Exploratory Data Analysis I

Roberta De Vito

Sapienza Università di Roma

Matrix form of the data

Matrix form of the data

- $(y_i, \underline{x}_i), \quad i = 1, \ldots, n$

Matrix form of the data

- $(y_i, \underline{x}_i), \quad i = 1, \ldots, n$
- $\underline{x}_i = (x_{i1}, \ldots, x_{ip})^\top$

Matrix form of the data

- $(y_i, \underline{x}_i), \quad i = 1, \dots, n$
- $\underline{x}_i = (x_{i1}, \dots, x_{ip})^\top$
- the matrix form for the data

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Matrix form of the data

Examples?

Matrix form of the data

- teens example

$$\begin{array}{ccc} y_1 & \dots & y_n \\ teendep_1 & \dots & teendep_n \end{array}$$

- variables

$$X = \begin{pmatrix} yes & yes & \dots & none \\ \vdots & \vdots & \vdots & \vdots \\ no & yes & \dots & 3/week \\ \vdots & \vdots & \ddots & \vdots \\ yes & no & \dots & 6/week \end{pmatrix}$$

- $i = 1, \dots, n$

Matrix form of the data

- teens example

$$\begin{array}{ccc} y_1 & \dots & y_n \\ teendep_1 & \dots & teendep_n \end{array}$$

- variables

$$X = \begin{array}{c} \begin{array}{cccc} \text{jail} & \text{act} & \dots & \text{smok} \end{array} \\ \begin{pmatrix} yes & yes & \dots & none \\ \vdots & \vdots & \vdots & \vdots \\ no & yes & \dots & 3/week \\ \vdots & \vdots & \ddots & \vdots \\ yes & no & \dots & 6/week \end{pmatrix} \end{array}$$

- $i = 1, \dots, n$

Statistical Settings

- input/predictors/covariates (X) and output/outcome/response (Y)

Statistical Settings

- input/predictors/covariates (X) and output/outcome/response (Y)
- main idea: link between data and phenomenon

Statistical Settings

- input/predictors/covariates (X) and output/outcome/response (Y)
- main idea: link between data and phenomenon
- unknown function f and error term ϵ, $y = f(x) + \epsilon$

Statistical Settings

- input/predictors/covariates (X) and output/outcome/response (Y)
- main idea: link between data and phenomenon
- unknown function f and error term ϵ, $y = f(x) + \epsilon$
- Need to estimate f to make prediction or to classify outcomes
  1. Need to draw inferences about the model
  2. Understand how it is X

What is statistics?

- It is not really an exact science

What is statistics?

- It is not really an exact science
- Prevision: manhole in New York

What is statistics?

- It is not really an exact science
- Prevision: manhole in New York
- From the disorder to order

Chicken, fridge, poetry... and statistics

Do you know what statistics is? A thing
that's used to record facts
of the newborn people, those with an ail,
the dead, the imprisoned and those marrying.

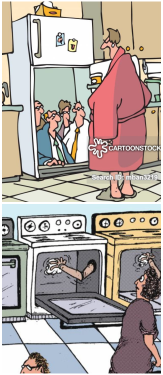
But for me the statistic that is funny
happens when percentage enters the game,
so that, the mean is always the same
even for the person who has no money.

That is, in accordance to current statistics
one person eats one chicken per year:
and, even if your budget doesn't allow you,
nevertheless the statistics allows it,
although someone else will eat two.

Alberti Salustri, known as Trilussa

Chicken, fridge, poetry... and statistics

I really don't trust statistics much. A man with his head in a hot oven and his feet in a freezer has statistically an average body temperature.



Charles Bukowski

Without deviation from the norm, progress is not possible.

Frank Zappa

Expected Value

- Discrete : $\sum_{r=1}^{k} x_r f_x(x_i;\theta) = \sum x_r p_r$

Expected Value

- Discrete : $\sum_{r=1}^{k} x_r f_x(x_i;\theta) = \sum x_r p_r$
- Continuous $\int x f_x(x;\theta)$

Expected Value

- Discrete : $\sum_{r=1}^{k} x_r f_x(x_i;\theta) = \sum x_r p_r$
- Continuous $\int x f_x(x;\theta)$
- Example: the toss

$$E[X] = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$$

- Example with freezer (-18) and oven (100)

$$E[X] = \frac{100 - 18}{2} = 41$$

Expected Value: The Concept

The mean is useful, but we need to understand what it does (and does not) tell us.

- The mean gives a general idea, but it does not show differences between individuals.
- It is like a blurred picture: it shows the overall view, but not the details.
- The mean helps us understand the "typical" value for a group, but it does not tell us who has more or who has less.

Variance

- (Trilussa's chicken joke)
  "If you eat two chickens and I eat none, statistics says we ate one each."

- The mean is 1.

- But be careful: one person is still hungry and the other is full.
- The values are very different from each other.
- Variance tells us how far the numbers spread out from the

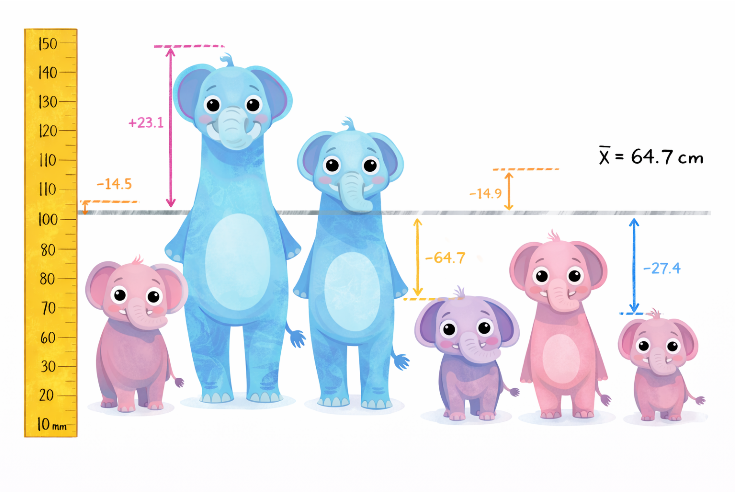
Variance

Variance is computed by:

1. Measuring the distance of each value from the mean (how far it deviates).
2. Squaring those distances (so negative and positive deviations don't cancel out).
3. Taking the average of these squared differences.

Result: a "spread meter" for your data.

Variance



150
140
130
120
110
110
100
80
80
70
60
30
20
10 mm
+23.1
-14.5
-14.9
$\bar{X}$ = 64.7 cm
-64.7
-27.4

The variance: how to compute

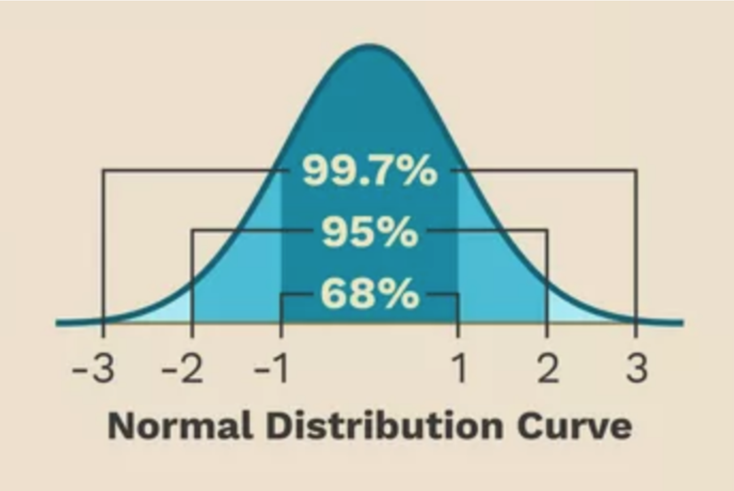
- $Var[X] = \sigma^2 = E[(X - \mu)^2]$
- $Var[X] = E[X^2] - E[X]^2$
- Example with freezer (-18) and oven (100)

$$V[X] = 6962$$

Var function in R

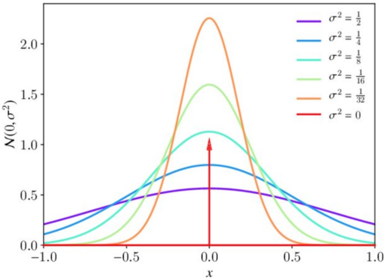
Standard deviation

$$s_x = \sigma = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$



The Normal Distribution

$$X \sim N(\mu, \sigma^2)$$



Correlation

Supervised techniques

- distinguish X and Y
- no matter if the relationship is linear or not linear
- example: depression

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Correlation

Unsupervised techniques

- all the variables have the same role (No outcome)
- Do they correlate?

$$r = \frac{\sum (x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sum (x_1 - \mu_1)^2 \sum (x_2 - \mu_2)^2}}$$

- example: diet with nutrients

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$