ПРАКТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ГОРНЫХ ИНЖЕНЕРОВ

1

С.Г. Баженова

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГОРНОЙ ПРОМЫШЛЕННОСТИ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ГОРНЫЙ УНИВЕРСИТЕТ

ПРАКТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ГОРНЫХ ИНЖЕНЕРОВ

1

С.Г. Баженова

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГОРНОЙ ПРОМЫШЛЕННОСТИ

2-е издание, исправленное

Рекомендовано Министерством образования Российской Федерации в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению «Горное дело»

МОСКВА

ИЗДАТЕЛЬСТВО МОСКОВСКОГО ГОСУДАРСТВЕННОГО ГОРНОГО УНИВЕРСИТЕТА

2001

УДК 662:658.5.001.57(075)
ББК 22.172:33
Б 16

Рецензенты:

- *Заместитель директора координационного центра «ИнтерАСУуголь», лауреат Государственной премии СССР, профессор М.А. БУРШТЕЙН*
- *Заведующий отделом ЦНИЭИуголь, кандидат технических наук Г.Г. ЯКУБСОН*

Баженова С.Г.

Б 16 Математико-статистические методы в горной промышленности: Учеб. пособие — 2-е изд., исправл. — М.: Издательство Московского государственного горного университета, 2001. — 99 с.

ISBN 5-7418-0100-5

Систематизированы основные характеристики статистической совокупности, большое внимание уделено математико-статистическим методам расчета, анализа и интерпретации полученных результатов. Каждый раздел представлен в виде локальной работы, снабжен краткими теоретическими сведениями и подробно рассмотренным примером расчета.

Для студентов, обучающихся по направлениям «Экономика», «Менеджмент» в горной промышленности.

УДК 662:658.5.001.57(075)
ББК 22.172:33

ISBN 5-7418-0100-5

ПРЕДИСЛОВИЕ

Предлагаемое читателю учебное пособие «Математико-статистические методы в горной промышленности» является второй книгой серии из четырех книг по статистике: «Практическая статистика», «Математико-статистические методы в горной промышленности», «Статистические методы оценивания показателей горного производства» и «Статистика. Термины и определения».

Это учебное пособие является естественным продолжением первой книги — «Практическая статистика», но может использоваться и самостоятельно.

Учебное пособие снабжено краткими теоретическими сведениями по обобщающим характеристикам статистической совокупности; приведены статистические методы расчета и полное описание примеров подобных расчетов.

Учебное пособие предназначено для студентов высшего горного образования, но может быть использовано специалистами и научными сотрудниками.

ОБЩИЕ ПОЛОЖЕНИЯ

Для приобретения навыков в работе по статистике весьма полезно выполнять отдельные части анализа как самостоятельные работы, поэтому ниже даются описания и рекомендации, отнесенные к отдельным темам, которые могут существовать отдельно от других.

Полный и систематический анализ наблюдений необходим для получения научно обоснованных выводов.

Отдельные темы выполняются как в процессе аудиторных занятий, так и в процессе самостоятельной работы.

К выполнению того или иного раздела задания разумно приступать после тщательной проработки соответствующего раздела теории.

Теоретической базой для выполнения работы являются знания, полученные при изучении дисциплины «Статистика».

Работа включает общее введение, где указываются цели данного статистического анализа, дается описание натурных наблюдений (откуда взяты, что фактически представляют, оценивается объем наблюдений и их достоверность, причины отбора факторов, взятых как натурные наблюдения и пр.).

Объектом исследования являются реальные данные горной промышленности.

При выполнении индивидуального задания следует руководствоваться следующим:

1. Законченная лабораторная или практическая работа, включающая расчеты, таблицы, графики должна сопровождаться краткой пояснительной запиской, где приводятся этапы решения в их последовательности.

Расчетная часть должна содержать формулы и все расчеты, в том числе и промежуточные.

2. Следует соблюдать правила точности вычислений. Точность вычислений не должна быть ниже той, которая соответствует начальным значениям, но и не выше требований задачи.

3. При выполнении расчетов, где можно предусмотреть вычислительный или смысловой контроль, следует провести контроль и убедиться в правильности расчетов.

4. При начальной обработке данных следует учесть размерности значений натурных наблюдений, привести их в соответствие и указать размерность при записи исходной информации.

5. При выполнении работы рекомендуется пользоваться микрокалькулятором и ЭВМ.

6. Законченная лабораторная или практическая работа оформляется согласно принятым правилам оформления и сдается на промежуточную проверку в заданные сроки.

В общем заключении дается конечный вывод проведенного полного статистического анализа и рекомендации по аналогичным статистическим данным.

Методические описания каждой работы включают постановку и алгоритм решения (цель работы и основные теоретические сведения), порядок выполнения работы, подготовку исходной информации, пример решения задачи.

Работа № 1

СТАТИСТИЧЕСКАЯ СОВОКУПНОСТЬ НАБЛЮДЕНИЙ. СБОР И ФОРМИРОВАНИЕ

Цель работы: Сформировать совокупность для статистического анализа.

1. Общие сведения

Статистические данные реальны, а потому они не могут быть полностью оторваны от реальной действительности. Но если даже построена математическая модель, то к результатам надо отнестись критически и проанализировать их.

Само же решение по полученной математической модели проводится автоматически, в полном отрыве от существа задачи.

Любое статистическое исследование выражается в следующих основных ступенях:

— массовое статистическое наблюдение. Сбор информации;

— сводка и группировка статистических данных. Формирование совокупности;

— построение статистических показателей и их анализ.

Статистическим наблюдением называется процесс получения данных о каких-либо явлениях путем регистрации их существенных признаков.

Наблюдения являются важнейшим звеном в статистическом исследовании, так как они дают материал, который в дальнейшем будет подвергнут обработке и анализу. Если материалы, собранные при статистиче-

ском наблюдении, будут неполными или неточными, это может повлечь за собой получение неправильного результата при анализе. Поэтому основными требованиями, предъявляемыми к наблюдению, являются правильность и безошибочность его проведения.

Организация наблюдения имеет две основные формы.

Первая — регистрация отдельных единиц наблюдения и запись результатов на специальных бланках. Таким образом могут формироваться «журнал наблюдений» или «таблица наблюдений» (рис. 1.1).

Статистическими данными можно назвать сведения о некотором числе объектов, обладающих теми или ины-

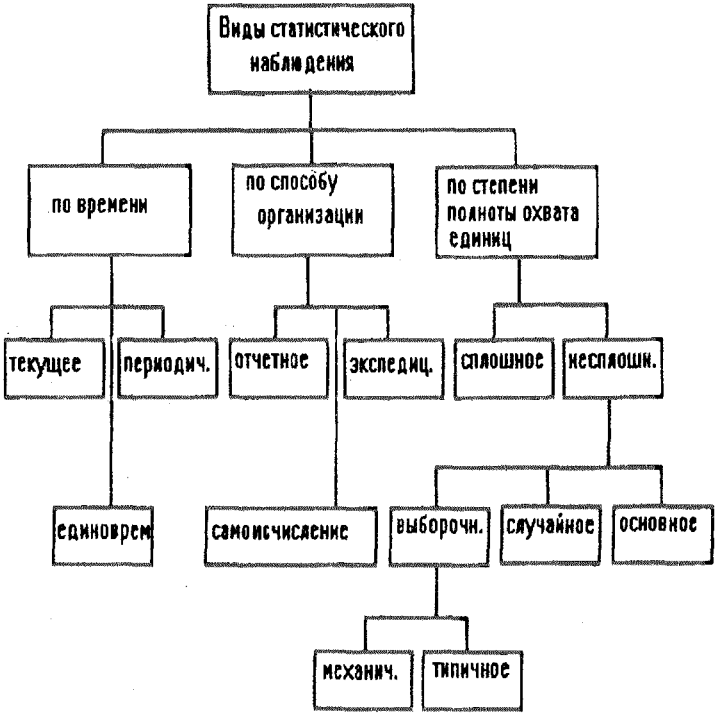

Рис. 1.1. Виды статистического наблюдения

ми признаками. Метод исследования статистических данных называется статистическим. Характерной особенностью настоящего этапа развития естественных и технических наук является широкое применение статистических методов и математики во всех областях знания. Но статистические методы в применении к различным областям настолько своеобразны, что их невозможно объединить в одну науку.

Рассмотрим применение математических методов исследования статистических данных для решения вопросов горного дела, экономики, организации и управления горным производством.

Статистика находит все большее применение в технике. Однако нельзя переоценивать статистические методы. Правильное применение математического анализа не может быть сведено к одним математическим приемам, а требует, прежде всего, предварительного теоретического анализа, хорошего знания физической сущности явления. Внедрение математической статистики в производственную практику поможет найти дополнительные пути к повышению культуры производства, эффективности использования оборудования и росту производительности труда, снижению себестоимости и увеличению рентабельности.

Вопросы рентабельности, себестоимости и т.д. рассматриваются в разделах и экономической и математической статистики.

Общими чертами **статистического метода** в различных областях знаний являются: сведения к подсчету числа объектов, входящих в те или иные группы; рассмотрение распределения количественных признаков; применение выборочного метода, когда детальное исследование всех объектов обширной совокупности (например массы полезного ископаемого) затруднительно; использование теории вероятностей при оценке достаточного числа наблюдений. Для некоторых наук вопросы приме-

нения вероятностно-статистических методов и машинной математики более или менее разработаны, для других, например, горного дела, еще очень много надо сделать.

Статистические методы опираются на массовые, повторяющиеся явления, в которых изменчивость обусловливается рядом причин. При этом задача статистики — установить характер этой изменчивости. При исследовании различных физических, химических, экономических, технических процессов часто встречаются явления, называемые случайными.

Как бы точно и подробно ни выполнялись условия отдельных экспериментов, невозможно достичь того, чтобы результаты полностью совпали. Результаты отдельных определений прочностных характеристик одной и той же породы еще больше будут отличаться друг от друга, если эти определения производятся в натуре, в различных участках горного массива, так как вариации факторов, сопровождающие эксперименты и порождающие различия результатов, будут еще более заметны.

Случайные отклонения неизбежно сопутствуют любому закономерному явлению, накладываясь на него.

В природе нет ни одного физического явления, в котором не присутствовали бы в той или иной мере случайности.

В ряде практических задач случайными элементами пренебрегают. При этом из бесчисленного множества признаков, влияющих на данное явление, выделяется главное, влиянием остальных, второстепенных, пренебрегают. Таким образом изучают некоторую модель явления. Затем применяется тот или иной аппарат (например, дифференциальные уравнения) для описания данного явления. Так можно выявить основную закономерность, свойственную данному явлению, и есть возможность предсказать результат при заданных условиях. Чем больше учтено признаков, влияющих на явление, тем точнее получаемый результат.

Эта схема решения пригодна лишь для задач, где исход опыта зависит от основного количества главных признаков, остающихся постоянными от опыта к опыту.

Множество сопутствующих качеств можно приблизительно представить в виде схемы, (рис. 1.2).

При изучении любого процесса, явления можно выделить совокупность однородных единиц. Такова совокупность отдельных объемов, блоков многоделимой массы горной породы. Такова совокупность единиц средств труда, с помощью которых осуществляется эксперимент: группа машин, приборов или деталей определенного типа.

Совокупность, состоящая из однородных единиц, обладающих количественной общностью, составляет **статистическую совокупность.**

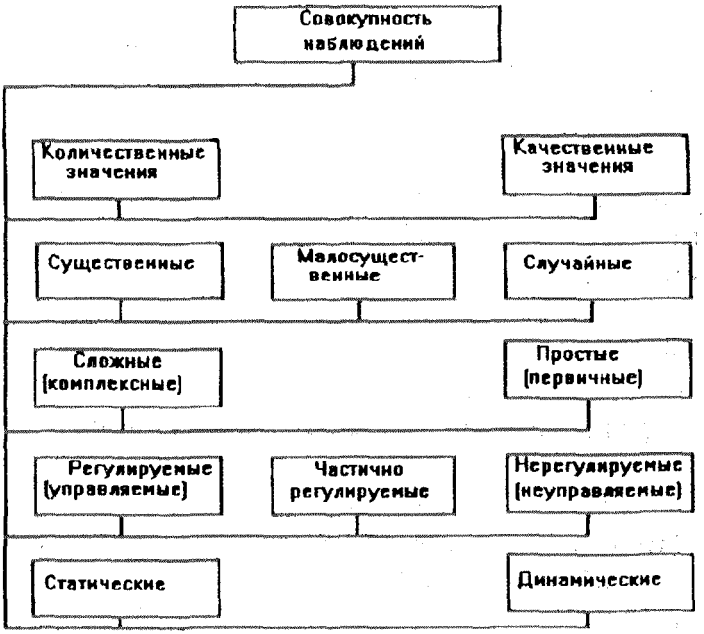

Рис. 1.2. Классификация наблюдений

Главное орудие статистики — обобщающие показатели, основанные на данных массового наблюдения. Для успешного применения методов математической статистики надо правильно разграничить совокупности, подлежащие обобщению. Такое разграничение производится методом группировки.

Группировка — важнейшее положение математической статистики. Необходимо отметить еще одно обстоятельство: математическая статистика тесно связана с теорией вероятностей. На предельных теоремах теории вероятностей базируется большинство выводов математической статистики.

2. Порядок работы

Для начала работы необходимо сформировать таблицу исходных значений, подлежащих анализу. Для этого в соответствии с индивидуальным заданием выбирается необходимая информация и заполняется таблица, в которой указываются наименования показателей предприятий, а в клетках таблицы — соответствующие значения показателей.

В качестве показателей, исследуемых в работе, как правило, используются:

1. Среднесуточная добыча угля, т.
2. Среднесписочная численность ППП по добыче, чел.
3. В том числе рабочих, чел.
4. Производительность труда за один отработанный чел.-день рабочего по добыче угля, кг.
5. Производительность труда за один отработанный чел.-день рабочего на очистных работах, кг.
6. Количество разрабатываемых шахтопластов, шт.

7. Средняя вынимаемая мощность пласта, м.

8. Максимальная глубина разработки, м.

9. Количество действующих очистных забоев, шт.

10. Длина действующих очистных забоев, м.

11. Количество комплексно-механизированных забоев, шт.

12. Добыча из действующих очистных забоев за год, всего, тыс. т.

13. В том числе из комплексно-механизированных забоев.

14. Среднедействовавшее за год количество очистных забоев, шт.

15. Среднедействовавшая за год линия очистных забоев, м.

16. Среднемесячное подвигание очистной линии забоев, м.

17. Среднесуточная добыча угля из одного действующего очистного забоя, т.

18. Проведение подготовительных выработок за год общее, м или другие показатели.

Перечисленные показатели могут быть изменены в соответствии с индивидуальным заданием. Формирование исходной таблицы сопровождается общим описанием работы предприятия, выделением максимальных и минимальных значений.

Исходная многомерная таблица показателей работы групп предприятий (табл. 1.1) подлежит анализу и исследованию.

Каждый столбец таблицы может рассматриваться как одномерная совокупность, попарно рассматриваемые столбцы — как двумерная совокупность. Совокупность, состоящая из трех и более столбцов — многомерная совокупность.

Работа № 1
Иванов И.И. группа _____
Таблица исходной информации
Группа предприятий №

Таблица 1.1

Предприятие	Показатели							
	1	2	3					*L*
№ 1								
№ 2								
⋮								
№ *N*								

3. Порядок выполнения работы

Каждый студент выполняет индивидуальное задание. Источником информации для выполнения работы являются сборники экономико-статистической отчетности горной промышленности за последние годы.

В соответствии с заданием заполняется таблица исходной информации (см. табл. 1.1). При этом из совокупности предварительно исключаются описки, промахи, резко выделяющиеся по величине варианты.

Полученная таблица исходной информации будет использована в работах № 3, № 4, № 5 и № 6.

В настоящей работе необходимо провести анализ данных по разбросу (оценить размах), построить графики и сделать выводы по предприятию в целом.

Данная работа является базовой и подготовительной для следующих, поэтому в ней отсутствуют контрольный пример и описание использования ЭВМ.

Работа № 2

ОПРЕДЕЛЕНИЕ НЕОБХОДИМОГО ОБЪЕМА НАБЛЮДЕНИЙ

Цель работы: Определить необходимый объем наблюдений при различных способах и схемах отбора единиц из генеральной совокупности в выборочную, с заданной вероятностью предельной ошибки.

1. Основные теоретические сведения

Как уже указывалось, под **статистическим наблюдением** понимают планомерный, научно организованный систематический сбор данных о явлениях и процессах путем регистрации заранее намеченных существенных признаков.

Статистическое наблюдение осуществляется в двух формах: представление отчетности и проведение специально организованных статистических наблюдений. При этом оно должно быть организовано таким образом, чтобы в результате были получены объективные, точные данные об изучаемом явлении, на основе которых можно правильно оценить и планировать развитие производства.

В зависимости от степени полноты охвата наблюдением изучаемого явления или объекта различают **сплошное и несплошное статистическое наблюдение** (рис. 2.1). При сплошном наблюдении обследованию подвергаются все без исключения единицы изучаемой совокупности, а при несплошном — только часть их. Различают следующие основные виды несплошного наблюдения: **выборочное** наблюдение, **монографическое** обследование и **метод основного массива.** Наиболее совершенным, науч-

Рис. 2.1. Оценка статистического наблюдения

но обоснованным способом несплошного наблюдения является выборочное наблюдение, при котором статистическому обследованию подвергаются не все единицы изучаемой совокупности, а лишь отобранные в определенном порядке и обеспечивающие получение данных, характеризующих всю совокупность в целом. К этому виду наблюдений прибегают в тех случаях, когда необходимо сэкономить силы и средства при проведении исследования, так как статистические исследования массовых совокупностей весьма трудоемки.

При выборочном наблюдении неизбежна некоторая свойственная ему погрешность, так как обследованию подвергается не вся совокупность, а только ее часть. Ошибки, свойственные выборочному наблюдению, называются ошибками **репрезентативности** (представительства). Ошибки характеризуют размер расхождения между данными выборочного наблюдения и всей совокупности в целом.

Различают **случайные и систематические** ошибки репрезентативности. **Случайные ошибки** обусловлены тем, что выбранная совокупность недостаточно точно воспроизводит всю совокупность. Их размеры и пределы заранее можно вычислить на основании закона больших чисел, а главное — довести до незначительных размеров путем включения в выборку достаточного количества единиц совокупности.

Систематические ошибки возникают при нарушении принципа случайности отбора единиц совокупности для наблюдения. Их можно избежать, осуществляя строгий объективный отбор единиц совокупности, при котором каждая из них имела бы абсолютно одинаковую возможность попасть в выборку.

В статистической литературе принято называть совокупность единиц, из которой производится отбор, генеральной совокупностью, а ее численность обозначают буквой N. Часть единиц, попавшая в выборку, называется выборочной совокупностью, а их численность обозначается буквой n.

Обобщающими характеристиками генеральной и выборочной совокупностей являются средняя и дисперсия: $\overline{X}$ и σ^2.

При формировании выборки применяют различные **виды, схемы и способы отбора. По виду отбор** единиц наблюдения подразделяют на индивидуальный (за прием отбирается одна единица), групповой (за один прием отбирается группа или серия единиц) и комбинированный.

В зависимости от **участия** отобранной единицы в дальнейшей выборке различают **повторную и бесповторную** схемы отбора. В первом случае отобранная однажды единица возвращается обратно в генеральную совокупность и снова участвует в выборке.

При **бесповторной схеме** отбора ранее отобранная единица не возвращается в генеральную совокупность и не может быть подвергнута повторному обследованию. Бесповторный отбор дает более точные результаты по

сравнению с повторным, так как при одном и том же объеме выборки наблюдение охватывает большее количество единиц изучаемой совокупности. Как правило, при выборочном наблюдении следует применять бесповторную схему отбора. И только в тех случаях, когда бесповторный отбор нельзя провести, применяется повторная схема отбора.

По способу формирования выборки различают собственно-случайный, механический, типический, серийный и комбинированный отбор. При **собственно-случайном способе** формирования выборки включение единиц в выборочную совокупность может осуществляться по схеме повторного или бесповторного отбора при помощи жеребьевки или таблицы случайных чисел. Такой способ формирования выборки является наиболее простым, но он уступает другим способам с точки зрения репрезентативности и точности результатов, а также из-за сложности в организационном отношении.

При **механическом способе** формирования выборки отбор единиц из генеральной совокупности производится механически через определенный интервал (например, выбирается каждая пятая, десятая и т. д. единица). Все единицы в изучаемой совокупности предварительно располагаются в определенном порядке (например, по алфавиту, местоположению и т. п.). Данный способ отбора является разновидностью собственно-случайного отбора, но имеет ряд организационных преимуществ и всегда бывает бесповторным. Поэтому средняя ошибка и необходимая численность механической выборки определяются по формулам собственно-случайной бесповторной выборки.

В тех случаях, когда генеральная совокупность неоднородна по показателям, подлежащим изучению, применяется **типический способ** отбора, при котором вся генеральная совокупность делится предварительно на группы по определенному типическому признаку, из которых в дальнейшем собственно-случайным или механическим способом формируется выборочная совокуп-

ность. Типический способ отбора также может быть как повторным, так и бесповторным.

Из всех типических групп можно отобрать число единиц, пропорциональное их численности, и непропорциональное. В зависимости от этого различают пропорциональный и непропорциональный типический отбор.

При **серийном способе отбора** вместо случайного отбора единиц совокупности осуществляется отбор групп (серий), внутри которых производится сплошное наблюдение. Точность серийной выборки зависит от межсерийной дисперсии (дисперсии групповых средних), а не от величины общей дисперсии.

При планировании выборочного наблюдения возникает вопрос о необходимой численности выборки. Последнюю можно определить, исходя из допустимой ошибки при выборочном наблюдении, вероятности, с которой нужно гарантировать величину устанавливаемой ошибки, меры колеблемости изучаемого признака и способа отбора.

Необходимая численность выборки определяется на основе предельной ошибки выборки. Если предельную ошибку выборки обозначить буквой σ, то последнюю можно определить из выражения

$$\sigma = T \times \mu,$$

где μ — средняя ошибка выборки; T — коэффициент, зависящий от вероятности, с которой гарантируется ошибка выборки (коэффициент доверия). Значения вероятности (P) для различных значений T приведены в таблице (табл. 2.3).

Предельная ошибка выборки зависит от трех факторов: степени колеблемости явления (σ^2), объема выборки (n) и от необходимой гарантированной вероятности (P). Формулы для вычисления предельных ошибок выборки при различных способах отбора приведены в таблице (табл. 2.1).

Таблица 2.1

Способ отбора	Схема отбора	Предельная ошибка выборки
Собственно случайный и механический отбор	Повторный	$\sigma = T \times \sqrt{\frac{\sigma^2}{n}}$
	Бесповторный	$\sigma = T \times \sqrt{\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)}$
Типический отбор	Повторный	$\sigma = T \times \sqrt{\frac{\overline{\sigma}_i^2}{n}}$
	Бесповторный	$\sigma = T \times \sqrt{\frac{\overline{\sigma}_i^2}{n}\left(1 - \frac{n}{N}\right)}$
Серийный	Повторный	$\sigma = T \times \sqrt{\frac{\overline{b}^2}{r}}$
	Бесповторный	$\sigma = T \times \sqrt{\frac{\overline{b}_i^2}{r}\left(1 - \frac{r}{R}\right)}$

В табл. 2.1 приняты следующие условные обозначения: $\overline{\sigma}_i^2$ — средняя из групповых дисперсий; $\overline{b}_i^2$ — межсерийная дисперсия; r — число отобранных серий; R — число серий в генеральной совокупности.

Величину $\overline{b}^2$ можно определить из выражения:

$$\overline{b}^2 = \frac{\sum\left(\widetilde{x}_i - \overline{X}\right)^2}{r},$$

где $\widetilde{x}_i$ — средняя в отдельных сериях; $\overline{X}$ — общая средняя для всей совокупности.

Проведя необходимые вычисления и пользуясь расчетными формулами (табл. 2.2), определить необходимую численность выборки при различных способах и схемах отбора.

Таблица 2.2

РАСЧЕТНЫЕ ФОРМУЛЫ ДЛЯ ОПРЕДЕЛЕНИЯ НЕОБХОДИМОЙ ЧИСЛЕННОСТИ ПРИ РАЗЛИЧНЫХ СПОСОБАХ И СХЕМАХ ОТБОРА

Способ отбора	Схема отбора	Предельная ошибка выборки
Собственно случайный и механический отбор	Повторный	$n = \frac{T^2 \times \sigma^2}{\delta^2}$
	Бесповторный	$n = \frac{T^2 \times \sigma^2 \times N}{\delta^2 \times N + T^2 \times \sigma^2}$
Типический отбор	Повторный	$n = \frac{T^2 \times \overline{\sigma}_i^2}{\delta^2}$
	Бесповторный	$n = \frac{T^2 \times \overline{\sigma}_i^2 \times N}{\delta^2 \times N + T^2 \times \overline{\sigma}_i^2}$
Серийный	Повторный	$r = \frac{T^2 \times \overline{b}^2}{\delta^2}$
	Бесповторный	$r = \frac{T^2 \times \overline{b}^2 \times N}{\delta^2 \times R + T^2 \times \overline{b}^2}$

Из приведенных в табл. 2.2 формул для определения необходимой численности выборки при различных способах и схемах отбора видно, что они зависят лишь от схемы отбора (повторный отбор или бесповторный). Способ отбора влияет следующим образом: в формуле для необходимой численности выборки при собственно-случайном или механическом способе вместо общей дисперсии σ^2 используют среднюю из внутригрупповых

дисперсий $\overline{\sigma}_i^2$ при типическом способе отбора и межсерийную дисперсию $\overline{b}^2$ при серийном способе. При этом в последнем случае вместо числа единиц в генеральной совокупности N используют число серий в генеральной совокупности R.

В связи с вышеизложенным в лабораторной работе при типическом и серийном способах отбора предварительно определяются значения $\overline{\sigma}_i^2$ или $\overline{b}_i^2$.

2. Порядок выполнения работы

1. Студент получает от преподавателя конкретный вариант исходных данных.

2. До выполнения расчетов на ЭВМ студенту необходимо предварительно выполнить следующую работу:

а) определить по табл. 2.3 минимальное, максимальное и заданное значение коэффициента доверия для заданного интервала значений вероятностей;

б) установить дополнительные значения предельной ошибки выборки;

в) при типическом и серийном способах отбора вычислить внутригрупповую $\overline{\sigma}_i^2$, и межсерийную $\overline{b}_i^2$ дисперсии, соответственно, и определить численность генеральной совокупности $\left(N = \sum N_i\right)$.

3. Ввод исходной информации в ЭВМ. Информация вводится в виде диалога путем ответов на вопросы ЭВМ. По введенным данным производятся вычисления, и выдается на экран дисплея и печать следующая информация:

— необходимая численность выборки при изменении вероятности предельной ошибки выборки в заданном интервале при повторной и бесповторной схемах отбора;

— необходимая численность выборки при заданных способе и схемах отбора, гарантирующая с заданной вероятностью предельную ошибку выборки;

— значения необходимой численности выборки при различных схемах отбора, если предельную ошибку выборки уменьшим или увеличим на заданную величину.

4. Построить графики зависимостей между вероятностью заданного результата и необходимой численностью выборки при повторной и бесповторной схемах отбора.

5. Анализ полученных результатов.

3. Подготовка исходной информации к расчетам на ЭВМ

Для решения задачи требуется следующая исходная информация: N — численность генеральной совокупности; R — предельная ошибка выборки, %; D — среднеквадратическое отклонение (дисперсия); $T1$ — минимальное значение коэффициента доверия, соответствующее нижней границе интервала вероятности гарантии необходимого результата, ($P1$); $T2$ — максимальное значение коэффициента доверия, соответствующее верхней границе интервала вероятности гарантии необходимого результата ($P2$); $T3$ — заданное значение коэффициента доверия; $R1$ — минимальное значение предельной ошибки выборки (если предельную ошибку уменьшить на заданную величину) в %; $R2$ — максимальное значение предельной ошибки выборки (если предельную ошибку выборки можно увеличить на заданную величину), %. Вероятность гарантии необходимого результата P связана с коэффициентом доверия T (табл. 2.3).

Таблица 2.3

ЗАВИСИМОСТЬ ВЕРОЯТНОСТИ P ОТ КОЭФФИЦИЕНТА ДОВЕРИЯ T

T	*P*	*T*	*P*	*T*	*P*	*T*	*P*	*T*	*P*
1,0	0,68269	1,5	0,86639	2,0	0,95450	2,5	0,98758	3,0	0,99730
1,1	0,72867	1,6	0,89040	2,1	0,96427	2,6	0,99068	3,1	0,99806
1,2	0,76986	1,7	0,91087	2,2	0,97219	2,7	0,99307	3,2	0,99863
1,3	0,80640	1,8	0,92814	2,3	0,97755	2,8	0,99489	3,3	0,99903
1,4	0,83849	1,9	0,94257	2,4	0,98360	2,9	0,99627		

4. Контрольный пример

На шахте планируется выборочное обследование процента выполнения норм выработки рабочими-сдельщиками. На стадии предварительного, в изучения вопроса установлено, что дисперсия по проценту выполнения норм составляет 200. Необходимо установить, какое количество рабочих следует обследовать в выборочном порядке из 1000 имеющихся при повторной и бесповторной схемах отбора, чтобы разность между средним процентом выполнения норм в выборочной и генеральной совокупностях не превысила 4 %, а результат можно было бы гарантировать с вероятностью 0,955. Определить, как изменится необходимая численность выборки, если предельную ошибку выборки уменьшить или увеличить на 2 %. Кроме этого, необходимо установить зависимость между вероятностью гарантии результата в пределах от 0,838 до 0,972 и численностью выборки при повторной и бесповторной схемах отбора.

В соответствии с принятыми обозначениями запишем нашу исходную информацию:

$N = 1000$ $\quad R2 = 6\,\%$

$D = 200$ $\quad P3 = 0{,}955$

$R = 4\,\%$ $\quad P1 = 0{,}838$

$R1 = 2\,\%$ $\quad P2 = 0{,}972.$

Определим по табл. 2.3 для $P1 = 0{,}838$; $P2 = 0{,}972$ и $P3 = 0{,}955$ соответствующие значения $T1$, $T2$, $T3$:

$$T1 = 1{,}4$$
$$T2 = 2{,}2$$
$$T3 = 2{,}0.$$

Введем исходную информацию в виде диалога в ЭВМ. После выполнения необходимых вычислений на печать выдается следующая информация (табл. 2.4).

Иванов И.И., группа_____ Работа № 2

Таблица 2.4

Коэффициент доверия	Доверительная вероятность	Необходимое число наблюдений	
		повтор. выборка	бесповтор.выборка
1	2	3	4
1,4	0,838	25	24
1,5	0,866	28	27
1,6	0,690	32	31
1,7	0,911	36	35
1,8	0,928	41	39
1,9	0,943	45	43
2,0	0,955	50	48
2,1	0,964	55	52
2,2	0,972	61	57

$T3 = 2{,}00$	$W = 50{,}00$	$G = 47{,}62$
$R1 = 2$	$W = 200$	$G = 166{,}67$
$R2 = 6$	$W = 22{,}22$	$G = 21{,}74$

W — необходимая численность выборки при повторной схеме отбора; G — необходимая численность выборки при бесповторной схеме отбора; $R1$ — минимальное значение предельной ошибки выборки; $R2$ — максимальное значение предельной ошибки выборки; $T3$ — заданное значение коэффициента доверия.

На основе полученных данных построим графики зависимости между вероятностью гарантии результата и необходимой численностью выборки при повторной и бесповторной схемах отбора (рис. 2.2).

Из графиков видно, что необходимый результат мы можем гарантировать с вероятностью 0,955 или 95,5 %, если обследуем 50 рабочих из 1000 при повторной схеме отбора и 48 рабочих из 1000 при бесповторной схеме отбора.

Если необходимо предельную ошибку выборки уменьшить на 2 %, то следует обследовать 200 рабочих из 1000 при повторной схеме отбора и 167 из 1600 при бесповторной.

Если нам нужно предельную ошибку выборки увеличить на 2 %, то необходимо обследовать 22 человека как при повторной, так и бесповторной схемах отбора.

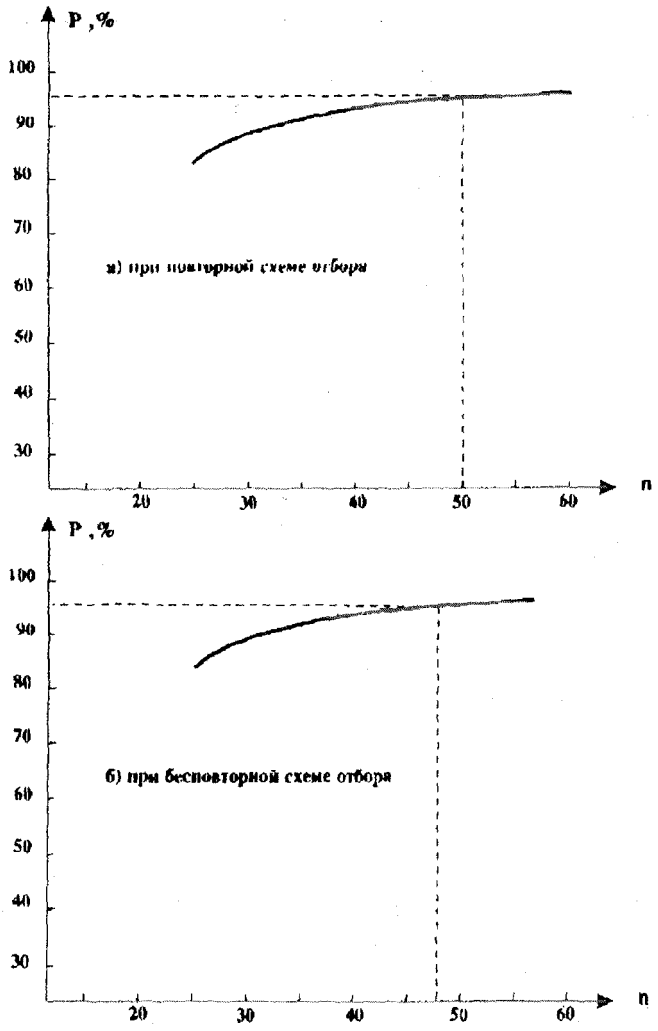

Рис. 2.2. Графики зависимости необходимой численности выборки от вероятности гарантии результата

5. Варианты заданий

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
1	1000	120	5	0,838	0,972	0,955	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
2	1100	200	5	0,838	0,972	0,911	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
3	2000	200	5	0,729	0,978	0,955	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
4	2000	350	4	0,729	0,978	0,955	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
5	2000	180	4	0,729	0,978	0,955	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
6	2200	200	5	0,683	0,955	0,866	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
7	2200	170	5	0,683	0,955	0,866	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
8	3000	350	6	0,683	0,988	0,972	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
9	3000	450	5	0,683	0,988	0,955	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
10	2500	200	4	0,683	0,988	0,972	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
11	2500	280	5	0,683	0,988	0,972	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
12	1500	220	4	0,866	0,988	0,955	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
13	1500	220	5	0,866	0,988	0,955	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
14	1500	180	4	0,866	0,987	0,911	±1,5

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
15	1800	250	4	0,866	0,983	0,942	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
16	6000	820	5	0,990	0,996	0,988	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
17	4000	650	5	0,890	0,991	0,955	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
18	4000	700	4	0,890	0,991	0,943	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
19	5000	810	5	0,770	0,972	0,911	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
20	5100	750	4	0,806	0,978	0,964	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
21	980	160	5	0,866	0,988	0,978	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
22	1500	200	5	0,866	0,988	0,972	±1

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
23	2000	400	4	0,806	0,955	0,928	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
24	2400	4400	5	0,866	0,997	0,977	±2

Номер варианта	Значения исходных данных						
	N	*D*	*R*	*P*1	*P*2	*P*3	Процент изменения *R*
25	13500	350	3	0,728	0,984	0,942	±1

Варианты 26—30. Для установления среднего заработка рабочих и служащих, занятых в топливной промышленности, планируется проведение типического пропорционального выборочного обследования.

Необходимо установить:

— зависимость между необходимой численностью выборки при повторной и бесповторной схемах отбора и вероятностью гарантии необходимого результата в заданных пределах;

— необходимую численность выборки при разных схемах отбора, чтобы разность между средней зарплатой в выборочной и генеральной совокупности не превысила заданного значения, а результат можно было бы гарантировать с заданной вероятностью;

— как изменится необходимая численность выборки, если предельную ошибку уменьшить и увеличить на заданную величину.

Исходные данные для расчетов приведены в таблицах.

Значения N_i приведены в тыс.чел., поэтому и необходимая численность выборки получится в таких же единицах.

Номер варианта	Наименование отрасли или отраслевой группы	Значения исходных данных						
		N_i	D_i	R	$P1$	$P2$	$P3$	Процент изменения R
26	Угольная	1500	1800	5	0,838	0,984	0,955	±1
	Нефтеперерабатывающая	1600	550					
	Нефтедобывающая	800	1600					
	Газовая	900	980					
	Сланцевая	80	400					
	Торфяная	120	670					

Номер варианта	Наименование отрасли или отраслевой группы	Значения исходных данных						
		N_i	D_i	R	$P1$	$P2$	$P3$	Процент изменения R
27	Угольная	1400	900	6	0,770	0,972	0,911	±2
	Нефтеперерабатывающая	500	275					
	Нефтедобывающая	1000	800					
	Газовая	910	490					
	Сланцевая	70	200					
	Торфяная	120	335					

Номер варианта	Наименование отрасли или отраслевой группы	Значения исходных данных						
		N_i	D_i	R	$P1$	$P2$	$P3$	Процент изменения R
28	Угольная	1450	1000	7	0,866	0,988	0,964	±2
	Нефтеперерабатывающая	550	300					
	Нефтедобывающая	900	700					
	Газовая	890	460					
	Сланцевая	70	210					
	Торфяная	140	330					

Номер варианта	Наименование отрасли или отраслевой группы	Значения исходных данных						
		N_i	D_i	R	$P1$	$P2$	$P3$	Процент изменения R
29	Угольная	1600	800	8	0,890	0,991	0,872	±3
	Нефтеперерабатывающая	500	300					
	Нефтедобывающая	850	875					
	Газовая	850	450					
	Сланцевая	50	250					
	Торфяная	150	325					

Номер варианта	Наименование отрасли или отраслевой группы	Значения исходных данных						
		N_t	D_t	R	$P1$	$P2$	$P3$	Процент изменения R
30	Угольная	1550	850	10	0,729	0,991	0,955	±3
	Нефтеперерабатывающая	550	350					
	Нефтедобывающая	900	750					
	Газовая	850	475					
	Сланцевая	50	275					
	Торфяная	100	300					

Работа № 3

ОДНОМЕРНАЯ СОВОКУПНОСТЬ НАБЛЮДЕНИЙ. ВАРИАЦИОННЫЙ РЯД

Цель работы: Построить интервальный вариационный ряд.

1. Основные теоретические сведения

Анализ работы горного предприятия (группы горных предприятий) начинается с анализа одного показателя. Как правило, в качестве первого показателя выбирается результирующий показатель, и для него проводится полный анализ. Эта одномерная совокупность представляется в виде вариационного ряда. Анализ проводится как вручную, так и на ЭВМ.

Конечная цель — установить вид распределения этой одномерной совокупности. Для этого высказывается гипотеза. Но гипотеза о том, что данная одномерная совокупность подчиняется выбранному закону распределения, требует статистического подтверждения и доказательства.

В качестве теоретического предположения принимаются: нормальный, логарифмически-нормальный и экспоненциальный виды распределения.

Для проверки выбранного вида распределения рассчитываются характеристики, по которым можно, с некоторой вероятностью, сделать вывод о правомерности данного вида распределения.

Будем считать, что предварительно проведены исследования статистической возможности использования данной совокупности.

Установлено, что совокупность достаточна по объему, репрезентативна, и в совокупности нет ошибок и промахов.

Все это дает основание для построения вариационного ряда.

Вариационным рядом называется ранжированная совокупность дискретных значений и соответствующая каждому значению частота. Такой ряд называется **дискретным.** Вариационный ряд может быть **дискретным и интервальным.** Вариационный ряд можно считать распределенным признаком.

Если совокупность очень велика по объему, или не имеет повторяющихся значений, или состоит из непрерывных значений, то представляется в виде интервального вариационного ряда. Интервальный вариационный ряд состоит не из конкретных значений совокупности, а из некоторых интервалов этих значений и соответствующих каждому интервалу частот. Другими словами, в интервальном вариационном ряду объединяются несколько значений совокупности как некоторый интервал. Интервалы могут быть разными или одинаковыми.

Рассчитываются значения **критерия Пирсона,** служащего основой для принятия (отрицания) гипотезы.

Ранжированный ряд (табл. 3.1) представляется как ряд исходных значений (вариант), расположенных в некотором порядке (убывания или возрастания) значений. Обычно значения располагают от меньшего к большему.

Таблица 3.1

Порядковый номер варианта	**1**	**2**	...	...	N
Значение варианта	x_1	x_2	...	...	x_N

Дискретный вариационный ряд (табл. 3.2) понимается как ранжированный ряд распределения, где каждому

значению варианта ставятся в соответствие его частота или частость. **Частота** — абсолютное число значений данного варианта в данном ряду, **частость** — относительное число значений данного варианта (отнесенное к общему числу наблюдений).

Таблица 3.2

Порядковый номер варианта	Признак X	Частота	Частость
1	x_1	m_1	m'_1
2	x_2	m_2	m'_2
⋮	⋮	⋮	⋮
i	x_i	m_i	m'_i
⋮	⋮	⋮	⋮
k	x'_k	m_k	m'_k

Необходимо проверить соотношения:

$$\sum_{i=1}^{k} m_i = N; \qquad \sum_{i=1}^{k} m'_i = 1,$$

где k — число различных значений вариант; i — текущее значение варианта (i = 1, 2, ..., k); m_i — частота i-го варианта; m'_i — частость i-го варианта, N — количество наблюдений.

Приведенные соотношения могут служить для проверки правильности построения дискретного вариационного ряда.

Результаты сводят в таблицу (см. табл. 3.2).

Для построения **интервального вариационного ряда** определяется **ширина интервала ряда распределения** (h).

Приближенное значение h вычисляется по эмпирической формуле Стерджесса:

$$h = (x_{max} - x_{min}) / (1 + 3{,}2 \lg N),$$

где x_{max} — наибольшее значение варианта в данном ряду; x_{min} — наименьшее значение варианта в данном ряду; N — общее число наблюдений в данном ряду или N — количество вариант (объем выборки).

За окончательное значение h принимается значение, близкое к расчетному, но округленное так, чтобы интервалы оказались удобными для расчетов.

Ширину интервала можно принимать одинаковой и разной для различных интервалов вариационного ряда.

В каждом интервале различают нижнюю и верхнюю границы.

Нижнюю границу (меньшее значение) первого интервала следует выбрать так, чтобы меньшее значение ряда было включено в первый интервал, и среднеинтервальное значение первого интервала было удобным для дальнейших расчетов. В конкретный интервал включаются все значения варианта, удовлетворяющие неравенству

$$(x_{min})_i \le x_j < (x_{max})_i,$$

где x_j — значение варианта ряда; j = 1, 2, ..., N; $(x_{min})_i$ — нижняя граница (меньшее значение) i-го интервала; $(x_{max})_i$ – – верхняя граница (большее значение) i-го интервала.

Значения $(x_{min})_i$ и $(x_{max})_i$, связаны соотношением

$$(x_{max})_i = (x_{min})_i + h.$$

Начальный (первый) и конечный (последний) интервалы можно сделать открытыми.

Интервальный вариационный ряд представлен таблицей (табл. 3.3).

Заполняя таблицу, следует иметь в виду принятые обозначения: n — число интервалов; i — порядковый

номер интервала ($i = 1, 2, ..., n$); $(x_{min})_i$ — нижняя граница интервала i; $(x_{max})_i$ — верхняя граница интервала i; m_i — статистическая частота интервала i.

Таблица 3.3

Порядковый номер интервала	$(x_{min} \div x_{max})_i$	$\bar{X}_i$	m_i	m'_i	M_i
1	$(x_{min} \div x_{max})_1$				
2	$(x_{min} \div x_{max})_2$				
⋮	⋮				
i	$(x_{min} \div x_{max})_1$				
⋮	⋮				
n	$(x_{min} \div x_{max})_1$				

m'_i — статистическая частость интервала i; $\bar{X}_i$ — среднеинтервальное значение, $\bar{X}_i = (x_{min} \div x_{max})_i / 2$; M_i — накопленная статистическая частота данного интервала; $M_i = M_{i-1} + m_i$.

Для первого интервала $M_1 = m_1$, для последнего интервала $M_n = N$. Последнее соотношение служит проверкой правильности построения интервального вариационного ряда.

Для расчета и построения интервального вариационного ряда на ЭВМ необходимо в диалоговом режиме задать: количество интервалов, размерность интервала (целое число или число с плавающей запятой).

ЭВМ разделит все наблюдения на указанное число интервалов, рассчитает величину интервала и построит интервальный вариационный ряд. Рекомендуется с помощью ЭВМ проварьировать количество интервалов и получить тем самым несколько вариантов решения.

2. Порядок выполнения работы

1. Дискретный вариационный ряд.
2. Интервальный вариационный ряд.
3. Дополнительные характеристики интервального вариационного ряда.
4. Расчет на ЭВМ вариантов вариационного ряда.
5. Выводы.

3. Контрольный пример

1. Для примера возьмем сменную производительность труда по добыче т/см. на 1 рабочего очистного забоя. Выпишем данные по мере поступления: 9249, 7050, 1609, 2004, 6237, 3615, 6237, 3615, 9249, 7050, 3615.

Всего наблюдений $N = 11$. Как уже указывалось, будем считать, что этого количества наблюдений достаточно для того, чтобы доверять полученным результатам. А грубые ошибки и промахи исключены из рассмотрения.

По этим данным трудно судить о средних характеристиках и вариации. Поэтому представим данные в виде вариационного ряда.

Ранжируем ряд. Расположим данные от меньшего к большему: 1609, 2004, 3615, 3615, 3615, 6237, 6237, 7050, 7050, 9249, 9249.

На практике, для небольшого ряда, как у нас, где $N = 11$, чаще всего не бывает повторяющихся значений. Каждое значение встречается только один раз, поэтому дискретный вариационный ряд не приводит к компактной записи и имеет вид (значения x_i даны произвольно) (табл. 3.4).

Таблица 3.4

x_i	3	2003	2493	2852	2908	8045	8644	5090	5567	6959	8644
m_i	1	1	1	1	1	1	1	1	1	1	1

Так как частота любого x_i, равна $m_i = 1$, то такой ряд не представляет интереса для статистического анализа и никакого обобщения и суждения не дает. Но некоторые характеристики получить можно, и об этом поговорим далее.

В нашем примере имеются повторяющиеся значения, поэтому дискретный вариационный ряд примет вид (табл. 3.5).

Таблица 3.5

x_i	1609	2004	3615	6237	7050	9249	N
m_i	1	1	3	2	2	2	11

Как видим, получилась запись совокупности, значительно короче исходной. Сокращение объема — сжатие информации — одна из целей представления совокупности в виде вариационного ряда.

2. Сформируем интервальный вариационный ряд. Для этого надо получить величину интервала. Как уже говорилось, величина интервала может быть назначена из каких-то соображений, а может быть рассчитана по эмпирической формуле:

$$h = \frac{x_N - x_1}{2},$$

где $x_N = x_{max}$ — наибольшее значение в совокупности; $x_1 = x_{min}$ — наименьшее значение в совокупности; N — количество наблюдений.

Для нашего примера

$$h = \frac{x_N - x_1}{1 + 3{,}2 \lg N} = \frac{9249 - 1609}{1 + 3{,}2 \lg N} = \frac{7640}{4{,}328} = 1765{,}25.$$

Для удобства расчетов примем $h = 2000$.

Теперь надо определить границы интервалов. Наименьшее значение совокупности 1609, поэтому имеем большой выбор для назначения нижней границы первого интервала. Это может быть $(x_{min})_1 = 0$, может быть$(x_{min})_1 = 1$, и т. д. до значения $(x_{min})_1 = 1609$.

Примем $(x_{min})_1 = 0$. Тогда первый интервал примет вид $(x_{min} \div x_{max})_1 = 0{,}000 \div 2000$, так как $(x_{max})_1$ формируется как $(x_{max})_1 = (x_{min})_1 + h$.

Аналогично, второй интервал примет вид

$$(x_{min} \div x_{max})_2 = 2000 \div 4000 \text{ и т. п.}$$

В результате получим первый столбец табл. 3.6 (вариант 2):

0000 ÷ 2000
2000 ÷ 4000
4000 ÷ 6000
6000 ÷ 8000
8000 ÷ 10000.

Но можно назначить начало первого интервала и другим, например, принять $(x_{min})_1 = 1000$. Тогда первый столбец примет вид

1000 ÷ 3000
3000 ÷ 5000
5000 ÷ 7000
7000 ÷ 9000
9000 ÷ 10000.

А можно назначить $(x_{min})_1 = 1500$, и первый столбец будет в виде

1500 ÷ 3500
3500 ÷ 5500
5500 ÷ 7500
7500 ÷ 9500.

Из всех этих вариантов пока предпочтительнее последний, так как для него совокупность получила наибольшее сокращение. Образовалось всего 4 интервала, а не 5, как в предыдущих вариантах ($n = 4$).

Тут следует учесть, что если в совокупности имеется значение, равное границе интервала, то оно включается в интервал, где x_{min} равно данному значению. Так, в нашем примере, если бы имелось значение совокупности равное 3500, то оно было бы включено во второй интервал: $(x_{min})_2 = 3500$.

Продолжая аналогично строить интервалы, можем заполнить первую и вторую колонки таблицы.

Как первый, так и последний интервалы можно представить в виде открытых интервалов. Так первый интервал принял бы вид:

< 3500, а последний: > 7500, что ровным счетом ничего не изменило бы в дальнейших расчетах.

2. Теперь можно рассчитать среднеинтервальные значения — $\bar{x}_i$.

Используя формулу $\bar{x}_i = \frac{(x_{max} + x_{min})_i}{2}$, заполним соответствующую колонку табл. 3.6 (вариант 1).

Так, $\bar{x}_1 = \frac{(x_{max} + x_{min})_1}{2} = \frac{3500 + 1500}{2} = 2500$;

$$\bar{x}_2 = \frac{(x_{max} + x_{min})_2}{2} = \frac{5500 + 3500}{2} = 4500 \text{ и т. д.}$$

Следующая колонка табл. 3.6 (вариант 1) — частоты интервалов. Для получения частоты m_1 надо обратиться к ранжированной совокупности и подсчитать количество значений, удовлетворяющих условию $(x_{min} \le x < x_{max})_i$, т. е. входящих в интервал 1500÷3500+3500. Таких значений в нашем примере два: 1609; 2004. Значит, можно записать: $m_1 = 2$. Аналогичным образом заполним всю колонку.

Проверим правильность нашей работы. Мы знаем, что $\sum_{i=1}^{n} m_i = N$, потому что мы произвели лишь группировку значений, и количество их должно сохраниться.

3. Теперь заполним следующую колонку — колонку (столбец) значений частостей (n'_i).

$$n'_1 = \frac{m_1}{N} = \frac{2}{11} = 0{,}18, \qquad n'_2 = \frac{m_2}{N} = \frac{3}{11} = 0{,}27 \text{ и т.д.}$$

Для проверки правильности расчетов необходимо проверить соотношение $\sum_{i=1}^{n} n'_i = 1$.

И осталось заполнить столбец значений накопленной частоты (M_i).

Согласно определению

$$M_1 = m_1 = 2;$$
$$M_2 = m_1 + m_2 = M_1 + m_2 = 2+3 = 5;$$
$$M_3 = m_1 + m_2 + m_3 = M_1 + m_3 = 5+4 = 9;$$
$$M_4 = M_3 + m_4 = 9+2 = 1.$$

Для проверки правильности существует соотношение $M_n = N$. В нашем примере $M_n = 1$.

В результате получаем табл. 3.6 в заполненном виде (табл. 3.6, вариант 1).

Таблица 3.6 (вариант 1)

1	$x_{min} \div x_{max}$	$\bar{x}_i$	m_1	n'_i	M_i
1	1500 ÷ 3500	2500	2	0,18	2
2	3500 ÷ 5500	4500	3	0,27	5
3	5500 ÷ 7500	6500	4	0,36	9
4	7500 ÷ 9500	8500	2	0,18	11

Очевидно, что назначая другую величину интервала, мы получим и другой интервальный вариационный ряд.

Другими словами, одну и ту же совокупность можно представить в виде различных интервальных вариационных рядов. Иногда, в интересах статистического анализа, бывает необходимым назначить количество интервалов, а, исходя из этого, определить величину интервала. Тогда величина интервала определится как

$$h = \frac{x_{\max} - x_{\min}}{n},$$

где n — количество интервалов.

Практика показывает, что рационально получить несколько интервальных рядов для одной совокупности. Тогда, конечно, полезно использовать соответствующие программы и ЭВМ.

Дополнения

Заполним (для будущего анализа) табл. 3.6, назначая разные значения $x_{\min}$. Это даст возможность научиться выбирать наилучший вариант.

Вариант 2. Назначим $x_{\min} = 0000$.

Таблица 3.6 (вариант 2)

1	$x_{\min} \div x_{\max}$	$\bar{x}_i$	m_i	n'_i	M_i
1	0000 ÷ 2000	1000	1	0,09	1
2	2000 ÷ 4000	3000	4	0,36	5
3	4000 ÷ 6000	5000	1	0,09	6
4	6000 ÷ 8000	7000	3	0,27	9
5	8000 ÷ 10000	9000	2	0,19	11

Вариант 3. Назначим $x_{min} = 1000$.

Таблица 3.6 (вариант 3)

1	$x_{min} \div x_{max}$	$\bar{x}_i$	m_1	n'_i	M_i
1	1000 ÷ 3000	2000	2	0,18	2
2	3000 ÷ 5000	4000	3	0,28	5
3	5000 ÷ 7000	6000	2	0,18	7
4	7000 ÷ 9000	8000	2	0,18	9
5	9000 ÷ 11000	10000	2	0,18	11

Работа № 4

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ СОВОКУПНОСТИ

Цель работы: вычислить статистические характеристики совокупности.

1. Основные теоретические сведения

Средние характеристики совокупности

Средняя — обобщающая количественная характеристика совокупности. Это позволяет одной величиной охарактеризовать признак.

В статистике различают много видов средних. Выбор той или другой средней определяется видом признака и целями исследования.

В данной работе будут рассмотрены средние оценки: средняя арифметическая, медиана и мода. Среднюю арифметическую называют параметрической средней. Средняя арифметическая может быть дискретной (или простой) и взвешенной.

Дискретная (или простая) средняя рассчитывается по формуле:

$$\overline{X}^{a}_{\text{дискр}} = \sum x_i / N,$$

где $\overline{X}^{a}_{\text{дискр}}$ — обозначение средней арифметической, дискретной, x_i — отдельные значения совокупности, ($i = 1, 2, ..., N$), N — количество значений в совокупности.

Если наблюдений в совокупности достаточно много, то некоторые значения повторяются. Тогда, представляя

совокупность в виде *дискретного вариационного ряда* можно вычислить среднюю арифметическую с помощью частот.

Частота выступает в виде веса данного значения совокупности, а общая формула примет вид

$$\overline{X}^{a}_{\text{взв}} = \sum_{i=1}^{k} x_i m_i \Big/ \sum_{i=1}^{k} m_i ,$$

где $\overline{X}^{a}_{\text{взв}}$ — средняя арифметическая взвешенная; $\overline{x}_i$ — отдельное значение совокупности; m_i — вес (частота) этого значения; i — 1, 2, ..., k — текущие значения; k — количество различных значений совокупности.

Для больших по объему совокупностей весь статистический анализ разумно вести, представив совокупность в виде интервального вариационного ряда.

Тогда средняя интервального вариационного ряда вычисляется по формуле

$$\overline{X}^{a}_{\text{инт}} = \sum_{i=1}^{n} x_i m_i \Big/ \sum_{i=1}^{n} m_i ,$$

где $\overline{X}^{a}_{\text{инт}}$ — средняя арифметическая интервального ряда; $\overline{x}_i$ — среднеинтервальное значение (середина интервала i); m_i — частота интервала (количество значений совокупности в i-ом интервале); i = 1, 2, ..., n — текущее значение интервала; n — количество интервалов.

Вспомним, что среднеинтервальное значение вычисляется как

$$\overline{x}_i = \frac{x_{\max} + x_{\min}}{2} ,$$

где $x_{\max}$ и $x_{\min}$ — максимальное и минимальное значения i-го интервала.

Очевидно, что соблюдаются соотношения

$$\overline{X}^{a}_{\text{дискр}} = \overline{X}^{a}_{\text{взв}} \neq \overline{X}^{a}_{\text{инт}}$$

Среди непараметрических средних значений рассмотрим медиану и моду. Медиана, так же как и среднеарифметическая, может быть дискретной и интервальной.

Медиана — срединное значение ранжированной совокупности. Поэтому, если рассмотреть ранжированную совокупность, то могут быть два пути определения медианы.

Если количество наблюдений нечетно, т. е. количество наблюдений может представляться в виде

$$N = 2k+1,$$

то $M_{\text{дискр}} = x_{k+1}$, а если $N = 2k$, т.е. наблюдений имеется четное количество, тогда

$$M_{\text{дискр}} = (x_k + x_{k+1}) / 2.$$

Для случая интервального вариационного ряда надо сначала определить медианный интервал, т. е. определить интервал, в который входит медиана.

Определяется **медианный интервал** по накопленной частоте M. Первый интервал, для которого выполняется соотношение $M \geq N / 2$ и является медианным.

Для него

$$Me_{\text{инт}} = (x_{\min})_k + h\frac{0{,}5\sum m_i - M_{k-1}}{m_k},$$

где $Me_{\text{инт}}$ — медиана интервального ряда; h — величина интервала интервального ряда; $(x_{\min})_k$ — нижняя граница медианного интервала; M_{k-1} — накопленная частота интервала, предшествующего медианному, m_k — частота медианного интервала, k — номер медианного интервала.

Из определений $Me_{\text{дискр}}$ и $Me_{\text{инт}}$ ясно, что их значения близки, но не совпадают.

Кроме того, очевидно, что значения $\overline{X}^a$ и Me достаточно близки по своим значениям, так как определяют среднюю и срединную части совокупности. Если совокупность достаточно однородна, то эти значения близки друг другу.

Другой непараметрической средней характеристикой является мода — Mo.

Mo — наиболее часто встречающееся значение совокупности.

Или, иначе, мода — значение совокупности с наибольшей частотой, $Mo = x_i$, при $m_i = \max\{m\}$. Различают совокупности одно, двухмодальные, трехмодальные и т. д.

Одномодальная совокупность имеет наблюдение с наибольшей частотой и характеризуется одним значением моды.

В двухмодальной совокупности есть два наблюдения с равными наибольшими частотами, т. е. совокупность характеризуется двумя значениями моды. В этом случае для дальнейшего исследования выбирают моду, близкую к среднеарифметическому значению.

Различно рассчитывают значение моды для дискретного и интервального ряда. И более того — для дискретного ряда не всегда можно определить значение моды, так как может существовать несколько наблюдений с равными и максимальными частотами. Поэтому часто необходимо определять моду, приведя совокупность к интервальному виду. Но тогда, сначала, как и в случае расчета медианы, определить модальный интервал. Модальный интервал имеет наибольшую частоту. А внутри этого интервала мода определяется как

$$Mo_{\text{инт}} = (x_{\min})_k + h\frac{m_k - m_{k-1}}{(m_k - m_{k-1}) + (m_k - m_{k+1})},$$

где $Mo_{инт}$ — мода интервального ряда; $(x_{min})_k$ — нижняя граница модального интервала; h — величина интервала; m_k — частота модального интервала; m_{k-1} — частота интервала, предшествующего модальному; m_{k+1} — частота интервала, следующего за модальным; k — номер модального интервала.

Для однородной совокупности характерна близость значений $\overline{X}^a$, *Me* и *Mo*.

Показатели колеблемости (вариации)

Средние величины характеризуют вариационный ряд одним числом. Но тогда эти характеристики не отражают изменчивость признака, не учитывают вариацию признака в данной совокупности.

В статистике принято несколько способов измерения вариации.

Самая простая оценка — вариационный размах.

***R* — вариационный размах** — определяется как разность между экстремальными значениями ранжированной совокупности

$$R = x_{max} - x_{min},$$

где x_{max} — наибольшее значение, x_{min} — наименьшее значение совокупности.

Размах во многом зависит от случайных обстоятельств, различен для разных выборок одного признака, а потому может быть применен как приблизительная, неустойчивая оценка вариации.

Более значимой оценкой является **простое среднее отклонение.**

Простое среднее отклонение является средним арифметическим отклонением (по абсолютной величине) отдельных значений (вариант) от общего среднеарифметического:

$$\Delta = \frac{\sum_{i=1}^{N} \left| x_i - \overline{X}_{\text{дискр}} \right|}{N},$$

где Δ — простое среднее отклонение; x_i — отдельное значение совокупности; $\overline{X}_{\text{дискр}}$, $\overline{X}_{\text{взв}}$ — среднеарифметические значения совокупности; N — количество наблюдений в совокупности.

Простое среднее отклонение может быть вычислено как дискретное (как показано выше) и как взвешенное

$$\Delta_{\text{взв}} = \frac{\sum \left| x_i - \overline{X}_{\text{взв}} \right| m_i}{\sum m_i},$$

Наиболее полной оценкой вариации признака является **средний квадрат отклонения** — дисперсия σ^2. **Дисперсия** — рассчитывается как средний квадрат отклонений отдельных значений от среднего арифметического.

Как и простое среднее отклонение, дисперсия может быть рассчитана как дискретная и как взвешенная

$$\sigma^2_{\text{дискр}} = \frac{\sum \left(x_i - \overline{X}_{\text{дискр}} \right)^2}{N}$$ — для дискретных значений;

$$\sigma^2_{\text{взв}} = \frac{\sum \left(x_i - \overline{X}_{\text{взв}} \right)^2 m_i}{\sum m_i}$$ — для взвешенных значений.

Эта оценка наиболее часто используется на практике как мера колеблемости признака. **Среднеквадратическое отклонение (или стандарт)** представляет собой квадратный корень из дисперсии. Так же как и предыдущие оценки, стандарт может рассчитываться как дискретный и взвешенный:

$$\sigma_{\text{дискр}} = \sqrt{\frac{\sum\left(x_i - \bar{X}_{\text{дискр}}\right)^2}{N}}$$ — для дискретных значений;

$$\sigma_{\text{взв}} = \sqrt{\frac{\sum\left(\bar{x}_i - \bar{X}_{\text{взв}}\right)^2 m_i}{\sum m_i}}$$ — для взвешенных значений.

Как правило, в статистическом анализе выполняются характеристики по интервальному вариационному ряду. Это вполне относится и к вычислению дисперсии и стандарта.

$$\sigma^2_{\text{инт}} = \frac{\sum\left(x_i - \bar{X}_{\text{инт}}\right)^2 m_i}{\sum m_i};$$

$$\sigma_{\text{инт}} = \sqrt{\frac{\sum\left(\bar{x}_i - \bar{X}_{\text{инт}}\right)^2 m_i}{\sum m_i}},$$

где $\bar{x}_i$ — среднеинтервальное значение интервала i; m_i — частота интервала i; $\sigma^2_{\text{инт}}$ — дисперсия интервального вариационного ряда; $\sigma_{\text{инт}}$ — стандарт интервального вариационного ряда.

Иногда статистический анализ использует и другие формулы расчета, но они за пределами нашего рассмотрения.

Как покажут дальнейшие исследования, стандартное отклонение необходимо учитывать при любом статистическом исследовании и анализе. Все эти оценки являются абсолютными величинами, их выражают в тех же единицах измерения, что и значение признака, и они характеризуют колеблемость признака. Но очень часто

используются и относительные показатели, и коэффициенты вариации.

Обычно используют один из показателей:

коэффициент осцилляции $R = \frac{R}{\overline{X}} \cdot 100\ \%$;

коэффициент по среднему отклонению $\Delta = \frac{\Delta}{\overline{X}} \cdot 100\ \%$;

коэффициент по стандарту $\sigma = \frac{\sigma}{\overline{X}} \cdot 100\ \%$.

Эти коэффициенты имеют смысл только при положительных значениях признака. Коэффициент вариации, величина которого превышает 30 %, свидетельствует о большой колеблемости значений признака в данной совокупности.

Дополнение

Стандартное отклонение часто используется при построении интервального вариационного ряда.

Учитывая, что чаще всего вариационный ряд укладывается в границы $\overline{X} \pm 3\sigma$, можно выбрать интервалы вариационного ряда равными σ или 2σ/3, или σ/2 и, соответственно, получить 6, 9 или 12 интервалов.

Если принять $h = \sigma$, то 6 интервалов примут вид:

$$\left(\overline{X} - 3\sigma\right) \div \left(\overline{X} - 2\sigma\right)$$

$$\left(\overline{X} - 2\sigma\right) \div \left(\overline{X} - \sigma\right)$$

$$\left(\overline{X} - \sigma\right) \div \left(\overline{X}\right)$$

$$\left(\overline{X}\right) \div \left(\overline{X} + \sigma\right)$$

$$\left(\overline{X} + \sigma\right) \div \left(\overline{X} + 2\sigma\right)$$

$$\left(\overline{X} + 2\sigma\right) \div \left(\overline{X} + 3\sigma\right).$$

Аналогичным образом можно построить 9 или 12 интервалов, если принять $h = 2\sigma/3$ или $h = \sigma/2$.

При этом практически все значения ($\approx$ 98 %) совокупности будут включены в интервальный вариационный ряд.

2. Порядок выполнения работы

Статистические характеристики вычисляются с помощью калькулятора.

Как принято — все расчеты в примерах даются очень подробно, хотя даже калькуляторы позволяют сделать это в одно действие.

Для расчета вариации признака используем исходную совокупность, а также табл. 3.6 (варианты 1, 2, 3).

По исходной или ранжированной совокупности можно рассчитать виды вариации дискретного вида.

3. Контрольный пример

Для примера используем ту же совокупность, что и в лабораторной работе № 1 — сменную производительность труда о добыче (т/см.) на одного рабочего очистного забоя.

Рассмотрим исходную совокупность (см. контрольный пример на стр. 42): 9249, 7050, 1609, 2004, 6237, 3615, 6237, 3615, 9249, 7050, 3615. Расположенная по рангу совокупность имеет вид: 1609, 2004, 3615, 3615, 3615, 6237, 6237, 7050, 7050, 9249, 9249. Этот вид совокупности позволяет рассчитать некоторые дискретные (или простые) средние характеристики: $\bar{X}$ — среднее, *Me* — медиану и *Mo* — моду.

1. Рассчитаем среднеарифметические значения совокупности.

Для расчета $\overline{X}_{\text{дискр}} = \sum_{i=1}^{N} x_i / N$ можно воспользоваться либо исходной совокупностью, либо ранжированной, так как нам надо просто сложить все значения совокупности и разделить полученную сумму на количество наблюдений

$$\overline{X}^{a}_{\text{дискр}} = \frac{\sum x_i}{N} =$$

$$= \frac{9249 + 7050 + 1609 + 2004 + 6237 + 3615}{11} +$$

$$+ \frac{6237 + 3615 + 9249 + 7050 + 3615}{11} =$$

$$= 5411{,}82 \cong 5412.$$

При больших объемах наблюдений такой расчет не рационален, так как существует большая вероятность ошибиться, напутать, пропустить значения и т. д. Поэтому разумнее воспользоваться совокупностью, представленной в более компактном виде, а именно — в виде дискретного вариационного ряда (табл. 4.1), где выписаны в порядке возрастания или убывания лишь различающиеся значения наблюдений и соответствующие им частоты.

Таблица 4.1

x_i	1609	2004	3615	6237	7050	9249	N
m_i	1	1	3	2	2	2	11

В этом случае можно рассчитать среднюю вариационного ряда, используя вес наблюдений (частоту). Такая средняя арифметическая носит название взвешенной

$$\overline{X}^{a}_{\text{взв}} = \frac{\sum x_i m_i}{\sum m_i} = \frac{1609 \cdot 1 + 2004 \cdot 1 + 3615 \cdot 3 + 6237 \cdot 2 + 7050 \cdot 2 + 9249 \cdot 2}{1+1+3+2+2+2} = 5412$$

Как видим, для вариационного ряда может упроститься расчет средней, но все-таки остается достаточно громоздким. Более компактная и наглядная запись совокупности представляется в виде интервального вариационного ряда (табл. 4.2).

Таблица 4.2

1	$x_{\min} \div x_{\max}$	$\overline{x}_i$	m_i	n_i'	M_i
1	1500 ÷ 3500	2500	2	0,18	2
2	3500 ÷ 5500	4500	3	0,27	5
3	5500 ÷ 7500	6500	4	0,36	9
4	7500 ÷ 9500	8500	2	0,18	11

Соответственно этой записи ведется и расчет средней характеристики — среднеарифметической средней интервального вариационного ряда.

$$\overline{X}^{a}_{\text{инт}} = \frac{\sum x_i m_i}{\sum m_i} = \frac{2500 \cdot 2 + 4500 \cdot 3 + 6500 \cdot 4 + 8500 \cdot 2}{2+3+4+2} =$$

$$= \frac{5000 + 13500 + 26000 + 17000}{11} = \frac{61500}{11} = 5591.$$

Очевидно, что среднее арифметическое, рассчитанное с использованием среднеинтервальных значений, практически никогда не совпадает со среднеарифметическим дискретным и взвешенным, так как использует не истинные значения $\overline{x}_i$, а уже усредненные — $\overline{x}_i$ — среднеинтервальные. В то же время $\overline{X}_{\text{дискр}} \equiv \overline{X}^{a}_{\text{взв}}$ и это соотношение может использоваться для проверки правильности расчетов.

2. Определим значение *Me*.

Me — срединное значение ранжированной совокупности — значение, которое как бы делит совокупность пополам. Еще одно определение — находится посередине совокупности. Определим $Me_{\text{дискр}}$ — дискретное значение медианы.

Всего значений совокупности $N = 11$, т. е. — нечетное количество. Поэтому $Me_{\text{дискр}} = x_{k+1}$, где k — определяется из соотношения $N = 2k + 1$. Значит $k = 5$, а

$$Me_{\text{дискр}} = x_{k+1} = x_6 = 6237.$$

Теперь получим $Me_{\text{инт}}$ — медиану интервального вариационного ряда.

Рассмотрим совокупность, представленную в виде интервального вариационного ряда. Воспользуемся табл. 3.6.

Определим медианный интервал. Всего наблюдений — $N = 11$, тогда $N/2 = 5{,}5$. Значит, срединное, шестое значение совокупности входит в интервал $i = k = 3$, так как для него выполняется соотношение $M_k \geq N/2$, а именно, $9 \geq 5{,}5$. Подставим значения этого интервала в формулу

$$Me_{\text{инт}} = (x_{\min})_k + h\frac{0{,}5\sum m_i - M_{k-1}}{m_k},$$

где $(x_{\min})_k = (x_{\min})_3 = 5500$; $h = 2000$; $M_{k-1} = M_{3-1} = M_2 = 5$; $m_k = m_3 = 4$.

Поэтому $Me_{\text{инт}} = 5500 + 2000\dfrac{0{,}5 \cdot 11 - 5}{4} = 5750$.

Как видим, значение $Me_{\text{дискр}}$ и $Me_{\text{инт}}$ не совпадают.

$Me_{\text{дискр}} = 6237$, а $Me_{\text{инт}} = 5750$.

3. Для определения моды — *Mo* — используем те же данные.

Дискретный вариационный ряд имеет наибольшую частоту.

$m_3 = 3$, а значение $Mo = x_3 = 3615$.

Для вычисления *Mo* интервального вариационного ряда выберем модальный интервал. Модальный интервал определяется по наибольшей частоте. Значит — это третий интервал, так как $m_k = m_3 = 4$.

Используем формулу

$$Mo_{\text{инт}} = (x_{\min})_k + \frac{m_k - m_{k-1}}{h(m_k - m_{k-1}) + (m_k - m_{k+1})},$$

где k — номер модального интервала, $k = 3$; $(x_{\min})_k$ — нижняя граница модального интервала, $(x_{\min})_k = 5500$; h — величина интервала, $h = 2000$; m_k — частота модального интервала, $m_3 = 4$; m_{k-1} — частота интервала, предшествующего модальному, $m_{3-1} = 3$; m_{k+1} — частота интервала, следующего за модальным, $m_{3+1} = 2$.

Тогда

$$Mo_{\text{инт}} = 5500 + 2000\frac{4-3}{(4-3)+(4-2)} = 6166{,}7 \cong 6167.$$

Как видим, значения моды и медианы различны.

Сведем вместе результаты наших расчетов. Получим табл. 4.3, которая и будет служить для дальнейшего анализа совокупности.

Размах $R = x_{\max} - x_{\min}$ — разница максимального и минимального значений совокупности $x_{\max} = 9249$, а $x_{\min} = 1609$.

Таким образом $R = 9249 - 1609 = 7640$.

Таблица 4.3

СРЕДНИЕ ХАРАКТЕРИСТИКИ СОВОКУПНОСТИ

Оценки	$\bar{X}^a$	*Me*	*Mo*
Дискретные	5412	6237	3615
Взвешенные	5412	6237	3615
Интервальные	5591	5750	6167

Размах во много раз превышает минимальное значение совокупности, поэтому можно ожидать, что и другие характеристики вариации будут большими. А это означает, что в совокупности значителен разброс данных, и трудно ожидать хороших результатов статистического анализа.

Простое среднее отклонение

$$\Delta_{\text{дискр}} = \frac{\sum |x_i - \bar{X}|}{N} =$$

$$= \frac{|1609-5412| + |2004-5412| + |3615-5412| + |3615-5412|}{11} +$$

$$+ \frac{|6237-5412| + |6237-5412| + |7050-5412| + |7050-5412|}{11} +$$

$$+ \frac{|9249-5412| + |9249-5412|}{11} =$$

$$= \frac{3803+3408+1797+1797+825+825+1638+1638+3837+3837}{11} =$$

$$= \frac{25202}{11} = 2473.$$

Простое среднее отклонение можно рассчитать и как взвешенное, что будет служить и проверкой правильности расчета, так как $\Delta_{\text{дискр}} = \Delta_{\text{взв}}$

$$\Delta_{взв} = \frac{|1609-5412|+|2004-5412|+|3615-5412|\cdot 3+|6237-5412|\cdot 2}{1+1+3+2+2+2} +$$

$$+\frac{|7050-5412|\cdot 2+|9249-5412|\cdot 2}{1+1+3+2+2+2} =$$

$$=\frac{3803+3408+1797\cdot 2+825\cdot 2+1638\cdot 2+3837\cdot 2}{11} = \frac{25202}{11} = 2473.$$

Если взять за исходную информацию совокупность в виде интервального вариационного ряда (см. табл. 4.3), то расчет простого среднего отклонения выполняется значительно проще, так как используются значения не истинные, а среднеинтервальные, и их частоты. И, соответственно, применяется $\overline{X} = \overline{X}_{инт}$

$$\Delta_{инт} = \frac{\sum |x_i - \overline{X}_{инт}| m_i}{\sum m_i} =$$

$$=\frac{|2500-5591|\cdot 2+|4500-5591|\cdot 3+|6500-5591|\cdot 4+|8500-5591|\cdot 2}{2+3+4+2} =$$

$$=\frac{3091\cdot 2+1091\cdot 3+909\cdot 4+2909\cdot 2}{11} = \frac{18909}{11} = 1719.$$

$$\Delta_{инт} = 1719.$$

Как видим, интервальные значения отличаются от дискретных. И это естественно, так как интервальные оценки являются более общими и усредненными. Простое среднее отклонение — одна из необходимых характеристик при проведении статистического анализа, и мы будем его использовать в дальнейшем.

Основные характеристики вариации — **дисперсия и стандарты.**

Эти характеристики также могут быть вычислены как дискретные, взвешенные и интервальные. Для вычисления дисперсии по дискретным значениям используется и среднеарифметическое дискретное $\left(\overline{X}_{\text{дискр}} = 5412\right)$.

$$\sigma^2_{\text{дискр}} = \frac{\sum\left(x_i - \overline{X}_{\text{дискр}}\right)}{N} =$$

$$= \frac{(1609 - 5412)^2 + (2004 - 5412)^2 + (3615 - 5412)^2}{11} +$$

$$+ \frac{(3615 - 5412)^2 + (3615 - 5412)^2 + (6237 - 5412)^2}{11} +$$

$$+ \frac{(6237 - 5412)^2 + (7050 - 5412)^2 + (7050 + 5412)^2}{11} +$$

$$+ \frac{(9249 - 5412)^2 + (9249 - 5412)^2}{11} =$$

$$= \frac{3803^2 + 3408^2 + 1797^2 + 1797^2 + 1797^2 + 825^2 + 825^2}{11} +$$

$$+ \frac{1638^2 + 1638^2 + 3837^2 + 3837^2}{11} =$$

$$= 10^6\left(\frac{14{,}46 + 11{,}61 + 3{,}23 + 3{,}23 + 3{,}23}{11} +\right.$$

$$\left. + \frac{0{,}68 + 0{,}68 + 2{,}68 + 2{,}68 + 14{,}72 + 14{,}72}{11}\right) =$$

$$= 10^6\left(\frac{14{,}46 + 11{,}61 + 9{,}69 + 1{,}36 + 5{,}37 + 29{,}44}{11}\right) =$$

$$10^6\left(\frac{71{,}93}{11}\right) = 10^6 \cdot 6{,}54,$$

$$\sigma^2_{\text{дискр}} = \sigma^2_{\text{взв}} = 6{,}54 \cdot 10^6.$$

Для вычисления $\sigma^2_{\text{инт}}$ — дисперсии для интервального вариационного ряда используются среднеинтервальные значения $\bar{x}_i$, среднеарифметическое значение $\bar{X}_{\text{инт}}$, вычисленное по интервальному ряду и соответствующие частоты — m_i

$$\sigma^2_{\text{инт}} = \frac{\sum\left(\bar{x}_i - \bar{X}_{\text{инт}}\right)^2 m_i}{\sum m_i} =$$

$$= \frac{(2500-5591)\cdot 2 + (4500+5591)^2 \cdot 3}{2+3+4+2} +$$

$$+ \frac{(6500+5591)^2 \cdot 4 + (8500+5591)^2 \cdot 2}{2+3+4+2} =$$

$$= \frac{(3091)^2 \cdot 2 + (1091)^2 \cdot 3 + (909)^2 \cdot 4 + (2909)^2 \cdot 2}{11} =$$

$$= 10^6\left(\frac{9{,}55\cdot 2 + 1{,}21\cdot 3 + 0{,}81\cdot 4 + 8{,}41\cdot 2}{11}\right) =$$

$$= 10^6\left(\frac{19{,}11 + 3{,}57 + 3{,}30 + 16{,}92}{11}\right) = 10^6 \cdot \frac{42{,}91}{11} = 3{,}9\cdot 10^6,$$

$$\sigma^2_{\text{инт}} = 3{,}9\cdot 10^6.$$

Как видим, дисперсии, вычисленные с разным усреднением, довольно резко отличаются друг от друга.

Значения дисперсии позволяют вычислить значение стандартного отклонения. Стандартное отклонение необходимо для дальнейшего статистического анализа.

$$\sigma_{\text{дискр}} = \sqrt{\sigma^2_{\text{дискр}}} = \sqrt{\frac{\sum\left(x_i - \bar{X}_{\text{дискр}}\right)^2}{N}} = \sqrt{6{,}54\cdot 10^6};$$

$$\sigma_{\text{дискр}} = 2{,}56 \cdot 10^3 = 2560;$$

$$\sigma_{\text{дискр}} = \sqrt{\sigma^2_{\text{инт}}} = \sqrt{\frac{\sum \left(x_i - \overline{X}_{\text{инт}}\right)^2 m_i}{\sum m_i}} = \sqrt{3{,}9 \cdot 10^6};$$

$$\sigma_{\text{инт}} = 1{,}97 \cdot 10^3 = 1970.$$

Для наглядности сведем все полученные результаты в табл. 4.4. Эта таблица будет нужна при выполнении следующих этапов анализа совокупности.

Таблица 4.4

ОЦЕНКИ ВАРИАЦИИ СОВОКУПНОСТИ

Оценки	R	Δ	σ^2	σ
Дискретные	7640	2473	$6{,}54 \cdot 10^6$	2560
Взвешенные	–	2473	$6{,}54 \cdot 10^6$	2560
Интервальные	–	1719	$6{,}54 \cdot 10^6$	1970

$A = 0;$ $E = 0;$ $\Delta = 1{,}25\,\sigma.$

Для проверки правильности расчетов следует графически проверить соблюдение (в пределах выбранной точности) следующих соотношений:

$Me_{\text{граф}} \cong Me_{\text{дискр}};$ $Mo_{\text{граф}} \cong Mo_{\text{дискр}};$

$\overline{X}_{\text{граф}} \cong X_{\text{дискр}};$ $\overline{X} - 3\sigma \cong x_{\min};$

$\overline{X} + 3\sigma \cong x_{\max}.$

Работа № 5

ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СОВОКУПНОСТИ

Цель работы: построить графики и оценить графически статистическую **совокупность.**

1. Основные теоретические сведения

Одномерная совокупность, представленная в виде вариационного ряда, может быть изображена в виде полигона, гистограммы, кумуляты, кривой Лоренца, огивы и т. п.

Полигональная ломаная, или **полигон,** или многоугольник распределения — строится в прямоугольной системе координат.

Полигон может быть построен для дискретного вариационного ряда и для интервального. Полигон для дискретного вариационного ряда строится следующим образом:

По оси абсцисс откладывают значения вариант, а по оси ординат — значения частот (или частостей).

Полученные на пересечении этих значений точки соединяют отрезками прямой. Такой график, как очевидно, можно построить только для случая часто повторяющихся вариантов в совокупности.

На практике же чаще всего случается, что наблюдений не слишком много, а потому и повторяющихся значений либо мало, либо вообще повторений нет. Все имеющиеся значения вариант совокупности встречаются только единожды. Частоты равны единице. График теряет смысл.

Поэтому для небольших по объему совокупностей рациональнее строить полигональную ломаную по интервальному вариационному ряду.

Но, как уже говорилось раньше, и для больших по объему совокупностей проще построить полигон по интервальному вариационному ряду.

Для этого по оси абсцисс откладываются значения середин интервала, — среднеинтервальные значения совокупности, а по оси ординат, как всегда, — частоты (или частости).

Гистограмма распределения строится только для совокупности, представленной в виде интервального вариационного ряда.

Гистограмма также строится в прямоугольной системе координат.

В отличие от полигона, для гистограммы на оси абсцисс откладываются отрезки, соответствующие интервалам значений. На каждом отрезке, как на основании строится прямоугольник, высотой которого служит значение частоты, соответствующей данному интервалу.

Получим как бы ступенчатую гистограмму. При таком построении допускается, что распределение вариант внутри интервала равномерно.

Можно представить себе, что при последовательном делении интервалов, ступенчатая гистограмма превратится в плавную кривую. Такая кривая носит название кривой распределения.

Кумулята или кумулятивная ломаная, выполняется в прямоугольной системе координат. По оси абсцисс откладываются значения признака (варианты), а по оси ординат — соответствующие накопленные частоты. Полученные точки пересечений соединяются отрезками прямой.

Кумулятивную ломаную можно построить как для дискретного вариационного ряда, так и для интервального.

Для интервального вариационного ряда по оси абсцисс откладываются среднеинтервальные значения. Нижней границе первого интервала соответствует частота равная нулю, а верхней границе последнего интервала — сумма всех частот или общее количество наблюдений.

При выборе соотношений между масштабами по осям абсцисс и ординат целесообразно использовать правило «золотого сечения». График располагается в прямоугольнике, размеры которого пропорциональны 5:8 или 3:4, и линии графика занимают всю площадь. Сравнение этих графиков показывает, что переход к интервальным значениям значительно сглаживает график и выявляет сущность совокупности.

2. Порядок выполнения работы

Вся совокупность, подлежащая анализу, представляется в виде интервального вариационного ряда.

Интервалы одинаковые и вычисляются по формуле

$$h \cong \frac{x_{\max} - x_{\min}}{1 + 3{,}2 \lg N},$$

где $x_{\max}$ — наибольшее значение совокупности; $x_{\min}$ — наименьшее значение совокупности; N — количество наблюдений.

Заполняется таблица вида табл. 5.1.

Таблица 5.1

i		$\overline{x}_i$	m_i	n_i
	$(x_{\max} - x_{\min})_1$			
	$(x_{\max} - x_{\min})_2$			
	$(x_{\max} - x_{\min})_n$			

Принятые обозначения: $(x_{max} - x_{min})$ — верхняя и нижняя границы i-го интервала; x_{min} — наименьшее значение совокупности входит в первый интервал; x_{max} — наибольшее значение совокупности входит в последний, n-й интервал; n — количество интервалов, i — текущий номер интервала, $i = 1, 2, ..., n$; $\bar{x}_i$ — среднее значение i-го интервала

$$\bar{x}_i = \frac{(x_{max} + x_{min})_i}{2};$$

m_i — частота i-го интервала, абсолютное количество наблюдений, входящих в i-й интервал.

Контроль правильности расчета

$$\sum_{i=1}^{n} m_i = N;$$

n_i — относительная частота или частость i-го интервала

$$n_i = \frac{m_i}{N}.$$

Контроль правильности расчетов

$$\sum n_i = 1.$$

В дальнейшем рассматривается вариационный ряд, состоящий из среднеинтервальных значений x_i, и соответствующих значений частоты m_i.

Нормальный закон распределения имеет вид:

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - \bar{X})^2 / 2\sigma^2},$$

где σ — среднеквадратическое отклонение

$$\sigma = \sqrt{\sum_{i=1}^{n}\left(\bar{x}_i - \bar{X}\right)^2 / \sum_{i=1}^{n} m_i} \ ;$$

$\bar{X}$ — среднеарифметическое значение.

$$\bar{X} = \sum \bar{x}_i m_i / N \ .$$

При этом σ и $\bar{X}$ — параметры нормального закона распределения.

Экспоненциальный нормальный закон распределения имеет вид:

$$\left(\bar{x}_i\right) = \left[e^{-x_i / \bar{X}}\right] / \bar{X} \ .$$

При этом x — один параметр экспоненциального закона распределения.

Логарифмически нормальный закон распределения описывается функцией

$$\Psi\left(\bar{x}_i\right) = \left[1 / \sigma_{\ln x} \cdot \sqrt{2\pi}\right] / e^{-\left(\ln \bar{x}_i - \bar{X}\right)^2 / 2\sigma^2_{\ln x}} \ .$$

Логарифмически нормальное распределение предполагает переход к новым переменным $z = \ln \bar{x}_i$.

Порядок расчета на ЭВМ

1. Рассчитываются

$$f\left(\bar{x}_i\right), Y\left(\bar{x}_i\right), \Psi\left(\bar{x}_i\right).$$

2. Рассчитываются

$$(P_i)_f = hf(\bar{x}_i);$$

$$(P_i)_Y = hY(\bar{x}_i);$$

$$(P_i)_\Psi = h\Psi(\bar{x}_i).$$

3. Рассчитываются $\chi^2 = \sum_{i=1}^{n} (m_i - NP_i)^2 / NP_i$ для каждой функции $f(\bar{x}_i), Y(\bar{x}_i), \Psi(\bar{x}_i)$.

3. Подготовка исходной информации

1. Вариационный ряд задается в виде:

$(x_{\min})_1$ — нижняя граница первого интервала; h — величина интервала; n — количество интервалов; m_1, m_2, ..., m_n — частоты первого, второго, ..., n-го интервала.

Интерпретация результатов расчетов на ЭВМ

В результате расчета на ЭВМ получаем 3 значения χ^2 — для нормального, экспоненциального и логарифмически-нормального закона распределения.

В качестве лучшего вида теоретического распределения выбирается вид, которому соответствует наименьшее значение $\chi^2_{ра}$.

Если необходимо более корректно выявить лучший вид распределения, то каждое значение $\chi^2_{ра}$ надо проверить по χ^2-критерию, определенному по таблицам, в зависимости от степеней свободы $f_1 = N - 1$, $f_2 = N - k - 1$, где k — количество параметров распределения.

Пример

Построим графики для совокупности, представленной в виде вариационного ряда. Воспользуемся таблицами данных: дискретный вариационный ряд — табл. 4.1, интервальный вариационный ряд — табл. 3.6 (вариант 1).

Таблица 4.1

x_i	1609	2004	3615	6237	7050	9249	N
m_i	1	1	3	2	2	2	11

Таблица 3.6 (вариант 1)

i	$x_{min} \div x_{max}$	$\bar{x}_i$	m_1	n'_i	M_i
1	1500 - 3500	2500	2	0,18	2
2	3500 - 5500	4500	3	0,27	5
3	5500 - 7500	6500	4	0,36	9
4	7500 - 9500	8500	2	0,18	11

Построим полигональную ломаную — «полигон». По оси абсцисс отложим значения вариант x_i, а по оси ординат — значения частот этих вариант m_i.

Из табл. 3.6 следует, что наименьшее $x_i = x_{min} = 1609$, а наибольшее $x_i = x_{max} = 9249$, поэтому на оси абсцисс отложим 1500 и 9500, т. е. значения, включающие min и max. И на полученном отрезке оси отметим точки, соответствующие значениям всех вариант таблицы 3.6 (вариант 1).

Как видим по таблице 4.1, наибольшее значение частоты $m_i = 3$.

Поэтому ось ординат достаточно разделить на 3 равных части (рис. 5.1). А масштаб графика выберем так, чтобы выдерживалось «золотое» соотношение: 5:8 или 3:4.

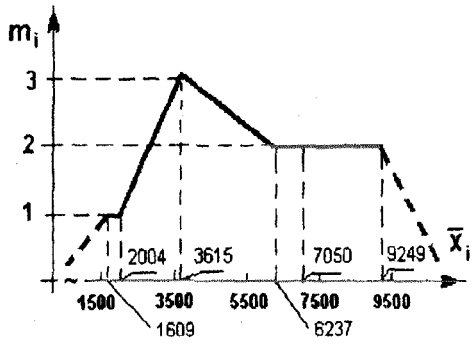

Рис. 5.1

Аналогично построим и полигон для совокупности, представленной в виде интервального вариационного ряда.

Используем данные табл. 3.6. По оси абсцисс графика откладывается значение столбца 3 из табл. 3.6 (вариант 1). Среднеинтервальные значения $\bar{x}_i$, варьируют в пределах 2500÷8500. Эти значения и должны служить границами графика (см. рис. 5.1). Наибольшая частота $m_{max} = 4$. Поэтому достаточно ось ординат разделить на 4 равных отрезка.

Нанесем точки на ось абсцисс: 2500, 4500, 6500 и 8500. На оси ординат отложим 2, 3, 4, 2.

На пересечениях этих значений отметим точки полигона и соединим точки отрезками прямой. Можно добавить в таблице 3.6 (вариант 1) две строки — в начале и в конце таблицы:

1	2	3	4	…
0	< 1500	500	0	…
5	> 9500	10500	0	…

Эти добавления дают нам возможность дополнить полигональную ломаную отрезками прямой до пересечения с осью абсцисс. Нанесем эти отрезки пунктиром.

Построим гистограмму.

Для этого используем интервальный вариационный ряд.

На оси абсцисс отложим отрезки, соответствующие интервалам вариационного ряда. На них, как на основании, построим прямоугольники (столбики), высотой, пропорциональной частоте.

Если представить, что интервалы последовательно и многократно делят на два, тогда столбики гистограммы становятся все тоньше и тоньше. И в пределе верхние отрезки столбиков превращаются в точки и получается плавная огибающая линия. Эта линия и носит название кривой распределения. Но этот процесс требует большого количества наблюдений.

В последующих частях анализа используются эти результаты.

По гистограмме (рис. 5.2) можно графически определить значение $Mo \cong 6100$. А ранее рассчитанные значения $Mo_{дискр} = 3615$ и $Mo_{инт} = 6167$.

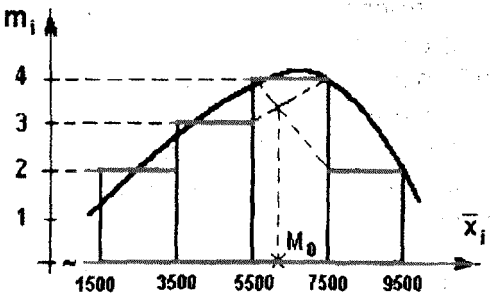

Рис. 5.2

Построим кумуляту. Кумулятивная ломаная (рис. 5.3) строится как для дискретного, так и для интервального ряда.

Не усложняя процесс анализа статистической совокупности, построим кумуляту только для интервального вариационного ряда.

Отложим по оси абсцисс значение середин интервала (среднеинтервальные), а по оси ординат — накопленные частоты и соединим точки пересечения отрезками прямой.

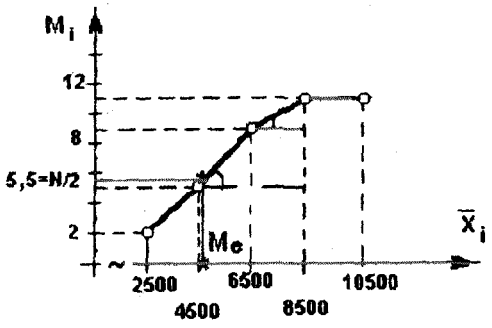

Рис. 5.3

Что можно сказать о совокупности по этому графику? Можно сказать, что основное количество наблюдений находится в средних интервалах, так как угол между осью абсцисс и отрезком кумуляты в этих интервалах больше, чем в последнем интервале. Ведь очевидно, что если бы интервал имел частоту, равную нулю, то соответствующий отрезок кумулятивной ломаной был бы параллелен оси абсцисс.

Так, если к нашей совокупности, соответственно интервальному ряду, добавить еще один интервал, то частота его будет равна 0.

На кумуляте можно графически определить значение *Me*.

Как известно, медиана — это значение признака, находящегося посредине совокупности. В нашем примере всего одиннадцать значений, $N = 11$, значит, $N/2 = 5{,}5$. Откладываем это значение на оси ординат, проводим горизонтальную линию (параллельную оси абсцисс) до пересечения с графиком кумуляты, из точки пересечения опускаем перпендикуляр на ось абсцисс. Это значение и есть значение медианы — $Me_{граф}$.

В данном примере $Me_{граф} \cong 5100$, а ранее рассчитанные значения $Me_{дискр} = 6237$, а $Me_{инт} = 5750$.

Разброс этих значений — свидетельство малого количества наблюдений и достаточно большого рассеяния значений в совокупности.

Выводы

По графику полигональной ломаной, построенной по дискретным значениям, нельзя сделать серьезного вывода, так как наблюдений немного и они имеют большой разброс, что и сказалось на графике. Несколько сглаженные интервальные значения позволили построить гистограмму достаточно симметричного вида.

Кумулята свидетельствует о том, что в построении вариационного ряда нет пустых интервалов, т. е. интервалов с частотой, равной нулю.

Таким образом, можно считать, что данная совокупность может быть включена в дальнейшее исследование.

Работа № 6

ПАРНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Цель работы: исследовать взаимосвязь двух признаков методами корреляции и регрессии.

1. Основные теоретические сведения

Пусть предварительно установлено, что изменчивость некоторого признака y зависит от изменчивости другого признака x.

Установлено, что признаки подчиняются нормальному закону распределения. Пусть вид аналитической зависимости заранее не определен.

Тогда цель анализа — установить вид зависимости признаков y и x, определить и оценить тесноту связи этих признаков и определить наилучшую форму регрессионного уравнения.

Аналитическую модель связи будем определять в виде: прямой линейной зависимости

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

обратной линейной зависимости

$$\hat{x} = b_0 + b_1 y; \tag{2}$$

параболы

$$\hat{y} = a_0 + a_1 x + a_2 x^2; \tag{3}$$

гиперболы

$$\hat{y} = a_0 + \frac{a_1}{x}. \tag{4}$$

Методы, которыми можно представить отношения в виде прямой, применимы и в случае кривых различного типа.

Типов кривых — бесконечное множество, однако в практике используются кривые нескольких основных видов, такие как параболы второго, третьего порядка, гиперболы.

Для получения количественных значений параметров этих уравнений применяется метод наименьших квадратов.

Однако вид уравнения заранее невозможно определить. Поэтому определяется несколько видов зависимостей и из них выбирается, согласно критериям, наилучший, который и принимается в качестве решения.

Для парных зависимостей исследуется несколько видов уравнений и сравниваются рассеяния результатов около линии регрессии (применяя остаточные дисперсии и критерий Фишера).

Выбранный вид уравнения регрессии служит указанием к выбору вида регрессии для множественных зависимостей. Принято определять параметры множественной регрессии, принимая порядок уравнения не ниже самого высокого из входящих парных зависимостей.

Среди самых простых видов — параболическая форма:

$$\hat{y} = a_0 + a_1 x + a_2 x^2 .$$

Другие формы зависимости:

— показательная;

— гиперболическая;

— экспоненциальная;

— обратно-экспоненциальная;

— экспоненциально-показательная;

— обратно-экспоненциально-показательная.

Исследователь, зная основные черты явления, может выбрать те или другие виды зависимости для расчета.

Остальные рассуждения: оценка тесноты связи, определение аргумента и функции, выбор вида уравнения

— все проводится в аналогии с линейной формой и потому не требует специального рассмотрения.

Весьма часто можно усмотреть известную связь между вариациями по различным признакам. В простейшем случае такая связь однозначна. Но это бывает редко. Если составить так называемую **корреляционную решетку,** то видна некоторая размазанность корреляции.

Собственно корреляционная решетка (табл. 6.1) включает частоты одного признака в конкретном интервале, распределенные по всем интервалам другого признака. Заполнение корреляционной решетки рекомендуется сделать следующим образом: выписать из табл. 6.2 все значения признака x, входящие в первый интервал, и соответствующие значения y. Затем распределить количество значений x и y по интервалам значений. Таким образом определяются условные частоты первой строки.

Другими словами, m_{11} — число значений y из первого интервала, которым соответствуют значения x из первого интервала; m_{12} — число значений y из первого интервала, которым соответствуют значения x из второго интервала и т. д.

Таблица 6.1

$\rightarrow j$ / i	$(x_H - x_b)$	$(x_H - x_b)_1$	$(x_H - x_b)_2$	...	$(x_H - x_b)$	m	$x\bar{y}_i$
$(y_H - y_b)$	$\bar{y}_i$ / $\bar{x}_i$	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$		
$(y_H - y_b)$	$\bar{y}_2$	m_{11}		...			
$(y_H - y_b)$	$\bar{y}_2$	m_{12}		...			
⋮	⋮	⋮	⋮				
$(y_H - y_b)$	$\bar{y}_k$						
	m	$\sum m_{1k}$	$\sum m_{2k}$		$\sum m_{kk}$		
	$\bar{y}_n$	$\bar{y}_{x1}$	$\bar{y}_{x2}$				

Аналогично следует поступить с другими интервалами. Разумеется, если просуммировать по строке значения условных частот, получим число значений y, входящих в соответствующий интервал по y, если просуммировать значения по столбцу, получим число значений x, входящих в соответствующий интервал по x_n, если просуммировать все значения условных частот в корреляционной решетке, получим общее число наблюдений N.

Таблица 6.2 представляет обобщенные результатов рассмотрения признаков x и y, т. е. — корреляционную решетку.

Так, первый столбец таблицы — столбец y — среднеинтервальные значения y, а верхняя строка таблицы — среднеинтервальные значения x.

На пересечениях столбцов и строк — частоты, соответствующие данным интервалам x и y.

Таблица 6.2.

y \ x	1500	1900	2300	2700	3100	3500
44						1
38			1	1	1	
32						
26		1	1			
20	4	5	10	3	1	
14	2	4		2	1	

Установление корреляции между признаками не дает оснований считать эти связи причинно-следственными. Вполне может быть, что эти признаки зависят еще от каких-то признаков.

Наглядное представление (рис. 6.1) можно получить, построив **корреляционное поле** (точечная диаграмма). Пусть, для примера имеем:

x 5,0	5,5	5,5	6,0	6,0	7,0	7,0	8,0
y 10,0	10,5	11,0	10,5	11,0	11,0	12,0	12,0

Вытянутость корреляционного поля по диагонали свидетельствует о несомненном наличии корреляции между признаками.

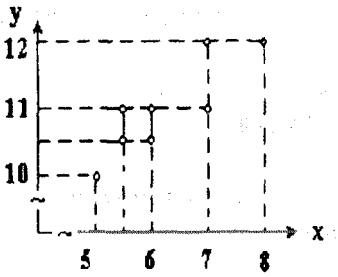

Рис. 6.1

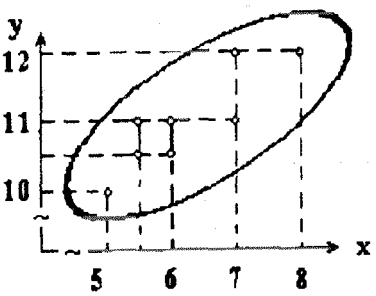

Рис. 6.2

Если число вариантов велико, то корреляционное поле имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии. Такой эллипс носит название **эллипс рассеяния** (рис. 6.2). Отклонение осей эллипса от координатных направлений указывает на наличие корреляции. Вытянутость говорит только о масштабах по осям.

Обычно на одно значение признака приходится несколько значений другого признака. Так, на одно значение суточной добычи приходится несколько значений ширины выработки. При этом рассчитывается условное или групповое среднее, так, $\bar{y}_x$ — среднее значение y при условии, что x — заданная величина.

Тогда о корреляции можно говорить только в случае, когда изменения $\bar{y}_x$ при переходе от x_i, к x носят систематический характер. Получим условные средние $\bar{y}^0_{x_0}$:

для

$x = 5{,}0$ $\quad y = 10{,}0$ $\quad \bar{y}_{x=5,0} = 10{,}0$

$x = 5{,}5$ $\quad y = \begin{cases} 10{,}5 \\ 11{,}0 \end{cases}$ $\quad \bar{y}_{x=5,5} = \dfrac{10{,}5 + 11{,}0}{2} = 10{,}75$

$x = 6{,}0$ $\quad y = \begin{cases} 10{,}5 \\ 11{,}0 \end{cases}$ $\quad \bar{y}_{x=6} = \dfrac{10{,}5 + 11{,}0}{2} = 10{,}75$

$x = 7{,}0$ $\quad y = \begin{cases} 11{,}0 \\ 12{,}0 \end{cases}$ $\quad \bar{y}_{x=7,0} = \dfrac{11{,}0 + 12{,}0}{2} = 11{,}5$

$x = 8{,}0$ $\quad y = 12{,}0$ $\quad \bar{y}_{x=8} = 12{,}0.$

А теперь рассчитаем условные средние x_y:
для

$y = 10{,}0$ $\quad x = \begin{cases} 5{,}0 \\ 5{,}5 \end{cases}$ $\quad \bar{x}_{y=10,0} = \dfrac{5{,}0 + 5{,}5}{2} = 5{,}25$

$y = 10{,}5$ $\quad x = \begin{cases} 5{,}5 \\ 6{,}0 \end{cases}$ $\quad \bar{x}_{y=10,5} = \dfrac{5{,}5 + 6{,}0}{2} = 5{,}75$

$y = 11{,}0$ $\quad x = \begin{cases} 6{,}0 \\ 7{,}0 \end{cases}$ $\quad \bar{x}_{y=11,0} = \dfrac{5{,}5 + 6{,}0 + 7{,}0}{2} = 6{,}17$

$y = 12{,}0$ $\quad x = \begin{cases} 7{,}0 \\ 8{,}0 \end{cases}$ $\quad \bar{x}_{y=12,0} = \dfrac{7{,}0 + 8{,}0}{2} = 7{,}5.$

Построим зависимости $(\bar{y}_x, x)$ и $(y, \bar{x}_y)$, обозначим их соответственно через (*) и (0) и назовем их эмпирическими линиями регрессии (рис. 6.3).

Линии не совпадают между собой, что является результатом размазанности корреляции. В общем случае линии — ломаные, но в простейшем случае — средние значения одного признака зависят от значений другого признака линейно. Такой случай называют линейной регрессией. Для простоты рассмотрим линейный случай.

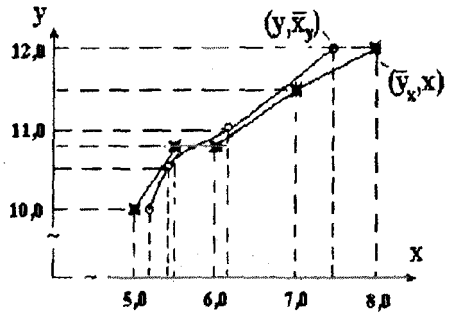

Рис 6.3

Точки, изображающие зависимость $\bar{y}_x$ от x и x_y от y, никогда не ложатся на одну прямую. Поэтому речь может идти только о том, чтобы найти такую прямую, которая проходила бы наиболее близко ко всем точкам. Смысл «близости» может быть разным:

а) наилучшая прямая та, при которой max эмпирического отклонения от расчетного — наименьший. Но тогда наиболее отклоняющаяся точка окажет большое влияние на определение прямой;

б) наилучшая прямая та, при которой площадь между ломаной эмпирических значений и прямой — наименьшая. Такую считать трудно;

в) в большинстве случаев целесообразен критерий, при котором требуется минимизировать сумму квадратов отклонений эмпирических точек от прямой (метод наименьших квадратов). При этом отклоняющаяся точка не имеет решающего значения.

Для практических целей могут представлять интерес оба уравнения регрессии. При размазанности корреляции коэффициенты уравнений не являются обратными.

$r_{xy} = r_{yx}$, если не размазаны данные.

Важной задачей теории корреляции является построение численного параметра, который давал бы количественное выражение степени или силы корреляции между признаками.

Корреляция тем более сильна, чем теснее точки корреляционного поля группируются около линии регрессии. И если корреляция полная (т. е. неучитываемых влияний нет), то имеем функциональную зависимость $r_{xy} = r_{yx} = 1$.

Если корреляция отсутствует (т. е. y в общем не зависит от x), то $r_{xy} = r_{yx} = 0$.

Корреляционную связь можно считать реальной, если полученный коэффициент корреляции значимо отличается от нуля. С этой целью используют таблицы r-распределения.

Если вычисленный коэффициент корреляции превосходит табличное значение для выбранного уровня значимости при числе степеней свободы $f = N - 2$, где N — число испытаний, то его можно считать значимо отличающимся от нуля.

Для выполнения анализа необходимо, чтобы коэффициенты корреляции были безусловно значимы.

Если $r_{yx} = 0$, то это только означает, что не может существовать линейная корреляционная связь, а криволинейная вполне может.

4. Контрольный пример

Находим параметры уравнения, определяющие зависимость y от x (1) и, наоборот, x от y (2).

Параметры уравнения (1) мы найдем, решая систему уравнений:

$$\begin{cases} Na_0 + a_1 \sum_{i=1}^{N} x = \sum_{i=1}^{N} y \\ a_o \sum_{i=1}^{N} x + a_1 \sum_{i=1}^{N} x^2 = \sum_{i=1}^{N} yx. \end{cases}$$

Пусть, решая данную систему уравнений относительно a_0 и a_1, получим:

$a_0 = 7{,}55$, $a_2 = 0{,}006$, т.е. другими словами, — получаем уравнение в виде

$$y = 7{,}55 + 0{,}06\,x. \qquad (1)$$

Параметры уравнения (2) мы найдем, решая систему уравнений:

$$\begin{cases} Nb_0 + b_1 \sum_{i=1}^{N} y = \sum_{i=1}^{N} x \\ b_0 \sum_{i=1}^{N} y + b_1 \sum_{i=1}^{N} y^2 = \sum_{i=1}^{N} yx. \end{cases}$$

Подставляя значения N, $\sum_{i=1}^{N} y$, $\sum_{i=1}^{N} x$ и т. д., решаем систему алгебраических уравнений относительно b_0 и b_1.

Получаем значения:

$$b_0 = 201{,}1;\ b_1 = 82{,}6. \qquad (2)$$

Уравнение принимает вид:

$$x = 201{,}1 + 82{,}6y.$$

Подставляя последовательно в уравнение (2) значения y, получим расчетные значения $\hat{x}$, а подставляя в уравнение (1) значения x, получим расчетные значения $\hat{y}$.

Для построения линейной зависимости $y(x)$ достаточно иметь две пары значений x_{min} и соответствующее значение $\hat{y}$; и x_{max} и соответствующее $\hat{y}$. В прямоуголь-

ных осях получим график линейной зависимости (рис. 6.4).

По рассчитанным значениям построим теоретические линии связи (рис. 6.5).

Определяем коэффициент корреляции:

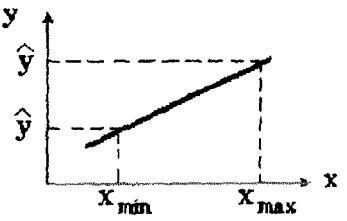

Рис. 6.4

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

где r — коэффициент парной корреляции.

$$\bar{x} = \frac{\sum_{i=1}^{N} xm}{\sum_{i=1}^{N} m} = 2226,$$

$$\bar{y} = \frac{\sum_{i=1}^{N} ym}{\sum_{i=1}^{N} m} = 20{,}9;$$

σ_x и σ_y — средние квадратичные отклонения, найденные по признаку x и признаку y:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} x^2 m}{\sum m} - \bar{x}^2 = 25281,$$

$$\sigma_y^2 = \frac{\sum_{i=1}^{N} y^2 m}{\sum_{i=1}^{N} m} - \bar{y}^2 = 82.$$

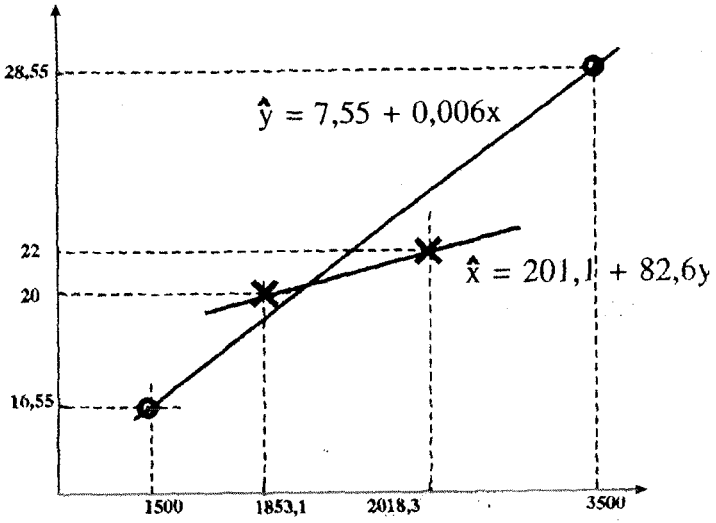

Рис. 6.5

Значение парного коэффициента корреляции — $r = 0{,}37$.

Определяем погрешность коэффициента корреляции:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} = \frac{1 - 0{,}37^2}{\sqrt{38}} = 0{,}1;$$

определяем надежность корреляции по формуле:

$$\mu = \frac{|r| \cdot \sqrt{N}}{1 - r^2}, \qquad \mu = \frac{|r| \cdot \sqrt{38}}{1 - r^2} = 2{,}5.$$

Так как $\mu<2,6$, то, согласно теореме Ляпунова, связь между признаками ненадежная, а доверительный интервал для r можно представить в виде:

$$r = 0,37 \pm \sigma_r = 0,37 \pm 0,1.$$

Для расчета параметров уравнения параболы (3) находят a_0, a_1, a_2.

Коэффициенты a_0, a_1, a_2 находятся, как решение системы уравнений

$$\begin{cases} Na_0 + a_1 \sum xm + a_2 \sum x^2 m = \sum ym \\ a_0 \sum xm + a_1 \sum x^2 m + a_2 \sum x^3 m = \sum yxm \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 = \sum yx^2. \end{cases}$$

Пусть получено $a_0 = 500$; $a_1 = -0,3$; $a_2 = 0,000007$.

В результате получим уравнение:

$$y = 500 - 0,3x + 0,000007x^2. \qquad (3)$$

Теперь последовательно подставим в это уравнение значения x, получим значения y:

x... 1500	1900	2300	2700	3100	3500
y... 20,7	18,27	18,0	20,0	24,2	30,7.

По этим значениям строим параболу (рис. 6.6).

$$\eta = \frac{\sigma_{xy}}{\sigma_y};$$

$$\sigma_{xy}^2 = \frac{1}{N}\sum m_x (y_x - \bar{y})^2;$$

$$\bar{y} = \frac{18 + 18{,}2 + 22 + 21 + 24 + 44}{6} \cong 24{,}5.$$

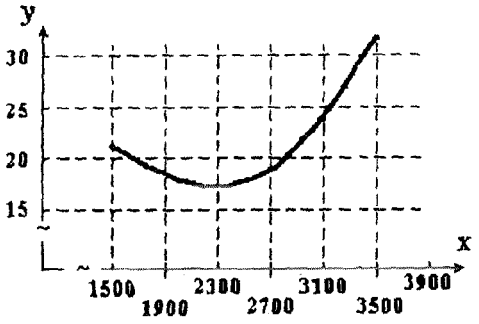

Рис. 6.6

Корреляционное отношение для параболы:

$$\eta = \frac{5{,}5}{7{,}2} = 0{,}76.$$

Найдем параметры уравнения связи, определяющие зависимость между y и x в виде гиперболы. Эти параметры мы найдем, решая систему уравнений:

$$\begin{cases} Na_0 + a_1 \sum \frac{1}{x} = \sum y \\ a_0 \sum \frac{1}{x} + a_1 \sum \frac{1}{x^2} = \sum \frac{y}{x.} \end{cases}$$

Пусть расчеты представлены в виде:

$$a_1 = 50000;$$

$$a_0 = -1{,}6.$$

Тогда уравнение гиперболы предстанет в виде:

$$y = -1{,}6 + \frac{50000}{x}.$$

Подставляя в уравнение связи значения x и получая значения y, построим гиперболу (рис. 6.7).

x...	1500	1900	2300	2700	3100	3500
y...	31,4	24,7	20,1	16,9	14,5	12,6.

По проведенным расчетам можно сделать выводы:

Корреляционная зависимость между x и y будет представлять собой параболическую зависимость: так как $\eta >> r$, то зависимость должна быть нелинейной, а исходя из логических соображений ясно, что с ростом x растет y, поэтому гипербола не может быть принята.

Доверительный интервал для коэффициента корреляции рассчитывается следующим образом. Вводят вспомогательную величину z:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Стандартная ошибка величины z:

$$\sigma_z = \frac{1}{\sqrt{N-3}}.$$

Далее можно воспользоваться табличными значениями z и σ_z.

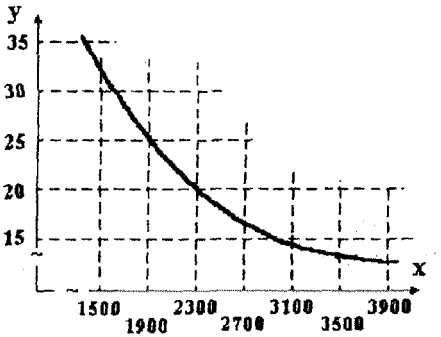

Рис. 6.7

Пример. При изучении корреляции (N = 500) получено r = 0,814.

Каковы доверительные границы для r?

$$\sigma_z = \frac{1}{\sqrt{500-3}} = \frac{1}{\sqrt{497}} = 0{,}045\,.$$

По таблицам значения $z(r) = -\ln\frac{1+r}{1-r}$ находим значения z. Доверительные интервалы для z рассчитываются по формуле $z_{нв} = \quad z \pm 2{,}58 \times \sigma_z$, где $z_н$, $z_в$ — нижняя и верхняя границы значения z.

Так, для значения r = 0,814 находим по таблицам: z = 1,139.

Тогда для $z_н = 1{,}139 - 2{,}58 \times 0{,}045 = 1{,}023$;

$$z_в = 1{,}139 + 2{,}58 \times 0{,}045 = 2{,}255$$

Обращаясь снова к тем же таблицам, находим

$$r_н = 0{,}771,\ r_в = 0{,}850.$$

Таким образом, доверительный интервал

$$r = 0{,}771 \div 0{,}850.$$

Проверка гипотезы об адекватности результатов наблюдений данным уравнения регрессии

Проверка того, что результаты наблюдений можно представить в виде полученного уравнения регрессии, осуществляется путем вычисления остаточной суммы S_R, которая представляет собой суммы квадратов отклонений экспериментальных (опытных) значений y_i от значений $\hat{y}_i$, найденных по данному уравнению регрессии, т.е.

$$S_R = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 .$$

Рассеяние результатов наблюдений вокруг линии регрессии характеризуется остаточной дисперсией σ_R^2,

$$\sigma_R^2 = \frac{S_R}{f_R},$$

где f_R — число степеней свободы, $f_R = N - K - 1 = f_2$, где N — число наблюдений, R — количество факторов = 2.

А дисперсия $\sigma^2(y)$, имеющая степень свободы $f_y = N - 1 + f_1$ характеризует рассеяние результатов наблюдения относительно $\bar{y}$.

Если сравнивать σ_R^2 и σ^2, то можно оценить, в какой степени предсказание по уравнению регрессии лучше, чем по среднему $\bar{y}$. Задача сравнения σ_R^2 и σ^2 решается при помощи F распределения (распределения Фишера), где вычисляется отношение (одно из нескольких, разработанных Фишером):

$$F_{\text{расч}} = \frac{\dfrac{\sum (\bar{y} - y_i)^2}{N-1}}{\dfrac{\sum (\hat{y}_i - y_i)^2}{N-k-1}},$$

которое сравнивается с табличным $F_{\text{табл}}$ для избранного уровня значимости и соответствующих степеней свободы числителя $f_i = N - l$ и знаменателя $f_2 = N - k - 1$.

Если $F_{\text{табл}} < F_{\text{расч}}$, то уравнение адекватно представляет результаты наблюдений, если $F_{\text{расч}} < F_{\text{табл}}$ — уравнение неадекватно наблюдениям.

2. Порядок выполнения работы

Предыдущий анализ устанавливает, что выбранные признаки подчиняются нормальному закону распределения, а по физическому смыслу очевидно, что взаимосвязь признаков существует.

Основная цель анализа: установить и оценить тесноту связи между выбранными признаками, найти и оценить аналитический вид зависимости этих признаков.

Пользуясь данными, а также результатами предыдущего задания, заполнить таблицу вида табл. 6.3, где приняты обозначения: N — общее число наблюдений; i — текущий номер наблюдения, $i = 1,2, ..., N$; y — один из заданных признаков, который в дальнейшем выступает и как функция, и как аргумент; x — один из заданных признаков, который в дальнейшем выступает и как функция, и как аргумент.

Таблица 6.3

i	y	x
1	y_1	x_1
	y_2	x_2
⋮	⋮	⋮
N	y_N	x_N

Естественно, что обозначения могут быть взяты и произвольно. Для получения парной зависимости (или ее еще называют двумерной, так как взято в рассмотрение два признака) анализ вести в следующей последовательности:

— построить корреляционное поле и эллипс рассеяния;

— заполнить корреляционную решетку;

— рассчитать параметры линейных зависимостей $y(x)$ и $x(y)$, так называемые прямые и обратные линейные и нелинейные зависимости на ЭВМ:

$$y = a_0 + a_1 x\,,$$

$$x = b_0 + b_1 y\,;$$

— рассчитать коэффициент Фишера (P) и оценить адекватность представления признаков;

— построить графики;

— рассчитать коэффициенты корреляции и корреляционного отношения;

— сделать вывод о тесноте связи выбранных признаков и о более вероятном аналитическом виде связи.

3. Подготовка к ЭВМ

Столбцы значений индивидуального задания (см. табл. 6.3) вводятся в ЭВМ в том порядке, как заданы в исходной информации (см. табл. 1.1). Размерность значений не должна превышать трех знаков перед запятой.

Информация вводится в режиме диалога. Результаты выдаются на печать в виде коэффициентов уравнения и табулированных значений.

Примерный вид печати результатов:

Среднее ряда $x\bar{x}$ = 1680,00000

Среднее квадр. откл. ряда x = 180,5550

Среднее ряда y = 17,10000

Среднее квадр. откл. ряда y = 2,1190

Коэффициент корреляции r = -0,6613

$$y_1 = (-0{,}0078)x + (30{,}1379)$$

Корреляционное отношение r = 0,6613

$y_2 = (0{,}0000)x\ \hat{}2 + (-0{,}0244)x + (43{,}8271)$

Корреляционное отношение $r = 0{,}6672$

$y_3 = (-0{,}0000)x\ \hat{}3 + (0{,}0000)x\ \hat{}2 + (-0{,}0323)x + (48{,}1732)$

Корреляционное отношение $r = 0{,}6872$.

Табуляция

x	y	y_1	y_2	y_3
2000.0000	15.00000	14.61660	15.01090	15.00110
1800.0000	18.00000	16.16870	16.09500	16.10160
1750.0000	17.00000	16.55680	16.42850	16.43280
1500.0000	19.00000	18.49690	18.47040	18.46170
1650.0000	20.00000	17.33280	17.17020	17.16770
1350.0000	20.00000	19.65100	19.99480	20.00270
1800.0000	15.00000	16.15870	16.09500	16.10160
1800.0000	15.00000	16.15870	16.09500	16.10160
1650.0000	14.00000	17.33280	17.17020	17.16770
1500.0000	18.00000	18.49690	18.47020	18.46170

СПИСОК ЛИТЕРАТУРЫ

1. *Аллен Р.* Экономические индексы. — М.: Статистика, 1980.
2. *Баженова С.Г.* Методические указания по выполнению расчетов и анализа двумерной совокупности на ЭВМ. — М.: МГИ, 1981.
3. *Баженова С.Г., Велесевич В.И.* Практикум по статистике горной промышленности. — М.: МГИ, 1982.
4. *Баженова С.Г.* Статистика-вероятностные модели горного производства. — М.: МГИ, 1984.
5. *Баженова С.Г.* Выравнивание рядов и интерпретация зависимостей. — М.: МГИ, 1984.
6. *Баженова С.Г.* Статистические методы в организации и управлении горной промышленностью. — М.: МГИ, 1977.
7. *Баженова С.Г., Велесевич В.И.* Математические методы статистики угольной промышленности. — М.: МГИ, 1986.
8. *Баженова С.Г.* Методические указания по выполнению аудиторно-домашних заданий по «Статистике». — М.: МГИ, 1988.
9. *Баженова С.Г., Лихтерман С.С.* Формирование производственных бригад на горных предприятиях. — М.: ЦНИИцветмет экономики и информ., 1989.
10. *Баженова С.Г.* Анализ работы горных предприятий. — М.: МГИ, 1992.
11. *Баженова С.Г.* Практическая статистика. — М.: Изд-во МГГУ, 1994.
12. *Боровков А.Л.* Теория вероятностей. — М.: Наука, 1976.
13. *Венецкий И.Г.* Вариационные ряды и их характеристики. — М.: Статистика, 1970.
14. *Гнеденко Б.В.* Курс теории вероятностей. — М.: Наука, 1969.
15. *Леман Э.* Проверка статистических гипотез. — М.: Наука, 1964.
16. *Казинец Л.С.* Теория индексов. — М: Госстатиздат, 1963.
17. *Колмогоров А.Н.* Основные понятия теории вероятностей. — М.: Наука, 1974.
18. *Крамер Г.* Математические методы статистики. — М.: Мир, 1975.

19. *Моррис У.* Наука об управлении. Байесовский подход. — М.: Мир, 1971.
20. *Мордэкэй Езекиэл* и *Карк А.Фокс.* Методы анализа корреляций и регрессий. — М.: Статистика, 1966.
21. *Немчинов В.С.* Избранные произведения. Т. 2. — М: Наука, 1967.
22. *Рыжов П.А.* Математическая статистика в горном деле. — М.: МИРГЭМ, 1965.
23. *Д.Тернер.* Вероятность, статистика и исследование операций. — М.: Статистика, 1976.
24. *Фишер И.* Построение индексов: Пер. с англ. — М.: ЦСУ, 1928.
25. Популярный экономико-статистический словарь. Словарь-справочник / Под ред. И.И. Елисеевой. — М.: Финансы и статистика, 1993.

СОДЕРЖАНИЕ

Предисловие 5

Общие положения 7

Работа № 1. Статистическая совокупность наблюдений. Сбор и формирование 9

Работа № 2. Определение необходимого объема наблюдений 17

Работа № 3. Одномерная совокупность наблюдений. Вариационный ряд 37

Работа № 4. Статистические характеристики совокупности 49

Работа № 5. Графическое представление совокупности 67

Работа № 6. Парная корреляция и регрессия 78

Список литературы 97

ПРАКТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ГОРНЫХ ИНЖЕНЕРОВ

Светлана Георгиевна **Баженова**

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГОРНОЙ ПРОМЫШЛЕННОСТИ

Режим выпуска «стандартный»

Редактор текста *Е.Н. Толстая*
Компьютерная верстка и подготовка оригинал-макета: *Т.Н. Абросимова*
Дизайн серии: *Е.Б. Капралова*
Полиграфическое производство: *Т.Д. Герасимова, Н.Д. Уробушкина, Г.Н. Потемкина*

Подписано в печать 31.08.2001. Формат 60×84/16. Бумага офсетная № 1. Гарнитура «Times». Печать трафаретная на цифровом дупликаторе. Уч.-изд. л. 6,28. Усл. печ. л. 5,81. Тираж 300 экз. Заказ 667

ИЗДАТЕЛЬСТВО МОСКОВСКОГО ГОСУДАРСТВЕННОГО ГОРНОГО УНИВЕРСИТЕТА
Лицензия на издательскую деятельность ЛР № 062809 от 30.06.98 г.
Код издательства 5Х7(03)

Отпечатано в типографии Издательства Московского государственного горного университета
Лицензия на полиграфическую деятельность ПЛР № 53-305 от 05.12.97 г.

119991, Москва, ГСП-1, Ленинский проспект, 6; Издательство МГГУ; тел. (095) 236-97-80; факс (095) 956-90-40