Анализ данных в политических науках. Описательные статистики. Визуализация

Н. В. Артамонов, И. А. Истомин

МГИМО МИД России

3 декабря 2017 г.

Что в основе

Два базовых понятия:

- Генеральная совокупность (population)
- Выборка (sample): ограниченная инфорамция из генеральной совокупности

Что в основе

Два базовых понятия:

- Генеральная совокупность (population)
- Выборка (sample): ограниченная инфорамция из генеральной совокупности

Цель

На основе выборочной информации сделать выводы и генеральной совокупности

Предварительный анализ

Вначале полезен “общий взгляд” на имеющиеся статистические данные:

1. описательные статистики
2. визуализация

Предварительный анализ

Вначале полезен “общий взгляд” на имеющиеся статистические данные:

1. описательные статистики
2. визуализация

Это позволит:

1. Получить полезную предварительную информацию
2. Выявить законометрности
3. Выявить “аномалии”. Например, выбросы и “нетипичные” наблюдения

Содержание

Пусть X – количественный фактор

Выборка[1] из фактора

$$X_1, X_2, \ldots, X_n$$

Далее n – объём выборки.

[1] или наблюдения

Пусть X – количественный фактор

Выборка[1] из фактора

$$X_1, X_2, \ldots, X_n$$

Далее n – объём выборки.

Определение

Статистика – функция от элементов выборки

[1] или наблюдения

Описательные статистики

Основные дескриптивные статистики для количественного фактора:

1. выборочное среднее
2. стандартное отклонение
3. min & max
4. медиана
5. квантили (квартиль и др)

Для парных наблюдений

1. коэффициент корреляции

Среднее и момент

Выборочное среднее

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

Выборочный момент порядка k

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}(X_1^k + \cdots + X_n^k)$$

Дисперсия и стандартное отклонение

Выборочная дисперсия

$$\mathrm{Var}(X) = \sigma_X^2 = \overline{X^2} - \left(\bar{X}\right)^2 \geq 0$$

Стандартное отклонение

$$s_X = \sqrt{\frac{n}{n-1}\,\mathrm{Var}(X)}$$

Их можно интерпретировать как “меры разброса” наблюдений вокруг среднего

Вариационный ряд

Определение

Вариационный ряд $X_{(i)}$*: элементы выборки упорядочены по возрастанию, т.е.*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

Вариационный ряд

Определение

Вариационный ряд $X_{(i)}$*: элементы выборки упорядочены по возрастанию, т.е.*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

По определению

- $X_{(n)}$ – максимальная статистика

$$X_{(n)} = \max\{X_1, \ldots, X_n\}$$

- $X_{(1)}$ – минимальная статистика

$$X_{(1)} = \min\{X_1, \ldots, X_n\}$$

Выборочная медиана

Два случая: n чётное и нечётное

Определение

Если $n = 2k + 1$, *то выборочная медиана*

$$\mathrm{MED} = X_{(k+1)}$$

(центральный элемент вариационного ряда)

Если $n = 2k$, *то выборочная медиана*

$$\mathrm{MED} = \frac{1}{2}(X_{(k)} + X_{(k+1)})$$

Выборочная квантиль

Пусть $p \in (0, 1)$.

Определение

Выборочная квантиль порядка p:

$$Q_p = X_{([pn]+1)}$$

где $[\cdot]$ – целая часть числа.

Выборочная квантиль

Пусть $p \in (0, 1)$.

Определение

Выборочная квантиль порядка p:

$$Q_p = X_{([pn]+1)}$$

где $[\cdot]$ – целая часть числа.

Частные случаи:

- $p = 1/4, 3/4$, то квартиль
- $p = k/10$, то дециль
- $p = k/100$, то перцентиль

Ковариация

Пусть имеем парные наблюдения количественных факторов

$$(X_1, Y_1) \cdots (X_n, Y_n)$$

Обозначим

$$\overline{XY} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i$$

Определение

Выборочный коэффициент ковариации

$$\operatorname{cov}(X, Y) = \overline{XY} - \bar{X} \cdot \bar{Y}$$

Коэффициент корреляции

Определение

Выборочный (парный) коэффициент корреляции

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

Коэффициент корреляции

Определение

Выборочный (парный) коэффициент корреляции

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

Свойства

1. $-1 \leq \text{corr} \leq 1$
2. $\text{corr} = \pm 1 \iff Y = \beta_0 + \beta_1 X \ (\beta_1 \neq 0)$
3. безразмерность: $\text{corr}(aX, bY) = \text{corr}(X, Y)$.

Коэффициент корреляции

Определение

Выборочный (парный) коэффициент корреляции

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

Свойства

1. $-1 \leq \text{corr} \leq 1$
2. $\text{corr} = \pm 1 \iff Y = \beta_0 + \beta_1 X \ (\beta_1 \neq 0)$
3. безразмерность: $\text{corr}(aX, bY) = \text{corr}(X, Y)$.

Интерпретация: степень зависимости между двумя факторами

Как вычислять?

Таблица: Описательные статистики в MS Excel & R

Статистика	MS Excel	R
Среднее	СРЗНАЧ	mean
Ст.отклонение	СТДОТКЛ	sd
min & max	МИН & МАКС	min & max
Медиана	МЕДИАНА	median
Квантиль	КВАНТИЛЬ	quantile
Корреляция	КОРРЕЛ	cor

Замечание:

1. в MS Excel находятся в разделе “Статистические”
2. в R функция `summary` сразу вычисляет основные дескриптивные статистики

Визуализация

Основные графики

1. Гистограмма
2. График рассеивания (точечная диаграмма)
3. График рассеивания со сглаживанием и форматированием

Базовые функции R

Гистограмма

```
hist(x,...)
```

Необязательные аргументы: col, main, xlab ,ylab

Базовые функции R

Гистограмма

```
hist(x,...)
```

Необязательные аргументы: col, main, xlab ,ylab

График рассеивания

```
plot(x, y,...)
```

Необязательные аргументы: col, main, xlab ,ylab

Пакет ggplot2

Гистограмма

```
qplot(data, x,...)
```

Необязательные аргументы: fill, main, xlab ,ylab

Пакет ggplot2

Гистограмма

```
qplot(data, x,...)
```

Необязательные аргументы: fill, main, xlab ,ylab

График рассеивания

```
qplot(data, x, y,...)
```

Необязательные аргументы: col, main, xlab ,ylab

Пакет ggplot2

Гистограмма

```
qplot(data, x,...)
```

Необязательные аргументы: fill, main, xlab ,ylab

График рассеивания

```
qplot(data, x, y,...)
```

Необязательные аргументы: col, main, xlab ,ylab

График рассеивания с “подогнанной прямой”

```
qplot(data, x, y,...)+
geom_smooth(method="lm")
```