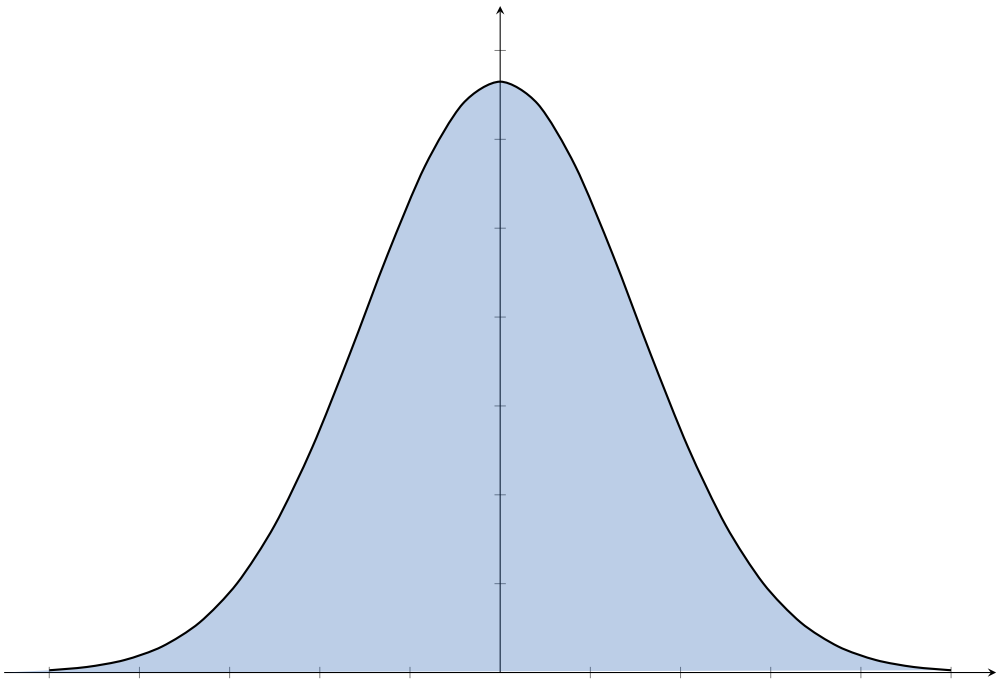
Dispense di BDA

A cura di Mattia Salvetti

Anno Accademico 2023/24



Indice

Business Data Analytics I

1 Introduzione

Lo scopo fondamentale di questo corso è quello di fornire lo studente con gli strumenti necessari a elaborare ed estrapolare informazioni utili da grandi moli di dati. Proprio la quantità di questi dati porta alla necessità di capire cose è importante e cosa no, e diventa fondamentale avere dei modelli per estrapolare informazioni automaticamente su basi statistiche.

1.1 Dati Multivariati

Possiamo rappresentare i dati in forma matriciale, avendo n unità statistiche di cui osserviamo p variabili. A questo punto possiamo scrivere la matrice dei dati come

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \in \mathbb{R}^{nxp} \tag{1}$$

dove x_{ij} è la j-esima variabile della i-esima unità statistica. Fatto ciò i dati possono essere rappresentati come una nuvola di punti appartenente a $\mathbb{R}^p$. Ricordiamo che al crescere di p l'analisi diventa più complicata, esistono quindi dei metodi di riduzione dimensionali che permetto di passare da p a z variabili con $z < p$.

2 Analisi Univariata

Possiamo vedere la nostra matrice dei dati come una serie di vettori riga

$$\mathbb{X} = [X_1, X_2, \dots, X_n]^T \tag{2}$$

Dove $X_1 = [x_{11}, x_{12}, \dots, x_{1p}]$, $X_2 = [x_{21}, x_{22}, \dots, x_{2p}]$ e così via. A questo punto, dai campioni si può fare statistica descrittiva o inferenziale. Partiamo da una rapida discussione della statistica descrittiva.

2.1 Statistica Descrittiva

La statistica descrittiva è una statistica che ci è fornita una descrizione della distribuzione dei dati. Alcuni indici per la statistica descrittiva sono:

- Media campionaria: $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \; \forall j$
- Varianza campionaria: $S_{jj} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ NB: spesso si usa $n-1$ al denominatore per ragioni di non distorsione.
- Deviazione standard: $\sqrt{S_{jj}}$

Un altro aspetto che tornerà utile per i nostri scopi è lavorare con variabili confrontabili. Per fare ciò è fondamentale che esse siano standardizzate. Standardizzare una variabile vuol dire renderla adimensionale e a media nulla, così facendo:

$$x^*_{\cdot j} = \frac{x_{\cdot j} - \bar{x}_j}{\sqrt{S_{jj}}} \tag{3}$$

2.2 Statistica Inferenziale

Da un campione casuale $x_1, \ldots, x_n$ possiamo, ipotizzando che i dati siano indipendenti ed identicamente distribuiti, ovvero che ogni osservazione sia indipendente da quella precedente e che esse seguano tutte la stessa distribuzione statistica $X_1, \ldots, X_n \sim \mathcal{F}(\theta)$, dove $\mathcal{F}(\theta)$ è la legge della variabile aleatoria, e θ è il parametro della famiglia di distribuzioni. Fatto ciò possiamo:

- Stima puntuale: usare i parametri per associare un singolo valore al parametro
- Stima intervallare: usare i parametri per associare un intervallo di valori al parametro con un certo intervallo di confidenza, ovvero costruisco un intervallo di confidenza con un certo livello
- Test d'ipotesi: usa i parametri della distribuzione per capire se l'ipotesi H_0 è vera o falsa.

2.2.1 Stima Puntuale