Текст 3. МАТЕМАТИКО-СТАТИСТИЧЕСКИЙ ИНСТРУМЕНТАРИЙ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

Прикладная статистика – самостоятельная научная дисциплина, обслуживающая широкий класс реальных задач статистического анализа данных:

- исследование динамики структуры состояний объектов (демографической или социальной структуры общества, структуры типологии потребительского поведения домашних хозяйств и т.п.);
- типологизация социально-экономических объектов (семей, фирм, предприятий, регионов, стран и т. п.);
- построение интегральных индикаторов качества или эффективности функционирования социально-экономической системы (уровня или качества жизни, качества населения, эффективности функционирования предприятия и т. п.);
- выявление скрытых (латентных) факторов, определяющих течение того или иного социально-экономического процесса;
- исследование и моделирование генезиса анализируемых статистических данных.

Два направления развития методов статистической обработки анализируемых данных:

- методы, предусматривающие возможность ***вероятностной интерпретации*** обрабатываемых данных и полученных в результате обработки статистических выводов – содержание математической статистики; предполагается вероятностная природа данных.
- методы, которые априори ***не опираются на вероятностную природу*** обрабатываемых данных (классификация и кластер-анализ, многомерного шкалирования, теории измерений и др.).

При разработке математико-статистического инструментария, а также при использованию разработанного метода в решении конкретной практической задачи. специалисту (математику-прикладнику, системному исследователю, методисту и пр.), приходится:

- глубоко вникать в ***содержательную сущность задачи***, адекватно «прилаживать» исходные модельные допущения (на которых строится любой математический метод) к выяснению сущности реальной задачи;

- решать задачу ***преобразования*** имеющейся ***исходной информации*** к стандартной (унифицированной) форме записи обрабатываемых статистических данных;

- ***разрабатывать*** практически реализуемые ***вычислительные алгоритмы и программное обеспечение*** с учетом специфики обрабатываемой статистической информации.

Прикладная статистика – самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов.

Под организацией ***сбора статистических данных имеется в виду*** лишь определение способа отбора подлежащих статистическому обследованию единиц (семей, предприятий, стран, пациентов и т.п.) из всей исследуемой совокупности. Сюда ***не включается*** разработка методологии измерителей анализируемых свойств отображаемого объекта: эта работа предполагает профессиональное (социологическое. экономическое и т. п.) изучение сущности задач, для решения которых требуется статистическая информация и относится к компетенции предметной статистики соответствующей области

Также используется термин «***анализ данных***», понимаемый в расширительном толковании.

Теория вероятностей и математическая статистика – основные «поставщики» математического инструментария для прикладной статистики и эконометрики. Ситуация применимости ***теоретико-вероятностного*** математического аппарата:

- мы находимся в условиях стационарного действия некоторого реального комплекса условий, включающего в себя «мешающее» влияние большого числа случайных (не поддающихся строгому учету и контролю) факторов, которые в свою очередь не позволяют делать полностью достоверные выводы о том, произойдет или не произойдет интересующее нас событие

- предполагается, что имеется принципиальную возможность (хотя бы мысленно реально осуществимая) многократного повторения эксперимента или наблюдения в рамках того же самого реального комплекса условий.

1. Возможные области применения – отдельные разделы экономики и социологии и в первую очередь задачи, связанные с исследованием поведения объекта (индивидуума, семьи или другой социально-экономической или производственной единицы) как представителя большой однородной совокупности подобных же объектов.

Традиционная область использования вероятностно-статистического аппарата является демография.

Важная общая черта – существенная многомерность обрабатываемой информации, характеризующей исследуемые явления или объекты – т. е. состояние или поведение каждого из этих объектов в любой фиксированный момент времени описывается набором соответствующих показателей.

Среди этих показателей могут быть:

- ***количественные*** (среднедушевой доход в семье, размер семьи, объем валовой продукции предприятия и т. д.);
- ***не количественные*** (качественные):
  - ***ранговые*** (классификация специалиста, сравнительная характеристика жилищных условий);
  - классификационные или ***номинальные*** (профессия, национальность, пол, причины миграции и т. п.).

Все эти показатели находятся в сложной взаимосвязи друг с другом – необходимость применять методы многомерного статистического анализа.

Второй категории возможных областей применения – допустимые вероятностно-статистические приложения. К ним относятся ситуации, характеризующиеся весьма значительными нарушениями требования сохранения неизменными условий эксперимента (стационарность).

К категории ***недопустимых вероятностно-статистических приложений*** относятся ситуации, характеризующиеся либо бессодержательностью идеи многократного повторения одного и того же эксперимента в неизменных условиях, сфор-

мулированной в требовании, либо полной детерминированностью изучаемого явления, т.е. – отсутствием «мешающего» влияния множества случайных факторов

В подобных ситуациях исследователь должен пользоваться методами анализа данных и не должен претендовать на вероятностную интерпретацию обрабатываемых данных и получаемых в результате их обработки выводов.

Строгих математических методов, позволяющих точно определять, находимся ли мы в условиях применимости вероятно-статистической «идеологии» не существует: любая вероятностная модель, так же как и любая математическая модель вообще, есть лишь некоторая аппроксимация исследуемой реальной действительности.

Концепция ***субъективных вероятностей*** в рамках которой правомерно говорить о вероятностной модели таких событий.

С помощью экспертов вместо действительной многократной реализации интересующего нас эксперимента в одних и тех же условиях ограничиваемся воображаемой прогонкой исследуемой ситуации «через сознание» многих экспертов.

Эксперт интерпретируется как некий измерительный прибор, работающий со случайной ошибкой. Точность работы этого «прибора», т. е. точность «прочтения» (в сознании эксперта) исхода интересующего нас события в будущем, очевидно, зависит как от степени объективного влияния «мешающих» случайных факторов, так и от степени осведомленности, компетентности и других субъективных качеств самого эксперта.

Единственным объективным судьей в подобных вопросах может быть лишь критерий практики. В связи с этим С.А.Айвазян приводит следующую цитату (Ф. Энгельс «Анти-Дюринг»): «... Математика, вообще столь строго нравственная, совершила грехопадение: она вкусила от яблока познания, и это открыло ей путь к гигантским успехам, но вместе с тем и к заблуждениям. Девственное состояние абсолютной значимости, неопровержимой доказанности всего математического навсегда ушло в прошлое; наступила эра разногласий, и мы дошли до того, что большинство людей дифференцирует и интегрирует не потому, что люди понимают, что они делают, а просто потому, что верят в это, так как до сих пор результат всегда получался правильным»

1. Введение в прикладной статистический анализ

Системность и ***технологичность*** – основные принципы формирования информационного обеспечения на всех этапах.

1.1. Назначение и содержание прикладной статистики

Теория вероятностей и математическая статистика по отношению к прикладной статистике – разработчики и поставщики существенной части математического аппарата.

Доводка же и развитие этого аппарата подчиняются требованиям и специфике приложенных областей и производятся в рамках дисциплины «прикладная статистика».

1.1.1. Два подхода к интерпретации и анализу исходных статистических данных

Вероятностно-статистический подход развивается в рамках классической математической статистики и предусматривает возможность вероятностной интерпретации анализируемых данных и получаемых в результате этого анализа статистических выводов.

При подобной (вероятностной) интерпретации исходных статистических данных исследователем рассматриваются:

- реально наблюдаемая, статистически представленная рядом наблюдений – ***выборка***;
- теоретически домысливаемая совокупность объектов – ***генеральная совокупность***.

Основные свойства и характеристики выборки – ***эмпирические*** (***выборочные***), могут быть проанализированы и вычислены по имеющимся эмпирическим (статистическим) данным.

Основные свойства и характеристики генеральном совокупности – теоретические – не известны исследователю.

Назначение математико-статистических методов: получить как можно более точное представление о теоретических свойствах и характеристиках по соответствующим свойствам и характеристикам выборок.

Проблемы планирования выборки и обработки выборочных данных.

Типы выборок: случайная выборка, случайная стратифицированная, квотная и т.п.

Если есть априорная вероятностная модель порождения данных (зависимостей между анализируемыми признаками), то она используется при выборе метода статистической обработки.

Принципиально иная ситуация:

- исследователь не располагает априорными сведениями о вероятностной природе анализируемых данных;
- эти данные вообще не могут быть интерпретированы как выборка из генеральной совокупности.

В этом случае исследователь должен опираться на соображения конкретно-содержательного плана: как именно получены анализируемые данные, какова конечная прикладная цель их анализа. Поскольку эти соображения основаны на обычной логике и реализуются, как правило, в рамках логико-алгебраического подхода.

Различие описываемых подходов проявляется в способе обоснования выбора критерия качества статистического вывода, а также в интерпретации самого критерия и получаемых статистических выводов.

После же выбор конкретного вида оптимизируемого критерия качества математические средства решения задачи статистического анализа и моделирования данных оказываются общими для обоих подходов (методы оптимизации).

На заключительном же этапе ***интерпретации*** полученных статистических каждый из подходов снова имеет свою специфику.

Общим для обоих описываемых подходов:

- наличие исходной статистической информации на «***входе***» задачи;
- необходимость наилучшего (оптимального для некоторого критерия качества) способа статистического анализа или моделирования этой информации с целью получения научных или практических выводов «***на выходе***».

Синтез двух указанных подходов к статистическому анализу и моделированию позволил прикладной статистике объединить как прикладные вероятностно-статистические методы многомерного статистического анализа, включая модели регрессии (опираясь на методы статистического оценивания и статистической проверки гипотез), так и логико-алгебраические (оптимизационные) методы анализа

данных исключающие использование в своих построениях вероятностных рассуждений и моделей.

1.1.2. Три центральные проблемы прикладной статистики

Перед тем как сформулировать три центральные проблемы прикладной статистики, следует остановиться на двух основных формах записи исходных статистических данных (***и.с.д.***). Первую, наиболее распространенную, форму представления и.с.д. обычно называют матрицей (таблицей) «объект-свойство», «объект-признак» и т.п.

$$(\text{и.с.д.})_1 = \begin{pmatrix} x_1^{(1)}(t) & x_1^{(2)}(t) & \cdots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & x_2^{(2)}(t) & \cdots & x_2^{(p)}(t) \\ \cdots & \cdots & \cdots & \cdots \\ x_n^{(1)}(t) & x_n^{(2)}(t) & \cdots & x_n^{(p)}(t) \end{pmatrix},$$

$$t = t_1, t_2, \ldots, t_N,$$

где $x_i^{(j)}(t)$ – значение j-го анализируемого признака, характеризующего состояние i-го объекта в момент времени t.

Пространственно-временная выборка: статистическому обследованию подвергаются n объектов, на каждом из объектов регистрируются значения p характеризующих его признаков в N последовательные моменты времени t.

Последовательность из N матриц «объект-свойство» – n реализаций p-мерного временного ряда $x_i(t), i = 1,\ldots,n$.

Одномоментные наблюдения: N=1 – пространственная статистика:

$$(\text{и.с.д.})_1 = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(p)} \\ \cdots & \cdots & \cdots & \cdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(p)} \end{pmatrix}$$

Если n=1 – единственная траектория p-мерного временного ряда.

В ряде ситуаций (исходные статистические данные получают с помощью специальных опросов, анкет, экспертных оценок) элементы первичного наблюдения является характеристика $\gamma_{i,j}(t)$ парного сравнения двух объектов (или признаков).

Мера сходства, различия, мера связи. Свойство симметричности. Коэффициент корреляции. Евклидово расстояние.

Матрицы ***парных сравнений*** объектов (размера n x n) или признаков (p x p).

$$(\text{и.с.д.})_2 = \begin{pmatrix} \gamma_{11}(t), & \gamma_{12}(t), & \dots, & \gamma_{1m}(t) \\ \gamma_{21}(t), & \gamma_{22}(t), & \dots, & \gamma_{2m}(t) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \gamma_{m1}(t), & \gamma_{m2}(t), & \dots, & \gamma_{mm}(t) \end{pmatrix} \begin{pmatrix} m = n & \text{или} & p; \\ t = t_1, & \dots, & t_N \end{pmatrix}.$$

Проблема 1. Статистическое исследование зависимостей – структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными.

Проблема 2. Разработка статистических методов классификации объектов и признаков – анализ матриц парных сравнений.

Всю анализируемая совокупность объектов, статистически представленную в виде матрицы данных (объект-свойство или парных сравнений) требуется разбить на сравнительно небольшое число (заранее известное или нет) однородных (смысле отношения сходства, близости и т.п.), групп или классов.

Для формализации этой проблемы анализируемые объекты рассматриваются как точки соответствующего признакового пространства.

Если исходные данные представлены в форме матрицы объект-свойство, то эти точки являются непосредственным геометрическим изображением многомерных наблюдений в ***p***-мерном пространстве.

Если исходные данные представлены в форме матрицы парных сравнений, исследователю не известны непосредственно координаты этих точек, но зато задана структура парных расстояний (близостей) между объектами.

Предполагается, что геометрическая близость двух или нескольких точек в этом пространстве означает близость «физических» состояний соответствующих объектов, их однородность.

Тогда проблема классификации состоит в разбиении анализируемой совокупности точек-наблюдений на сравнительно небольшое число классов таким образом, чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга.

Полученные в результате разбиения классы – кластеры, таксоны, образы; методы их нахождения – кластер-анализ, таксономией, распознаванием образов.

Cluster (англ.) – гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством

Taxon (англ.) – систематизированная группа любой категории (термин биологического происхождения)

В зависимости от конечных прикладных целей исследования используются процедуры кластер-анализа (метод k-средних), методы дискриминантпного анализа и т.п.

Проблема 3. Снижение размерности исследуемого признакового пространства с целью лаконичного (компактного, минимального и т.п.) объяснения природы анализируемых многомерных данных

Возможность лаконичного описания основана допущении: существует относительно небольшое число признаков-детерминант (главных компонент, общих факторов, наиболее информативных объясняющих переменных), достаточно точно описывающие как сами наблюдаемые переменные (все элементы исходных матриц данных), так и определяемые этими переменными свойства (характеристики) анализируемой совокупности.

Признаки-детерминанты (главных компонент, общих факторов, наиболее информативных объясняющих переменных) могут находиться среди исходных признаков, а могут быть латентными, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным.

Пример практической реализации этой идеи – периодическая система элементов Менделеева: в этом случае роль единственного признака-детерминанта играет заряд атомного ядра.

Несколько типовых задач

Отбор наиболее информативных показателей (включая выявление латентных факторов)

Речь идет об отборе из исходного (априорного) множества признаков или о построении в качестве некоторых комбинации исходных признаков относительно небольшого числа переменных, которые обладали бы свойством наибольшей информативности:

- достижение максимальной точности регрессионного прогноза некоторого результирующего количественного показателя у по известным значениям предикторных переменных

- наивысшая точность решения задачи отнесения объекта к одному из классов по значениям его описательных признаков – построение системы типообраэующих признаков в задачах классификации или выявление и интерпретация некоторой сводной (латентной) характеристики изучаемого свойства

- максимальная автоинформативность новой системы показателей – максимально точное воспроизведение всех исходных признаков по сравнительно небольшому числу вспомогательных переменных – наилучший автопрогноз –модели факторного анализа.

Сжатие массивов обрабатываемой и хранимой информации.

Этот тип задач требует в качестве одного из основных приемов решения построения экономной системы вспомогательных признаков, обладающих наивысшей автоинформативностью – наилучшим автопрогнозом. Для подобных задач используется сочетание методов классификации и снижения размерности. Методы классификации позволяют подчас перейти от массива, содержащего информацию по всем статистически обследованным объектам, к соответствующей информации только по эталонным образцам, где в качестве эталонных образцов берутся специальным образом отобранные наиболее типичные представители классов, полученных в результате операции разбиения исходного множества объектов на однородные группы.

Визуализация (наглядное представление) данных. Проблема проецирования анализируемых многомерных данных из исходного пространства на прямую, на плоскость, в крайнем случае – в трехмерное пространство. Причем так, чтобы интересующие нас специфические особенности исследуемой совокупности (например, ее расслоенность на кластеры), сохранились бы и после проецирования.

Снижении размерности анализируемого признакового пространства, подчиненного некоторым критериям адекватности. При условии, что размерность редуцированного пространства не должна превышать трех.

Построение условных координатных осей (многомерное шкалирование, латентно-структурный анализ)

В данном типе задач снижение размерности осуществляется за счет анализа матриц отношения близости.

1.2. Основные этапы прикладного статистического анализа

Этап 1. Исходный (предварительный) анализ исследуемой реальной системы. **Концептуальное моделирование**

В результате этого анализа определяются:

- цели исследования на неформализованном, содержательном, уровне;
- совокупность единиц, представляющая предмет статистического исследования;
- перечень отобранных из представленного специалистами априорного набора показателей, характеризующих состояние (поведение) каждого из обследуемых объектов, который предполагается использовать в данном исследовании;
- степень формализации соответствующих записей при сборе данных;
- общее время и трудозатраты, отведенные на планируемые работы, и коррелированные с ними временная протяженность и объем необходимого статистического обследования;
- моменты, требующие предварительной проверки перед составлением детального плана исследования (например, не всегда априори ясна возможность идентификации единиц наблюдения);
- формализованная постановка задачи, включающая (если возможно) вероятностную модель изучаемого явления и природу статистических выводов, к которым должен (или может) прийти исследователь в результате переработки массива исходных данных;
- формы, используемые для сбора первичной информации и для формирования тематических баз данных.

По затратам сил наиболее квалифицированного персонала трудоемкость первого этапа работы бывает сравнима с суммарной трудоемкостью всех остальных этапов при условии, что обработка проводится с помощью подходящего пакета программ.

Методы компьютерного ассистирования в проведении этой части работы.

Этап 2. Составление детального плана сбора исходной статистической информации.

При составлении этого плана необходимо учитывать полную схему дальнейшего статистического анализа.

При планировании особого внимания заслуживают случаи, когда:

- используется аппарат теории выборочных обследований, т. е. определяется, какой должна быть выборка – случайной, пропорциональной, расслоенной и т. п.;
- «организационно-методическая подготовка»; вопросы разработки методологии определения априорной системы показателей, характеризующих исследуемый объект или процесс отнесены к области предметной статистики.

Этап 3. Сбор исходных статистических данных и формирование тематической БД исследования.

Разработка инструментария для формирования тематических рабочих файлов (проблемных файлов исследования файлов частных задач «problem files») – таблиц «объект-признак», «объект-объект», «признак-признак».и пр. свободных таблиц.

Одновременно формируется метабаза исследования – определения используемых терминов в различных форматах (для автоматического формирования таблиц и т.п.), формулы группировок и категорий, методические материалы и т.п.

Этап 4. Первичная статистическая обработка данных.

Решаются задачи:

- отображение текстовых переменных в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу – составление и реализация справочников показателей;
- статистическое описание исходных совокупностей с определением пределов варьирования переменных;
- анализ резко выделяющихся наблюдений (больших отклонений);
- анализ и восстановление пропущенных значений;
- проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных;

- унификация типов переменных, когда с помощью различных приемов добиваются унифицированной записи всех переменных;

- экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений –процессом составления сводки и группировки.

Кроме того, вычислительная реализация решения следующих вопросов:

- учет размерности и алгоритмической сложности задачи и одновременно возможностей используемого вычислительного средства;
- формулировку задачи на входном языке используемого программного обеспечения и т.п.

Анализ резко выделяющихся наблюдений

Даже беглый предварительный просмотр исходных данных может вызвать сомнения в истинности (или правомерности) отдельных наблюдений, слишком резко выделяющихся на общем фоне.

Вопрос: обнаруженные резкие отклонения в исходных данных (аномальные выбросы):

- обычные случайные колебания выборки

- существенные искажения условий сбора статистических данных или ошибки регистрации (записи)?

В последнем случае «подозрительные» наблюдения следует исключить из дальнейшего рассмотрения.

Определение критического (порогового) уровня аномальности.

Восстановление пропущенных (стертых) наблюдений

В матрицах исходных статистических данных (9.1) или (9.2) по разным причинам (в том числе и в результате исключения аномальных данных) могут быть пропуски отдельных элементов или каких-то частей строк или столбцов.

Исключать по этой причине из дальнейшего рассмотрения весь объект слишком расточительно. Возникает задача наилучшего восстановления пропущенных данных.

Методы восстановления пропущенных значений.

Проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных

Применение ряда статистических методов является правомерным лишь в ситуациях, когда справедливо допущение о статистической независимости обрабатываемого ряда наблюдений.

Поэтому, перед тем как подвергнуть имеющиеся результаты наблюдения основной статистической обработке, необходимо выяснить, являются ли они статистически независимыми или их следует рассматривать как последовательности взаимозависимых величин.

Унификация типа переменных

Среди компонент анализируемого многомерного признака могут быть показатели трех разных типов: количественные, качественные (порядковые, ординальные) и классификационные (номинальные).

Группировка количественных признаков.

«Оцифровка» неколичественных переменных (ранговых и номинальных).

Экспериментальный анализ закона распределения исследуемой генеральной совокупности

Вычисление основных числовых характеристик распределения:

- среднего значения,
- дисперсии,
- среднеквадратического отклонения,
- минимальное и максимальное значения,
- коэффициентов асимметрии и эксцесса,
- элементов выборочной ковариационной матрицы,

Численный и графический анализ одномерных законов распределения рассматриваемых показателей: построение соответствующих гистограмм, эмпирических функций распределения, проверка гипотезы о нормальности распределения.

$$\Phi(x;m,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}e^{-\frac{(x-m)^2}{2\sigma^2}}dx$$

Этап 5. Составление детального плана вычислительного анализа материала

Справка по собранному материалу и результатам предварительного анализа.

Определение основных групп, для которых будет проводиться дальнейший анализ.

Пополнение и уточнение тезауруса содержательных понятий.

Блок-схема анализа с указанием привлекаемых методов. Задание оптимизационных критериев для выбора метода (методов) основной статистической обработки анализируемых данных.

Этап 6. Вычислительная реализация основной части статистической обработки данных

.Основная задача – эффективное управление вычислительным процессом: формулирование задания обработки и описание данных на входном языке используемого программного обеспечения.

Этап 7. Подведение итогов исследования

Построения формального статистического отчета о проведенном исследовании.

Затем результаты исследования, его основные выводы формулируются в содержательных терминах. Если исследование проводилось в рамках математико-статистических методов и моделей, то его выводы формулируются в терминах оценок неизвестных параметров анализируемой системы или в виде ответа на вопрос о справедливости проверяемой гипотезы и сопровождаются гарантируемыми количественными оценками степени их достоверности.

Если же исследование осуществлялось средствами анализа данных (т. е. в рамках логико-алгебраического подхода), то его выводы не претендуют на вероятностную интерпретацию.

В заключение проверяется, в какой мере достигнуты намеченные на этапе 1 содержательные цели работы, и если достигнуты не все из них, то объясняется, почему. Работа завершается содержательной формулировкой новых задач, вытекающих из проведенного исследования.

2. Производные показатели – индикаторы

При анализе использовались производные показатели – первичные индикаторы, которые рассчитывались по значениям первичных статистических показателей. Эти индикаторы разбиваются на следующие группы.

2.1. Стандартизированные индикаторы – Z-индикаторы, Z-оценки

Процедуры стандартизации: вычитание из значения показателя для каждого объекта наблюдения его среднего значения (***СР***) и делении полученного результата на величину среднеквадратического отклонения (***СКО***) рассматриваемого показателя от его среднего значения.

- Характеристическое свойство: СР =0, СКО =1.
- Шкалы всех стандартизированных индикаторов в качестве «нуля» имеют СР, а в качестве масштаба («единицы») используется «единица» разброса – СКО.
- Позволяет сравнивать значения различных индикаторов как попарно, так и в составе группы.

2.2 Масштабные структурные индикаторы (S-индикаторы)

Эти индикаторы характеризуют масштабность тех или иных явлений, происходящих во всей совокупности (генеральной совокупности) в целом.

Пример.

- Для показателя «Объем инновационной продукции (работ, услуг), млн. руб.»
- S-индикатор «Доля инновационной продукции, произведенной в субъекте РФ в общем объеме инновационной продукции, произведенной в РФ».

Сумма значений этого показателя по всем субъектам РФ равна 100%.

2.3. Линейные структурные индикаторы (R-индикаторы)

Для нескольких показателей рассматриваемого объекта, характеризующих различные аспекты общего свойства, определяют их сравнительные доли на уровне объекта.

Пример. Для показателей:

- Численность исследователей, докторов наук (чел.).

- Численность исследователей, кандидатов наук (чел.).

R-индикаторы:

- Доля докторов наук в общей численности исследователей с ученой степенью в субъекте РФ, %.
- Доля кандидатов наук в общей численности исследователей с ученой степенью в субъекте РФ, %.

В этом случае для каждого субъекта РФ сумма значений его R- индикаторов равна 100%.

3. Метод главных компонент

3.1. Продуктовые инновации

- bl27 Объем инновационной продукции (работ, услуг) (млн.руб.)
- bl28 % инновационной продукции в общем объеме отгруженной
- bs27 Доля СРФ в общем объеме инновационной продукции (работ, услуг) в РФ

3.2. Пример построения факторных моделей: блок показателей «Продуктовые инновации»

ФМ 1. Входные параметры для построения факторной модели

- N ***Количество факторизуемых показателей*** – первичных индикаторов N = 2
- X ***Множество объектов***, для которых при факторизации рассматриваются значения частных критериев. В анализируемой ситуации, множество X – совокупность состояний отобранных субъектов РФ по годам.
- ПИ=$\{ПИ_i\}_{i=1,...N}$ ***Вектор первичных индикаторов*** для блока «Продуктовые инновации», рассматриваемых на всей совокупности субъектов РФ

ФМ 2. Параметры модели, – результаты моделирования

St[***P***] – стандартизированная переменная (Z-оценка) для P.

- M ***Количество выделенных факторов*** – интегральных индикаторов (M=1)

- ИИ=$\{ИИ_i\}_{i=1,...M}$ ***Вектор факторов*** – интегральных индикаторов блока «Продуктовые инновации»
- $ИИ_1$ Fa1
- FW= $\{FW_{i,j}\}_{i=1...M;\ j=1,...N}$ ***Матрица факторных весов*** (нагрузок) $FW_{i,j}$ фактора i для первичного индикатора j
- FS= $\{FS_{i,j}\}_{i=1...M;\ j=1,...N}$ ***Матрица коэффициентов индивидуальных оценок факторов*** $FS_{i,j}$ фактора i для показателя j

Основные формулы

$$ИИ_j(x) = \sum_{i=1,...,N} FS_{i,j}\ St[ПИ_i(x)] \quad \forall\ j = 1,...,N,\ x \in X,$$

$$St[ПИ_i(x)] = \sum_{j=1,...,M} FW_{i,j}\ ИИ_j(x) \quad \forall\ j = 1,...,N,\ x \in X,$$

Обозначения (Ср(Р) – среднее значение, а СКО(Р) – среднеквадратическое отклонение от среднего значения для переменной Р):

$$St[P](x) = \frac{P(x) - Cp(P)}{СКО(P)}, \quad P(x) = Cp(P) + СКО(P) \cdot St[P](x).$$

$$\text{Ср}(P) = \frac{1}{|X|}\sum_{x \in X} P(x);\ \text{СКО}(P) = \left(\frac{1}{|X|-1}\sum_{x \in X}(P(x) - Cp(P))^2\right)^{1/2}.$$

Свойства. Вектор построенных факторов – ортонормальная система

$$Cp(ИИ_j) = 0, \qquad СКО(ИИ_j) = 1;$$

все факторы попарно ортогональны (корреляционно независимы)

$$\sum_{x \in X} ИИ_j(x) \cdot ИИ_l(x) = 0.$$

Выбор параметра М – ***количества отобранных в модель факторов*** – ответственный этап описания модели:

от него непосредственно зависит суммарный коэффициент информативности всех факторов, отобранных в модель.

Суммарный коэффициент информативности – интегральная оценка адекватности модели – доля общей дисперсии факторизованных переменных, объясняемой моделью.

Максимальное значение (100%) достигается в случае, когда отбираются все факторы – их количество числу исходных переменных.

Компромисс в выборе количества исходных факторов: чем больше их число, тем выше адекватность модели, но тем сложнее использование моделей для решения прикладных задач.

Идеальная ситуация для прикладных целей достигается отбором минимального количества факторов (одного фактора), особенно когда нас интересует критерий сравнения рассматриваемых объектов, что не всегда допустимо, исходя из требований адекватности модели.

Обычно для отбора количества факторов устанавливается некоторое пороговое значение информативности, например 0,7.

Количество отобранных в модель факторов определяется системой исходных показателей и системой индикаторов, построенной по исходным переменным, а также сложностью решаемой задачи.

Факторная модель для блока «Продуктовые инновации»

Код и наименование показателей		Ср	СКО	Fa1	Fa2
Информативность фактора, %				**86,4**	**13,6**
Кумулятивная информативность, %				86,4	100,0
Среднее значение (Ср)				0,0	0,0
Среднеквадратическое отклонение (СКО)				1,0	1,0
Коэффициенты индивидуальных оценок факторов					
bs27	Доля инновационной продукции СРФ в ее общефедеральном объеме	1,84	3,36	**0,535**	**1,407**
bl28	% инновационной продукции в общ. объеме отгруженной	4,65	5,07	**0,535**	**-1,407**
Факторные веса показателей					
bs27	Доля инновационной продукции СРФ в ее общефедеральном объеме	1,84	3,37	**0,96**	**0,36**
bl28	% инновационной продукции в общ. объеме отгруженной	4,65	5,07	**0,96**	**-0,36**

Основные формулы – интегральные индикаторы ***по матрице коэффициентов индивидуальных оценок факторов***

$$\boldsymbol{Fa}1(\boldsymbol{x}) = 0{,}535 \cdot \boldsymbol{St}[\boldsymbol{bs}27](\boldsymbol{x}) + 0{,}535 \cdot \boldsymbol{St}[\boldsymbol{bl}28](\boldsymbol{x}),$$
$$\boldsymbol{Fa}2(\boldsymbol{x}) = 1{,}407 \cdot \boldsymbol{St}[\boldsymbol{bs}27](\boldsymbol{x}) - 1{,}407 \cdot \boldsymbol{St}[\boldsymbol{bl}28](\boldsymbol{x}).$$

Расчетные формулы – модель факторов – ***по матрице факторных весов***

$$\boldsymbol{St}[\boldsymbol{bs}27](\boldsymbol{x}) = 0{,}96 \cdot \boldsymbol{Fa}1(\boldsymbol{x}) + 0{,}36 \cdot \boldsymbol{Fa}2(\boldsymbol{x}),$$
$$\boldsymbol{St}[\boldsymbol{bl}28](\boldsymbol{x}) = 0{,}96 \cdot \boldsymbol{Fa}1(\boldsymbol{x}) - 0{,}36 \cdot \boldsymbol{Fa}2(\boldsymbol{x}).$$

$$\boldsymbol{St}[\boldsymbol{bs}27](\boldsymbol{x}) = (\boldsymbol{bs}27(\boldsymbol{x}) - 1{,}84) / 3{,}36,$$
$$\boldsymbol{bs}27(\boldsymbol{x}) = 1{,}84 + 3{,}36 \cdot \boldsymbol{St}[\boldsymbol{bs}27](\boldsymbol{x}).$$

$$St[bl28](x) = (bl28(x) - 4{,}65 / 5{,}07),$$
$$bl28(x) = 4{,}65 + 5{,}07 \cdot St[bl28](x).$$

Интерпретация интегральных индикаторов (факторов):

Fa1 – объединенный (суммарный) процент инновационной продукции СРФ на федеральном и субфедеральном уровнях.

Fa2 – сбалансированный процент инновационной продукции СРФ на федеральном и субфедеральном уровнях.

Первый интегральный индикатор рассматривается как рейтинговой – по нему строится первый частный рейтинг для блока «Продуктовые инновации».

Второй рейтинговый индикатор ***Рей2*** определяется формулой:

$$Рей2 = bs27 \,/\, bl28$$

и интерпретируется следующим образом:

Рей2 – эффективность продуктовых инноваций в СРФ – на сколько % увеличится вклад субъекта РФ в федеральный объем инновационной продукции при увеличении доли инновационной продукции в субъекта РФ на 1%/

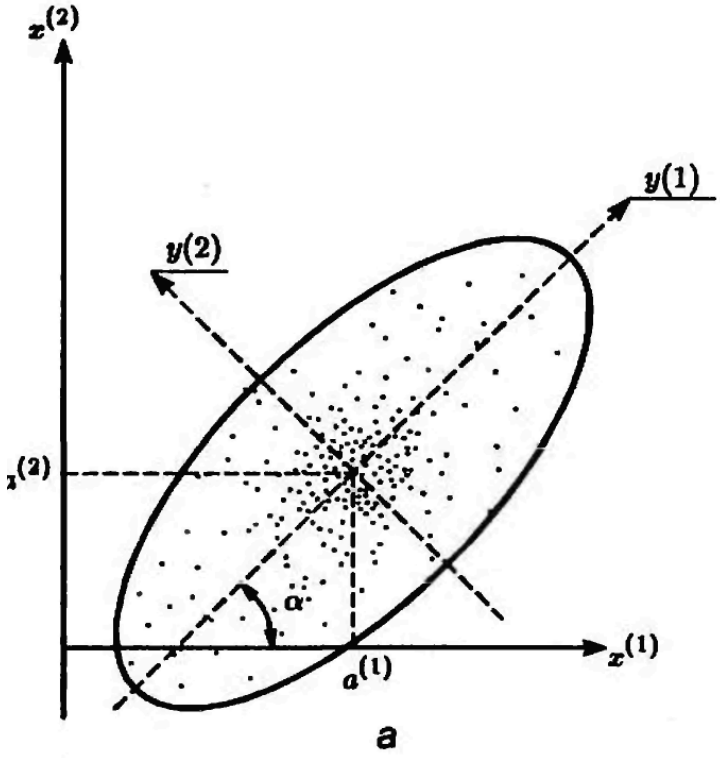
3.3 Геометрическая интерпретация метода главных компонент

Первой главной компонентой *y(1)* исследуемой системы показателей $X = (x^{(1)},..., x^{(p)})$ называется такая стандартизованная линейная комбинация этих показателей, которая среди прочих стандартизированных линейных комбинаций переменных $x^{(1)},..., x^{(p)}$ обладает наибольшей дисперсией.

k-й главной компонентой $y(k)$ $(k = 2,3,..., p)$ исследуемой системы показателей $X = (x^{(1)},..., x^{(p)})$ называется такая стандартизированная линейная комбинаций этих показателей, которая не коррелированна с ***k-1*** предыдущими главными компонентами и среди всех прочих стандартизированных и некоррелированных с предыдущими ***k-1*** главными компонентами линейных комбинаций переменных $x^{(1)},..., x^{(p)}$ обладает наибольшей дисперсией.

Система инвариантов для метода ***ГК***:

- ***ковариационная матрица*** – преобразования центрирования исходных показателей;
- ***корреляционная матрица*** – преобразования стандартизации исходных показателей.



а

Эллипс рассеяния исследуемых наблюдений и направление координатных осей главных компонент ***y***(1) и ***y***(2). Вырожденный (одномерный) случай: отсутствует разброс точек в направлении второй главной компоненты ***y***(2).


Структурно-функциональные ТИПЫ МОДЕЛЕЙ
Типологическая:
«Интегральный показатель»
Переменные состояния латентные
Входные переменные
Факторная:
«Латентный фактор»
Переменные состояния латентные
Входные переменные
Регрессионная:
«Восстановление зависимости»
Результирующие переменные
Входные переменные
Результирующие переменные
Переменные состояния латентные
Входные переменные
Регрессионно-типологическая модель