Precorso di Statistica per le Lauree Magistrali

Gianni Betti

5-6 Ottobre 2023 - Ore 12-14

Diagrammi a nuvola di punti, covarianza e correlazione campionaria (3.7 Stock & Watson)

- Questa ultima parte del precorso di Statistica è di fatto una introduzione al modello di regressione lineare e quindi all'econometria.
- Rivediamo alcune relazioni tra variabili, denotate da X e Y (es. X = età e Y = retribuzione).
- Alla domanda: quale è la relazione tra età e retribuzione (o tra retribuzione ed età) ? E' possibile passare in rassegna tre metodi per riassumere tale legame:

1. **Il diagramma a nuvola di punti**
2. **La covarianza campionaria**
3. **Il coefficiente di correlazione campionario**

Il diagramma a nuvola di punti

- Un diagramma a nuvola di punti, è un grafico delle n osservazioni campionarie su X_i e Y_i, nel quale ciascuna osservazione è rappresentata dal punto (X_i, Y_i).
- Per esempio, nella Figura 3.2 a pagina 71 del libro di testo, è rappresentata la nuvola di punti (anche *scatter diagram* in inglese) dell'età (X) e della retribuzione (Y) di un campione di 200 manager tratto dall'indagine CPS.

Il diagramma a nuvola di punti

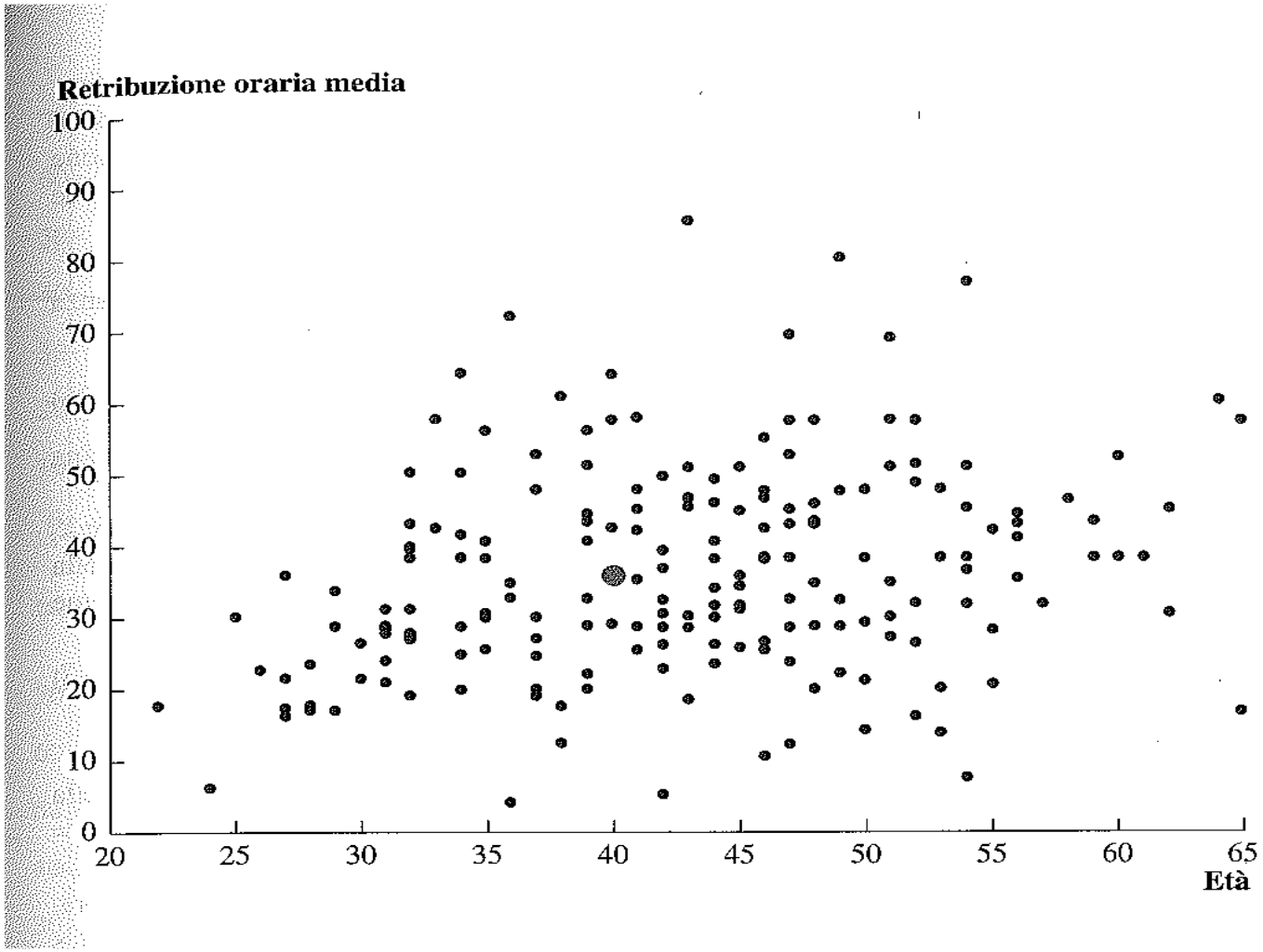


Figura 3.2

Grafico a nuvola della retribuzione oraria media sull'età.

Ogni punto del grafico rappresenta l'età e la retribuzione media di uno dei 200 lavoratori nel campione. Il punto in evidenza corrisponde a un lavoratore di 40 anni che guadagna 35,78$ all'ora. I dati sono relativi a manager dell'industria informatica e sono tratti dal CPS del marzo 2009.

Il diagramma a nuvola di punti

Per esempio, uno dei lavoratori nel campione ha 40 anni e guadagna (in media) 37,78$ all'ora. L'età e la retribuzione di questo lavoratore sono evidenziati dal cerchio più ampio.

Il grafico stesso presenta una relazione positiva tra età e retribuzione per questo campione; ciò ha ovviamente anche una interpretazione economica: i lavoratori più anziani, con un curriculum e una esperienza maggiore, tendono a guadagnare di più di quelli giovani (a parità di altre caratteristiche).

Infatti, questa non è una relazione esatta; ovvero, conoscendo solo l'età, non è possibile prevedere esattamente la retribuzione corrispondente.

Covarianza campionaria

- La covarianza (su tutta la popolazione) è stata introdotta nel Paragrafo 2.3 come una proprietà della distribuzione di probabilità congiunta delle variabili casuali X e Y. Poiché la distribuzione della popolazione è in realtà ignota, in pratica è ignota anche la covarianza ed è quindi necessario stimarla attraverso il campione, sfruttando l'insieme delle coppie (X_i, Y_i).
- La covarianza campionaria è indicata con $s_{XY,}$ ed è definita nella relazione (3.24) di pagina 71:

Covarianza campionaria

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$ • (3.24)

- Come per la varianza campionaria, la sommatoria nella (3.24) è divisa per (n-1) e non per n; anche qui la differenza è dovuta all'usa dello stimatore di X e Y medio invece che del valore vero.
- Concettualmente, entra di nuovo in gioco il concetto dei "*gradi di libertà*".

Coefficiente di correlazione campionario

- Il coefficiente di correlazione campionario si indica con r_{XY}, ed è dato dal rapporto della covarianza campionaria (s_{XY}) e le deviazioni standard della X e Y (scarti quadratici medi campionari):

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

- (3.25)

Coefficiente di correlazione campionario

La correlazione campionaria misura la forza dell'associazione (relazione) **lineare** esistente tra la variabile casuale X e la Y in un campione di n osservazioni. Come per la correlazione nella popolazione, la correlazione campionaria varia tra -1 e +1.

La correlazione sarà tanto più vicina ad 1 in valore assoluto, quanto più le coppie (X_i, Y_i) giacciono su una retta nel piano.

Diagrammi a nuvola di punti e correlazione campionaria

E’ interessante analizzare il concetto di correlazione e quindi di correlazione campionaria attraverso la nuvola di punti o “*scatter plot*”.

La Figura 3.3 di pagina 73 riporta quattro esempi di relazioni, nei quali in due casi vi è una forte relazione lineare, mentre negli altri due casi non vi è (correlazione pressoché nulla).

Diagrammi a nuvola di punti e correlazione campionaria

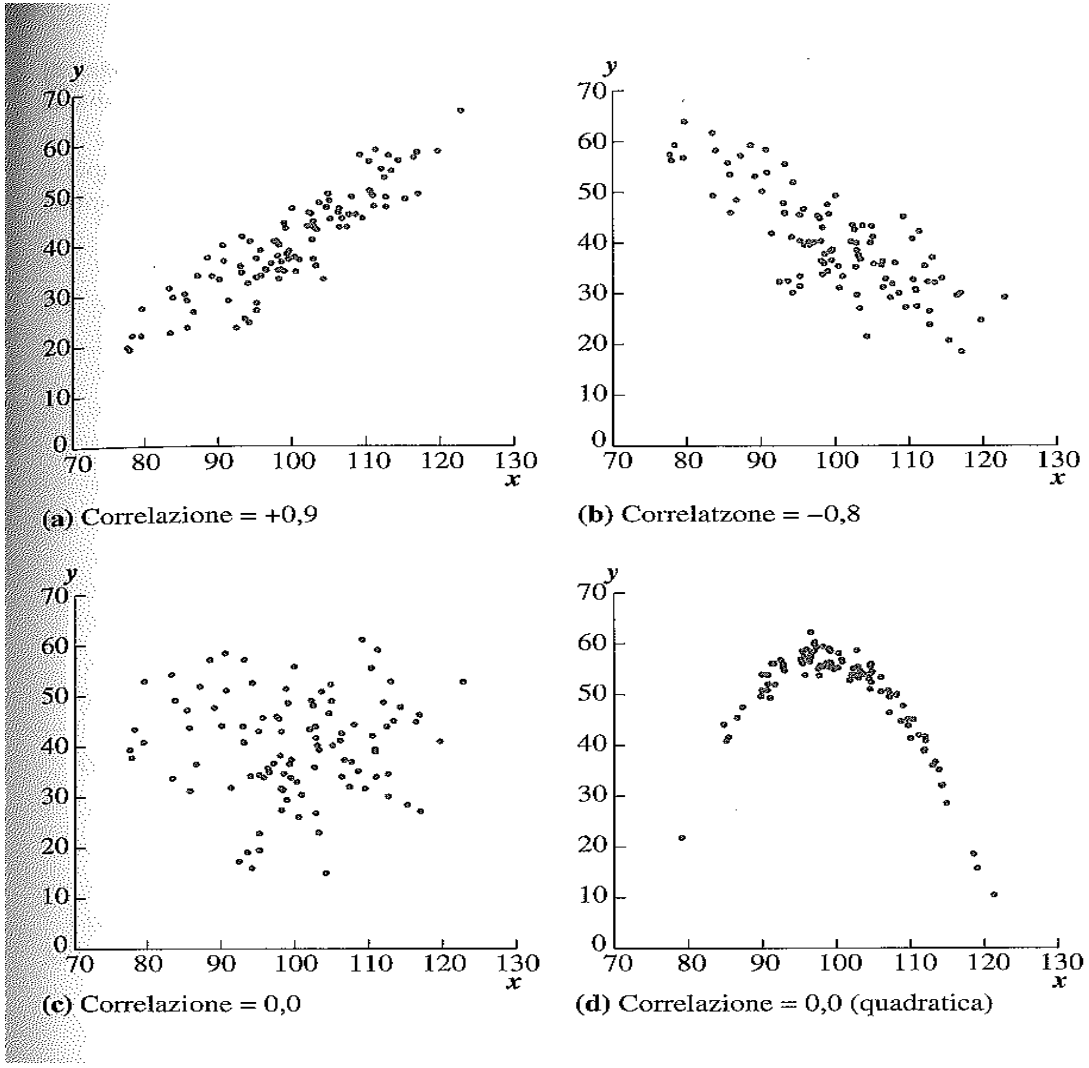


Figura 3.3
Diagramma a nuvola per quatro insiemi di dati ipotetici.

I diagramma a nuvola delle Figure 3.3a e 3.3b mostrano relazioni lineari forti tra *X* e *Y*. Nella Figura 3.3c, *X* è indipendente da *Y* e le due variabili sono incorrelate. Anche le due variabili nella Figura 3.3d sono incorrelate, benché siano legate non linearmente.

Diagrammi a nuvola di punti e correlazione campionaria

- La Figura 3.3a mostra una forte relazione lineare positiva tra le variabili, con una correlazione campionaria di +0,9. La Figura 3.3b mostra una forte relazione negativa, con una correlazione campionaria pari a -0,8.
- La Figura 3.3c mostra un diagramma a nuvola senza una relazione ben definita, con una correlazione nulla.
- Infine, la Figura 3.3d mostra una relazione ben definita, ma sicuramente **non lineare**: anche qui la correlazione è nulla. Questo esempio finale mette in evidenza un punto importante: il coefficiente di correlazione è una misura di associazione **lineare**.

Consistenza della covarianza e della correlazione campionaria

Come la varianza campionaria, anche la covarianza campionaria gode della proprietà della consistenza:

$$S_{XY} \xrightarrow{p} \sigma_{XY}. \qquad (3.26)$$

In altre parole, per grandi campioni, la covarianza campionaria tende (ovvero è con alta probabilità vicina) alla covarianza nella popolazione.

Come preparazione al corso, è possibile leggere l'Appendice 3.3 di pagina 81 per una più semplice dimostrazione della consistenza della varianza campionaria.

Esercizio

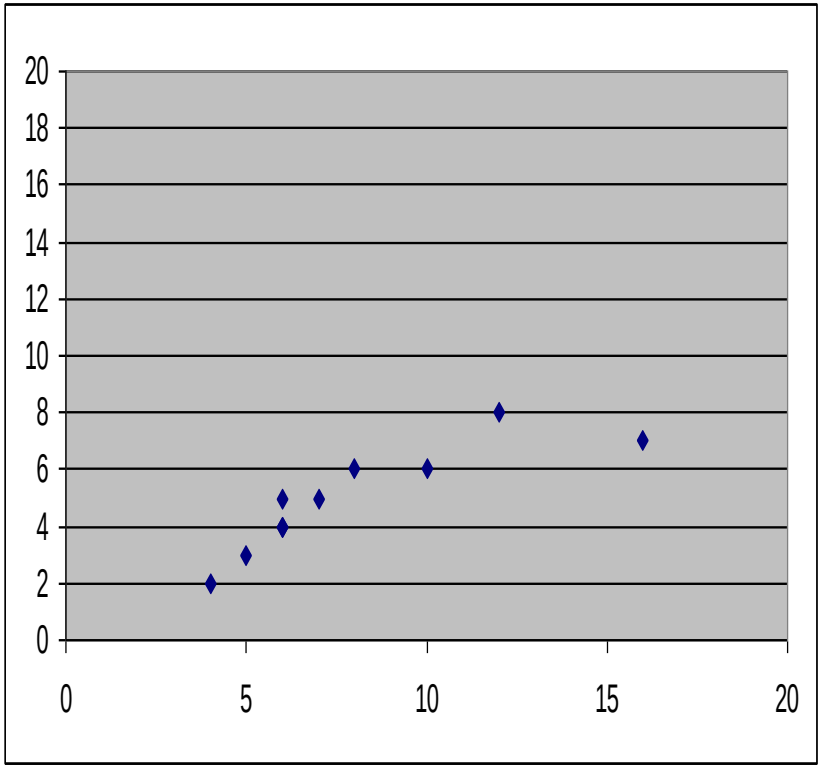
- In una indagine campionaria sono state estratte 10 aziende per le quali sono state misurate le unità di input (X) e le unità di output (Y)
- Le informazioni disponibili sono le seguenti:

Azienda	Y = Output	X = Input
1	8	12
2	5	6
3	4	6
4	5	7
5	6	8
6	6	10
7	2	4
8	3	5
9	4	6
10	7	16

- Costruire il diagramma a nuvola di punti, la covarianza e il coefficiente di correlazione campionari

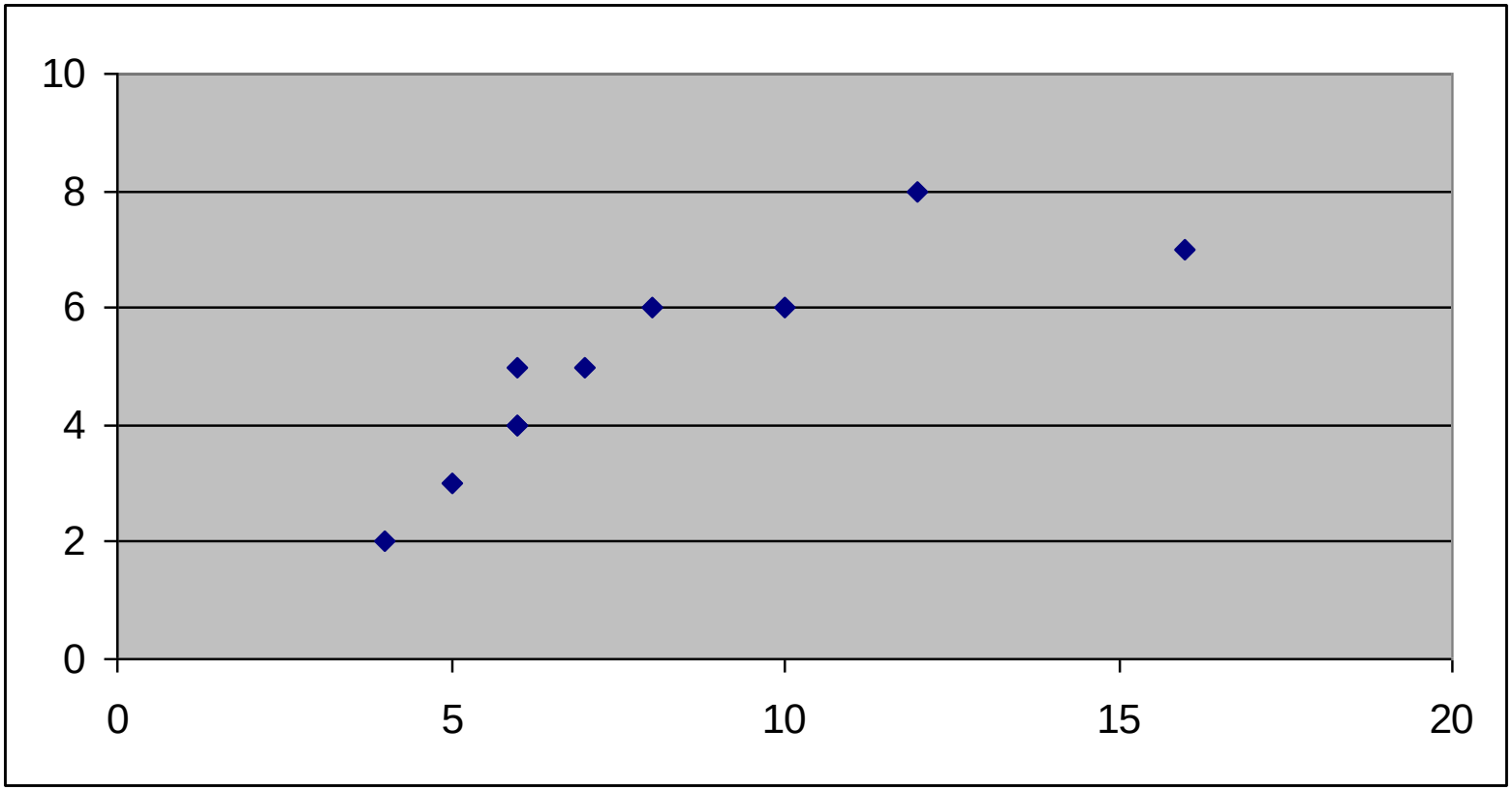
Soluzione esercizio

- Diagramma a nuvola di punti (assi in proporzione)



Soluzione esercizio

- Diagramma a nuvola di punti (assi non in proporzione)



Covarianza campionaria

- Il primo passo consiste nella calcolo delle medie e degli scarti dalle medie delle variabili X e Y:

Azienda	X = Input	Y = Output	Y-Ymed	X-Xmed
1	12	8	3	4
2	6	5	0	-2
3	6	4	-1	-2
4	7	5	0	-1
5	8	6	1	0
6	10	6	1	2
7	4	2	-3	-4
8	5	3	-2	-3
9	6	4	-1	-2
10	16	7	2	8
	8	**5**	**0**	**0**

Covarianza campionaria

- Per stimare il numeratore della covarianza campionaria è necessario calcolare i prodotti tra le X e le Y in forma di scarto:

Azienda	X = Input	Y = Output	Y-Ymed	X-Xmed	y*x
1	12	8	3	4	12
2	6	5	0	-2	0
3	6	4	-1	-2	2
4	7	5	0	-1	0
5	8	6	1	0	0
6	10	6	1	2	2
7	4	2	-3	-4	12
8	5	3	-2	-3	6
9	6	4	-1	-2	2
10	16	7	2	8	16
	8	**5**	**0**	**0**	**52**

- Dividendo il numeratore per (n-1) = 10-1 = 9, otteniamo
- **52/9 = 5,78** approssimato alla seconda cifra decimale.

Coefficiente di correlazione campionario

- Per il calcolo del coefficiente di correlazione campionario è necessario stimare le varianza campionarie di X e Y, e i corrispondenti scarti quadratici medi campionari (deviazioni standard).

Azienda	Y-Ymed	X-Xmed	x*x	y*y
1	3	4	16	9
2	0	-2	4	0
3	-1	-2	4	1
4	0	-1	1	0
5	1	0	0	1
6	1	2	4	1
7	-3	-4	16	9
8	-2	-3	9	4
9	-1	-2	4	1
10	2	8	64	4
	0	**0**	**122**	**30**

- Le varianze sono pari a **122/9=13,56** e **30/9=3,33**
- Gli scarti quadratici medi campionari pari a **3,68** e **1,82**

Coefficiente di correlazione campionario

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{5,78}{3,68 * 1,82} = \frac{5,78}{6,72} = 0,86$$