Федеральное агентство по образованию
Иркутский государственный университет

А. Ю. Филатов

Конспект лекций по многомерным статистическим методам

учебное пособие

Иркутск 2007

Печатается по решению редакционно-издательского совета
Иркутского государственного университета

УДК 330.43
ББК 65в6я73

Рецензенты: д-р техн. наук, проф. Зоркальцев В. И.
(зав. кафедрой математической экономики ИМЭИ ИГУ);
канд. физ.-мат. наук Тюрнева Т. Г.
(доцент кафедры теории вероятностей и дискретной математики ИМЭИ ИГУ)

Филатов А. Ю. Конспект лекций по многомерным статистическим методам: учеб. пособие / А. Ю. Филатов. – Иркутск: Иркут. ун-т, 2007. – 37 с.
ISBN 978-5-9624-0190-4

Охватывает основные разделы многомерных статистических методов. В частности, даны основы корреляционного анализа количественных, порядковых и категоризованных переменных, статистических методов классификации объектов и методов снижения размерности признакового пространства. Изложение сопровождается задачным материалом, приводятся модели из области микро- и макроэкономики. Издание содержит типовые задания для контрольных работ и вопросы к экзамену.

Предназначено для студентов, изучающих многомерные статистические методы, в качестве дополнения к лекционному курсу и рекомендуемой литературе.

ISBN 978-5-9624-0190-4

От автора

Несмотря на то, что курс многомерных статистических методов читается во многих вузах, до сих пор является незаполненной столь востребованная студентами ниша «конспектов лекций», выполненных преподавателями. Есть много хороших как российских, так и переводных учебников (среди них особо отметим двухтомник С. А. Айвазяна и В. С. Мхитаряна «Прикладная статистика и основы эконометрики», на следование которому, в первую очередь, и ориентирован данный курс). Есть совсем небольшие «шпаргалки», содержащие минимум информации и основные формулы. В то же время имеется насущная необходимость в чем-то среднем – а именно, в пособии, где студент, прослушавший курс лекций, быстро найдет необходимую формулу, свойство или разобранный пример, не утонув при этом в потоке излишней информации. Данное пособие может быть полезно для подготовки к экзаменам, для быстрого восстановления в памяти нужного материала, при решении практических задач.

При написании автор руководствовался принципом максимального приближения стиля изложения к конспекту лекций: тезисно, иногда обрывочными фразами (а именно так пишутся конспекты!) дать основную информацию; по возможности, разложить ее по пунктам; выделить самое главное (для этого использованы жирный и курсивный шрифты, шрифты большего размера, обрамление); привести основные формулы, а также соответствующие команды из Excel (базовые навыки работы в Excel необходимы для проведения расчетов). Поскольку численные примеры существенно облегчают восприятие, учебное пособие содержит примеры и задачи, основанные на моделях микро- и макроэкономики. Все примеры обозначаются знаком «##».

Материал, содержащийся в пособии, разбит на 15 лекций, каждая из которых занимает две страницы на одном развороте. Сделана попытка полностью соответствовать изложению в рамках курса лекций, который автор читает в ИМЭИ ИГУ. Для этого произведен отход от некоторых принятых стандартов. В частности, список литературы не вынесен на отдельную страницу, а дан (как и при чтении лекций студентам) в начале курса. В содержании указаны не названия лекций, а приведена краткая информация о темах, в них содержащихся. В заключительной части пособия приведены типовые задания для контрольных работ и вопросы к экзамену.

Издаваемый «конспект лекций» является вторым в серии изданных пособий. Ему предшествовало издание конспекта лекций по эконометрике. Также планируется издание конспекта по третьему эконометрическому курсу, в первую очередь посвященному анализу временных рядов.

Литература

1. Айвазян С. А. Прикладная статистика и основы эконометрики / Айвазян С. А., Мхитарян В. С. – М.: Юнити, 2002.
2. Айвазян С. А. Прикладная статистика в задачах и упражнениях. / Айвазян С. А., Мхитарян В. С. – М.: Юнити, 2001.
3. Сошникова Л. А. Многомерный статистический анализ в экономике. / Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шеффер М. – М.: Юнити, 1999.

1. Вероятностно-статистический подход (классическая математическая статистика) – статистические требования хотя бы приблизительно выполняются.

Выборка (реально наблюдаемая) эмпирические св-ва и характ-ки	$\Rightarrow$	***Ген. совокупность*** (теор. домысливаемая) теоретические св-ва и характ-ки

Цель: найти как можно точнее

Важны:

1) правильный выбор модели;
2) правильный метод статистической обработки данных.

2. Логико-алгебраический подход (эконометрика) – нет никаких сведений о вероятностной природе анализируемых данных или данные не могут быть интерпретированы как выборка из генеральной совокупности, есть только соображения конкретно-содержательного плана.

!!! Математ. средства одинаковы! Инструментарий – прикладная статистика !!!

Основные формы записи исходных данных (что на входе)

1. Матрица «объект-свойство»

$$\begin{pmatrix} x_1^{(1)}(t) & \dots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & \dots & x_2^{(p)}(t) \\ \dots & \dots & \dots \\ x_n^{(1)}(t) & \dots & x_n^{(p)}(t) \end{pmatrix}$$

$i = 1,\dots,n$ – объекты
$j = 1,\dots,p$ – свойства
$t = \{t_1, t_2, \dots, t_N\}$

– пространственно-временная выборка.

Часто равноотстоящие моменты времени $t_2 - t_1 = t_3 - t_2 = \dots = t_N - t_{N-1} = \Delta t \Rightarrow t = \{1,2,\dots,N\}$.

Предприятия / число сотрудников, средняя зарплата, объем фондов…

Частные случаи:

1) $N = 1$ – одномоментное наблюдение (сравнение цен на различные товары по регионам в конкретный момент времени);
2) $n = 1$ – анализ единственной траектории p-мерного временного ряда (динамика экономических показателей России);
3) $p = 1,\ n = 1$ – анализ одного временного ряда (динамика курса доллара).

2. Матрица парных сравнений

$\gamma_{ij}(t)$ – попарное сравнение объектов (или признаков) в момент времени t.

мера сходства, поток продукции $i \to j$ (экспорт, импорт), расстояние, отношение предпочтения, коэффициент корреляции признаков $x^{(i)}$ и $x^{(j)}$.

$$\begin{pmatrix} \gamma_{11}(t) & \dots & \gamma_{1n}(t) \\ \dots & \dots & \dots \\ \gamma_{n1}(t) & \dots & \gamma_{nn}(t) \end{pmatrix} \text{ или } \begin{pmatrix} \gamma_{11}(t) & \dots & \gamma_{1p}(t) \\ \dots & \dots & \dots \\ \gamma_{p1}(t) & \dots & \gamma_{pp}(t) \end{pmatrix}.$$

Часто $\gamma_{ij} = \gamma_{ji}$.

Основные задачи прикладной статистики

1. Статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными.
Корреляционный анализ, регрессионный анализ, анализ временных рядов.

2. Разработка статистических методов классификации объектов.
Всю совокупность n объектов разбить на небольшое число однородных групп.

3. Снижение размерности исследуемого признакового пространства.
p – исходное число анализируемых признаков,
$p' < p$ – итоговое число признаков (наиболее информативных).

Основные этапы прикладного статистического анализа

1. Предварительный анализ исследуемой реальной системы.
1) Определение основных целей исследования;
2) отбор признаков $x^{(1)}, x^{(2)}, ..., x^{(p)}$;
3) определение степени формализации записей при сборе данных, возможна ли идентификация единиц измерения;
4) определение форм, используемых для сбора первичной информации.

2. Составление детального плана сбора исходной статистической информации, определение выборки.

3. Сбор данных и ввод в компьютер.

4. Первичная обработка данных:
1) отображение переменных, описанных текстом (номинальная шкала с n градациями, порядковая шкала, небольшое число категорий);
2) унификация типов переменных (количественных, порядковых, категоризованных);
3) статистическое описание исходных совокупностей с определением пределов варьирования переменных;
4) обработка резко выделяющихся наблюдений (исключение, меньший вес при превышении порога, преобразование данных);
5) восстановление пропущенных данных;
6) проверка однородности порций данных: $(p \times n_1) + (p \times n_2) + ... + (p \times n_k) = (p \times n)$;
7) проверка статистической независимости последовательности наблюдений, составляющих массив данных;
8) экспериментальный анализ закона распределения генеральной совокупности (сводка и группировка):
выборочное среднее, дисперсия, асимметрия, эксцесс;
элементы выборочной корреляционной матрицы;
численный и графический анализ показателей;
использование априорных сведений о природе данных.

5. Составление детального плана вычислительного анализа, определение методов, формирование оптимизационного критерия.

6. Вычислительная реализация.

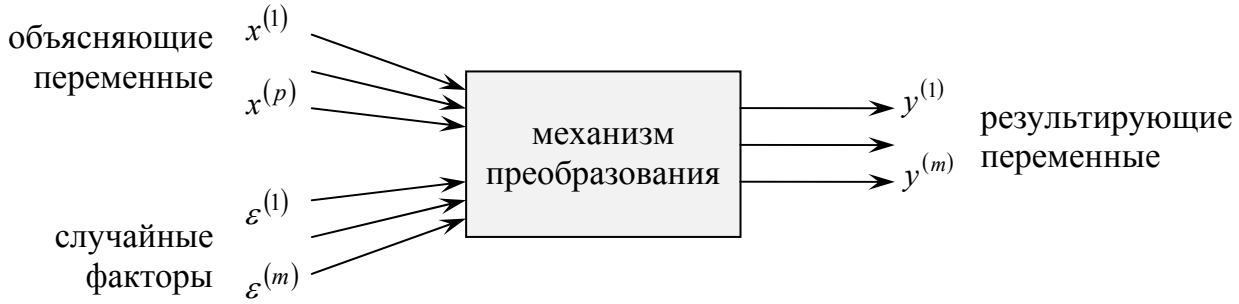
7. Подведение итогов и выводы.

Статистическое исследование зависимостей

Законы природы – взаимосвязи между изучаемыми явлениями.
Взаимосвязи выявляются на основании статистического наблюдения по выборке, частные наблюдения → общая закономерность.

$x^{(1)}, x^{(2)}, ..., x^{(p)}$ – ***объясняющие переменные*** (независимые, экзогенные), описывающие условия функционирования системы, часто поддаются регулированию.
$y^{(1)}, y^{(2)}, ..., y^{(m)}$ – ***результирующие переменные*** (зависимые, эндогенные), описывающие результат функционирования системы.
$\varepsilon^{(1)}, \varepsilon^{(2)}, ..., \varepsilon^{(m)}$ – ***латентные*** (скрытые) остаточные компоненты, отражающие влияние неучтенных факторов и случайных ошибок в измерениях.

Анализируемая система



Общая задача статистического исследования зависимостей:
На основе n измерений $\left\{x_i^{(1)},...,x_i^{(p)};\ y_i^{(1)},...,y_i^{(m)}\right\}$, $i = 1,...,n$ построить вектор-функцию

$$f\left(x^{(1)},...,x^{(p)}\right) = \begin{pmatrix} f^{(1)}\left(x^{(1)},...,x^{(p)}\right) \\ \\ f^{(m)}\left(x^{(1)},...,x^{(p)}\right) \end{pmatrix},$$

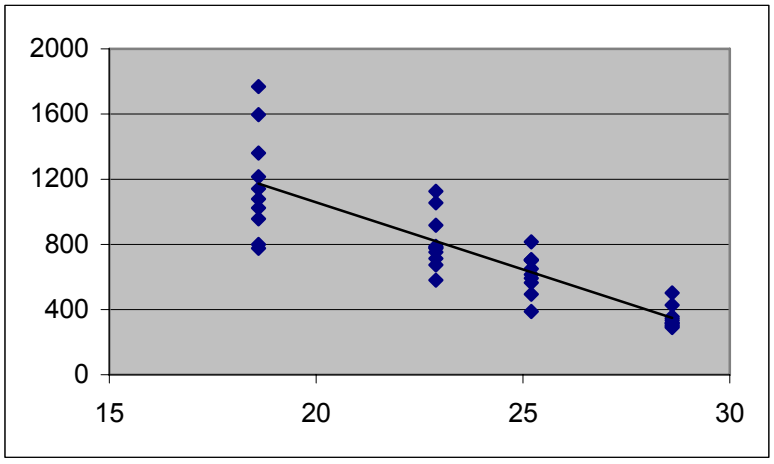
позволяющую наилучшим образом восстанавливать значения Y по заданным X.

1. Какова структура модели?
2. Каков критерий качества аппроксимации?
3. Какова прикладная цель исследования?

!!! Далее рассматриваем случай единственного результирующего показателя !!!

Анализируется поведение двумерной случайной величины (y, x) – цена пива «Brahma» в гипермаркете «Бонус» (руб.), суточный объем продаж (бут.). По каждой из 4 цен продажи проводились в течение 10 дней.

$$y = f(x) + \varepsilon$$



цена	объем продаж	средний объем продаж	среднекв. отклон., коэф. вариации
$x_1 = x_2 = ... = x_{10} = x_1^0 = 28{,}6$	$y_1 = 335$ $y_6 = 429$ $y_2 = 299$ $y_7 = 306$ $y_3 = 346$ $y_8 = 318$ $y_4 = 292!$ $y_9 = 357$ $y_5 = 501!!$ $y_{10} = 316$	$\bar{y}(x_1^0) = \frac{1}{10}\sum_{i=1}^{10} y_i = 350$	$\bar{s}(x_1^0) = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(y_i - 350)^2} = 66$, $\hat{V}(x_1^0) = \frac{66}{350} = 18{,}9\%$
$x_{11} = x_{12} = ... = x_{20} = x_2^0 = 25{,}2$	$y_{11} = 615$ $y_{16} = 388!$ $y_{12} = 564$ $y_{17} = 707$ $y_{13} = 707$ $y_{18} = 494$ $y_{14} = 650$ $y_{19} = 594$ $y_{15} = 817!!$ $y_{20} = 702$	$\bar{y}(x_2^0) = \frac{1}{10}\sum_{i=11}^{20} y_i = 624$	$\bar{s}(x_2^0) = \sqrt{\frac{1}{9}\sum_{i=11}^{20}(y_i - 624)^2} = 122$, $\hat{V}(x_2^0) = \frac{122}{624} = 19{,}6\%$
$x_{21} = x_{22} = ... = x_{30} = x_3^0 = 22{,}9$	$y_{21} = 582!$ $y_{26} = 753$ $y_{22} = 916$ $y_{27} = 781$ $y_{23} = 1125!!$ $y_{28} = 676$ $y_{24} = 778$ $y_{29} = 714$ $y_{25} = 790$ $y_{30} = 1056$	$\bar{y}(x_3^0) = \frac{1}{10}\sum_{i=21}^{30} y_i = 817$	$\bar{s}(x_3^0) = \sqrt{\frac{1}{9}\sum_{i=21}^{30}(y_i - 817)^2} = 168$, $\hat{V}(x_3^0) = \frac{168}{817} = 20{,}6\%$
$x_{31} = x_{32} = ... = x_{40} = x_4^0 = 18{,}6$	$y_{31} = 1141$ $y_{36} = 1078$ $y_{32} = 1024$ $y_{37} = 955$ $y_{33} = 1596$ $y_{38} = 799$ $y_{34} = 1770!!$ $y_{39} = 1362$ $y_{35} = 776!$ $y_{40} = 1217$	$\bar{y}(x_4^0) = \frac{1}{10}\sum_{i=31}^{40} y_i = 1172$	$\bar{s}(x_4^0) = \sqrt{\frac{1}{9}\sum_{i=31}^{40}(y_i - 1172)^2} = 325$, $\hat{V}(x_4^0) = \frac{325}{1172} = 27{,}7\%$

Класс функций: гипотеза о линейной функции спроса.

$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$, $\hat{y}_i = \theta_0 + \theta_1 x_i$, θ_0, θ_1 – неизвестные параметры модели.

Критерий качества: МНК, взвешенный МНК, обобщенный МНК, сумма модулей.

$\sum \varepsilon_i^2 = \sum (y_i - \theta_0 - \theta_1 x_i)^2 \to \min\limits_{\theta_0, \theta_1}$, если остатки $\varepsilon_i \in N(0, \sigma^2)$.

$\hat{\theta}_0 = 2702$, $\hat{\theta}_1 = -82{,}3$, $\hat{y}_i = 2702 - 82{,}3 x_i$.

Возможные цели исследования:

1. Установление факта наличия/отсутствия связи.
   Найти численное значение соответствующего показателя тесноты связи. Из всех показателей тесноты связи выбирается максимальный, он определяет вид связи.
2. Прогноз неизвестных значений.
   Найти точечную $\hat{Y} = \hat{f}(X)$ и интервальную $Y \in \left[\hat{f}(X) - \varepsilon_P(X,n);\ \hat{f}(X) + \varepsilon_P(X,n)\right]$ оценку прогнозируемого значения. Вид связи определяется минимумом ошибки прогноза.
3. Выявление причинных связей между объясняющими и результирующими переменными, управление значениями Y путем регулирования X. Проникновение в физический механизм связи. Важно правильное определение структуры модели.
   ## Истинная зависимость спроса y от цены $x^{(1)}$ и цены конкурента $x^{(2)}$:
   $y = 1000 - 10x^{(1)} + 12x^{(2)} + \varepsilon$.
   Нашли: $\hat{f}(X) = 1000 + 2x^{(1)}$, при этом установили, что $x^{(1)} \approx x^{(2)}$.
   У $\hat{f}(X)$ хорошие прогностические свойства, при том, что разные функции и даже разные знаки. Однако управление спросом невозможно.

Типовые задачи эконометрического моделирования

1. Нормирование

$y = f\left(x^{(1)},...,x^{(p)};\Theta\right)+\varepsilon$.

Необходимо найти вектор параметров Θ.

Решение задач массового обслуживания (супермаркет, такси).

Оценивание величины постоянных и средних переменных издержек.

2. Прогноз, планирование, диагностика

$y_i,\ x_i^{(1)},...,x_i^{(p)},\ i=1,...,n$ – значения в прошлом или на аналогичных объектах.

Необходимо оценить y_{n+1} по известным $x_{n+1}^{(1)},...,x_{n+1}^{(p)}$.

Прогнозирование спроса.

Диагностика эффективности рекламы.

Прогнозирование курсов акций и курсов валют.

3. Оценка труднодоступных для наблюдения параметров системы

Восстановление возраста археологической находки по косвенным показателям.

Оценка денежных сбережений по доходу.

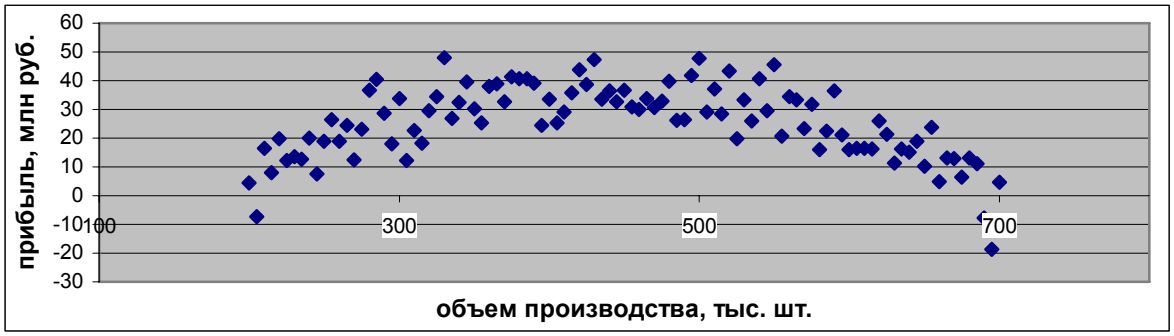
4. Оценка не подлежащих измерению латентных показателей

Ранжирование стран по уровню жизни (индекс человеческого развития и т. д.).

Оценивание эффективности работы предприятий.

5. Оптимальное управление параметрами системы

Поиск оптимального объема производства и цены, максимизирующих прибыль.



Регулирование макроэкономических показателей: структурные преобразования, фискальная и монетарная политика, инвестиционная активность государства.

Типы зависимостей между количественными переменными

1. Зависимость между неслучайными переменными

Нет необходимости в вероятностно-статистической теории.

Выручка фирмы при фиксированной цене: $TR = pq$.

2. Регрессионная зависимость случайного показателя от неслучайных

$y = f(X)+\varepsilon(X),\quad M\varepsilon(X)\equiv 0.$

1) y измеряется с ошибками;

2) y зависит от других, не учтенных в модели факторов.

Зависимость экономического роста от макроэкономических показателей (денежной массы, ставки рефинансирования, налоговых ставок и т. д.).

3. Корреляционно-регрессионная зависимость между случайными переменными *y* и *X*

Большая часть экономических моделей.

4. Зависимость структурного типа (конфлюэнтный анализ)

Известен закон $y = f(X,\theta)$, но y и X наблюдаются с погрешностями. Цель: найти истинные значения переменных по наблюдениям

$$\left.\begin{aligned} \hat{x}_i^{(k)} &= x_i^{(k)} + \varepsilon_{x_i}^{(k)},\ k = 1,...,p, \\ \hat{y}_i &= y_i + \varepsilon_{y_i}, \end{aligned}\right\} \quad i = 1,...,n.$$

Уравнение денежного обмена Ньюкомба-Фишера.

Модель электроэнергетической системы, связанной законами Кирхгофа.

Выбор общего вида функции регрессии

1. **Линейные:** $f(X,\theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)}$.
2. **Степенные:** $f(X,\theta) = \theta_0 \left(x^{(1)}\right)^{\theta_1} \times ... \times \left(x^{(p)}\right)^{\theta_p}$.
3. **Алгебраические полиномы степени** $m \geq 2$**:**

$$f(X,\theta) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)} + \sum_{k_1=1}^{p}\sum_{k_2=1}^{p} \theta_{k_1 k_2} x^{(k_1)} x^{(k_2)} + ... + \sum_{k_1=1}^{p} ... \sum_{k_m=1}^{p} \theta_{k_1 k_2 ... k_m} x^{(k_1)} x^{(k_2)} ... x^{(k_m)}.$$

Рекомендации при выборе функции *f*(*X*)

!!! Главный принцип: от простого к сложному ⇒ начинать с линейной модели !!!

1. Не следует гнаться за чрезмерной сложностью функции

Можно построить полином (n–1)-степени, проходящий через n точек, однако в этом случае мы реагируем на случайные отклонения и при переходе к другой выборке увидим явное рассогласование.

При увеличении числа параметров падает точность оценивания, в частности, увеличивается ширина доверительного интервала.

2. Максимальное использование априорной информации о содержательной (физической, экономической) сущности анализируемой зависимости

1) Будет ли монотонной или имеет экстремум;
2) стремится ли при $x^{(k)} \to \infty$ к асимптотам (## насыщение при потреблении блага);
3) аддитивное или мультипликативное воздействие объясняющих переменных;
4) не должен ли график априори проходить через какие-то точки пространства.

3. Предварительный анализ геометрической структуры данных

1) Построение корреляционных полей в количестве $p(p+1)/2$: $\left(x^{(j)}, x^{(k)}\right)$ и $\left(x^{(j)}, y\right)$;
2) визуальное прослеживание характера вытянутости каждого поля (линейное, нелинейно-монотонное, с наличием одного или нескольких экстремумов и т. д.);
3) изучение поведения условных средних (диапазон значений переменной по оси абсцисс разбивают на интервалы группировки).

!!! Итог: одна или несколько рабочих гипотез, проверка и выбор наиболее адекватной из которых производится статистическими методами !!!

Статистический критерий, основанный на группировке данных

Минус всех статистических критериев проверки гипотезы о виде функции регрессии – они не сообщают, является ли проверяемый вид наилучшим (единственно верным), а только говорят о непротиворечивости вида функции и исходных данных либо отвергают гипотетическую форму зависимости как не соответствующую.

Пусть высказана гипотеза $H_0:\ M(y:X)=f(X,\theta_1,...,\theta_k)$, и получены (например, по МНК) оценки параметров $\hat{\theta}_1,...,\hat{\theta}_k$. Число интервалов группировки $s>k$.

1. Задаем уровень значимости α – вероятность отвергнуть истинную гипотезу.

2. Вычисляем эмпирическое значение критерия:

$$F_{\text{эмп}}=\frac{\frac{1}{s-k}\sum_{j=1}^{s}n_j\left(\bar{y}_j-f\left(\bar{X}_j,\hat{\theta}\right)\right)^2}{\frac{1}{n-s}\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(y_{ji}-\bar{y}_j\right)^2}.$$

Здесь $\bar{X}_j$ – середина j-интервала группировки,

n_j – число наблюдений, в него попавших,

$f\left(\bar{X}_j,\hat{\theta}\right)$ – значение функции регрессии в середине j-интервала,

$\bar{y}_j$ – условное среднее из ординат j-интервала,

y_{ji} – i-значение ординаты из числа попавших в j-интервал.

Числитель дроби – мера рассеивания вокруг регрессионной поверхности.
Знаменатель дроби – мера рассеивания вокруг условных средних (независимая от выбора вида регрессии).

3. Вычисляем критические точки:

$$F_{\text{крит}\,\min}=F\left(1-\frac{\alpha}{2};s-k;n-s\right),\quad F_{\text{крит}\,\max}=F\left(\frac{\alpha}{2};s-k;n-s\right).$$

4. Делаем вывод:

$F_{\text{эмп}}<F_{\text{крит}_{\min}}\Rightarrow$ функция реагирует на случайные отклонения (завышено число k),

$F_{\text{эмп}}\in\left\lfloor F_{\text{крит}\,\min};F_{\text{крит}\,\max}\right\rfloor\Rightarrow$ вид функции не противоречит исходным данным,

$F_{\text{эмп}}>F_{\text{крит}_{\max}}\Rightarrow$ функция недостаточно гибкая, ее нужно усложнить, k слишком мало.

Некоторые общие рекомендации

1. При обнаружении нелинейности в парных статистических связях анализируемых переменных $x^{(j)}$ и y нужно попытаться применить линеаризующие преобразования

$y=\theta_0x^{\theta_1}\ \Leftrightarrow\ \tilde{y}=\tilde{\theta}_0+\theta_0\tilde{x}$, где $\tilde{y}=\ln y,\ \tilde{x}=\ln x,\ \tilde{\theta}_0=\ln\theta_0$.

2. Желательно найти модель, наиболее устойчивую к варьированию состава выборочных данных, на основании которых она оценивается.

Если общий вид зависимости $y=f\left(X,\theta\right)$ «угадан» правильно, то результаты оценивания параметра θ по различным подвыборкам будут мало отличаться друг от друга.

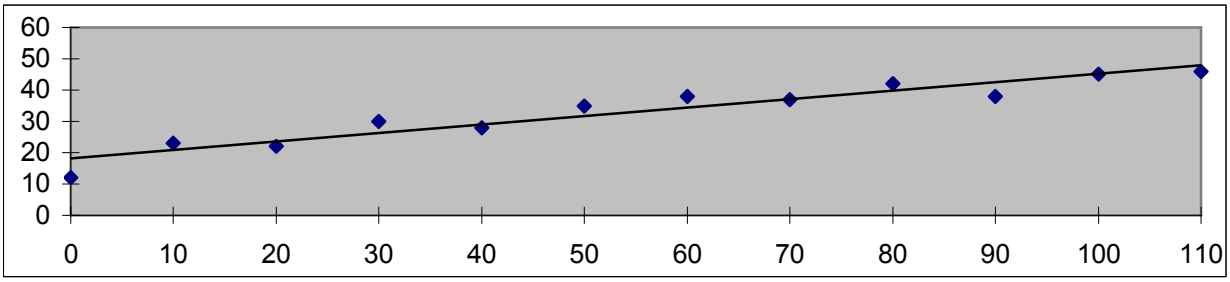
Пример проверки гипотезы о виде функции регрессии

Зависимость урожайности зерновых (y, ц/га) в некотором хозяйстве от объема использования минеральных удобрений (x, кг/га).

Исходные данные:

x_i	0	10	20	30	40	50	60	70	80	90	100	110
y_i	12	23	22	30	28	35	38	37	42	38	45	46

График:



Предполагаемый вид функции регрессии – линейная, $\hat{y} = \theta_0 + \theta_1 x$.

Оценки параметров, полученные с помощью МНК:

$\hat{\theta}_0$ = ИНДЕКС(ЛИНЕЙН(ячейки y_1:y_{12}; ячейки x_1:x_{12}); 2) = 18,12,

$\hat{\theta}_1$ = ИНДЕКС(ЛИНЕЙН(ячейки y_1:y_{12}; ячейки x_1:x_{12}); 1) = 0,271,

$\hat{y} = 18{,}12 + 0{,}271x$.

x_i	0	10	20	30	40	50	60	70	80	90	100	110
y_i	12	23	22	30	28	35	38	37	42	38	45	46
$\bar{x}_j$		10			40			70			100	
$\bar{y}_j$		19			31			39			43	
$f(\bar{x}_j, \hat{\theta})$		20,8			28,9			37,1			45,2	
$(\bar{y}_j - f(\bar{x}_j, \hat{\theta}))^2$		3,32			4,24			3,77			4,75	
$(y_{ji} - \bar{y}_j)^2$	49	16	9	1	9	16	1	4	9	25	4	9

Эмпирическое значение критерия:

$$F_{\text{эмп}} = \frac{\frac{1}{4-2}(3\times 3{,}32 + 3\times 4{,}24 + 3\times 3{,}77 + 3\times 4{,}75)}{\frac{1}{12-4}(49+16+9+1+9+16+1+4+9+25+4+9)} = \frac{24{,}11}{19} = 1{,}269.$$

Критические точки при уровне значимости $\alpha = 0{,}05$:

$F_{\text{крит min}}$ = FРАСПОБР(0,975; 4-2; 12-4) = 0,025,

$F_{\text{крит max}}$ = FРАСПОБР(0,025; 4-2; 12-4) = 6,059,

$F_{\text{эмп}} \in \left[F_{\text{крит}_{\min}}; F_{\text{крит}_{\max}}\right] \Rightarrow$ вид функции регрессии не противоречит исходным данным при уровне значимости $\alpha = 0{,}05$.

Квадратичная функция:

$\hat{y} = 14{,}84 + 0{,}47x - 0{,}002x^2$, $F_{\text{эмп}} = 0{,}119$,

вид функции регрессии не противоречит исходным данным при $\alpha = 0{,}05$.

Корреляционный анализ количественных переменных

1. **Выбрать подходящий измеритель статистической связи** (коэффициент корреляции, корреляционное отношение и т. д.).
2. **Оценить (с помощью точечной и интервальной оценок) его числовое значение по выборочным данным.**
3. **Проверить гипотезу о том, что полученное числовое значение действительно свидетельствует о наличии статистической связи** (корреляционная характеристика значимо отлична от нуля).

Рассматривается статистическая зависимость:

$y(X) = f(X) + \varepsilon(X)$.

$X = \left(x^{(1)}, \ldots, x^{(p)}\right)$ – объясняющие переменные, y – результирующая переменная.

$Dy = Df + D\varepsilon$ – связь безусловных характеристик.

Теснота связи – максимальна, если по заданному значению X можно восстановить $y(X)$ без всякой случайной ошибки

$\varepsilon(X) \equiv 0$, $D\varepsilon = 0$, $Dy = Df$.

Теснота связи – минимальна, если значения X не несут никакой информации об y

$f(X) \equiv c = \text{const}$, $Df = 0$, $Dy = D\varepsilon$.

Коэффициент детерминации

Коэффициент детерминации y по X – универсальный показатель степени тесноты статистической связи.

$$K_d(y,X) = \frac{Df}{Dy} = 1 - \frac{D\varepsilon}{Dy} \in [0;1]$$ –

Коэффициент детерминации отражает долю общей вариации y, объясненную функцией регрессии $f(X)$.

$K_d(y,X) = 0$, если $Df = 0$, $Dy = D\varepsilon$ – полное отсутствие связи.

$K_d(y,X) = 1$, если $D\varepsilon = 0$ – функциональная зависимость $y = f(X)$.

Выборочное значение коэффициента детерминации:

$$\hat{K}_d(y,X) = 1 - \frac{s_\varepsilon^2}{s_y^2}$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2, \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

$$s_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(X_i\right)\right)^2,$$ если $\hat{f}\left(X_i\right)$ – статистически оцененное значение функции регрессии в точке X_i.

$$s_\varepsilon^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}_j\right)^2$$, если есть группировка.

Зависимость спроса на пиво «Brahma» от цены

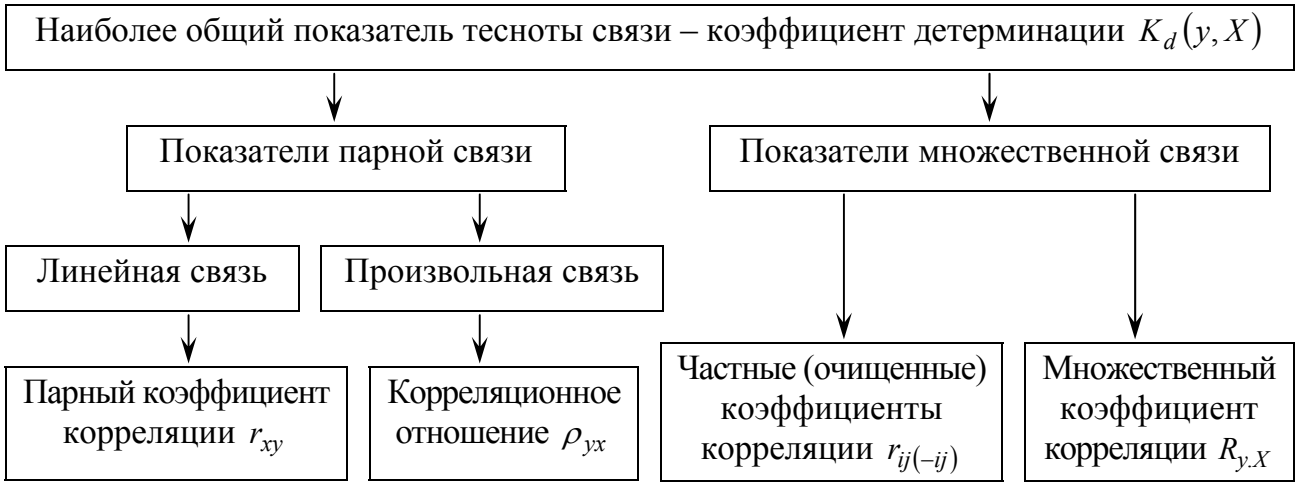
$$\bar{y} = \frac{1}{40}(335 + 299 + \ldots + 1217) = 740{,}65,$$

$$s_y^2 = \frac{1}{40}\left((335 - 740{,}65)^2 + (299 - 740{,}65)^2 + \ldots + (1217 - 740{,}65)^2\right) = 124013,$$

$$s_\varepsilon^2 = \frac{1}{40}\left((335 - 349{,}9)^2 + (299 - 349{,}9)^2 + \ldots + (1217 - 1171{,}8)^2\right) = 34494,$$

$$\hat{K}_d(y,X) = 1 - \frac{s_\varepsilon^2}{s_y^2} = 1 - \frac{34494}{124013} = 0{,}722 = 72{,}2\%$$ – теснота чуть выше среднего уровня.

Основные показатели тесноты статистической связи



Парные корреляционные характеристики

Парные корреляционные характеристики измеряют тесноту связи без учета опосредованного или совместного влияния других показателей, только на основе наблюдения значений двух переменных.

Парный коэффициент корреляции

Парный коэффициент корреляции измеряет тесноту парной линейной связи:

$$r_{xy} = \frac{M\big((x - Mx)(y - My)\big)}{\sigma_x \sigma_y}.$$

Если *x* и *y* распределены по нормальному закону, то функция регрессии имеет линейный вид. Кроме того, выполняются следующие свойства:

1. $r_{xy} \in [-1;1]$.

   Если $r_{xy} > 0$, то положительная (монотонно возрастающая) парная связь;

   если $r_{xy} < 0$, то отрицательная (монотонно убывающая) парная связь.
2. Если *x* и *y* статистически независимы, то $r_{xy} = 0$.
3. $|r_{xy}| = 1$ тогда и только тогда, когда имеется функциональная линейная связь.
4. Коэффициент корреляции – симметричная характеристика: $r_{xy} = r_{yx}$.
5. Если $r_{xy} = 0$, то *x* и *y* статистически независимы.
6. $K_d(y,x) = r_{xy}^2$.

Свойства 1–4 выполняются и **в общем случае парной линейной зависимости**, однако близость коэффициента корреляции нулю не означает статистическую независимость x и y. Возможно, что исследуемые переменные даже связаны функциональным нелинейным соотношением (т. е. $K_d(y,x) = 1$).

Выборочный коэффициент корреляции:

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \text{КОРРЕЛ}(x_1 : x_n;\ y_1 : y_n).$$

Проверка гипотезы об отсутствии парной линейной связи

Вопрос: какую величину выборочного коэффициента корреляции можно считать достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными?
Величина зависит от размерности, поскольку с уменьшением объема выборки ослабевает надежность статистических характеристик.

Гипотеза о статистической независимости *x* и *y* $H_0: r_{xy} = 0$.

Статистика $\hat{t} = \dfrac{\hat{r}_{xy}\sqrt{n-2}}{\sqrt{1-\hat{r}_{xy}^2}} \sim t(n-2)$ – закон распределения Стьюдента.

1. Выбираем уровень значимости α.

2. Сравниваем эмпирическое и критическое значение критерия

$t_{\text{эмп}} = \dfrac{|\hat{r}_{xy}|\sqrt{n-2}}{\sqrt{1-\hat{r}_{xy}^2}}$, $t_{\text{крит}} = t(\alpha; n-2) = \text{СТЬЮДРАСПОБР}(\alpha; n-2)$.

Если $t_{\text{эмп}} > t_{\text{крит}}$, то гипотеза H_0 отвергается при заданном уровне значимости; между переменными имеется связь, близкая к линейной.

Связь между ростом денежной массы и индексом цен $\hat{r}_{xy} = 0{,}8;\ n = 11;\ \alpha = 0{,}05$.

$t_{\text{эмп}} = \dfrac{0{,}8\sqrt{11-2}}{\sqrt{1-0{,}8^2}} = 4$, $t_{\text{крит}} = \text{СТЬЮДРАСПОБР}(0{,}05; 9) = 2{,}262$.

$4 > 2{,}262 \Rightarrow$ гипотеза H_0 отвергается при $\alpha = 0{,}05$; имеется линейная связь.

Если $\alpha = 0{,}001$, то $t_{\text{крит}} = 2{,}262 \to 4{,}781$, гипотеза H_0 принимается; линейной связи нет.
Если $n = 38$, $\alpha = 0{,}001$ то $t_{\text{эмп}} = 4 \to 8$, $t_{\text{крит}} = 4{,}146 \to 3{,}582$, H_0 отвергается; связь есть.

Построение доверительного интервала для истинного значения коэффициента корреляции

Доверительный интервал для теоретического значения коэффициента корреляции r_{xy} – асимметричен и смещен относительно оценки $\hat{r}_{xy}$.

1. Выбираем доверительную вероятность γ.

2. Убираем асимметричность преобразованием Фишера $\hat{z} = \dfrac{1}{2}\ln\dfrac{1+\hat{r}}{1-\hat{r}} = \text{ФИШЕР}(\hat{r})$.

3. Убираем смещение $\tilde{z} = \hat{z} - \dfrac{\hat{r}}{2(n-1)}$.

4. Находим доверительный интервал для переменной *z*

$$z \in [z_1; z_2] = \left[\tilde{z} + \frac{u_{(1-\gamma)/2}}{\sqrt{n-3}};\ \tilde{z} + \frac{u_{(1+\gamma)/2}}{\sqrt{n-3}}\right],$$

$u_{(1-\gamma)/2} = \text{НОРМСТОБР}((1-\gamma)/2)$ и $u_{(1+\gamma)/2} = \text{НОРМСТОБР}((1+\gamma)/2)$ – – квантили нормального стандартного распределения.

5. Возвращаемся в исходные координаты обратным преобразованием Фишера

$r = \dfrac{e^z - e^{-z}}{e^z + e^{-z}} = \text{ФИШЕРОБР}(z)$.

6. Истинное значение *r* с доверительной вероятностью γ лежит в интервале

$r \in [\text{ФИШЕРОБР}(z_1);\ \text{ФИШЕРОБР}(z_2)]$.

По данным $n = 50$ предприятий получен коэффициент корреляции $\hat{r} = -0{,}654$, характеризующий тесноту связи между себестоимостью продукции y и производительностью труда x. Построить интервальную оценку для r, задавшись 95%-ной доверительной вероятностью.

$\hat{z} = \text{ФИШЕР}(\hat{r}) = \text{ФИШЕР}(-0{,}654) = -0{,}7823$ – убираем асимметричность,

$\tilde{z} = \hat{z} - \dfrac{\hat{r}}{2(n-1)} = -0{,}7823 - \dfrac{-0{,}654}{2\times(50-1)} = -0{,}7756$ – убираем смещение,

$z_1 = \tilde{z} + \dfrac{u_{(1-\gamma)/2}}{\sqrt{n-3}} = -0{,}7756 + \dfrac{\text{НОРМСТОБР}(0{,}025)}{\sqrt{50-3}} = -1{,}0615$ – левая граница интервала для z,

$z_2 = \tilde{z} + \dfrac{u_{(1+\gamma)/2}}{\sqrt{n-3}} = -0{,}7756 + \dfrac{\text{НОРМСТОБР}(0{,}975)}{\sqrt{50-3}} = -0{,}4897$ – правая граница интервала для z,

$r_1 = \text{ФИШЕРОБР}(z_1) = \text{ФИШЕРОБР}(-1{,}0615) = -0{,}7862$ – левая граница интервала для r,

$r_2 = \text{ФИШЕРОБР}(z_2) = \text{ФИШЕРОБР}(-0{,}4897) = -0{,}4540$ – правая граница интервала для r,

$r \in [-0{,}7862;\ -0{,}4540]$ с вероятностью 95%.

!!! При уменьшении доверительной вероятности или увеличении объема выборки интервал сужается, а при увеличении доверительной вероятности или сокращении объема выборки – расширяется !!!

$r \in [-0{,}7454;\ -0{,}5289]$ с вероятностью 80%.

$r \in [-0{,}8182;\ -0{,}3798]$ с вероятностью 99%.

$r \in [-0{,}7260;\ -0{,}5656]$ при $n = 200$.

$r \in [-0{,}9027;\ -0{,}0051]$ при $n = 10$.

Влияние ошибок измерения анализируемых переменных на величину коэффициента корреляции

Если переменные x и y измерены с ошибками ε_x и ε_y, то экспериментальные данные (x_i, y_i) – это выборочные значения искаженной случайной величины (x', y')

$x' = x + \varepsilon_x,\ \ y' = y + \varepsilon_y$.

Если ошибки ε_x и ε_y – независимы между собой, не зависят от x и y, распределены по нормальному закону с нулевыми математическими ожиданиями и дисперсиями σ_1^2 и σ_2^2, то искаженное значение коэффициента корреляции r_{xy} можно скорректировать по следующей формуле:

$$r'_{xy} = \frac{r_{xy}}{\sqrt{\left(1 + \dfrac{\sigma_1^2}{\sigma_x^2}\right)\left(1 + \dfrac{\sigma_2^2}{\sigma_y^2}\right)}}.$$

!!! Любые ошибки измерения ослабляют исследуемую корреляционную связь между переменными. Это искажение тем меньше, чем меньше отношение дисперсий ошибок к дисперсиям самих исходных переменных !!!

Исследование парных нелинейных связей
Корреляционное отношение

!!! Если исследуемая зависимость отклоняется от линейного вида, то парный коэффициент корреляции *r* теряет смысл как характеристика степени тесноты связи !!!

Двумерные выборочные данные: $(x_1 y_1), (x_2 y_2), ..., (x_n y_n)$.
По переменной *x* производится разбиение на *s* интервалов группировки.

Корреляционное отношение *y* по *x*:

$$\hat{\rho}^2_{yx} = \frac{s^2_{\bar{y}(x)}}{s^2_y} = \frac{\frac{1}{n}\sum_{j=1}^{s} n_j \left(\bar{y}_j - \bar{y}\right)^2}{\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}\right)^2}$$ – оценка коэффициента детерминации.

$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}$ – среднее для *j*-интервала группировки.

$\bar{y} = \frac{1}{n}\sum_{j=1}^{s} n_j \bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_i$ – общее среднее.

Свойства корреляционного отношения:

1. $\rho_{yx} \in [0; 1]$, хотя иногда для монотонных зависимостей приписывают их знак.
2. $|\rho_{yx}| = 1$ тогда и только тогда, когда имеется функциональная зависимость.
3. $\rho_{xy} = 0$ тогда и только тогда, когда $\bar{y} = \bar{y}_j = \text{const}$, т. е. полное отсутствие корреляционной связи.
4. Корреляционное отношение – асимметричная характеристика: $\rho_{yx} \neq \rho_{xy}$.
   ## $y = x^2$.

x	-1	0	1
y	1	0	1

   $\rho_{yx} = 1$, $\rho_{xy} = 0$
5. $\rho_{yx} \geq |r(x, y)|$, если линейная зависимость, то значения близки.

Из свойства 5 следует, что величину $\hat{\rho}^2_{yx} - \hat{r}^2(x, y)$ можно рассматривать как меру отклонения регрессионной зависимости от линейного вида.

Проверка гипотезы об отсутствии нелинейной связи

Гипотеза о статистической независимости *x* и *y* $H_0: \rho^2_{yx} = 0$.

1. Выбираем уровень значимости α.

2. Сравниваем эмпирическое и критическое значение критерия:

$$F_{\text{эмп}} = \frac{\hat{\rho}^2_{yx}}{1 - \hat{\rho}^2_{yx}} \frac{n - s}{s - 1}, \quad F_{\text{крит}} = \text{FРАСПОБР}(\alpha; s - 1; n - s).$$

Если $F_{\text{эмп}} > F_{\text{крит}}$, то гипотеза об отсутствии связи произвольного вида отвергается с вероятностью ошибки α $\Rightarrow$ связь есть.

Доверительный интервал для истинного значения корреляционного отношения ρ_{yx}

1. Выбираем доверительную вероятность γ.

2. Вычисляем точечную оценку $\hat{\rho}_{yx}^2$.

3. Вычисляем вспомогательное число степеней свободы ν^*.

$$\nu^* = \frac{\left(s-1+n\hat{\rho}_{yx}^2\right)^2}{s-1+2n\hat{\rho}_{yx}^2}.$$

4. Вычисляем критические точки распределения Фишера

$$F_{\frac{1-\gamma}{2}} = F\left(\frac{1-\gamma}{2}; \nu^*; n-s\right), \quad F_{\frac{1+\gamma}{2}} = F\left(\frac{1+\gamma}{2}; \nu^*; n-s\right).$$

5. Вычисляем доверительный интервал для истинного значения ρ_{yx}

$$\rho_{yx}^2 \in \left[\frac{(n-s)\hat{\rho}_{yx}^2}{n\left(1-\hat{\rho}_{yx}^2\right)F_{\frac{1-\gamma}{2}}} - \frac{s-1}{n}; \frac{(n-s)\hat{\rho}_{yx}^2}{n\left(1-\hat{\rho}_{yx}^2\right)F_{\frac{1+\gamma}{2}}} - \frac{s-1}{n}\right].$$

Связь между прибылью и объемом производства: $\hat{\rho}=0{,}6$, $n=50$, $s=10$.

1) Проверить гипотезу об отсутствии нелинейной связи при $\alpha=0{,}05$;
2) построить доверительный интервал для истинного значения ρ_{yx} при $\gamma=0{,}95$.

1) $F_{\text{эмп}} = \frac{0{,}6^2}{1-0{,}6^2}\frac{50-10}{10-1} = 2{,}5$, $F_{\text{крит}} = \text{FРАСПОБР}\left(0{,}05; 10-1; 50-10\right) = 2{,}124$,

$F_{\text{эмп}} > F_{\text{крит}} \Rightarrow H_0$ отвергается, связь между переменными присутствует.

2) $\nu^* = \frac{\left(10-1+50*0{,}6^2\right)}{10-1+2*50*0{,}6^2} = 16{,}2$,

$F_{\frac{1-0{,}95}{2}} = F_{0{,}025} = \text{FРАСПОБР}\left(0{,}025; 16{,}2; 50-10\right) = 2{,}154$,

$F_{\frac{1+0{,}95}{2}} = F_{0{,}975} = \text{FРАСПОБР}\left(0{,}975; 16{,}2; 50-10\right) = 0{,}399$,

$\rho_{\min}^2 = \frac{(50-10)\times 0{,}6^2}{50\times\left(1-0{,}6^2\right)\times 2{,}154} - \frac{10-1}{50} = 0{,}029$, $\rho_{\max}^2 = \frac{\left(50-10\right)*0{,}6^2}{50*\left(1-0{,}6^2\right)*0{,}399} - \frac{10-1}{50} = 0{,}949$,

$\rho_{yx} \in \left[\sqrt{0{,}029}; \sqrt{0{,}949}\right]$,

$\rho_{yx} \in \left[0{,}17; 0{,}97\right]$.

Оценка $\hat{\rho}_{yx}$ – асимметричная. Значения правого или левого конца доверительного интервала могут выходить за пределы 0 или 1 из-за неточности аппроксимации F-распределением. Однако использованный подход является более точным, чем необоснованное применение нормального закона

$$\rho \sim N\left(\hat{\rho}; \frac{1-\hat{\rho}^2}{\sqrt{n}}\right).$$

Исследование линейной зависимости результирующей переменной от нескольких объясняющих переменных

Парные коэффициенты корреляции $r\left(y, x^{(i)}\right)$ не учитывают влияние на эту связь других переменных $x^{(j)}, j \neq i$ $\Rightarrow$ **необходим измеритель связи, очищенный от опосредованного влияния других переменных**, т. е. дающий оценку тесноты связи между y и $x^{(j)}$ при условии, что значения остальных переменных зафиксированы на некотором постоянном уровне.

Частные (очищенные) коэффициенты корреляции

Приведенные формулы справедливы для многомерного нормального закона и приближенно для линейных множественных связей $y = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)} + \varepsilon(X)$. Обозначим для удобства $y \equiv x^{(0)}$.

$r_{ij(-ij)} = \dfrac{-R_{ij}}{\left(R_{ii} R_{jj}\right)^{1/2}}$ – частный коэффициент корреляции между переменными $x^{(i)}$ и $x^{(j)}$ при фиксированнных значениях всех остальных переменных.

R_{kl} – алгебраическое дополнение для r_{kl} в определителе корреляционной матрицы

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \ldots & r_{0p} \\ r_{10} & 1 & r_{12} & \ldots & r_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{p0} & r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}.$$

$R_{kl} = (-1)^{k+l} \det A_{kl}$, матрица A_{kl} получена из R вычеркиванием k-строки и l-столбца.

\#\# $r_{01(2)} = \dfrac{r_{01} - r_{02} r_{12}}{\sqrt{\left(1 - r_{02}^2\right)\left(1 - r_{12}^2\right)}}$ – формула, примененная к трехмерному признаку.

Свойства (проверка гипотез, доверительные интервалы) **частных коэффициентов корреляции** k-порядка (исключение влияния k переменных) такие же, как у парных коэффициентов корреляции с поправкой: **объем выборки уменьшается на k**.

\#\# $n = 37$ – число исследованных предприятий легкой промышленности,
$x^{(0)} \equiv y$ – качество ткани (в баллах),
$x^{(1)}$ – среднемесячное число профилактических наладок автоматической линии,
$x^{(2)}$ – среднемесячное число обрывов нити.
$\hat{r}_{01} = 0{,}105,\ \hat{r}_{02} = 0{,}024,\ \hat{r}_{12} = 0{,}996$.

$$R = \begin{pmatrix} 1 & 0{,}105 & 0{,}024 \\ 0{,}105 & 1 & 0{,}996 \\ 0{,}024 & 0{,}996 & 1 \end{pmatrix}.$$

$$r_{01(2)} = \frac{0{,}105 - 0{,}024 \times 0{,}996}{\sqrt{\left(1 - 0{,}996^2\right)\left(1 - 0{,}024^2\right)}} = 0{,}9079,$$

$$r_{02(1)} = \frac{-\left(0{,}105 \times 0{,}996 - 0{,}024\right)}{\sqrt{\left(1 - 0{,}996^2\right)\left(1 - 0{,}105^2\right)}} = -0{,}9068.$$

Связь имеется, что согласуется с профессиональными представлениями.
Найдем доверительный интервал для истинного значения $r_{01(2)}$.
Исключаем одну переменную $\Rightarrow$ $n = 37 - 1 = 36$.

$$z = \text{ФИШЕР}(0{,}9079) - \frac{0{,}9079}{2(36-1)} = 1{,}5022.$$

$$z \in \left[1{,}5022 - \frac{1{,}96}{\sqrt{36-3}};\ 1{,}5022 + \frac{1{,}96}{\sqrt{36-3}}\right],\quad z \in [1{,}1610;\ 1{,}8434],\quad r \in [0{,}8214;\ 0{,}9511].$$

$n = 20$ – число лет наблюдений за погодой,
$x^{(0)} \equiv y$ – урожайность кормовых трав,
$x^{(1)}$ – весеннее количество осадков,
$x^{(2)}$ – накопленная за весну сумма активных (выше +5,5^{0}C) температур.
$\hat{r}_{01} = 0{,}80,\ \hat{r}_{02} = -0{,}40,\ \hat{r}_{12} = -0{,}56$.

$$R = \begin{pmatrix} 1 & 0{,}80 & -0{,}40 \\ 0{,}80 & 1 & -0{,}56 \\ -0{,}40 & -0{,}56 & 1 \end{pmatrix}.$$

$$r_{01(2)} = \frac{0{,}80 - 0{,}56 \times 0{,}40}{\sqrt{\left(1 - 0{,}56^2\right)\left(1 - 0{,}40^2\right)}} = 0{,}759,$$

$$r_{02(1)} = \frac{-\left(0{,}40 - 0{,}80 \times 0{,}56\right)}{\sqrt{\left(1 - 0{,}56^2\right)\left(1 - 0{,}80^2\right)}} = 0{,}097.$$

$r_{01(2)} \in [0{,}448;\ 0{,}898],\ \ r_{02(1)} \in [-0{,}376;\ 0{,}526]$.

Множественный коэффициент корреляции

Множественный коэффициент корреляции – это коэффициент корреляции между y и линейной функцией регрессии y по x, т. е. между y и линейной комбинацией $x^{(1)}, \ldots, x^{(p)}$, для которой значение коэффициента корреляции максимально.

$$R_{y.X} = r\left(y, f(X)\right) = r\left(y,\ \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}\right).$$

Свойства множественного коэффициента корреляции (МКК):

1. $K_d(y;X) = R^2_{y.X} = 1 - \dfrac{D\varepsilon}{Dy}$.
2. Вычисление МКК по корреляционной матрице: $R^2_{y.X} = 1 - \dfrac{|R|}{|R_{00}|}$.
3. Вычисление МКК по частным КК: $R^2_{y.X} = 1 - \left(1 - r^2_{01}\right)\left(1 - r^2_{02(1)}\right)\left(1 - r^2_{03(12)}\right)\ldots\left(1 - r^2_{0p(123\ldots(p-1))}\right)$.

$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{01}\right)\left(1 - \hat{r}^2_{02(1)}\right) = 1 - \left(1 - 0{,}105^2\right)\left(1 - (-0{,}9068)^2\right) = 0{,}8243,$

$$\hat{R}^2_{y.\left(x^{(1)}x^{(2)}\right)} = 1 - \left(1 - \hat{r}^2_{02}\right)\left(1 - \hat{r}^2_{01(2)}\right) = 1 - \left(1 - 0{,}024^2\right)\left(1 - 0{,}9079^2\right) = 0{,}8243,$$

$$\hat{R}_{y.\left(x^{(1)}x^{(2)}\right)} = \sqrt{0{,}8243} = 0{,}9079.$$

4. МКК мажорирует все парные и частные КК, характеризующие статистическую связь y: $R^2_{y.X} \geq r^2_{0j\left(I_j\right)}$, где I_j – любое подмножество $\{1\ldots p\}$, не содержащее j.
5. Присоединение новой переменной не может уменьшить величины R (вне зависимости от порядка присоединения): $R^2_{y.x^{(1)}} \leq R^2_{y.\left(x^{(1)},x^{(2)}\right)} \leq R^2_{y.\left(x^{(1)},x^{(2)},x^{(3)}\right)} \leq \ldots \leq R^2_{y.\left(x^{(1)},x^{(2)},\ldots x^{(p)}\right)}$.

Проверка гипотезы об отсутствии множественной линейной связи

Гипотеза о статистической независимости y и $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ $\quad H_0 : R_{y.X} = 0$.

1. Выбираем уровень значимости α.

2. Сравниваем эмпирическое и критическое значение критерия:

$$F_{\text{эмп}} = \frac{\hat{R}^2_{yx}}{1 - \hat{R}^2_{yx}} \frac{n - p - 1}{p}, \quad F_{\text{крит}} = \text{FРАСПОБР}(\alpha;\ p;\ n - p - 1).$$

Если $F_{\text{эмп}} > F_{\text{крит}}$, то гипотеза об отсутствии множественной линейной связи отвергается с вероятностью ошибки $\alpha \Rightarrow$ связь есть.

Корреляционный анализ порядковых переменных

$x^{(1)}, x^{(2)},..., x^{(p)}$ – порядковые переменные
(порядковое место в ряду, упорядоченному по свойству $x^{(k)}$).

!!! Если есть неразличимые по некоторому свойству объекты, то всем приписывается ранг, равный среднему арифметическому !!!

Типовые задачи

1. Анализ структуры упорядочений

1) Точки разбросаны равномерно ⇒ нет согласованности между переменными;
2) часть из $(p+1)$ переменных близки друг к другу ⇒ согласованные переменные;
3) некоторые из n объектов близки между собой.

2. Анализ совокупной согласованности переменных

Исследование степени согласованности мнений экспертов.

3. Построение единого группового упорядочения объектов,
ранжировки $x^{(0)}$, наименее удаленной от $x^{(1)}, x^{(2)},..., x^{(p)}$.

Ранговый коэффициент корреляции Спирмена

$$\hat{r}_{kj}^{(S)} = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(x_i^{(k)} - x_i^{(j)}\right)^2 .$$

$\hat{r}_{kj}^{(S)} = 1$, если $x_i^{(k)} = x_i^{(j)},\ i = 1,..., n$,

$\hat{r}_{kj}^{(S)} = -1$, если $x_i^{(k)} = n - x_i^{(j)} + 1,\ i = 1,..., n$.

В остальных случаях $\left|\hat{r}_{kj}^{(S)}\right| < 1$.

Если присутствуют объединенные ранги:

$$\hat{r}_{kj}^{(S)} = \frac{1/6\left(n^3 - n\right) - \sum_{i=1}^{n}\left(x_i^{(k)} - x_i^{(j)}\right)^2 - T^{(k)} - T^{(j)}}{\sqrt{\left(1/6\left(n^3 - n\right) - 2T^{(k)}\right)\left(1/6\left(n^3 - n\right) - 2T^{(j)}\right)}},\quad T^{(k)} = \frac{1}{12}\sum_{t=1}^{m^{(k)}}\left(\left(n_t^{(k)}\right)^3 - n_t^{(k)}\right),$$

$m^{(k)}$ – число групп объединенных рангов, $n_t^{(k)}$ – число элементов в каждой группе.

##1. 10 инвестиционных проектов, проранжированных 2 экспертами:

$$\begin{matrix} x^{(1)} & x^{(2)} \\ 1 & 2 \\ 2 & 3 \\ 3 & 1 \\ 4 & 4 \\ 5 & 6 \\ 6 & 5 \\ 7 & 9 \\ 8 & 7 \\ 9 & 8 \\ 10 & 10 \end{matrix}$$

$$\hat{r}_{12}^{(S)} = 1 - \frac{6}{990}(1+1+4+...+0) = 1 - \frac{84}{990} = 0{,}915.$$

##2. 10 стран, проранжированных по уровню жизни и уровню использования интернета:

$$\begin{matrix} x^{(1)} & x^{(2)} \\ 1 & 1{,}5 \\ 2{,}5 & 1{,}5 \\ 2{,}5 & 4{,}5 \\ 4{,}5 & 4{,}5 \\ 4{,}5 & 4{,}5 \\ 6{,}5 & 4{,}5 \\ 6{,}5 & 8 \\ 8 & 8 \\ 9{,}5 & 8 \\ 9{,}5 & 10 \end{matrix}$$

$$T^{(1)} = \frac{1}{12}\left(2^3 - 2\right)\times 4 = \frac{24}{12} = 2,$$

$$T^{(2)} = \frac{1}{12}\left(\left(2^3 - 2\right) + \left(4^3 - 4\right) + \left(3^3 - 3\right)\right) = 7{,}5,$$

$$\hat{r}_{12}^{(S)} = \frac{\frac{990}{6} - (0{,}25+1+...+0{,}25) - 2 - 7{,}5}{\sqrt{\left(\frac{990}{6} - 2\times 2\right)\left(\frac{990}{6} - 2\times 7{,}5\right)}} = 0{,}911.$$

Недостатки коэффициента корреляции Спирмена:

1. Недостаточная изученность статистических свойств.
2. Невозможность построения частных коэффициентов корреляции.
3. Необходимость полного пересчета при добавлении нового объекта.

Ранговый коэффициент корреляции Кендалла

$\hat{r}_{kj}^{(K)} = 1 - \dfrac{4\nu\left(x^{(k)}, x^{(j)}\right)}{n(n-1)}$, $\nu\left(x^{(k)}, x^{(j)}\right)$ – **минимальное число обменов** соседних элементов переменной $x^{(j)}$ для ее приведения к виду $x^{(k)}$.

$\hat{r}_{kj}^{(K)} = 1$, если $x_i^{(k)} = x_i^{(j)}$, $i = 1,...,n$,

$\hat{r}_{kj}^{(K)} = -1$, если $x_i^{(k)} = n - x_i^{(j)} + 1$, $i = 1,...,n$, $\nu = n(n-1)/2$.

В остальных случаях $\left|\hat{r}_{kj}^{(K)}\right| < 1$.

$\nu\left(x^{(k)}, x^{(j)}\right)$ – **число инверсий** (разложенных в разном порядке пар элементов из $x^{(k)}$ и $x^{(j)}$).

!!! Удобно произвести сортировку данных по переменной $x^{(k)}$!!!

$$\begin{pmatrix}4\\3\\2\\1\end{pmatrix}\to\begin{pmatrix}4\\3\\1\\2\end{pmatrix}\to\begin{pmatrix}4\\1\\3\\2\end{pmatrix}\to\begin{pmatrix}1\\4\\3\\2\end{pmatrix}\to\begin{pmatrix}1\\4\\2\\3\end{pmatrix}\to\begin{pmatrix}1\\2\\4\\3\end{pmatrix}\to\begin{pmatrix}1\\2\\3\\4\end{pmatrix}\quad\Leftrightarrow\quad\begin{pmatrix}1\\2\\3\\4\end{pmatrix}\begin{pmatrix}4\\3\\2\\1\end{pmatrix}\begin{matrix}3\\2\\1\\0\end{matrix}$$

$$\boxed{\nu = 6,\ \hat{r}_{12}^{(K)} = 1 - \frac{4\times 6}{4\times 3} = -1}.$$

Если присутствуют объединенные ранги:

$$\hat{r}_{kj}^{*(K)} = \frac{\hat{r}_{kj}^{(K)} - \dfrac{2\left(u^{(k)} + u^{(j)}\right)}{n(n-1)}}{\sqrt{\left(1 - \dfrac{2u^{(k)}}{n(n-1)}\right)\left(1 - \dfrac{2u^{(j)}}{n(n-1)}\right)}}, \qquad u^{(k)} = \frac{1}{2}\sum_{t=1}^{m^{(k)}} n_t^{(k)}\left(n_t^{(k)} - 1\right).$$

##1

$$\begin{pmatrix}1\\2\\3\\4\\5\\6\\7\\8\\9\\10\end{pmatrix}\begin{pmatrix}2\\3\\1\\4\\6\\5\\9\\7\\8\\10\end{pmatrix}\begin{matrix}1\\1\\0\\0\\1\\0\\2\\0\\0\\0\end{matrix}$$

$\nu\left(x^{(1)}, x^{(2)}\right) = 5,$

$\hat{r}_{12}^{(K)} = 1 - \dfrac{4\times 5}{90} = 0{,}778.$

##2

$$\begin{pmatrix}1\\2{,}5\\2{,}5\\4{,}5\\4{,}5\\6{,}5\\6{,}5\\8\\9{,}5\\9{,}5\end{pmatrix}\begin{pmatrix}1{,}5\\1{,}5\\4{,}5\\4{,}5\\4{,}5\\4{,}5\\8\\8\\8\\10\end{pmatrix}$$

$\hat{r}_{12}^{(K)} = 1,$

$u^{(1)} = \dfrac{1}{2}(4\times 2\times 1) = 4,$

$u^{(2)} = \dfrac{1}{2}(2\times 1 + 4\times 3 + 3\times 2) = 10,$

$\hat{r}_{12}^{(K)} = \dfrac{1 - \dfrac{2\times(4+10)}{90}}{\sqrt{\left(1 - \dfrac{2\times 4}{90}\right)\left(1 - \dfrac{2\times 10}{90}\right)}} = 0{,}818.$

Приближенное соотношение: $\hat{r}_{kj}^{(S)} \approx 1{,}5\,\hat{r}_{kj}^{(K)}$, $n > 10$, $|r| \neq 1$.

Проверка гипотезы о наличии связи

Связь есть, если $\left|\hat{r}^{(S)}\right| > t\left(\dfrac{\alpha}{2}; n-2\right)\sqrt{\dfrac{1 - \left(\hat{r}^{(S)}\right)^2}{n-2}}$ **или** $\left|\hat{r}^{(K)}\right| > u\left(1 - \dfrac{\alpha}{2}\right)\sqrt{\dfrac{2(2n+5)}{9n(n-1)}}$.

$0{,}915 > \text{СТЬЮДРАСПОБР}(0{,}025; 8)\sqrt{\left(1 - 0{,}915^2\right)/8}$, $0{,}915 > 0{,}392$

$0{,}778 > \text{НОРМСТОБР}(0{,}975)\sqrt{\dfrac{2\times(2\times 10+5)}{9\times 10(10-1)}}$, $0{,}778 > 0{,}487$

} Связь есть при $\alpha = 0{,}05$

Доверительный интервал для коэффициента Кендалла

$$r_{kj}^{(K)} \in \left[\hat{r}_{kj}^{(K)} - u\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{2}{n}\left(1 - \left(\hat{r}_{kj}^{(K)}\right)^2\right)};\ \hat{r}_{kj}^{(K)} + u\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{2}{n}\left(1 - \left(\hat{r}_{kj}^{(K)}\right)^2\right)}\right]$$

– интервал приближенный и используется только для больших выборок.

Связь между несколькими порядковыми переменными
Коэффициент конкордации

$$\hat{W}(m)=\frac{12}{m^2\left(n^3-n\right)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}x_i^{(k_j)}-\frac{m(n+1)}{2}\right)^2,$$

n – число объектов, m – число переменных, $k_1,k_2,...,k_m$ – номера переменных.

$$\hat{W}(m)=\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m}x_i^{(k_j)}-\frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2\left(n^3-n\right)-m\sum_{j=1}^{m}T^{(k_j)}}$$ – если присутствуют объединенные ранги.

$\hat{W}(m)\in[0;1]$,
$\hat{W}(m)=1$, когда переменные одинаковы,
$\hat{W}(m)=0$, когда распределение случайно.

Нет отрицательных значений,
т. к. начиная с $m=3$
$x_i^{(2)}=n-x_i^{(1)}+1,\ x_i^{(3)}=n-x_i^{(1)}+1\Rightarrow x_i^{(2)}=x_i^{(3)}$.

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$\sum$	$\sum-16{,}5$
1	2,5	2	5,5	−11
4,5	1	1	6,5	−10
2	2,5	4,5	9	−7,5
3	4,5	4,5	12	−4,5
4,5	4,5	4,5	13,5	−3
7,5	8	4,5	20	3,5
6	9	8	23	6,5
9	6,5	8	23,5	7
7,5	10	8	25,5	9
10	6,5	10	26,5	10

2 2 | $T^{(1)}=\frac{1}{12}\left(2^3-2\right)\times 2=1$,
2 2 2 | $T^{(2)}=\frac{1}{12}\left(2^3-2\right)\times 3=1{,}5$,
4 3 | $T^{(3)}=\frac{1}{12}\left(\left(4^3-4\right)+\left(3^3-3\right)\right)=7$.

$$\hat{W}(3)=\frac{\left((-11)^2+(-10)^2+(-7{,}5)^2+...+10^2\right)}{\frac{1}{12}\times 3^2\times\left(10^3-10\right)-3\times(1+1{,}5+7)}=0{,}828.$$

Проверка гипотезы о наличии связи

$m(n-1)\hat{W}(m)>X^2(\alpha;n-1)\Rightarrow$ связь есть при уровне значимости α.

$X^2_{\text{эмп}}=3\times 9\times 0{,}828=22{,}35$, $X^2_{\text{крит}}=ХИ2ОБР(0{,}05;9)=16{,}92$,
$22{,}35>16{,}92\Rightarrow$ связь есть при $\alpha=0{,}05$.

$m=28,\ n=13,\ \hat{W}(28)=0{,}08$.
$X^2_{\text{эмп}}=28\times 12\times 0{,}08=26{,}88$, $X^2_{\text{крит}}=ХИ2ОБР(0{,}05;12)=21{,}03$,
$26{,}88>21{,}03\Rightarrow$ связь есть при $\alpha=0{,}05$.

Корреляционный анализ категоризованных переменных

$x^{(1)},x^{(2)}$ – категоризованные переменные (описываемые конечным числом состояний).
пол, социальная страта, жилищные условия, фирма-производитель и т. д.

Таблица сопряженности:

	1	2	...	m_2		
1	n_{11}	n_{12}	...	n_{1m_2}	$n_{1\bullet}=\sum n_{1j}$	$w_{1\bullet}=n_{1\bullet}/n$
2	n_{21}	n_{22}	...	n_{2m_2}	$n_{2\bullet}=\sum n_{2j}$	$w_{2\bullet}=n_{2\bullet}/n$
...	...	...	...	...	...	...
m_1	$n_{m_1 1}$	$n_{m_1 2}$	...	$n_{m_1 m_2}$	$n_{m_1\bullet}=\sum n_{m_1 j}$	$w_{m_1\bullet}=n_{m_1\bullet}/n$
	$n_{\bullet 1}=\sum n_{j1}$	$n_{\bullet 2}=\sum n_{j2}$	...	$n_{\bullet m_2}=\sum n_{jm_2}$	n	
	$w_{\bullet 1}=n_{\bullet 1}/n$	$w_{\bullet 2}=n_{\bullet 2}/n$	...	$w_{\bullet m_2}=n_{\bullet m_2}/n$		

Статистическая независимость переменных: $\frac{n_{ij}}{n} \approx \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n}$, $w_{ij} \approx w_{i\bullet}w_{\bullet j}$.

Чем больше отклонение, тем больше показатель связи; $\delta_{ij} = w_{ij} - w_{i\bullet}w_{\bullet j}$.

$x^{(1)}$ – пол (м/ж), $x^{(2)}$ – удовлетворенность зарплатой (да/нет), $n = 100$.

$$\begin{pmatrix} \boxed{\begin{matrix}50 & 0\\ 0 & 50\end{matrix}} & \begin{matrix}50\\ 50\end{matrix} & \begin{matrix}0{,}5\\ 0{,}5\end{matrix} \\ \begin{matrix}50 & 50\end{matrix} & 100 & \\ \begin{matrix}0{,}5 & 0{,}5\end{matrix} & & \end{pmatrix}, \quad \Delta = \begin{pmatrix} 0{,}25 & -0{,}25 \\ 0{,}25 & 0{,}25 \end{pmatrix},$$

максимально тесная связь, знание значения одной переменной позволяет восстановить значение другой.

$$\begin{pmatrix} \boxed{\begin{matrix}25 & 25\\ 25 & 25\end{matrix}} & \begin{matrix}50\\ 50\end{matrix} & \begin{matrix}0{,}5\\ 0{,}5\end{matrix} \\ \begin{matrix}50 & 50\end{matrix} & 100 & \\ \begin{matrix}0{,}5 & 0{,}5\end{matrix} & & \end{pmatrix}, \quad \Delta = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

полное отсутствие связи, знание значения одной переменной не позволяет сделать никаких выводов о значении другой.

$$\begin{pmatrix} \boxed{\begin{matrix}42 & 28\\ 18 & 12\end{matrix}} & \begin{matrix}70\\ 30\end{matrix} & \begin{matrix}0{,}7\\ 0{,}3\end{matrix} \\ \begin{matrix}60 & 40\end{matrix} & 100 & \\ \begin{matrix}0{,}6 & 0{,}4\end{matrix} & & \end{pmatrix}, \quad \Delta = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

полное отсутствие связи, знание значения одной переменной не позволяет сделать никаких выводов о значении другой.

Характеристика тесноты связи – квадратичная сопряженность

$$\hat{X}^2 = n\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\delta_{ij}^2}{w_{i\bullet}w_{\bullet j}} = n\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 1\right), \quad \hat{X}^2 \in [0; +\infty).$$

Проверка гипотезы о наличии связи

$\hat{X}^2 > X^2(\alpha; (m_1 - 1)(m_2 - 1)) \Rightarrow$ связь есть при уровне значимости α.

Коэффициент Крамера

Недостаток квадратичной сопряженности: при $n \to \infty$ $X^2 \to \infty$.

$$\hat{C} = \sqrt{\frac{\hat{X}^2}{n\min\{m_1 - 1; m_2 - 1\}}} \in [0; 1].$$

Зависимость оплаты труда (низкая; средняя; высокая) от образования (неполное среднее; среднее; среднее специальное; высшее; высшее со степенью), $n = 300$.

$$\begin{pmatrix} \boxed{\begin{matrix}14 & 28 & 13 & 18 & 3\\ 11 & 48 & 35 & 51 & 17\\ 1 & 10 & 7 & 19 & 25\end{matrix}} & \begin{matrix}76\\ 162\\ 62\end{matrix} & \begin{matrix}0{,}25\\ 0{,}54\\ 0{,}21\end{matrix} \\ \begin{matrix}26 & 86 & 55 & 88 & 45\end{matrix} & 300 & \\ \begin{matrix}0{,}09 & 0{,}29 & 0{,}18 & 0{,}29 & 0{,}15\end{matrix} & & \end{pmatrix}$$

Равномерное распределение

$$\begin{pmatrix} 7 & 22 & 14 & 22 & 11 \\ 14 & 46 & 30 & 48 & 24 \\ 5 & 18 & 11 & 18 & 9 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} 0{,}025 & 0{,}021 & -0{,}013 & -0{,}014 & -0{,}028 \\ -0{,}010 & 0{,}005 & 0{,}018 & 0{,}012 & -0{,}024 \\ -0{,}015 & -0{,}026 & -0{,}015 & 0{,}003 & 0{,}052 \end{pmatrix} \quad \frac{\delta_{ij}^2}{w_{i\bullet}w_{\bullet j}} = \begin{pmatrix} 0{,}028 & 0{,}006 & 0{,}000 & 0{,}003 & 0{,}021 \\ 0{,}002 & 0{,}000 & 0{,}003 & 0{,}001 & 0{,}007 \\ 0{,}012 & 0{,}011 & 0{,}006 & 0{,}000 & 0{,}088 \end{pmatrix}$$

$\hat{X}^2 = 300 \times 0{,}188 = 56{,}48$, $\hat{C} = \sqrt{\frac{56{,}48}{300\min\{4; 2\}}} = 0{,}307$.

$X^2_{\text{крит}} = \text{ХИ2ОБР}(0{,}001; 4 \times 2) = 26{,}12$, $56{,}48 > 26{,}12 \Rightarrow$ связь есть при $\alpha = 0{,}001$.

Распознавание образов и классификация объектов

Ранее: объекты одинаковы только при точном совпадении по всем признакам.
1. Для количественных признаков ситуация крайне редкая
2. Даже если разбивать на интервалы, классов очень много: $p = 5,\ m_j = 3,\ k = 3^5 = 243$.

Теперь: необходимо разбить все множество n объектов (матрица объект-свойство размерности $n \times p$) на небольшое количество k классов ($k << n$) и отнести каждый объект к одному из них.

Бюджетное обследование семей:

Общественная/национальная принадлежность, образование, квалификация, характер труда, демографический тип, размер и возраст семьи, тип населенного пункта, жилищные условия, размер и структура имущества, уровень доходов	$\Rightarrow$	Уровень и структура потребления, объем и содержание использования свободного времени, интенсивность изменений социального, трудового и демографического статуса

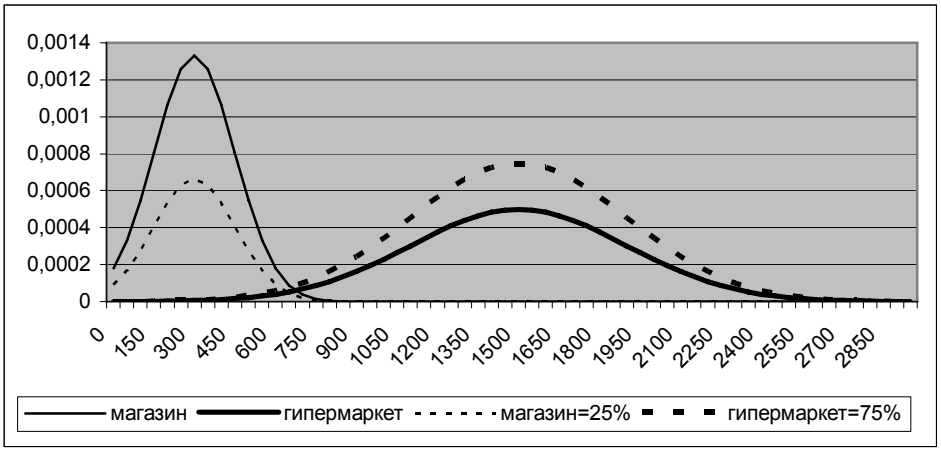
Типологизация математических постановок задач

Определяющий момент: какая имеется априорная информация.
1. Априорная информация о классах.
2. Априорная статистическая информация (обучающие выборки $X_{j1},...,X_{jn_j}$, $j = 1,...,k$, в каждую из которых входит n_j объектов, принадлежащих j-классу).

Априорные сведения о классах	**Предварительная выборочная информация**	
	Нет информации	**Есть обучающие выборки**
Сведения практически отсутствуют	Кластер-анализ	Непараметрические методы дискриминантного анализа
Классы заданы в виде законов распределения с неизвестными параметрами	Расщепление смеси нескольких генеральных совокупностей	Параметрические методы дискриминантного анализа
Классы заданы однозначным описанием законов	Различение статистических гипотез	Обучающие выборки не нужны

Общая идея методов классификации

Вероятность покупки в магазине или гипермаркете в зависимости от суммы.



Решение в пользу класса, для которого больше ордината функции плотности вероятности.

В случае нескольких точек – в пользу класса, для которого больше произведение ординат.

Метод максимального правдоподобия

$$\hat{\Theta}_{ММП} = \arg\max_{\Theta} L(X,\Theta) = \arg\max_{\Theta} \prod_{i=1}^{n} f(X_i,\Theta).$$

Одномерный нормальный закон: $f\left(x,a,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-a)^2/2\sigma^2}$.

$$L(x,a,\sigma) = \frac{1}{(2\pi)^{n/2}\left(\sigma^2\right)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2} \to \max_{a,\sigma^2},$$

$$l(x,a,\sigma) = \ln L(x,a,\sigma) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2 \to \max_{a,\sigma^2},$$

$$\frac{\partial l}{\partial a} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-a) = 0, \qquad \boxed{\hat{a}_{ММП} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}},$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-a)^2 = 0, \qquad \boxed{\hat{\sigma}^2_{ММП} = \frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

Дискриминантный анализ

!!! Суммарные потери от ошибочной классификации минимальны !!!

$$\sum_{\substack{i=1\\ i\neq j}}^{k} \pi_i f_i(X_0)\, c(j:i) \to \min_j$$

π_i – априорная вероятность отнесения объекта к i-классу,
$f_i(X_0)$ – функция плотности вероятности для i-класса,
$c(j:i)$ – потери от неправильной классификации (к j-классу отнесли объект i-класса).

Если случай равных потерь $c(j:i) = c_0 = \text{const}$, задача упрощается: $\pi_j f_j(X_0) \to \max_j$.

Если случай одинаковых вероятностей $\pi_1 = \ldots = \pi_k = 1/k$, задача: $f_j(X_0) \to \max_j$.

Если неизвестны параметры законов распределения, находим их из обучающих выборок, и задача сводится к предыдущей. Если обучающие выборки взяты случайным образом из генеральной совокупности, априорные вероятности $\pi_j = n_j / \sum n_j$.

Данные по 20 фирмам, уклоняющимся и не уклоняющимся от уплаты налогов:

уклоняются

$x_{1i}^{(1)}$	$x_{1i}^{(2)}$
740	680
670	600
560	550
540	520
590	540
590	700
560	540

$A_1 = (\ 607 \quad 590\ )$

$\Sigma_1 = \begin{pmatrix} 4449 & 2929 \\ 2929 & 4543 \end{pmatrix}$

$\boxed{X_0 = (\ 560 \quad 600\ )}$

не уклоняются

$x_{2i}^{(1)}$	$x_{2i}^{(2)}$
750	590
360	600
720	750
540	710
570	700
520	670
590	790
670	700
620	730
690	840
610	680
550	730
590	750

$A_2 = (\ 598 \quad 711\ )$

$\Sigma_2 = \begin{pmatrix} 9351 & 1970 \\ 1970 & 4346 \end{pmatrix}$

Многомерный нормальный закон:

$$f(X,A,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-A)^T\Sigma^{-1}(X-A)}$$

$\pi_1 = 7/20 = 0{,}35$, $\pi_2 = 13/20 = 0{,}65$.

$\pi_1 f_1(X_0) = 0{,}35 \times 0{,}000026 = 0{,}000009$,

$\pi_2 f_2(X_0) = 0{,}65 \times 0{,}000006 = 0{,}000004$,

$\pi_1 f_1(X_0) > \pi_2 f_2(X_0) \Rightarrow$ фирму $X_0 = (560\ 600)$ **следует подозревать** в уклонении от уплаты налогов.

Всего 3 из 20 фирм обучающих выборок (выделены в таблице) при такой проверке будут отнесены к неверному классу.

Расщепление смеси нескольких генеральных совокупностей

Если законы распределения заданы, но их параметры неизвестны, а обучающие выборки отсутствуют, то решается задача расщепления смеси распределений:

Главная задача – определение неизвестных параметров распределений методом максимального правдоподобия или методом моментов:

$$f_1(x,a_1,\sigma_1)=\frac{2}{\sqrt{2\pi}\sigma_1}e^{-(x-a_1)^2/2\sigma_1^2},$$

$$f_2(x,a_2,\sigma_2)=\frac{2}{\sqrt{2\pi}\sigma_2}e^{-(x-a_2)^2/2\sigma_2^2},$$

$$f(x)=\pi_1 f_1(x)+\pi_2 f_2(x).$$

$\prod_{i=1}^{n}(\pi_1 f_1(x_i,a_1,\sigma_1)+\pi_2 f_2(x_i,a_2,\sigma_2))\to \max_{a_1,\sigma_1,a_2,\sigma_2,\pi_1,\pi_2}$, далее задача сводится к предыдущей.

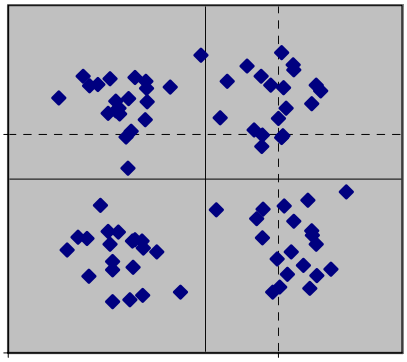
Классификация без обучения. Кластер-анализ

Есть набор наблюдений, который необходимо автоматически классифицировать, однако про классы ничего не известно, иногда даже их число.

Полученные классы, объекты из которых обладают сходными свойствами, называют **кластерами** (cluster) или **таксонами** $\Rightarrow$ **кластер-анализ = численная таксономия**.

2 типа задач:

1. Найти простое разбиение на такие классы, что объекты одного класса достаточно близки друг к другу.
   Решение есть всегда – разбиение на интервалы по каждой переменной.
2. Найти естественное разбиение исходных наблюдений на четко выраженные кластеры – «сгустки», лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удаленные части.
   Решения может не быть – один общий кластер.



Расстояние между отдельными объектами

1. **Евклидово расстояние** $d_E(X_i,X_j)=\sqrt{\sum_{k=1}^{p}\left(x_i^{(k)}-x_j^{(k)}\right)^2}$ **:**
   1) геометрическая близость объектов ($p=1,2,3$);
   2) переменные $x^{(1)},\ldots,x^{(p)}$ однородны по смыслу и одинаково важны;
   3) переменные независимы и обладают примерно одинаковой дисперсией.
2. **Взвешенное евклидово расстояние** $d_{BE}(X_i,X_j)=\sqrt{\sum_{k=1}^{p}w_k\left(x_i^{(k)}-x_j^{(k)}\right)^2}$ **:**
   w_k – вес, пропорциональный степени важности переменной и обратно пропорциональный среднеквадратическому отклонению $w_k\sim 1/\sigma_k$.
3. **Хэммингово расстояние** $d_H(X_i,X_j)=\sum_{k=1}^{p}\left|x_i^{(k)}-x_j^{(k)}\right|$ **:**
   число несовпадений значений, используется для булевых переменных («да/нет»).

Расстояние между классами

S_l – l-класс объектов,
$\overline{X}(l)$ – среднее арифметическое точек, входящих в l-класс (центр тяжести кластера).

1. **Расстояние по принципу ближнего соседа** $\rho_{\min}(S_l, S_m) = \min\limits_{X_i \in S_l,\, X_j \in S_m} d(X_i, X_j)$.
2. **Расстояние по принципу дальнего соседа** $\rho_{\max}(S_l, S_m) = \max\limits_{X_i \in S_l,\, X_j \in S_m} d(X_i, X_j)$.
3. **Расстояние по центрам тяжести** $\rho(S_l, S_m) = d\left(\overline{X}(l), \overline{X}(m)\right)$.
4. **Расстояние по принципу средней связи** $\rho_{cp}(S_l, S_m) = \dfrac{1}{n_l n_m} \sum\limits_{X_i \in S_l} \sum\limits_{X_j \in S_m} d(X_i, X_j)$.
5. **Обобщенное расстояние (Колмогоров)** $\rho_\tau(S_l, S_m) = \left(\dfrac{1}{n_l n_m} \sum\limits_{X_i \in S_l} \sum\limits_{X_j \in S_m} d^\tau(X_i, X_j) \right)^{1/\tau}$.

$\tau \to -\infty \Rightarrow \min$, $\tau \to +\infty \Rightarrow \max$, $\tau = 1 \Rightarrow$ среднее арифметическое,
$\tau = 0 \Rightarrow$ среднее геометрическое, $\tau = -1 \Rightarrow$ среднее гармоническое.

Функционалы качества разбиения на классы

$S = (S_1, ..., S_k)$ – разбиение.

При заданном числе классов *k*:

1. **Взвешенная сумма внутриклассовых дисперсий**

$$Q_1(S) = \frac{1}{n} \sum_{l=1}^{k} \sum_{X_i \in S_l} d^2\left(X_i, \overline{X}(l)\right) = \frac{1}{n}(n_1 D_1 + ... + n_k D_k) \to \min .$$

2. **Взвешенная сумма квадратов попарных внутриклассовых расстояний**

$$Q_2(S) = \frac{1}{n_1^2 + ... + n_k^2} \sum_{l=1}^{k} \sum_{X_i, X_j \in S_l} d^2(X_i, X_j) = \frac{(\Sigma_1 + ... + \Sigma_k)}{n_1^2 + ... + n_k^2} \to \min .$$

$$Q_3(S) = \frac{1}{n} \sum_{l=1}^{k} \frac{1}{n_l} \sum_{X_i, X_j \in S_l} d^2(X_i, X_j) = \frac{1}{n}\left(n_1 \frac{\Sigma_1}{n_1^2} + ... + n_k \frac{\Sigma_n}{n_k^2} \right) = \frac{1}{n}\left(\frac{\Sigma_1}{n_1} + ... + \frac{\Sigma_n}{n_k} \right) \to \min .$$

3. **Обобщенная внутриклассовая дисперсия**
уместна, когда главная задача – сокращение числа переменных p.

$Q_4(S) = \det\left(\sum\limits_{l=1}^{k} n_l \hat{\Sigma}_l \right)$, $Q_5(S) = \prod\limits_{l=1}^{k} \left(\det \hat{\Sigma}_l\right)^{n_l}$, $\hat{\Sigma}_l$ – выборочная ковариационная матрица.

При неизвестном числе классов *k*:

$Q(S)$ – **мера внутриклассового рассеяния** (при $k \uparrow$ $Q(S) \downarrow$).
$Z(S)$ – **мера концентрации** (при $k \uparrow$ $Z(S) \downarrow$).

$Z_\tau(S) = \left(\dfrac{1}{n} \sum\limits_{i=1}^{n} \left(\dfrac{n(X_i)}{n} \right)^\tau \right)^{1/\tau}$, $n(X_i)$ – число точек в кластере, содержащем точку X_i.

$Z_{-1}(S) = \dfrac{1}{k}$, $Z_1(S) = \dfrac{1}{n^2} \sum\limits_{l=1}^{k} n_l^2$, $Z_{-\infty}(S) = \min\limits_{l=1,...,k} \dfrac{n_l}{n}$, $Z_{+\infty}(S) = \max\limits_{l=1,...,k} \dfrac{n_l}{n}$.

При любом τ для $k = n$ $Z_\tau(S) = 1/n$, самая низкая концентрация, все точки раздельно;
$k = 1$ $Z_\tau(S) = 1$, самая высокая концентрация, все точки вместе.

Двухкритериальная задача: $Q(S) \to \min$, $Z(S) \to \max$.

1. **Свертка:** $\alpha Q(S) - \beta Z(S) \to \min$, $\alpha Q(S) + \beta / Z(S) \to \min$, $(Q(S))^\alpha / (Z(S))^\beta \to \min$.
2. **Выделение главного критерия:** $Q(S) \to \min, Z(S) \ge Z_0$ или $Z(S) \to \max, Q(S) \le Q_0$.

Задачи кластер-анализа

(1) – небольшое количество наблюдений, n – несколько десятков (страны, регионы, города, фирмы, технологические процессы и т. д.);

(2) – большое количество наблюдений, n – сотни, тысячи (люди, изделия и т. д.).

(А) – число классов задано;

(Б) – число классов необходимо найти;

(В) – число классов неизвестно, но его не нужно искать – требуется построить иерархическое дерево: оптимальное разбиение для любого количества классов.

Виды процедур кластер-анализа

1. **Иерархические процедуры** – В1, иногда А1, Б1.
2. **Параллельные процедуры** – А1, Б1. Используют практически всю имеющуюся информацию, предполагают распараллеливание.
3. **Последовательные процедуры** – иногда А2, Б2. Используют малую часть информации, а также результат разбиения на предыдущем шаге.

Иерархические процедуры

1. **Агломеративные процедуры** – сначала n кластеров, на каждой итерации объединение ближайших кластеров и пересчет матрицы расстояний.
2. **Дивизимные процедуры** – сначала один кластер, затем его разделение.

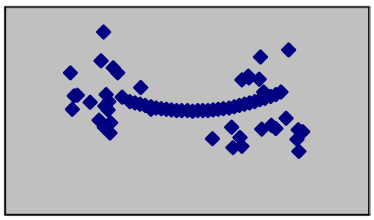
Для задач (А) продолжаем, пока число классов не станет равно k.

Для задач (Б) останавливаемся по одному из критериев качества разбиения на классы.

Для задач (В) осуществляем расчеты до конца.

\+ Самый полный анализ
Наглядность

– Громоздкость (только при малом числе наблюдений)
Далеко не всегда оптимальное решение



Для агломеративной процедуры «ближайшего соседа» возможен «цепочечный эффект», который можно нивелировать ограничением на максимальное расстояние между точками одного кластера.

Для процедур «дальнего соседа», «средней связи», «обобщенного расстояния» – свои недостатки.

Агломеративные пороговые процедуры

Задана монотонная последовательность порогов $c_1,...,c_t$. На j-шаге объединяются все объекты, расстояние между которыми не превышает c_j, $j = 1,...,t$.

Параллельные процедуры

!!! Любые разбиения сравнимы между собой с помощью одного из критериев, однако полный перебор невозможен $\Rightarrow$ сокращение числа вариантов !!!

1. **Процедуры последовательного переноса точек из класса в класс** – каждый объект X_i последовательно перемещается во все кластеры и остается там, где минимально значение критерия качества $Q(S)$. Завершить, когда улучшений больше нет.
2. **Процедуры эталонных множеств** – экспертно строится k эталонных множеств (точек) $E_1^{(0)},...,E_k^{(0)}$. Каждый объект X_i помещается в кластер, для которого является типичным. Из полученных кластеров $S_1^{(1)},...,S_k^{(1)}$ выбираются наиболее типичные эталонные представители $E_1^{(1)},...,E_k^{(1)}$. Завершить, когда $S^{(j)} = S^{(j-1)}$.

!!! Параллельно запускать с разных начальных разбиений и эталонных множеств !!!

Последовательные процедуры

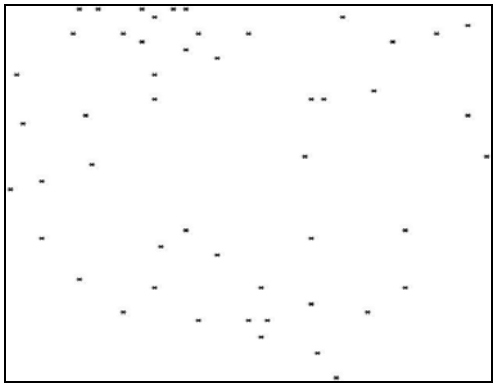
Процедура *k*-средних (упрощенный вариант процедуры эталонных множеств).
При известном числе классов:
$S_i^{(0)} = X_i,\ i = 1,...,k$ – первые k точек определяют k кластеров, $w_i^{(0)} = 1,\ i = 1,...,k$ – веса.
Каждый объект $X_i,\ i = k+1,...,n$ помещается в кластер, к центру тяжести которого он находится ближе всего. Пересчитывается центр тяжести, вес увеличивается на 1.

При неизвестном числе классов:
Задаются константы: φ_0 – **мера грубости**, ψ_0 – **мера точности**, $\varphi_0 < \psi_0$.
Если расстояния между центрами тяжести меньше φ_0, кластеры объединяются;
если расстояние от очередного объекта до ближайшего кластера больше ψ_0, он образует новый кластер.

После прохождения всех n объектов процесс зацикливается на 2,3-й и т. д. круг. Завершение при отсутствии изменений по итогам очередного круга.

Имеется набор из 50 точек на координатной плоскости, который необходимо разбить на некоторое количество кластеров, используя процедуру любого типа.
k – число кластеров, n_i – число точек в i-кластере, D_i – дисперсия i-кластера.

Критерий качества: $\boxed{D + 30k = \frac{1}{n}\sum_{i=1}^{k} n_i D_i + 30k \to \min}$.



3;40	49;30	50;37	63;44	40;45
25;14	52;27	65;14	51;6	35;42
26;19	7;27	59;11	54;3	30;48
7;20	32;10	50;12	20;11	28;48
13;48	50;20	42;8	13;15	25;40
65;21	60;38	43;10	78;30	20;45
2;26	70;45	30;21	75;35	25;47
14;35	4;34	35;18	32;45	23;48
12;45	40;10	75;46	30;43	23;44
15;29	42;14	55;47	25;37	16;48

Иерархическая процедура:
Шаг 1. (52;37) и (50;37) → (51;37), $k = 49$. Шаг 2. (30;48) и (28;48) → (29;48), $k = 48$.

Последовательная процедура $\varphi_0 = 15$, $\varphi_0 = 20$:
Шаг 1. $(3;40) \in S_1,\ w_1 = 1$.
Шаг 2. $d((3;40);(25;14)) > 20 \Rightarrow (25;14) \in S_2,\ w_2 = 1$.
Шаг 3. $d((26;19);(25;14)) < 20 \Rightarrow (26;19) \in S_2,\ w_2 = 2,\ \overline{X}(2) = (25{,}5;\ 16{,}5)$.
Шаг 4. $d((7;20);(25{,}5;\ 16{,}5)) < 20 \Rightarrow (7;20) \in S_2,\ w_2 = 3,\ \overline{X}(2) \approx (19{,}33;\ 17{,}67)$.
Шаг 5. $d((13;48);(3;40)) < 20 \Rightarrow (13;48) \in S_1,\ w_1 = 2,\ \overline{X}(1) = (8;44)$.
Шаг 6. $d((65;21);(8;44)) > 20$, $d((65;21);(19{,}33;\ 17{,}67)) > 20 \Rightarrow (65;21) \in S_3,\ w_3 = 1$.
……

Результат после 1-го круга:
$D_1 = 84{,}19;\ D_2 = 146{,}88,\ D_3 = 185{,}42,\ D_4 = 104{,}69,\ D_5 = 48{,}62$, $\boxed{D + 30k = 264{,}72}$.
$n_1 = 8,\ n_2 = 12,\ n_3 = 11,\ n_4 = 6,\ n_5 = 13$.

Результат после окончания процедуры:
$D_1 = 82{,}05;\ D_2 = 78{,}50,\ D_3 = 106{,}72,\ D_4 = 131{,}60,\ D_5 = 67{,}43$, $\boxed{D + 30k = 239{,}82}$.
$n_1 = 8,\ n_2 = 10,\ n_3 = 8,\ n_4 = 9,\ n_5 = 15$.

Снижение размерности признакового пространства

$x^{(1)},...,x^{(p)} \quad \rightarrow \quad z^{(1)},...,z^{(p')}, \ p' < p$.

Причины снижения размерности:

1. Наглядное представление данных ($p = 1,2,3,...$).
2. Лаконизм моделей $\Rightarrow$ упрощение счета и интерпретации.
3. Сжатие объемов хранимой информации.

!!! Новые переменные $z^{(1)},...,z^{(p)}$ – из числа исходных $x^{(1)},...,x^{(p)}$ или определяются по некоторому правилу по совокупности исходных (линейная комбинация) !!!

Требования к новым показателям:

1. Максимальная информативность.
2. Взаимная некоррелированность.
3. Минимальное искажение геометрической структуры исходных данных.

Ситуации, в которых снижение размерности осуществить легко:

1. **Дублирование информации** (сильно взаимосвязанные показатели).
   ## тест MMPI – сбор «лишних» данных для идентификации невалидности.
2. **Наличие неинформативных переменных** (переменных, практически не меняющихся при переходе от объекта к объекту).
   ## владение основами работы в интернете клиентов интернет-магазина.
3. **Агрегирование** (или простое суммирование) однотипных переменных.
   ## зарплата по основному месту работы, по совместительству, прочие доходы;
   ## индексы цен и объемов.

Критерии:

1. **Внешней информативности** – максимальная точность восстановления y.
2. **Автоинформативности** – максимальная точность восстановления не только результирующего показателя y, но и исходных переменных $x^{(1)},..., x^{(p)}$.

Метод главных компонент

1. **Подготовительный этап:**
   1) **центрирование переменных** – переход к $x_i^{(j)} - \bar{x}^{(j)}$.
   2) **нормирование переменных** – переход к $\left(x_i^{(j)} - \bar{x}^{(j)}\right) \Big/ \sqrt{\sigma_j}$ – особенно важно, если показатели измеряются в различных единицах или имеется существенно отличающийся разброс значений;
   3) **вычисление матрицы ковариаций**

$$\sum = \begin{pmatrix} \hat{\sigma}_{11} & ... & \hat{\sigma}_{1p} \\ & ... & \\ \hat{\sigma}_{p1} & ... & \hat{\sigma}_{pp} \end{pmatrix}, \ \hat{\sigma}_{kj} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(k)} - \bar{x}^{(k)}\right)\left(x_i^{(j)} - \bar{x}^{(j)}\right) = \text{КОВАР}\left(x_1^{(k)},...,x_n^{(k)};x_1^{(j)},...,x_n^{(j)}\right).$$

2. **Решение характеристического уравнения** $\left|\sum - \lambda E\right| = 0$**:**
   1) **нахождение собственных чисел** $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$
   для симметричной положительно определенной матрицы $\sum$ все корни – положительные вещественные числа;
   2) **нахождение собственного вектора** $l^{(k)}$ для каждого корня характеристического уравнения λ_k с помощью системы линейных однородных уравнений $\left(\sum - \lambda_k E\right) l^{(k)} = 0, \ \left\| l^{(k)} \right\| = 1$.

3. **Переход к новым переменным** $Z = XL$

$z^{(k)} = X l^{(k)}$, $k = 1,..., p'$ – новые переменные, **«главные компоненты»**.

!!! МГК максимально сохраняет вариабельность исходных переменных !!!

$I_{p'} = \dfrac{\lambda_1 + ... + \lambda_{p'}}{\lambda_1 + ... + \lambda_p}$ – доля дисперсии, вносимой первыми p' главными компонентами.

!!! По величине $I_{p'}$ определяем количество p' главных компонент !!!

Зависимость спроса на пиво «Brahma» ($\tilde{x}^{(2)}$,бут.) от цены ($\tilde{x}^{(1)}$, руб.)

исходные переменные

$\tilde{x}^{(1)}$	$\tilde{x}^{(2)}$
28,6	349,9
25,2	623,8
22,9	817,1
18,6	1171,8

$\bar{\tilde{x}}^{(1)} = 23,8$, $\bar{\tilde{x}}^{(2)} = 740,7$.

$\tilde{\sigma}_1 = 3,6$, $\tilde{\sigma}_2 = 299,2$.

центрир. и нормир. переменные

$x^{(1)}$	$x^{(2)}$
1,314	−1,306
0,378	−0,391
−0,254	0,256
−1,437	1,441

$$\sum = \begin{pmatrix} 1 & -0,99997 \\ -0,99997 & 1 \end{pmatrix}$$

$$\begin{vmatrix} 1-\lambda & -0,99997 \\ -0,99997 & 1-\lambda \end{vmatrix} = 0, \quad (1-\lambda)^2 - 0,99997^2 = 0, \quad \lambda_1 = 1,99997, \ \lambda_2 = 0,00003.$$

$$p' = 1 \ \Rightarrow \ I_1 = \frac{1,99997}{1,99997 + 0,00003} = 0,99999 = 99,999\%.$$

Найдем $z^{(1)}$ по $\lambda_1 = 1,99997$:

$$\begin{cases} \begin{pmatrix} 1-1,99997 & -0,99997 \\ -0,99997 & 1-1,99997 \end{pmatrix} \begin{pmatrix} l_1^{(1)} \\ l_2^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \left\| l^{(1)} \right\| = 1 \end{cases} \Rightarrow l_2^{(1)} = -l_1^{(1)} \Rightarrow 2\left(l_1^{(1)}\right)^2 = 1 \Rightarrow \boxed{l^{(1)} = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix} \approx \begin{pmatrix} 0,707 \\ -0,707 \end{pmatrix}}$$

$z^{(1)} = 0,707x^{(1)} - 0,707x^{(2)}$.

Аналогично найдем $z^{(2)}$ по $\lambda_2 = 0,00003$:

$$\begin{cases} \begin{pmatrix} 1-0,00003 & -0,99997 \\ -0,99997 & 1-0,00003 \end{pmatrix} \begin{pmatrix} l_1^{(2)} \\ l_2^{(2)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \left\| l^{(1)} \right\| = 1 \end{cases} \Rightarrow l_2^{(2)} = l_1^{(2)} \Rightarrow 2\left(l_1^{(2)}\right)^2 = 1 \Rightarrow \boxed{l^{(2)} = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} \approx \begin{pmatrix} 0,707 \\ 0,707 \end{pmatrix}}$$

$z^{(2)} = 0,707x^{(1)} + 0,707x^{(2)}$.

Матрица перехода: $L = \begin{pmatrix} 0,707 & 0,707 \\ -0,707 & 0,707 \end{pmatrix}$.

Автопрогноз: $Z = XL \quad \Rightarrow \quad X = ZL^{-1} = ZL^T$.

$X \approx z^{(1)}\left(l^{(1)}\right)^T = z^{(1)}\left(l_1^{(1)} \, l_2^{(1)}\right) = \left(0,707z^{(1)}; \ -0,707z^{(1)}\right)$, $p' = 1$ – приближенно.

$X = \left(z^{(1)}; z^{(2)}\right)\begin{pmatrix} \left(l^{(1)}\right)^T \\ \left(l^{(2)}\right)^T \end{pmatrix} = \left(0,707z^{(1)} + 0,707z^{(2)}; \quad 0,707z^{(1)} - 0,707z^{(2)}\right)$, $p' = 2$ – точно.

исходные		центрир. и нормир.		новая		восстановленные			восст. исход.	
$\tilde{x}^{(1)}$	$\tilde{x}^{(2)}$	$x^{(1)}$	$x^{(2)}$	$z^{(1)}$		$\hat{x}^{(1)}$	$\hat{x}^{(2)}$		$\hat{\tilde{x}}^{(1)}$	$\hat{\tilde{x}}^{(2)}$
28,6	349,9	1,314	−1,306	−1,852		1,310	−1,310		28,59	348,8
25,2	623,8	0,378	−0,391	0,544		0,384	−0,384		25,22	625,6
22,9	817,1	−0,254	0,256	−0,361		−0,255	0,255		22,90	816,9
18,6	1171,8	−1,437	1,441	−2,035		−1,439	1,439		18,59	1171,3

Зависимость месячных объемов продаж компакт-дисков ($\widetilde{x}^{(3)}$, тыс. шт.) от цены ($\widetilde{x}^{(1)}$, руб.) и рекламных вложений ($\widetilde{x}^{(2)}$, тыс. руб.)

исходные переменные

$\widetilde{x}^{(1)}$	$\widetilde{x}^{(2)}$	$\widetilde{x}^{(3)}$
80	25	15
100	40	18
90	0	10
75	10	17
120	60	14
85	80	26
100	10	11
70	0	25
120	15	6
75	5	12

$\bar{x}^{(1)} = 91{,}5,$
$\bar{x}^{(2)} = 24{,}5,$
$\bar{x}^{(3)} = 15{,}4.$

$\sigma_1 = 17{,}2,$
$\sigma_2 = 25{,}8,$
$\sigma_3 = 6{,}0.$

центрир. и нормир. переменные

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
−0,669	0,019	−0,066
0,495	0,600	0,431
−0,087	−0,948	−0,895
−0,960	−0,561	0,265
1,659	1,374	−0,232
−0,378	2,149	1,756
0,495	−0,561	−0,729
−1,251	−0,948	1,590
1,659	−0,368	−1,557
−0,960	−0,755	−0,563

$$\sum = \begin{pmatrix} 1 & 0{,}440 & -0{,}536 \\ 0{,}440 & 1 & 0{,}431 \\ -0{,}536 & 0{,}431 & 1 \end{pmatrix}.$$

Решение характеристического уравнения:

$$\begin{vmatrix} 1-\lambda & 0{,}440 & -0{,}536 \\ 0{,}440 & 1-\lambda & 0{,}431 \\ -0{,}536 & 0{,}431 & 1-\lambda \end{vmatrix} = 0, \quad \lambda_1 = 1{,}551, \quad \lambda_2 = 1{,}325, \quad \lambda_3 = 0{,}124.$$

Матрицы прямого и обратного перехода:

$$L = \begin{pmatrix} -0{,}605 & 0{,}543 & 0{,}583 \\ 0{,}225 & 0{,}818 & -0{,}529 \\ 0{,}764 & 0{,}189 & 0{,}617 \end{pmatrix}, \quad L^{-1} = L^T = \begin{pmatrix} -0{,}605 & 0{,}225 & 0{,}764 \\ 0{,}543 & 0{,}818 & 0{,}189 \\ 0{,}583 & -0{,}529 & 0{,}617 \end{pmatrix}.$$

$$p' = 1 \quad \Rightarrow \quad I_1 = \frac{1{,}551}{1{,}551 + 1{,}325 + 0{,}124} = 0{,}517 = 51{,}7\%.$$

$$\boxed{p' = 2 \quad \Rightarrow \quad I_2 = \frac{1{,}551 + 1{,}325}{1{,}551 + 1{,}325 + 0{,}124} = 0{,}959 = 95{,}9\%.}$$

Переход к новым переменным:

$$z^{(1)} = -0{,}605x^{(1)} + 0{,}225x^{(2)} + 0{,}764x^{(3)}, \quad z^{(2)} = 0{,}543x^{(1)} + 0{,}818x^{(2)} + 0{,}189x^{(3)}.$$

Восстановление исходных переменных:

$$\hat{x}^{(1)} \approx -0{,}605z^{(1)} + 0{,}543z^{(2)}, \quad \hat{x}^{(2)} \approx 0{,}225z^{(1)} + 0{,}818z^{(2)}, \quad \hat{x}^{(3)} \approx 0{,}764z^{(1)} + 0{,}189z^{(2)}.$$

новые переменные

$z^{(1)}$	$z^{(2)}$
0,358	−0,360
0,165	0,841
−0,844	−0,992
0,657	−0,931
−0,871	1,981
2,054	1,884
−0,982	−0,328
1,758	−1,155
−2,275	0,306
-0,019	−1,246

$\Rightarrow$

восстановлен. переменные

$\hat{x}^{(1)}$	$\hat{x}^{(2)}$	$\hat{x}^{(3)}$
−0,412	−0,214	0,206
0,357	0,725	0,285
−0,029	−1,002	−0,833
−0,903	−0,614	0,326
1,603	1,425	−0,291
−0,218	2,003	1,925
0,416	−0,490	−0,813
−1,691	−0,549	1,125
1,542	−0,262	−1,681
−0,665	−1,023	−0,250

$\Rightarrow$

восстановлен. исходные

$\hat{\tilde{x}}^{(1)}$	$\hat{\tilde{x}}^{(2)}$	$\hat{\tilde{x}}^{(3)}$
84	19	17
98	43	17
91	−1	10
76	9	17
119	61	14
88	76	27
99	12	10
62	10	22
118	18	5
80	−2	14

Матрица нагрузок главных компонент на исходные признаки

$A \in R^{p \times p'}$, $A = L\Lambda^{1/2}$, $\Lambda^{1/2} = diag\left\{\sqrt{\lambda_j}\right\}$.

!!! Элементы матрицы нагрузок имеют смысл, когда анализируемые переменные $x^{(1)},...,x^{(p)}$ не только центрированы, но и нормированы !!!

Свойства:

1. a_{ij} – парные коэффициенты корреляции между $x^{(i)}$ и $z^{(j)}$.

2. $\boxed{\sum_{i=1}^{p} a_{ij}^2 = a_{1j}^2 + a_{2j}^2 + ... + a_{pj}^2 = \lambda_j}$ Сумма квадратов элементов j-столбца матрицы нагрузок равна дисперсии j-главной компоненты λ_j.

$$A^T A = \left(L\Lambda^{1/2}\right)^T \left(L\Lambda^{1/2}\right) = \Lambda^{1/2} L^T L \Lambda^{1/2} = \Lambda^{1/2} L^{-1} L \Lambda^{1/2} = \Lambda^{1/2}\Lambda^{1/2} = \Lambda.$$

3. $\sum_{j=1}^{p} a_{ij}^2 = a_{i1}^2 + a_{i2}^2 + ... + a_{ip}^2 = 1$ Сумма квадратов элементов i-строки матрицы нагрузок равна дисперсии i-исходной переменной, т. е. единице.

Интерпретация главных компонент:

$A = \begin{pmatrix} 0{,}753^* & 0{,}625^* & 0{,}205 \\ 0{,}280 & 0{,}942^* & 0{,}186 \\ 0{,}952^* & 0{,}218 & 0{,}217 \end{pmatrix} \Rightarrow$ $z^{(1)}$ – тесно связана с $x^{(1)}$ и $x^{(3)}$, $z^{(2)}$ – тесно связана с $x^{(1)}$ и $x^{(2)}$.

Наблюдения – помесячные данные

$x^{(1)}$ – число торговых точек, где распространяется продукция, шт.
$x^{(2)}$ – расходы на рекламу, руб.
$x^{(3)}$ – доля новинок в ассортименте, %
$x^{(4)}$ – средний месячный доход на душу населения, руб.
$x^{(5)}$ – количество праздников, шт.

главная компонента $z^{(j)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$
собственное число λ_j	3,09	1,36	0,43	0,10	0,02
суммарный вклад, %	61,8	89,0	97,6	99,6	100,0

$$I_1 = \frac{3{,}09}{3{,}09+1{,}36+0{,}43+0{,}10+0{,}02} = 0{,}618 = 61{,}8\%, \quad I_2 = \frac{3{,}09+1{,}36}{3{,}09+1{,}36+0{,}43+0{,}10+0{,}02} = 0{,}89 = 89\%.$$

$\boxed{p' = 2}$ – на 2 главные компоненты $z^{(1)}$ и $z^{(2)}$ приходится 89% суммарной вариации.

Для интерпретации главных компонент построим матрицу нагрузок A, обратим внимание на максимальные по абсолютной величине элементы каждого столбца, особенно $\left|a_{ij}\right| > 0{,}6$ – наиболее тесная связь между $x^{(i)}$ и $z^{(j)}$ (отметим их звездочками):

$$A = \begin{matrix} z^{(1)} & z^{(2)} \end{matrix} \begin{pmatrix} 0{,}95^* & -0{,}19 \\ 0{,}97^* & -0{,}17 \\ 0{,}94^* & -0{,}28 \\ 0{,}24 & 0{,}88^* \\ 0{,}56 & 0{,}67^* \end{pmatrix} \begin{matrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \\ x^{(5)} \end{matrix} \Rightarrow$$ $z^{(1)}$ – тесно связана с $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$, $z^{(2)}$ – тесно связана с $x^{(4)}$ и $x^{(5)}$.

$z^{(1)}$ – можно интерпретировать в модели зависимости объема продаж y от $x^{(1)},...,x^{(p)}$ как переменную, объясняющую влияние эффекта замещения (наша фирма получает большую или меньшую долю на рынке).

$z^{(2)}$ – можно интерпретировать как переменную, объясняющую влияние эффекта дохода (изменяется суммарный спрос на рынке).

Типовые задания для контрольных работ

Задание 1

Компания-производитель сотовых телефонов в течение года наращивала выпуск продукции x и ежемесячно отслеживала издержки y. Данные представлены в таблице:

Выпуск, тыс. шт.	400	425	450	475	500	525	550	575	600	625	650	675
Издержки, млн руб.	612	671	747	792	845	981	1220	1341	1539	1975	2126	2308

1. Статистически проверить гипотезы о линейной и квадратичной зависимости издержек от выпуска с помощью критерия, основанного на группировке данных (данные разбить на 4 интервала в зависимости от объема выпуска). Уровень значимости принять равным 0,05.
2. Найти оптимальный выпуск, если отпускная цена модели составляет 2900 руб.

Задание 2

В условиях предыдущей задачи

1. Подсчитать выбочный коэффициент корреляции. Проверить гипотезу о наличии линейной связи при уровне значимости 0,05. Найти интервальную оценку коэффициента корреляции с доверительной вероятностью 95%.
2. Определить значение корреляционного отношения. Проверить гипотезу о наличии связи при уровне значимости 0,05. Группировать данные по 4 интервалам в зависимости от объема выпуска.

Задание 3

10 стран ранжированы по 2 показателям: темпам роста ВВП и уровню иностранных инвестиций. В таблице приведены места каждой из стран:

Темпы роста ВВП	1	2–3	2–3	4	5	6–7	6–7	8–10	8–10	8–10
Иностр. инвестиции	2–3	1	4–6	2–3	7–10	7–10	4–6	7–10	7–10	4–6

Подсчитать ранговые коэффициенты корреляции Спирмена и Кендалла, проверить гипотезу о наличии связи при уровне значимости 0,05.

Задание 4

В целях определения наиболее предпочтительных направлений отдыха в зависимости от сезона был проведен опрос 200 респондентов. Его результаты сведены в таблице:

	Турция	**Египет**	**Таиланд**	**ОАЭ**
лето	44	12	7	4
весна, осень	15	24	12	11
зима	3	35	19	14

Вычислить значение характеристики X^2 квадратичной сопряженности и определить наличие связи при уровне значимости 0,001. Подсчитать коэффициент Крамера.

Задание 5

В инвестиционный портфель входят акции 40 компаний. Известны доходность и риск каждого вида ценных бумаг, выраженные в виде математического ожидания и среднеквадратического отклонения прибыли:
(7;2), (9;7), (8;6), (11;3), (15;8), (12;4), (16;11), (4;2), (6;3), (10;1), (15;11), (7;3), (9;3), (12;2), (12;7), (5;6), (18;12), (10;3), (14;5), (16;3), (6;1), (8;2), (13;3), (14;3), (15;7), (8;7), (5;2), (6;4), (9;5), (10;2), (10;7), (16;10), (12;5), (17;7), (9;4), (17;11), (14;6), (16;5), (9;9), (7;8).
Разбить набор акций на кластеры, используя процедуру кластер-анализа любого типа.

Критерий качества: $\frac{1}{n}\sum_{i=1}^{k} n_i D_i + k \to \min$, где k – количество кластеров, n_i – число точек, попавших в i-кластер, D_i – дисперсия в i-кластере.

Вопросы к экзамену

1. Основные этапы статистического анализа.
2. Исходные данные и возможные результаты исследования.
3. Виды функций регрессии и рекомендации при их выборе.
4. Проверка гипотезы о виде функции регрессии.
5. Корреляционный анализ количественных переменных. Коэффициент детерминации.
6. Измеритель линейной связи – парный коэффициент корреляции.
7. Измеритель нелинейной связи – корреляционное отношение.
8. Частные коэффициенты корреляции.
9. Множественный коэфициент корреляции.
10. Корреляционный анализ порядковых переменных. Ранговые коэффициенты корреляции.
11. Связь между несколькими порядковыми переменными. Коэффициент конкордации.
12. Корреляционный анализ категоризованных переменных.
13. Распознавание образов. Типологизация постановок задач.
14. Дискриминантный анализ.
15. Расщепление смеси нескольких генеральных совокупностей.
16. Кластер-анализ. Расстояние между объектами. Расстояние между кластерами.
17. Функционалы качества разбиения на классы.
18. Типы процедур кластер-анализа.
19. Снижение размерности признакового пространства.
20. Метод главных компонент.

Содержание

Александр Юрьевич Филатов,
e-mail: fial@irlan.ru, ICQ 10793366

Другие авторские разработки в области математической экономики выложены на сайтах
http://polnolunie.baikal.ru/me/mat_ec.htm
http://matec.isu.ru
http://fial_.livejournal.com

КОНСПЕКТ ЛЕКЦИЙ ПО МНОГОМЕРНЫМ СТАТИСТИЧЕСКИМ МЕТОДАМ
учебное пособие

Редактор Г. А. Никифорова

Темплан 2007. Поз. 67.

Подписано в печать 1.08.2007. Формат 60×84 1/16.
Бумага писчая белая. Печать трафаретная. Уч.-изд.л. 2,4.
Тираж 125 экз.

Редакционно-издательский отдел
Иркутского государственного университета
664003, Иркутск, бул. Гагарина, 36