Exploratory Data Analysis I

Roberta De Vito

Brown Public Health

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

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

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

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

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

- $i = 1, \dots, n$

Matrix form of the data

- There is no Y
- variables

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

The variables can be continuous or categorical

Supervised VS Unsupervised

Supervised

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

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

Supervised VS Unsupervised

Unsupervised

- 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

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

Supervised VS Unsupervised

	Unsupervised	Supervised
Continuous	Clustering Dimension Reduction • SVD • PCA • K-mean	Regression • Linear • Polynomial Decision tree Random Forest
Categorical	Association Analysis Hidden Markov Model	Logistic Regression SUM

Expected Value

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)

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

The variance

The variance

- $Var[X] = \sigma^2 = E[(X-\mu)^2]$

The variance

- $Var[X] = \sigma^2 = E[(X - \mu)^2]$
- $Var[X] = E[X^2] - E[X]^2$

The variance

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

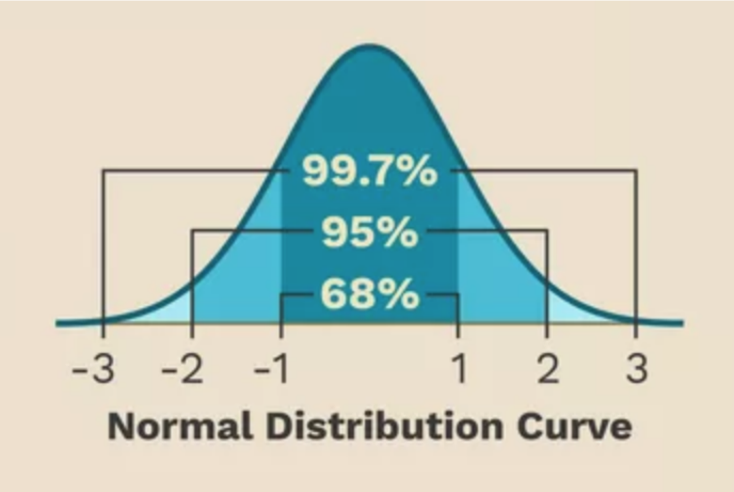
$$V[X] = 6962$$

Var function in R

Standard deviation

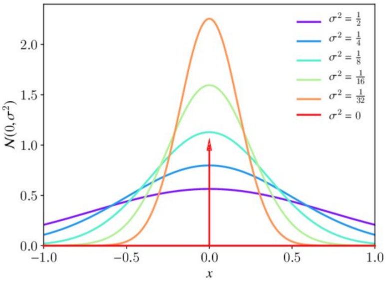
Standard deviation

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



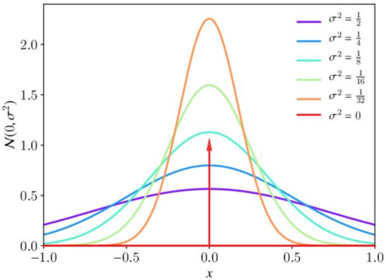
The Normal Distribution

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



The Normal Distribution

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



Central limit theorem: the average of many samples converges to a normal distribution as the number of samples increases

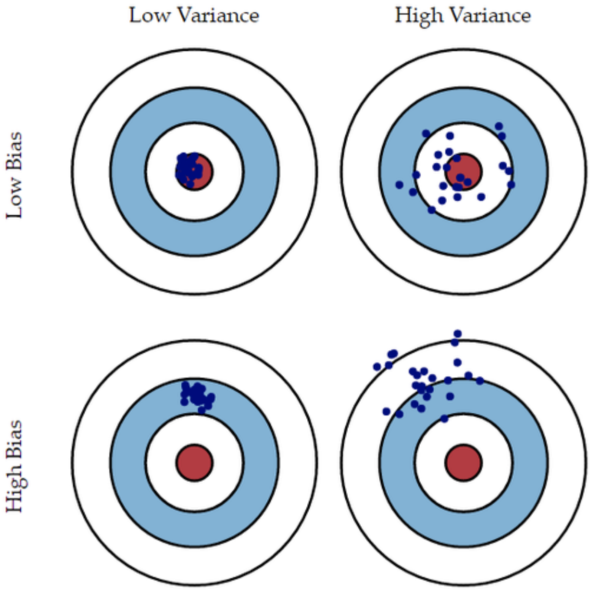
1. Prediction

For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$Y - \hat{Y} = f(x) + \epsilon - \hat{f}(x) = \underbrace{(f(x) - \hat{f}(x))}_{reducible} + \underbrace{\epsilon}_{irreducible}$$

$$E[(Y - \hat{Y})^2 | X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{reducible} + \underbrace{Var(\epsilon)}_{irreducible}$$

1. Prediction



2. Inference

- What is inference?

2. Inference

- What is inference?
- from the population to the sample

2. Inference

- What is inference?
- from the population to the sample
- AIM: to arrive at the true
- $X_1, \ldots, X_n \rightarrow x_1, \ldots, x_n$
- estimate θ giving a function f
- example: the average of the students' age at Brown University

PROBLEM: we will never discover the true θ, we can discover the law f

Inferential problem

1. sample
2. function

Example: Bernoulli

$$\{x_1, x_2, x_3 \ldots, x_1 0\} = \{1, 1, 1, \ldots, 0\}, \quad 4C, 6H$$

the law

$$f(x_1, \ldots, x_n) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} 0 \leq \theta \leq 1$$

Inferential problem

1. Exact estimation

$$\frac{\sum x_i}{n} = \frac{4}{10}$$

2. confidence intervals

$$(0.4 - \delta_1; 0.4 + \delta_1)$$

3. Hypothesis testing

$$\begin{array}{ll} H_0: & \theta = \frac{1}{2} \\ H_1: & \theta = \frac{3}{4} \end{array}$$

Considerations

- When interested in prediction, want flexible model; interpretability less important
- When interested in inference, want interpretable (often less flexible) model

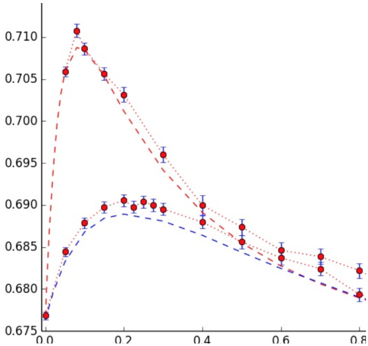
Parametric VS Nonparametric

Parametric

- assumption on f, then fit the model
- Pros: easy to estimate, simple
- Cons: far from the true f, our estimates can be poor, more parameters, overfitting

Parametric VS Nonparametric

Nonparametric



- no assumption on f
- problem of overfitting