Анализ данных в политических науках. Введение

Н. В. Артамонов, И. А. Истомин

МГИМО МИД России

28 ноября 2017 г.

О лекторах

Артамонов Никита Вячеславович: зав.каф. МЭИТ, к.ф.–м.н., доц.
email: artamonov@inno.mgimo.ru

Истомин Игорь Александрович: ст.преп. каф. ПАМП, к.полит.н.
email: i.istomin@inno.mgimo.ru

Web: meit.mgimo.ru

Содержание

Почему работа со статистикой?

Современная тенденция: широкое использование анализа статистических данных:

- в экономическое науке
- в бизнесе/финансах
- в социальных науках
- в политических науках

Некоторые причины:

- Доступность статистических данных, в т.ч. открытых
- Доступность вычислительных мощностей (PC, серверы/облака, суперкомьютеры)
- Современный вероятносто-статистический аппарат

Что это даёт?

Какие преимущества даёт анализ статистических данных:

1. Получение новых знаний
2. Проверка теорий на “согласованность с реальностью”
3. Поддержка принятия решений (Data Driven Decision)
4. Количественные оценки зависимостей между факторами
5. Тестирование гипотез
6. Практические выводы (прогнозирование, оценка последствий от принятия решений etc)
7. Дополнительное конкурентное преимущество
8. Многое другое

Недостаток информации

Проблемы прикладного анализа данных

Существенный недостаток информации (непреодолимый). Невозможно учесть всё! Например, многие факторы ненаблюдаемы.

Недостаток информации

Проблемы прикладного анализа данных

Существенный недостаток информации (непреодолимый). Невозможно учесть всё! Например, многие факторы ненаблюдаемы.

Пример (Функция спроса)

Линейная функция спроса $Q = \beta_0 + \beta_1 P$ не подходит для описания реальной зависимости, т.к. спрос не определяется только ценой!

Недостаток информации. Что делать?

Вопрос: Как работать при неопределённости/недостатке информации?

Недостаток информации. Что делать?

Вопрос: Как работать при неопределённости/недостатке информации?

Ответ: Стандартно используется аппарат теории вероятностей и мат. статистики.

Пример

Влияние ненаблюдаемых/неучтённых факторов моделируем случайной величиной.

Недостаток информации. Что делать?

Вопрос: Как работать при неопределённости/недостатке информации?

Ответ: Стандартно используется аппарат теории вероятностей и мат. статистики.

Пример

Влияние ненаблюдаемых/неучтённых факторов моделируем случайной величиной.

Пример (Линейная функция спроса)

Более реалистична модель спроса $Q = \beta_0 + \beta_1 P + u$, *где* u *– ненаблюдаемая случайная величина, описывающая влияние невключённых (например, ненаблюдаемых) факторов.*

Какие бывают переменные

Два вида переменных:

- Количественные
- Качественные или факторные:
  - упорядоченные
  - неупорядоченные

Пример

GDP, метод огранки, гендерный фактор, цвет

Типы статистических данных

Важно!

Статистические методы и модели зависят от вида статистических данных.

Типы статистических данных

Важно!

Статистические методы и модели зависят от вида статистических данных.

Четыре основных вида статистических данных:

1. перекрёстные данные,
2. временные ряды,
3. панельные данные,
4. объединённые выборки.

Перекрёстные данные

Перекрёстные данные (cross-sectional data) – данные, полученные в один период времени (или близкие периоды времени, так что фактором время можно пренебречь).

Также называются кросс-секционными данными или кросс-секцией.

Временные ряды

Временные ряды или исторические данные (time series) – данные об одном факторе, полученные (как правило) через равные промежутки времени

Панельные данные

Панельные данные (panel data) – по каждому элементу кросс-секции есть исторические данные.

Объединённые выборки

Объединённые выборки (pooled data) – объединение перекрёстных выборок, полученных в близкие промежутки времени.

Статистические модели

Основные модели/методы:

1. Базовые распределения (нормального etc)
2. Кластерный анализ и задача классификации
3. Корреляционный анализ
4. Метод главных компонент
5. Дисперсионный (ANOVA) и Факторный анализ
6. Регрессия (линейная, бинарного выбора etc)
7. Анализ временных рядов
8. ML/AI (нейронные сети, глубокое обучение etc)

Статистические методы

Статистические модели:

- параметрические
- непараметрические

Статистические методы:

- параметрические
- робастные
- непараметрические

Выбор зависит от целей и предположений

Корреляция и причинность

Важно!

Формальные выводы о зависимости, основанные на анализе данных, не означают причинно-следственную связь.

«Корреляция»[a] (correlation) не означает причинность (causality)!

---

[a] «Корреляция» – количественное выражение статистической связи между факторами.

Корреляция и причинность

Важно!

Формальные выводы о зависимости, основанные на анализе данных, не означают причинно-следственную связь.

«Корреляция»[a] (correlation) не означает причинность (causality)!

---

[a] «Корреляция» – количественное выражение статистической связи между факторами.

Пример

1. *Предновогодние распродажи и наступление Нового года.*
2. *Кривая Филипса*

Корреляция и причинность

Важно!

Для обоснования причинно-следственной связи необходимы дополнительные аргументы (a priori).
Прикладное исследование не сводится к формальной работе с данными!

Корреляция и причинность

Важно!

Для обоснования причинно-следственной связи необходимы дополнительные аргументы (a priori).
Прикладное исследование не сводится к формальной работе с данными!

Пример

Две теории бизнес-цикла: RBC[a] vs New-Keynesian.
Колебания GDP vs колебания M2: кто на кого влияет.

[a] Finn E. Kydland & Edward C. Prescott, Nobel Prize 2004

Ложная корреляция

Что это такое?

Ложная корреляция (spurious correlation): формально данные показывают, что зависимость есть, но логически зависимости не должно быть.

Ложная корреляция

Что это такое?

Ложная корреляция (spurious correlation): формально данные показывают, что зависимость есть, но логически зависимости не должно быть.

Пример (http://www.tylervigen.com)

Расходы на науку в US и число самоубийст (удушение) $(r = 0.99)$

Ложная корреляция

Что это такое?

Ложная корреляция (spurious correlation): формально данные показывают, что зависимость есть, но логически зависимости не должно быть.

Пример (http://www.tylervigen.com)

Расходы на науку в US и число самоубийст (удушение) $(r = 0.99)$
Число утонувших в бассейне и фильмов с Николасом Кейджем $(r = 0.66)$

Ложная корреляция

Что это такое?

Ложная корреляция (spurious correlation): формально данные показывают, что зависимость есть, но логически зависимости не должно быть.

Пример (http://www.tylervigen.com)

Расходы на науку в US и число самоубийст (удушение) $(r = 0.99)$

Число утонувших в бассейне и фильмов с Николасом Кейджем $(r = 0.66)$

Процент разводов в штате Мэн и подушевое потребление маргарина $(r = 0.99)$

Ложная корреляция

Что это такое?

Ложная корреляция (spurious correlation): формально данные показывают, что зависимость есть, но логически зависимости не должно быть.

Пример (http://www.tylervigen.com)

Расходы на науку в US и число самоубийст (удушение) $(r = 0.99)$
Число утонувших в бассейне и фильмов с Николасом Кейджем $(r = 0.66)$
Процент разводов в штате Мэн и подушевое потребление маргарина $(r = 0.99)$
Возраст мисс Америка и число смертей от пара и горячих предметов $(r = 0.87)$.

Программное обеспечение

K.O.: Современный анализ данных невозможен без компьютера.

[1] WYSIWYG = What You See Is What You Get

Программное обеспечение

K.O.: Современный анализ данных невозможен без компьютера.

Хорошая новость: Всё уже запрограммировано и сделано до нас. Есть много специализированных программ для анализа данных, хороших и разных.

Software:

- Свободное vs Проприетарное
- Общего назначения vs Специализированное
- WYSIWYG[1] vs “программирование”

[1] WYSIWYG = What You See Is What You Get

Программное обеспечение

Software общего назначения

- Табличные процессоры: MS Excel, Google Tables, LibreOffice Calc, gnumeric etc

Специализированное Software (непромышленное)

- свободное: gretl, PSPP,
- проприетарное: SPSS, Eviews, STATA, Statistica, MatLab

Программирование:

- R, Python, Julia etc

Как хранить статистические данные?

Важно!

Как правило статистические данные представляются в виде таблиц (по столбцам).

Табличные процессоры: данные, форматирование, графики, формулы etc.

Специализированное Software: данные, описание данных, скрипты etc

Текстовый формат `.csv` или `.txt` (свободный): данные в текстовом формате

Формат CSV

Формат CSV представления таблиц[2]: таблица хранится в текстовом формате по строкам

Особенности:

1. в каждой строке ячейки разделяются одним из специальных разделителей: пробел “,” “;” “\t”
2. один из десятичных разделителей: “.” “,”
3. совместим со всеми табличными процессорами, статистическим программами и языками программирования.

Также могут иметь расширение `.txt`

[2] CSV = Comma Separated Values

Другие форматы

Soft	Формат
MS Excel	`.xls .xlsx`
SPSS	`.sav`
STATA	`.dta .do`
EViews	`.wf1`
grel	`.gdt`
R	`.R`
LibreOffice	`.ods`