокупности; *сказуемое* – цифры, при помощи которых характеризуются выделенные в подлежащем единицы или группы.

По статистическим таблицам для графической иллюстрации строят *гистограммы* (для равноинтервальных рядов распределения), *полигоны* (для неравноинтервальных рядов распределения) и *кумуляты* (кривые накопленных частот или частостей). Кумуляты строятся как для равноинтервальных, так и для неравноинтервальных рядов распределения.

$\bar{x}$ Для характеристики рядов распределения определяются показатели центра распределения и показатели вариации. В качестве показателей центра распределения используется среднее арифметическое всех значений совокупности (), мода (Mo) и медиана (Me).

Среднее арифметическое (простое и взвешенное) рассчитывают по формулам (1.3)–(1.5):

$$\bar{x} = \sum x_i / N, \quad (1.3)$$

$$\bar{x} = \sum x_i f_i / \sum f_i, \quad (1.4)$$

$$\bar{x} = \sum x_i q_i / \sum q_i, \quad (1.5)$$

где x_i – индивидуальные значения наблюдения i или значение варианты i; N – объем совокупности; f_i – частота варианты i; q_i – частость варианты i.

Мода (Mo) – значение признака, которое наиболее часто встречается в совокупности (в ряду распределения). Мода для интервального ряда распределения рассчитывается по формуле (1.6):

$$\mathrm{Mo} = x_0 + i\frac{f_{\mathrm{Mo}} - f_{\mathrm{Mo}-1}}{(f_{\mathrm{Mo}} - f_{\mathrm{Mo}-1}) + (f_{\mathrm{Mo}} - f_{\mathrm{Mo}+1})}, \quad (1.6)$$

где x_0 – нижняя граница модального интервала; i – величина интервала; f_{Mo} – частота в модальном интервале; $f_{\mathrm{Mo}-1}$ – частота в предыдущем интервале; $f_{\mathrm{Mo}+1}$ – частота в следующем интервале за модальным.

Медиана (Me) делит совокупность наблюдений на две равные по объему части так, чтобы число элементов в совокупности с индивидуальными

значениями, меньше медианы, было равно числу элементов в совокупности с индивидуальными числами, больше медианы.

Медиана для интервального ряда определяется по формуле (1.7):

$$\mathrm{Me} = x_0 + i \cdot \frac{N/2 - F_0}{f_{\mathrm{Me}}}, \quad (1.7)$$

где x_0 – нижняя граница интервала, в котором находится медиана; i – величина интервала; F_0 – накопленная частота в интервале, предшествующем медианному; f_{Me} – частота в медианном интервале.

Если число единиц совокупности (N) нечетное, то номер медианы определяется не как $N/2$, а как $(N + 1)/2$.

Для дискретного ряда медиана определяется следующим образом. Так, если совокупность содержит нечетное число единиц, то все элементы совокупности располагаются в порядке возрастания значения признака и за медиану принимается значение признака, стоящего в середине такого ранжированного ряда. Положение медианы определяется ее номером: $N_{\mathrm{Me}} = (N + 1)/2$. Если совокупность содержит четное число единиц ($N = 2k$, $k = N/2$), то в этом случае за медиану условно принимают значение $\mathrm{Me} = 1/2(x_k + x_{k+1})$, так как в ряду нет члена, который делил бы совокупность на две равные по объему группы.

Кроме того, медиану для дискретного ряда распределения можно определить по накопленным частотам. Для этого сначала определяется медианный номер: для четного числа единиц совокупности $N_{\mathrm{Me}} = N/2$; для нечетного числа единиц совокупности $N_{\mathrm{Me}} = (N + 1)/2$. Затем накапливаются частоты до тех пор, пока накопленная частота не превысит медианный номер. Значение признака, которому соответствует первая накопленная частота, превысившая медианный номер, является медианой.

$\bar{x}$ Если совпадают среднее арифметическое (), мода (Mo) и медиана (Me), то ряд распределения является симметричным. Если Mo > , то ряд будет иметь левостороннюю асимметрию, а если Mo < , то правостороннюю асимметрию. В умеренных рядах соотношение между указанными показателями выражается следующим образом (1.8):

$$|\mathrm{Mo} - \bar{x}| \le 3|\mathrm{Mo} - \bar{x}|, \quad (1.8)$$

Мода и медиана могут быть определены графически с помощью гистограммы и кумуляты. Гистограмма создается для равноинтервального

ряда распределения, для чего на оси абсцисс строится ряд сомкнутых прямоугольников, у каждого из которых основанием служит величина интервала признака, а высотой является частота каждого интервала. Для определения моды по гистограмме правую вершину модального прямоугольника соединяют с правым верхним углом предыдущего прямоугольника, а левую вершину модального прямоугольника – с левым верхним углом последующего прямоугольника. Из точки пересечения этих прямых опускают перпендикуляр на ОХ. Абсцисса этой точки и будет модой распределения. Для рядов с неравными интервалами в качестве высоты прямоугольников принимается плотность распределения или строится полигон. Для графического определения медианы используется кумулята. Для этого из верхней границы каждого интервала на оси абсцисс восстанавливается перпендикуляр, соответствующий по высоте накопленной частоте с начала ряда по данный интервал. Соединив последовательно вершины перпендикуляров, получают кривую, называемую кумулятой. Из точки на оси ординат, соответствующей половине всех частот (порядковому номеру медианы), проводят прямую, параллельную оси абсцисс, до пересечения ее с кумулятой. Опустив из этой точки перпендикуляр на ось абсцисс, находят значение медианы (Me). Пользуясь кумулятой, можно определить значение признака у любой единицы ранжированного ряда.

σ_x^2 Кроме показателей центра распределения, совокупность данных характеризуется их разбросом относительно этого центра. Для этого рассчитывают следующие показатели вариации: размах (R), дисперсия (), среднее квадратическое отклонение () и коэффициент вариации (v).

σ^2 Размах (R) представляет собой разность между максимальным и минимальным значениями варианты. Мера рассеивания значений показателя относительно среднего арифметического задается дисперсией () – средним значением квадрата отклонения вариантов среднего арифметического (1.9)–(1.11):

$$\sigma_x^2 = \sum (x_i - \bar{x})^2 / N, \quad (1.9)$$

$$\sum (x_i - \bar{x})^2 \quad (1.10)$$

$$\sigma_x^2 = \sum (x_i - \bar{x})^2 \quad (1.11)$$

σ_x Наряду с дисперсией в качестве меры рассеивания широко используется среднее квадратическое отклонение (СКО) (), рассчитываемое как корень квадратный из дисперсии (1.12):

$$\sigma_x = \sqrt{\sigma_x^2}. \qquad (1.12)$$

Следующим важным показателем, характеризующим степень разброса значений варианты от среднего значения, является коэффициент вариации ν, определяемый как отношение СКО к величине среднего арифметического, умноженное на 100 % (1.13):

$$\nu = \sigma_x / \bar{x} \cdot 100 \quad \% . \qquad (1.13)$$

Если значение показателя ν < 33 %, то можно сделать вывод о том, что совокупность однородна, степень отклонения индивидуальных значений варианты от средней величины незначительна и средняя арифметическая типична. Следовательно, подобные результаты анализа целесообразно распространять на всю совокупность.

Задание

1. Определить цель и задачи данной работы, отобрать из множества факторов три-четыре, в наибольшей степени характеризующие объекты с точки зрения выбранной цели.
2. По каждому фактору построить равно- и неравноинтервальные ряды распределения, обосновав количество интервалов (по формуле Стерджесса).
3. Построить две структурные группировки (равноинтервальую и неравноинтервальную) по одному (двум) из количественных признаков, построить гистограмму (полигон) и кумуляты, обобщить полученные результаты группировки и сводки данных. Рассчитать показатели центра распределения: среднюю арифметическую, моду и медиану, сделать выводы относительно симметричности ряда распределения. Определить моду и медиану графически. Рассчитать показатели вариации и сделать выводы по коэффициенту вариации. Рассчитать показатели центра распределения: среднее арифметическое, моду и медиану, сделать выводы относительно симметричности ряда распределения. Определить моду и медиану графически. Рассчитать показатели вариации и сделать выводы по коэффициенту вариации.
4. Построить типологическую группировку по качественному признаку. Обосновать выбор признака, сделать выводы по результатам группировки.
5. Отобрать два наиболее взаимосвязанных признака, определить признак-фактор и признак-результат. Построить две аналитические группировки (равноинтервальную и неравноинтервальную). Выявить закономерность

развития совокупности и связь признаков, характеризующих ее. Обобщить полученные результаты группировки. Построить гистограмму (полигон) и кумуляты. Сравнить их с точки зрения отражения ими структуры распределения. Рассчитать показатели центра распределения: среднее арифметическое, моду и медиану, сделать выводы относительно симметричности ряда распределения. Определить моду и медиану графически. Рассчитать показатели вариации и сделать выводы по коэффициенту вариации.

6. Построить комбинационную группировку на основе развития и видоизменения одной из аналитических группировок, построенных в п. 5 данного задания. Выявить тип и характер связи между анализируемыми признаками. Сделать выводы.

Лабораторная работа № 2. СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ВЗАИМОСВЯЗИ

В данной лабораторной работе использование дисперсионного анализа рассматривается на примере количественных факторов. Общая степень изменчивости признака-результата определяется его общей дисперсией. Сила зависимости результата (y) от фактора (x) характеризуется той частью общей дисперсии, которая вызывается (объясняется) изменчивостью этого фактора. Для выполнения лабораторной работы следует использовать аналитические группировки, построенные в первой лабораторной работе. Начинать анализ связей признаков следует с изучения влияния признака-фактора на признак-результат. Группировка должна быть равно- и неравноинтервальной. В каждой из групп необходимо определить среднее значение результата и его внутригрупповую дисперсию как характеристику силы влияния на результат всех остальных факторов (учтенных и бесконечного числа неучтенных в условии задания). Результаты обработки данных во всех группах следует свести в отчете по лабораторной работе в табл. 2.1 (гипотетический пример).

Таблица 2.1

Результаты дисперсионного анализа статистических данных

Показатели		x_i	y_i	f_i	$\sigma_i^2 f_i$	$y_i - \overline{y}$	$(y_i - \overline{y}) f_i$
Интер	1	1,90	279,3	10	54 334,10	–247,96	61 484,2
	2	4,45	413,6	10	32 614,40	–113,66	12 918,6
	3	7,25	555,1	10	148 619,00	27,84	775,1

	4	9,80	705,3	9	514 152,00	178,04	31 698,2
	5	12,70	683,0	11	157 420,00	155,74	24 255,0
Сумма		$x_i f_i$	$z_i f_i$	$N = 50$	907 147,90	$y_i - \bar{y}$	$(y_i - \bar{y})^2 f_i$
Среднее		7,22	527,26	–	181 142,79	0,00	26 226,0

Окончание табл. 2.1

Показатели	x_i	y_i	f_i	$\sigma_i^2 f_i$	$y_i - \bar{y}$	$(y_i - \bar{y}) f_i$
Обозначение среднего	$\bar{x} = \frac{\sum x_i f_i}{N}$	$\bar{y} = \frac{\sum y_i f_i}{N}$	–	$\bar{\varepsilon}^2 = \frac{\sum \sigma_i^2 f_i}{\sum f_i}$	$\frac{\sum (y_i - \bar{y}) f_i}{N}$	$\delta^2 = \frac{(y_i - \bar{y})^2 f_i}{N}$

N, $\sum$ – объем совокупности; x_i – i-й признак-фактор; y_i – i-й признак-результат; f_i – частота i-й группы (численность i-й группы); $\bar{x}$ – среднее значение признака-фактора; $\bar{y}$ – среднее значение признака-результата; $\bar{\varepsilon}^2$ – средняя внутригрупповая дисперсия признака-результата; δ^2 – межгрупповая дисперсия признака-результата.

В столбцах x_i и y_i этой таблицы указаны пары среднегрупповых уровней фактора и результата. Соединив точки (x_i; y_i) на графике отрезками прямых, получаем ломаную, которая называется *эмпирической линией регрессии* и служит для определения класса функции взаимосвязи признака-результата с признаком-фактором.

Основной результат построения эмпирической регрессии предполагает выбор класса функции и оценку тесноты взаимосвязи между признаками, то есть расчет коэффициента детерминации (2.1) и корреляционного (дисперсионного) отношения (2.2):

$$\eta^2 = \delta^2 / (\delta^2 + \bar{\varepsilon}^2), \quad (2.1)$$

$$\eta = \sqrt{\delta^2 / (\delta^2 + \bar{\varepsilon}^2)}, \quad (2.2)$$

где δ^2 – межгрупповая дисперсия (дисперсия групповых средних) признака-результата; $\bar{\varepsilon}^2$ – средняя внутригрупповая дисперсия признака-результата.

Корреляционное отношение (η) количественно измеряет тесноту взаимосвязи между признаком-результатом и признаком-фактором. Чем больше корреляционное отношение, тем теснее взаимосвязь между признаками. При анализе взаимосвязи возможны следующие ситуации: $\eta = 1$ – взаимосвязь носит функциональный характер; $\eta = 0$ – взаимосвязь отсутствует; $0 < \eta < 1$ –взаимосвязь носит корреляционный характер.

Качественную оценку тесноты корреляционной связи между признаками можно представить в следующем виде (табл. 2.2).

Таблица 2.2

Качественная оценка тесноты связи между признаками

η	0	0…0,2	0,2…0,3	0,3…0,5	0,5…0,7	0,7…0,9	0,9…0,99	1
Связь	отсут-ствует	очень слабая	слабая	умерен-ная	заметная	тесная	очень тесная	функцио-нальная

δ_0^2 При определении коэффициента корреляционного отношения (η) рассчитывают три дисперсии: *среднюю внутригрупповую* (), *межгрупповую* () и *общую* ().

$\bar{\varepsilon}^2$ Средняя внутригрупповая дисперсия (дисперсия групповых средних () характеризует случайную вариацию результата, возникающую под влиянием других, неучтенных факторов, и не зависит от условия (признака-фактора), положенного в основу группировки. В корреляционном анализе она называется остаточной дисперсией и определяется по формуле (2.3):

$$\bar{\varepsilon}^2 = \sum \sigma_i^2 n_i / \sum n_i, \quad (2.3)$$

где σ_i^2 – дисперсия признака-результата по отдельной *i*-й группе. Ее можно определить по формуле (2.4):

$$\sigma_i^2 = \sum (y_{ij} - \bar{y}_i)^2 f_i / \sum f_i, \quad (2.4)$$

где y_{ij} – *j*-значение признака-результата в *i*-й группе, $\bar{y}_i$ – среднее значение признака-результата в *i*-й группе.

δ^2 Межгрупповая дисперсия (дисперсия групповых средних) отражает систематическую вариацию результата, то есть те различия в величине изучаемого признака, которые возникают под влиянием фактора, положенного в основу группировки. Межгрупповая дисперсия называется объясненной и определяется по формуле (2.5):

$$\delta^2 = \sum (\bar{y}_i - \bar{y}_o)^2 n_i / \sum n_i, \quad (2.5)$$

где $\bar{y}_i$ – среднее значение признака-результата по отдельной *i*-й группе; n_i – число единиц в *i*-й группе; $\bar{y}_o$ – общее среднее значение признака-результата (по всей совокупности).

Указанные дисперсии взаимосвязаны между собой определенным равенством. Величина общей дисперсии равна сумме межгрупповой дисперсии признака-результата и средней внутригрупповой дисперсии признака-результата (2.6):

$$\sigma_0^2 = \delta^2 + \bar{\varepsilon}^2. \quad (2.6)$$

Это тождество отражает закон (правило) сложения дисперсий. Опираясь на это правило, можно определить, какая часть общей дисперсии складывается под влиянием признака-фактора, положенного в основу группировки.

r_{xy} При линейной форме связи для измерения тесноты связи кроме корреляционного отношения используется также другой показатель, который называется коэффициентом линейной корреляции (). Он определяется по формуле (2.7):

$$r_{xy} = ((\overline{xy}) - \overline{x} \cdot \overline{y}) / \sigma_x \sigma_y. \quad (2.7)$$

Коэффициент линейной корреляции может принимать значения от –1 до +1. Отрицательные значения указывают на наличие обратной (убывающей) линейной зависимости, положительные – прямой (возрастающей) линейной зависимости. Если коэффициент линейной корреляции равен нулю, то можно сделать вывод, что линейная связь отсутствует.

При построении аналитической регрессии от уравнения, описывающего связь, не требуется, чтобы ему удовлетворяли значения признаков для всех элементов совокупности. Требуется лишь, чтобы функция, описывающая уравнение связи, была «ближайшей» к рассматриваемой корреляционной зависимости.

При решении этой задачи большое распространение получил метод наименьших квадратов (МНК), суть которого состоит в способе измерения «расстояния» между исследуемой корреляционной зависимостью и некоторой функцией. В качестве такой меры в методе наименьших квадратов принят средний квадрат отклонения индивидуальных значений признака-результата (y) от соответствующих значений функции $f(x)$.

Если считать, что исследуемая связь между признаками – линейная, то нужно определить параметры линейного уравнения регрессии (2.8):

$$\overline{y}_x = a + b\overline{x}. \quad (2.8)$$

На основе системы нормальных уравнений можно получить (2.9):

$$\begin{cases} \overline{y}_x = a + b\overline{x} \\ (\overline{xy}) = a\overline{x} = b\overline{x}^2 \end{cases} \quad (2.9)$$

Параметры 2 линейного уравнения регрессии a и b можно следующим образом (2.10) и (2.11):

$$b = r_{xy}\sigma_y / \sigma_x; \quad (2.10)$$

$$a = \overline{y} - b\overline{x}. \quad (2.11)$$

Задание

1. Определить цель статистического исследования, отобрать из множества признаков-факторов, присущих объекту наблюдения (совокупности), два в наибольшей степени (априори) влияющие на признак-результат (для

выполнения данной работы целесообразно воспользоваться аналитическими группировками, выполненными в предыдущей работе). Для каждой пары признаков построить корреляционное поле.

2. По каждому признаку-фактору построить эмпирическую линию регрессии, оценить тесноту связи между признаками и определить класс функции взаимосвязи. Для этого рассчитывают коэффициент корреляционного отношения (η) и коэффициент линейной корреляции ().

3. Отыскать аналитическое выражение функции взаимосвязи групповых средних, для которого взаимосвязь признака-фактора с признаком-результатом имеет более тесный характер. Построить аналитическую линию регрессии в той же системе координат, что и эмпирическая линия регрессии. Дать сравнительную оценку полученных результатов.

4. Определить параметры линейного уравнения регрессии и дать интерпретацию коэффициентов *a* и *b*.

Лабораторная работа № 3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ, МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Если в процессе анализа формы эмпирической линии регрессии или исходя из внутреннего содержания исследуемых явлений установлено, что исследуемую связь допустимо считать линейной, возможности количественного анализа существенно расширяются за счет применения развитого линейного аппарата моделирования взаимосвязей. Сила линейной взаимосвязи двух признаков измеряется коэффициентом корреляции, применение которого к изучению нелинейных связей недопустимо. Применение корреляционного анализа, линейных моделей необходимо обосновывать. Линейные модели широко применяются для аналитического описания зависимостей признаков (при допустимой их точности). Иногда переход к линейным методам анализа взаимосвязей обеспечивается использованием качественно обоснованных нелинейных преобразований исходных переменных.

В общем случае многофакторная линейная регрессия имеет вид (3.1):

$$\bar{y}_x = a_0 + \sum a_i \bar{x}_i, \qquad (3.1)$$

где $\bar{y}_x$ – расчетное значение признака-результата при фиксированном значении признака-фактора $x = (x_1, x_2, x_3, \ldots, x_k)$; a_0 – свободный член; $a_1, a_2, \ldots, a_k$ – коэффициенты регрессии; $x_1, x_2, \ldots, x_k$ – факторные признаки; $\bar{x}_i$ – среднее

значение признака-результата. При фиксированном x_i вариация бесчисленного множества остальных факторов (неучтенных) в частном измерении может приводить к отличиям от y.

Параметры уравнения множественной регрессии также рассчитываются МНК. Решается система нормальных уравнений с $k + 1$ неизвестным (3.2):

$$\begin{cases} a_0 n + a_1 \sum x_{i1} + a_2 \sum x_{i2} + \ldots + a_k \sum x_{ik} = \sum y_i \\ a_0 \sum x_{i1} + a_1 \sum x_{i1}^2 + a_2 \sum x_{i1} x_{i2} + \ldots + a_k \sum x_{i1} x_{ik} = \sum y_i x_{i1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \sum x_{ik} + a_1 \sum x_{ik} x_{i1} + a_2 \sum x_{ik} x_{i2} + \ldots + a_k \sum x_{ik}^2 = \sum y_i x_{ik} \end{cases} \quad (3.2)$$

где x_{ij} – значение j-го факторного признака в i-м наблюдении; y_i – значение результативного признака в i-м наблюдении.

В данной лабораторной работе используется модель связи факторов вида (3.1), устанавливающая соотношения абсолютных приростов результирующей и факторной переменных относительно средних арифметических соответствующих переменных.

Коэффициенты определяют знак и тесноту связи соответствующих пар признаков.

Чем ближе коэффициент корреляции по модулю к единице, тем ближе связь двух признаков к линейной функциональной. При функциональной линейной связи результата и фактора каждому значению фактора соответствует единственное значение результата. Для статистической связи при одном и том же уровне данного фактора могут быть различные значения результата. На значение результата при этом оказывают влияние и другие факторы. Для достаточно большого числа наблюдений статистическая связь между фактором x и результатом y проявляется «в среднем». Это означает, что при данном x можно с помощью регрессионного уравнения предсказывать, каким в среднем будет значение признака y при всем возможном многообразии не-учтенных факторов. Чем больше по модулю коэффициент корреляции конкретных факторов, тем в большей мере они достойны быть включенными в регрессионную модель.

Для расчета указанных коэффициентов корреляции необходимо кроме средних уровней самих признаков и их произведений знать среднеквадратические отклонения этих признаков. Формула расчета коэффициента линейной корреляции имеет вид (3.3):

$$r = ((\overline{xy}) - \overline{x} \cdot \overline{y}) / \sigma_x \sigma_y. \quad (3.3)$$

Анализ показателей линейной корреляции позволяет отбирать в регрессионную многофакторную модель признаки-факторы. В модель должны быть включены коэффициенты линейной корреляции, существенно влияющие на признак-результат. В данной работе необходимо рассчитать с помощью ПК параметры многофакторного регрессионного уравнения «классического вида» (3.1). В уравнении коэффициент a_0 находится через коэффициенты a_i и другие параметры переменных следующим образом (3.4):

$$a_0 = \bar{y}_x - \sum a_i \bar{x}_i, \tag{3.4}$$

где $\beta_i = \beta_i \sigma_{yi} / \sigma_{xii}$.

В свою очередь коэффициенты регрессии β_i определяются решением системы линейных уравнений с коэффициентами, равными найденным коэффициентам корреляции (3.5):

$$\begin{cases} r_{x1y} = \beta_1 + r_{x1x2}\beta_2 + \ldots + r_{x1xn}\beta_n \\ r_{x2y} = r_{x2x1}\beta_1 + \beta_2 + \ldots + r_{x2xn}\beta_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{xny} = r_{xnx1}\beta_1 + r_{xnx2}\beta_2 + \ldots + \beta_n \end{cases} \tag{3.5}$$

Коэффициент β_i показывает на какую часть сигмы (σ_y) изменилось бы значение результата, если бы соответствующий *i*-й фактор изменился на сигму (σ_{xi}), а прочие факторы остались без изменения.

Теоретическую точность уравнения определяют через коэффициент множественной корреляции *R*, рассчитываемый как (3.6):

$$R = \sqrt{r_{yx1}\beta_1 + r_{yx2}\beta_2 + \ldots + r_{yxn}\beta_n}. \tag{3.6}$$

Доля неучтенных факторов, остаточная дисперсия (σ^2) определяется в (3.7):

$$\sigma^2 = 1 - R^2. \tag{3.7}$$

Этот показатель тесноты связи факторов и результата всегда положителен ($0 \le R \le 1$). При *R*, стремящемся к 1, моделируемая связь становится функциональной. Ошибка (среднеквадратическая) уравнения определяется как (3.8):

$$E = \sigma_y \sqrt{1 - R^2}. \tag{3.8}$$

Далее определяются частные и совокупный коэффициенты эластичности, которые позволяют оценить изменение признака-результата при изменении фактора (факторов) на 1 %. Частные коэффициенты эластичности рассчитываются по формуле (3.9):

$$Э_j = \Delta x_j / \bar{x}_j : \Delta y / \bar{y} = a_j \bar{x}_j / \bar{y}, \tag{3.9}$$

где $\bar{y}_j$ – среднее значение j-го признака-фактора; Δx_j – изменение j-го признака-фактора; – среднее значение признака-результата; Δy – изменение признака-результата; a_j – коэффициент регрессии при j-м признаке-факторе. $Э_j$ показывает, на сколько процентов следует ожидать изменения результативного признака при изменении фактора j на 1 % и неизменном значении других факторов.

Совокупная эластичность позволяет оценить эластичность признака-результата в целом при совокупном изменении факторов (3.10):

$$Э_c = \sum Э_j. \qquad (3.10)$$

Задание

1. Определить цель статистического исследования, отобрать из множества признаков-факторов, присущих объекту наблюдения, 3–4 в наибольшей степени (априори) влияющих на признак-результат. При выполнении данной работы необходимо обосновать выбор существенных признаков-факторов, влияющих на признак-результат.
2. Рассчитать коэффициенты линейной корреляции признаков-факторов и признака-результата и свести их в корреляционную матрицу. Анализируя матрицу коэффициентов линейной корреляции и экспериментируя с построением модели многофакторной линейной регрессии с различным количеством и сочетанием факторов, составить многофакторную регрессионную модель уравнения, которая содержит 2–3 признака-фактора, существенно влияющих на признак-результат. Определить коэффициенты регрессии (β_i) и проинтерпретировать смысл этих коэффициентов.
3. Определить коэффициент множественной корреляции (R) и среднеквадратическую ошибку (E), обобщить полученные результаты.
4. Рассчитать частные и совокупный коэффициенты эластичности ($Э_i$), ($Э_c$). Сделать выводы.

Список рекомендуемой литературы

Елисеева И. И., Юзбашев М. М. Общая теория статистики: Учеб. 5 изд., перераб. и доп. М.: Финансы и статистика, 2003.

Экономическая статистика: Учеб. / Под ред. Ю. Н. Иванова. 2-е изд., доп. М.: ИНФРА-М, 2000.

Ефимова М. Р., Петрова Е. В., Румянцев В. Н. Общая теория статистики: Учеб. 2-е изд., испр. и доп. М.: ИНФРА-М, 2000.

Макарова Н. В., Трофимец В. Я. Статистика в Excel: Учеб. пособие. М.: Финансы и статистика, 2002.

Практикум по общей теории статистики: учеб. пособие / М.Р. Ефимова, О.И. Ганченко, Е.В. Петрова. – 3-е изд., перераб. и доп. – М.: Финансы и статистика, 2008. – 368 с.

Статистика: методические указания к практическим занятиям/ сост.: О.Г. Алексеева. --- СПб.: СПбГЭТУ «ЛЭТИ», 2009. – 79 с.

Статистика: методические указания по выполнению лабораторных работ/ сост.: А.Э. Сулейманкадиева. – СПб.: СПбГЭТУ «ЛЭТИ», 2006. – 20 с.

http//www.ibooks.ru

http://www.elibrary.ru

http://uisrussia.msu.ru/is4/main.jsp

http://www.alleng.ru