ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ – ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ФАКУЛЬТЕТ ЭКОНОМИКИ
КАФЕДРА СТАТИСТИКИ

О. И. ОБРАЗЦОВА

Статистические методы в изучении предпринимательства

Летняя школа «Предпринимательство в России: теория и практика, методология исследований»

Звенигород, «Солнечная Поляна», 5 – 10 июля 2010

Статистика – совокупность методов, которые дают нам возможность принимать решение в условиях неопределённости.

Абрам Вальд

Основные проблемы анализа предпринимательства в странах постсоветского пространства :

- Ограниченность данных государственной статистики предпринимательства
- Низкий уровень готовности экспертов и предпринимателей к сотрудничеству
- Пропуски в данных
- Широкий круг непараметрических данных
- Неоднородность данных альтернативной статистики, малые выборки

Источники данных о предпринимательстве

**Статистика – позитивная наука
Она… занимается тем, "что есть",
а не тем, что "должно быть"
Кейнс**

- **Статистика – язык экономической науки, инструмент функциональной диагностики живого экономического организма**
- **Статистика принципиально нейтральна, независима от какой-либо этической позиции или нормативных суждений.**
- **Конечная цель - формулирование и проверка гипотезы, которая дает правильные и значимые (т.е. не являющиеся трюизмами) предсказания относительно пока ещё не наблюдавшихся или в принципе не поддающихся наблюдению явлений**

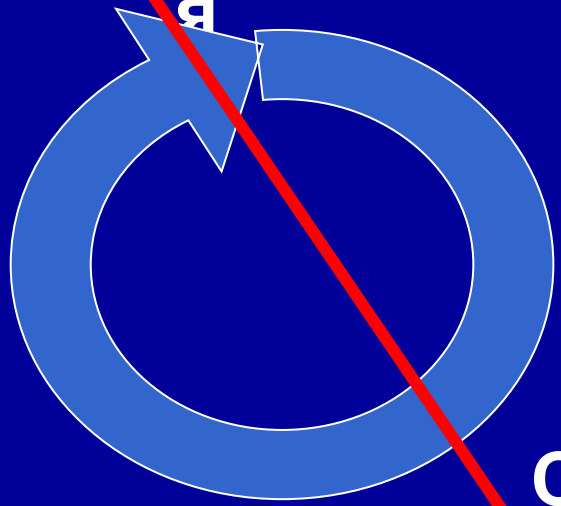
Колесо знаний Уоллеса

Знания, не рождённые опытом, матерью всякой достоверности, бесплодны и полны ошибок.

Леонардо да Винчи

Эволюция теорий

$Pi \rightarrow TT \rightarrow EE$



Теория

Гипотеза

Статистическая конкретизация

Наблюдение

Эмпирический анализ

Для стран постсоветского пространства западные теории предпринимательства не подтверждаются эмпирическими данными

?

Этапы статистического исследования

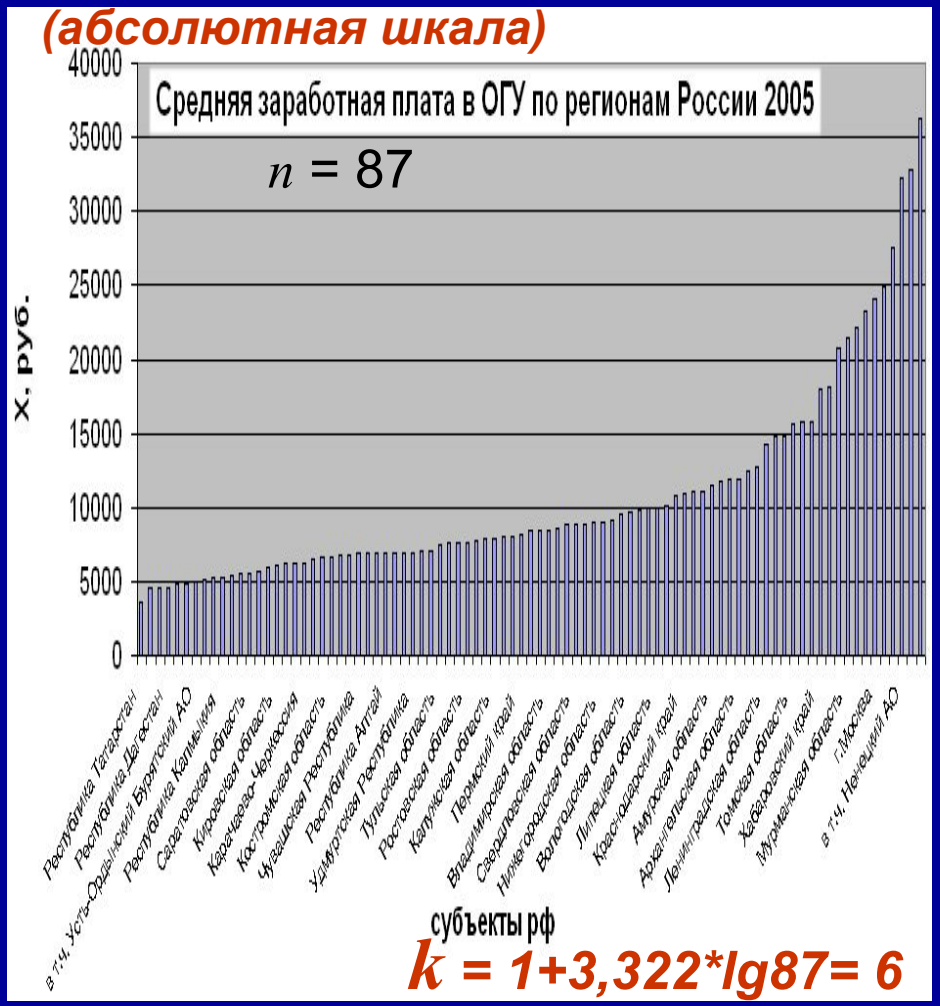

Классификация объектов и многокритериальный выбор
Измерение эффекта воздействия одного или нескольких факторов на результат
Pi
Прогноз развития ситуации
Экспертные оценки
Статистические данные
Ситуации

Ограничения в оценке зависимостей

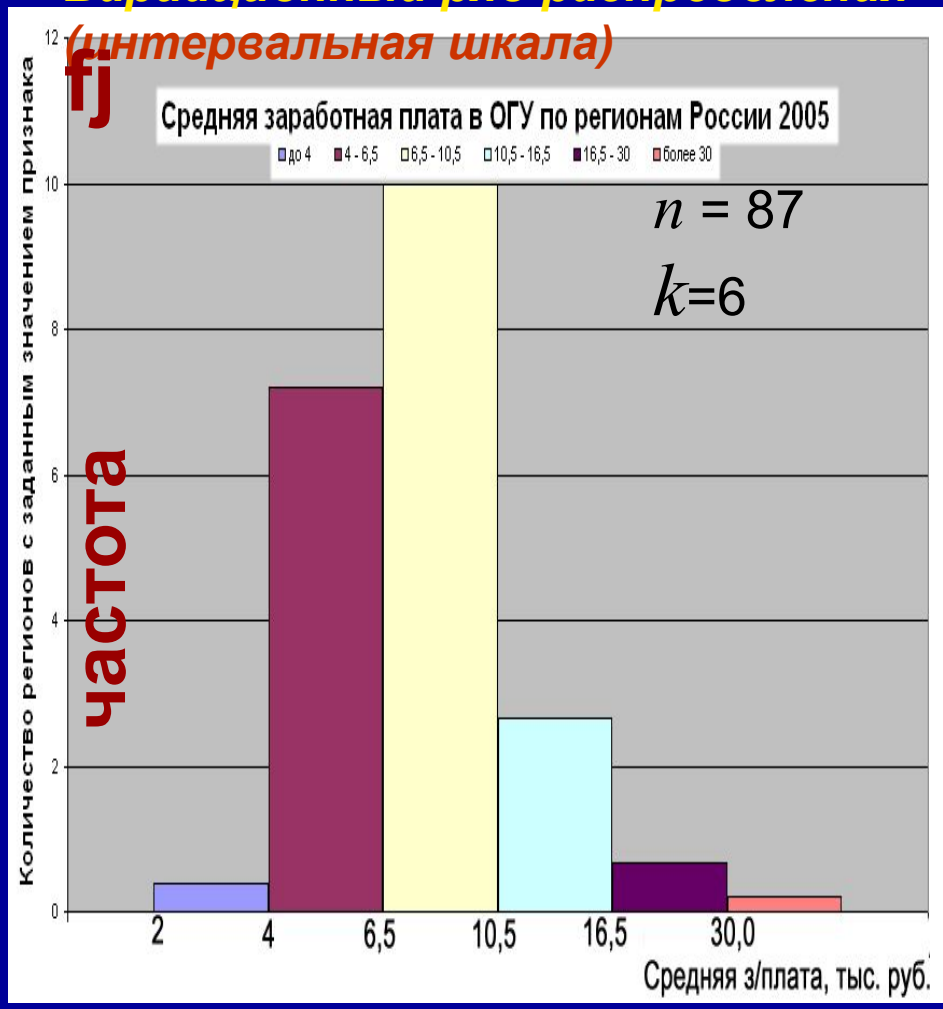
Шкала измерения влияющих переменных	Шкала измерения зависимых переменных	Применяемые методы
Интервальная или отношений	Интервальная или отношений	Регрессионный и корреляционный анализ
Времени	Интервальная или отношений	Анализ временных рядов
Номинальная или порядковая	Интервальная или отношений	Дисперсионный анализ
Смешанная ситуация	Интервальная или отношений	Ковариационный и регрессионный анализ
Номинальная или порядковая	Номинальная или порядковая	Анализ ранговых корреляций и таблиц сопряженности
Номинальная или порядковая	Интервальная или отношений	Кластерный анализ, дискриминантный анализ, таксономия

Результаты группировки повышают информационную силу статистических данных

Упорядоченный ряд наблюдения (абсолютная шкала)

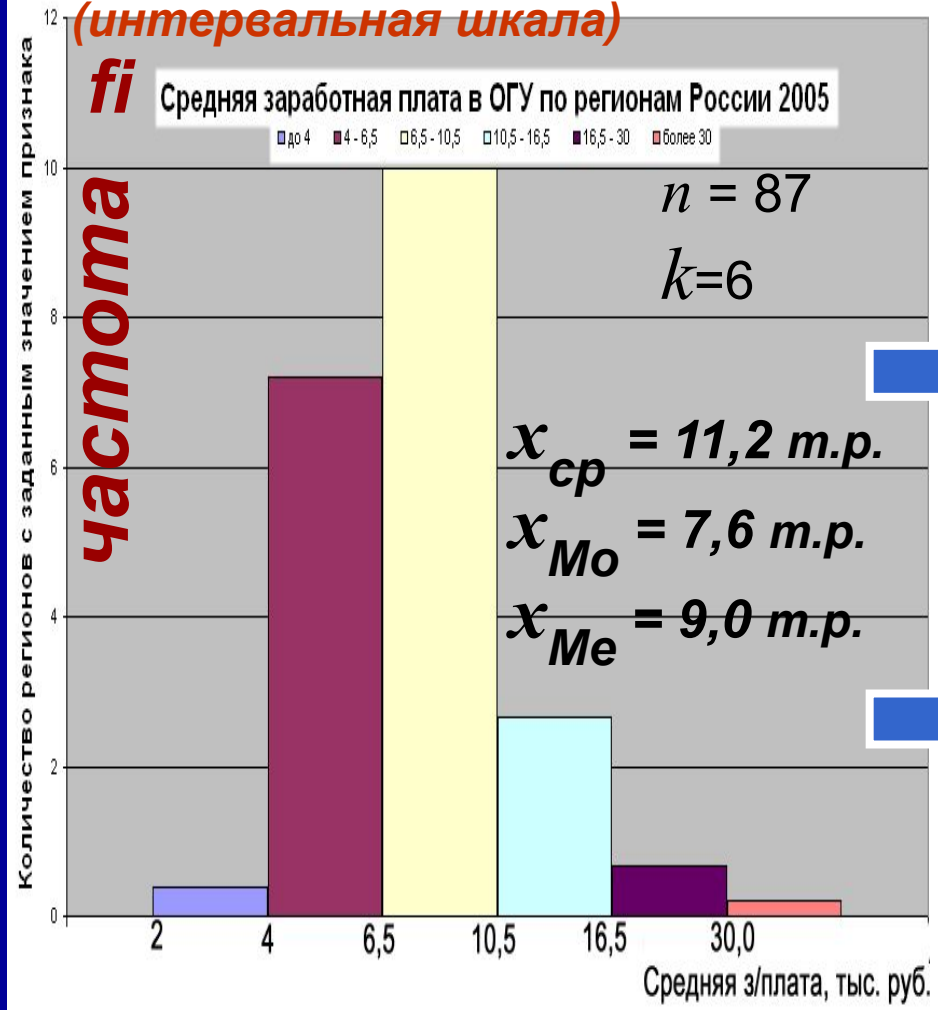


Вариационный ряд распределения (интервальная шкала)



Группировка позволяет оценить структурные закономерности в форме обобщающих показателей распределения

Вариационный ряд распределения (интервальная шкала)



$n = 87$

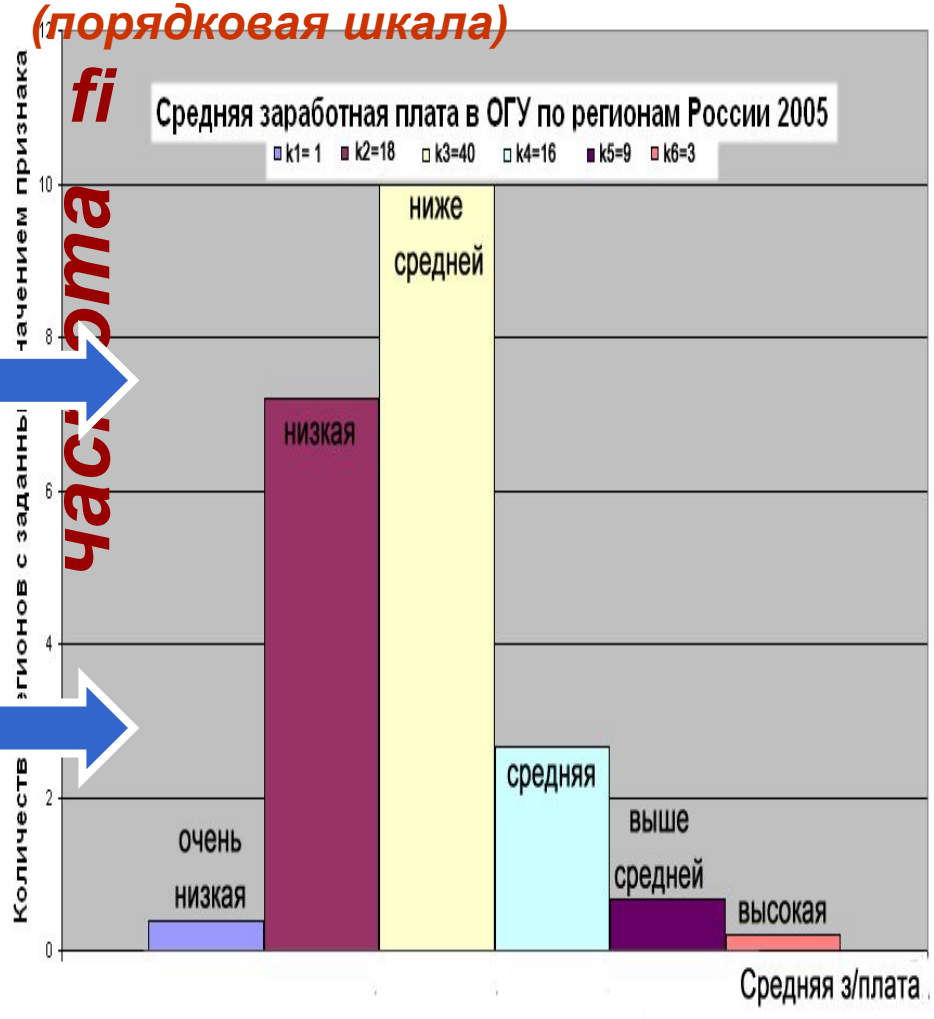
$k=6$

$x_{ср} = 11{,}2$ т.р.

$x_{Mo} = 7{,}6$ т.р.

$x_{Me} = 9{,}0$ т.р.

Атрибутивный ряд распределения (порядковая шкала)



Измерение эффекта воздействия одного или нескольких факторов на результат

Как влияет *образование* на предпринимательскую активность индивида?

→ Дисперсионный анализ

Как влияет *возраст* на предпринимательскую активность индивида?

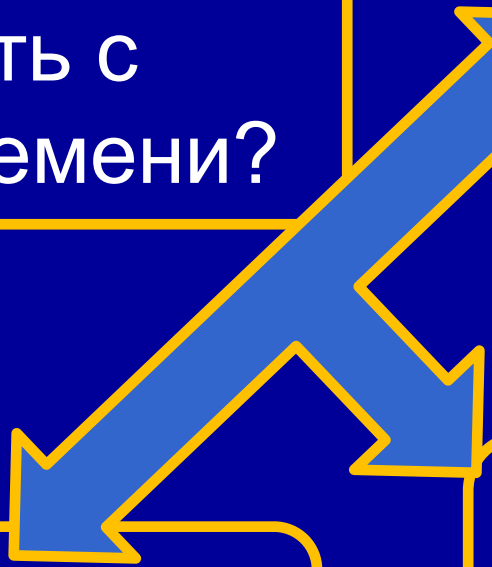
→ Регрессионный анализ

Как влияют признаки индивида (возраст, доход, ресурсы, экономическая нагрузка в семье, продолжительность безработицы, профессиональный стаж и т.п.) на предпринимательскую активность

→ Факторный и компонентный анализ

Прогноз развития ситуации

Как изменится предпринимательская активность с течением времени?



Экстраполяция динамического ряда

Автокорреляционная функция

Корреляция рядов динамики (с лагом или без)

Классификация объектов и многокритериальный выбор

Ab haedis segregare oves.

Евангелие от Матфея 25, 32

Какие группы стран GEM можно выделить по уровню предпринимательской активности?

Кластерный анализ (таксономия)

Чем определяются различия между группами, если они значимы?

Дискриминантный анализ

Распознавание образов

К какой из выделенных групп следует отнести страну, не участвующую в GEM?

Кластерный анализ - разбиение множества объектов на однородные группы на основе изучения вариации классифицирующей переменной

- Количество кластеров может быть известно или неизвестно заранее
- Отсутствуют обучающие выборки
- Разрыв пространства существования фактора может возникать также и при определенной комбинации независимых переменных
- Агломеративная процедура (сначала объединяют самые близкие объекты, затем к ним присоединяют более дальние)

Алгоритм кластеризации

КА

Анализ и интерпретация результатов

Формирование кластеров

Выбор и расчет меры сходства между наблюдениями

Выбор классифицирующих признаков

Формирование выборки для анализа

Мера сходства - расстояние $d_{ij}(O_i, O_j)$ между объектами Oi и Oj: чем меньше расстояние, тем более похожими считаются наблюдения

- *Евклидово расстояние* $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$
- *Хеммингово расстояние (*городских кварталов, *Манхэттэнское,* путь таксиста)

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

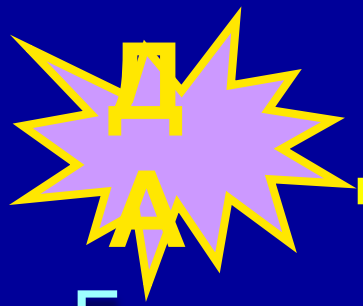
Частные случаи расстояния Махаланобиса (симметричного, монотонного в признаном пространстве, минимального к самому себе)

$$d_{ij} = \sqrt{(X_i - X_j)^T \Lambda^T \sum^{-1} (X_i - X_j)}$$

Расстояние между кластерами

- «ближайшего соседа» (одиночная связь)
- «дальнего соседа» (полная связь)
- между «центроидами»
- по «средней связи»

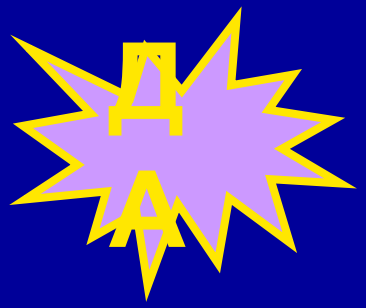
Разные процедуры КА для одних и тех же данных могут давать различное разбиение на кластеры. Только метод k-средних имеет строгое статистическое обоснование!



ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Группа экспертов исследует возможность переговоров с террористами, захватившими заложников. Их интересуют те особенности ситуации, при которых возможно безопасное освобождение заложников, даже если требования террористов не выполнены. ... Дискриминантный анализ может обеспечить получение необходимых данных.

Клекка У. Р. Дискриминантный анализ



Классы – значения классифицирующей переменной (шкала не сильнее порядковой)

Ситуация 1:

Классифицирующая переменная зависит от дискриминантных → аналог многофакторного регрессионного анализа для отклика в атрибутивной шкале

Ситуация 2:

Дискриминантные переменные зависят от классифицирующей → аналог обобщенного многомерного дисперсионного анализа

Назначение: изучение различий между двумя и более классами объектов по комбинации описывающих переменных → получение по обучающей выборке правил (цензов, формул) для определения групповой принадлежности объекта

Задача 1:

Интерпретация → определение количества и значимости дискриминантных функций и границ их значений для объяснения различий между классами

Задача 2:

Классификация → определение класса, к которому принадлежит новый объект

Предпосылки:

- Наблюдения принадлежат к двум или более классам
- В каждом классе есть как минимум два объекта
- Количество дискриминантных переменных не более чем (N – 2)
- Дискриминантные переменные измерены в шкале интервалов или шкале отношений
- Дискриминантные переменные линейно независимы
- Дискриминантные переменные, измеренные в абсолютной шкале, распределены по многомерному нормальному закону распределения (каждая распределена нормально при фиксированных прочих переменных)
- Ковариационные матрицы классов можно считать равными между собой

Совет: будьте внимательны при формировании обучающих выборок!

Типичная ошибка: эти выборки не содержат переменных, по которым фактически происходит классификация объектов → классификация невозможна.

Проверка: объедините классы обучающей выборки в один и попробуйте разделить их с помощью кластерного анализа. Если исходной классификации не получилось, то подбор переменных выполнен неправильно

Алгоритм анализа для k классов, объекты характеризуются *p* переменными (обучающие выборки $X^{(j)}$, объемом n_j)

1. Рассчитываются средние значения по каждой переменной для каждого класса

$$\bar{x}_j^{(l)} = \frac{1}{n_l}\sum_{i=1}^{nl} x_{ij}^{(l)}$$

$$t_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(X_{jki} - \bar{\bar{X}}_j\right)\left(X_{jki} - \bar{\bar{X}}_l\right)$$

2. Определяются оценки ковариационных матриц для каждого класса S_i

$$s_{mj}^{(l)}(x) = \frac{1}{n_l}\sum_{i=1}^{nl}\left(x_{ij} - \bar{x}_j\right)\left(x_{im} - \bar{x}_m\right)$$

$$W_{jl} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(X_{jki} - \bar{\bar{X}}_{jk}\right)\left(X_{jki} - \bar{\bar{X}}_{lk}\right)$$

3. Рассчитывается несмещенная оценка объединенной ковариационной матрицы

$$S = \frac{1}{\sum_{i=1}^{k} n_i - k - 1}\sum_{i=1}^{k} n_i S_i$$

ДА

Алгоритм анализа для k классов, объекты характеризуются *p* переменными (обучающие выборки $X^{(j)}$, объемом n_j)

4. Рассчитываются векторы оценок коэффициентов дискриминантной функции (независимость исходных переменных!)

$$A^{(l)} = S^{-1}\overline{X}^{(l)}$$

5. Оцениваются дискриминантные константы (собственные значения) и каноническая корреляция

$$\lambda_l = \frac{1}{2}\overline{X}^{(l)}\left(S^{-1}\overline{X}^{(l)}\right)$$

$$r_i^* = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

6. Определяется принадлежность новых объектов к классу на основе дискриминантной функции Z_j

ДА

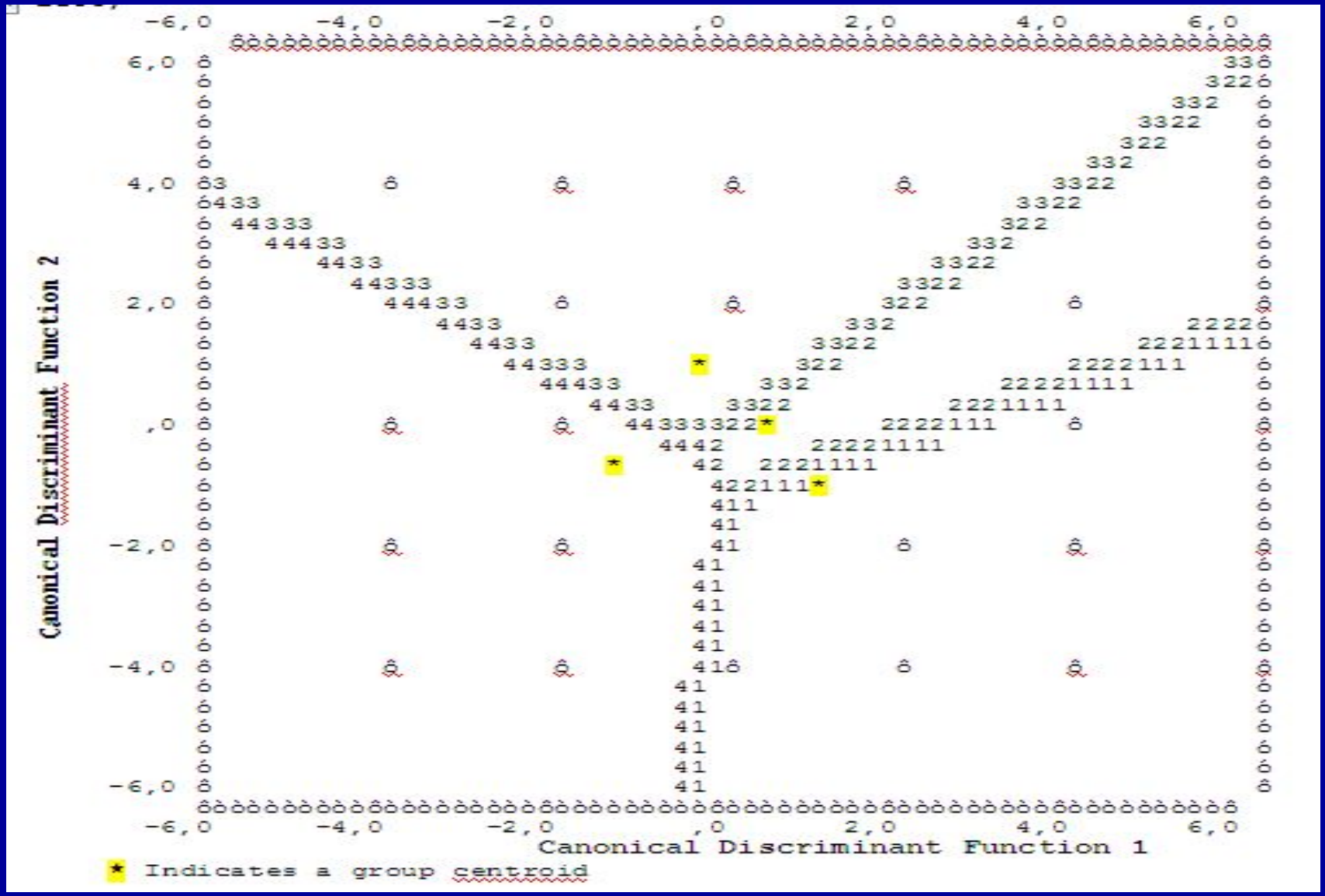
Если необходима классификация…

Канонические дискриминантные функции (независимы, центроиды различаются):

$$f_{ki} = u_o + \sum_{j=1}^{p} u_j X_{jki} ,$$

- f_{ki} — значение канонической дискриминантной функции для i-го объекта в *k*-м классе
- u_j — нестандартизованные коэффициенты дискриминантной функции
- X_{jki} — значение дискриминантной переменной X_j для *i*-го объекта в классе *k*.
- $k^x i$ *минимально* (лямбда Уилкса) и не превышает (k – 1) или дискриминантных переменных j_{max} (в зависимости от того, какая из величин меньше)

Территориальная карта



ДА

Интерпретация (дискриминация): переход к стандартизованным к-там и стандартизованным функциям

$$c_i = u_i \sqrt{\frac{W_{ii}}{n - K}}$$

- *n* — общее число наблюдений,
- *K* — число классов (групп),
- W_{ii} — диагональный элемент матрицы оценки рассеивания

Вклад стандартизованного коэффициента в дискриминантную функцию пропорционален его величине

Распознавание образов: классификация без интерпретации

- Основа классификации – каноническая дискриминантная функция
- Критерий отнесения наблюдения к определённому классу – квадрат расстояния Махаланобиса (до центроида)
- $D^2(X,G_k)=(n-K)\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}(X_i-\bar{X}_{il})(X_j-\bar{X}_{jk})$ или $D^2_{корр}=\frac{n-p-3}{n-2}D^2-\left[\sum_{i=1}^{K}\frac{p}{n_i}\right]$
- Для групп с разной наполненностью:

$$D^{*2}(X,G_k)=D^2(X,G_k)-2\ln P_{apriori,k}$$

Что ещё почитать?

- Миллс Ф. Статистические методы – М.:Госстатиздат. 1958
- Плюта В. Сравнительный многомерный анализ в эконометрическом моделировании. - М.: ФиС. 1989
- Прикладная статистика: классификация и снижение размерности: справ. изд. / Айвазян С.А., Бухштабер В. М., Енюков И.С., Мешалкин Л.Д. - М.: ФиС. 1989
- Сошникова Л.А. и соавт. Многомерный статистический анализ в экономике. – М.: ЮНИТИ-ДАНА, 1999
- Факторный, дискриминантный и кластерный анализ: Пер с англ. - М.: ФиС. 1989
- Хейс Д. Причинный анализ в статистических исследованиях – М.: Финансы и статистика, 1981
- Статистический анализ в экономике / Под ред. Громыко Г.Л.. – М.: Изд-во МГУ, 1992
- Общая теория статистики: Учебник / Боярский А.Я., Ясин Е.Г. – М.: МГУ, 1977

Благодарю за внимание!