RICHIAMI DI STATISTICA

- La statistica è la scienza che permette di conoscere il mondo intorno a noi attraverso i dati.

- Quale è la media della distribuzione del reddito dei neolaureati?

- Per rispondere dovremmo condurre un'approfondita indagine sulla popolazione dei lavoratori studiandone la distribuzione del reddito.
  - Costi alti
  - È impossibile intervistare tutti i membri della popolazione

- L’idea fondamentale è che si possono ottenere informazioni sulla distribuzione di una popolazione scegliendo un campione casuale di tale popolazione.

- L’econometria fa uso di tre tipologie di metodi statistici
  - La stima
  - La verifica di ipotesi
  - Gli intervalli di confidenza

La Stima

- Implica il calcolo di un valore numerico che sia la "migliore congettura" ricavabile dal campione circa una caratteristica ignota della distribuzione della popolazione.
- Supponiamo di voler conoscere il reddito medio delle donne neolaureate Y (μ_Y).
- Per fare ciò possiamo calcolare la media campionaria $\overline{Y}$ ottenendola da n osservazioni i.i.d.

La Stima

- Stimatore: funzione di un campione di dati estratti casualmente da una popolazione
- Stima: valore numerico di uno stimatore quando è calcolato usando i dati di campione specifico.
- La stessa caratteristica di una popolazione può essere studiata attraverso diversi stimatori alternativi.
  - Come facciamo a scegliere il migliore?

La Stima

- Intuitivamente vorremmo che la distribuzione campionaria del nostro stimatore sia concentrata attorno al valore ignoto della caratteristica della popolazione che stiamo studiando.

- Quali sono le caratteristiche che il “migliore” stimatore deve avere?

La Stima

- Non distorsione
  - La media della distribuzione dello stimatore deve essere uguale a μ_Y
- Consistenza
  - Maggiore è il campione, minore è l'effetto distorsivo di deviazioni casuali
- Efficienza
  - Il nostro stimatore deve avere la minima varianza

La Stima

- La media campionaria $\overline{Y}$ è il migliore adattamento ai dati in quanto la differenza quadratica media tra le osservazioni e tale media è la più piccola tra tutti i possibili stimatori.

(1) $$\sum_{i=1}^{n}\left(Y_i - m\right)^2$$

- Tale formula fornisce una misura dell'errore di previsione del corretto valore di Y_i.
- Lo stimatore *m* che minimizza la (1) è detto stimatore dei minimi quadrati.

Campionamento Casuale

- Perché un corretto campionamento casuale è fondamentale per effettuare stime corrette?

- Esempi
  - elezioni presidenziali americane 1936
  - Tasso di disoccupazione

Verifica di Ipotesi

- Il primo passo da compiere nella verifica di un'ipotesi è la specificazione dell'ipotesi stessa.
  - Ipotesi nulla (H_0)

- Usiamo i dati per confrontare l'ipotesi nulla con una seconda ipotesi detta ipotesi alternativa (H_1), che è valida se la nulla non lo è.

Verifica di Ipotesi

- Ad esempio l'ipotesi nulla che la media di Y nella popolazione, E(Y), assuma un valore specifico, $\mu_{Y,0}$ viene indicata
  - H_0:E(Y)= $\mu_{Y,0}$
- L'ipotesi alternativa più generale è che
  - H_1:E(Y)≠$\mu_{Y,0}$ è detta bilaterale in quanto E(Y) può essere sia minore sia maggiore di $\mu_{Y,0}$
- Il problema che gli statistici (ed anche voi) affrontano è quello di utilizzare l'evidenza empirica derivante da un campione selezionato casualmente per stabilire se accettare l'ipotesi nulla oppure se scegliere l'ipotesi alternativa.

Valore-p del test

- Dato un campione è assai improbabile che il valore della media campionaria coincida con il valore ipotizzato.
- Tale differenza può essere dovuta sia al fatto che l'ipotesi nulla sia falsa sia al fatto che, anche se vera, vi siano delle piccole differenze a causa del campionamento casuale.
- Queste due cause sono impossibili da distinguere con certezza.
- Per questo si può operare un calcolo probabilistico per sottoporre a verifica l'ipotesi nulla tenendo conto dell'effetto del campionamento.
- Tale calcolo è chiamato valore-p dell'ipotesi nulla.

Valore-p del test

- Il valore-p (livello di significatività osservato) è la probabilità di ottenere una misura della media campionaria che sotto l'ipotesi nulla sia lontano dalle code della distribuzione almeno quanto la media campionaria effettivamente calcolata.
- Se nel campione di studenti neolaureati la retribuzione media è di 22,24$, il valore-p è la probabilità di osservare un valore della media campionaria che, a causa del campionamento causale, sia diverso da 20$ (media della popolazione) almeno quanto il valore osservato (22,24$) assumendo che l'ipotesi nulla sia vera.
- Se il valore-p è piccolo (0.5%) allora è inverosimile che tale campione venga estratto qualora fosse vera l'ipotesi nulla e, quindi, possiamo concludere che tale ipotesi è falsa.

Valore-p del test

- Matematicamente abbiamo che

$$valore - p = \Pr_{H_0}\left[\left|\overline{Y} - \mu_{Y,0}\right| > \left|\overline{Y}^{act} - \mu_{Y,0}\right|\right]$$

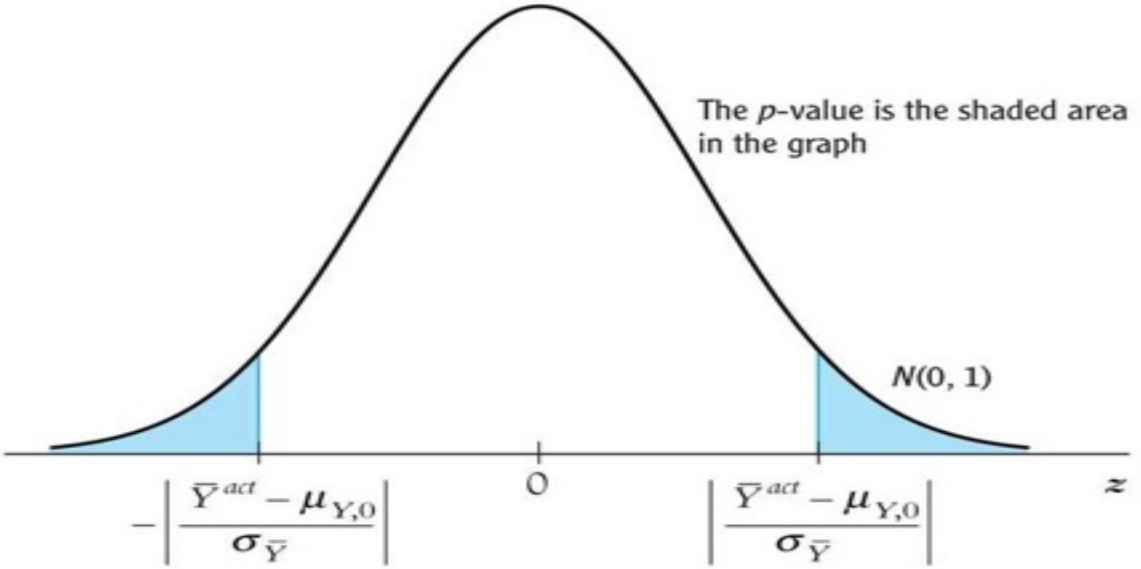
- Il valore-p è l'area nelle code della distribuzione della media campionaria.
- Se il valore-p è elevato allora il valore osservato è coerente con l'ipotesi nulla.
- Per calcolare il valore-p bisogna conoscere la distribuzione campionaria di $\overline{Y}$ sotto l'ipotesi nulla

Valore-p del test

- Tale calcolo è problematico per piccoli campioni. Diversamente lo stesso problema non si pone nel caso di grandi campioni
  - Perché?
- Possiamo approssimare la distribuzione della media campionaria ad una normale

$$N(\mu_{Y,0}, \sigma^2_{\bar{Y}}), \text{ con } \sigma^2_{\bar{Y}} = \frac{\sigma^2_Y}{n}$$

Valore-p del test



- Standardizzando, $\left(\bar{Y}-\mu_{Y,0}\right)/\sigma_{\bar{Y}}$ si distribuisce come una N(0,1).
- Il valore-p è la probabilità ombreggiata sulle code della distribuzione normale standard.

Varianza campionaria

- La varianza campionaria è

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \text{ e lo stimatore di } \sigma_{\overline{Y}} \text{ è } \frac{s_Y}{\sqrt{n}}$$

- Due modifiche rispetto la varianza della popolazione
  - μ_Y è sostituita con $\overline{Y}$ in quanto usiamo lo stimatore della media della popolazione
  - Al denominatore abbiamo n-1 invece di n per correggere una distorsione introdotta dalla stima della media della popolazione (correzione per i gradi di libertà)
- La varianza campionaria è uno stimatore consistente, cioè tende alla varianza della popolazione per n grande

Statistica t

- Nella verifica delle ipotesi, un ruolo fondamentale è giocato dalla media campionaria standardizzata che prende il nome di statistica t

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

- Una statistica test è una statistica usata per la verifica di ipotesi.

Statistica t

- Nel caso di grandi campioni la varianza campionaria è prossima a quella della popolazione.
- La statistica t ha approssimativamente la stessa distribuzione di $\left(\overline{Y} - \mu_{Y,0}\right)/\sigma_{\overline{Y}}$ e, (perché?), di una normale standard.
- Quindi per n grande, il valore-p è

$$valore - p = 2\Phi(-\left|t^{act}\right|), \text{ dove } t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,o}}{SE(\overline{Y})}$$

Esempio

- n=200 neolaureati
- H_0:E(Y)=20$/h
- La retribuzione media del campione è 22.64$/h e la deviazione std. è 18.14$/h

L'errore standard di $\bar{Y}$ è $\frac{s_Y}{\sqrt{n}} = \frac{18.14}{\sqrt{200}} = 1.28$ ed il valore della statistica t

è $t^{act} = \frac{(22.64 - 20)}{1.28} = 2.06$; dalla tavola otteniamo che

il valore - p è $2\Phi(-2.06) = 0.039$

- In altre parole, supponendo che l'ipotesi nulla sia vera la probabilità di ottenere una media campionaria distante da zero almeno quanto quella calcolata sui dati è 3.9%

Statistica t

- Quando la popolazione si distribuisce come una normale, la statistica t si distribuisce secondo una t di Student con n-1 gradi di libertà.
- Non useremo la distribuzione t di Student in quanto le differenze con una normale standard sono molto ridotte specialmente per n grandi.

Verifica di Ipotesi

- Supponiamo di decidere che l'ipotesi nulla venga rifiutata se il valore-p è < 5%.
- Sapendo che l'area delle code della distribuzione normale al di fuori dell'intervallo ±1.96 è 5%, otteniamo la seguente regola,

$$\text{rifiutare } \mathrm{H}_0 \text{ se } \left|t^{act}\right| > 1.96$$

  - Cioè si rifiuta l'ipotesi nulla se il valore della statistica t calcolata sul campione è maggiore di 1.96 in valore assoluto.
  - Se la statistica t si distribuisce come una N(0,1) la probabilità di rifiutare erroneamente l'ipotesi nulla è 5%.

Terminologia della verifica di ipotesi

- Livello di significatività del test
  - La probabilità prefissata di rifiutare l'ipotesi nulla quando è vera
- Valore critico della statistica
  - Valore per il quale il test passa dal non rifiuto al rifiuto dato un certo livello di significatività
- Regione di rifiuto (accettazione)
  - Insieme dei valore della statistica per i quali il test (non) rifiuta l'ipotesi nulla

Terminologia della verifica di ipotesi

- Livello minimo del test
  - La probabilità che il test porti al rifiuto dell'ipotesi nulla quando questa è vera
- Potenza del test
  - La probabilità che il test rifiuti correttamente l'ipotesi nulla qundo è vera l'alternativa
- Il valore-p è il livello di significatività più basso per il quale si può rifiutare l'ipotesi nulla dato il valore osservato della statistica test

Esempio

- Il livello di significatività del test è 5%
- Il valore critico è 1.96
- La regione di rifiuto comprende tutti i valori assunti dalla statistica t che siano al di fuori dell'intervallo ±1.96.
- Se il test rifiuta con un livello di significatività del 5%, si dice che la media della popolazione è statisticamente diversa da quella del campione al livello di significatività del 5%.
- La t statistica è 2.06. Tale valore è maggiore di 1.96 e quindi l'ipotesi è rifiutata al livello 5%

Quale livello di significatività?

- Solitamente viene utilizzato un livello di significatività pari al 5%
- In alcuni casi è consigliabile usare livelli più conservatori (1%, 0.1%).
- Costo: minore è il livello di significatività, più grande è valore critico e più difficile è rifiutare l'ipotesi nulla quando è falsa.
- Più basso è il livello di significatività, minore è la potenza del test.

Verifica ipotesi $E(Y)=\mu_{Y,0}$ contro $E(Y)\neq\mu_{Y,0}$

1. Si calcola l'errore standard della media campionaria, $SE(\bar{Y}) = \hat{\sigma}_{\bar{Y}} = s_Y / \sqrt{n}$

2. Si calcola la statistica t, $t = (\bar{Y} - \mu_{Y,0}) / SE(\bar{Y})$

3. Si calcola il valore-p $= 2\Phi(-|t^{act}|)$. Si rifiuta l'ipotesi nulla al livello di significatività del 5%, se il valore-p è minore di 0.05

Alternative Unilaterali

- Ipotesi alternativa unilaterale $H_1: E(Y) > \mu_{Y,0}$
- L'approccio al calcolo del valore-p ed alla verifica di ipotesi è uguale al caso di alternative bilaterali, con la differenza che il test rifiuta solo quando il valore della statistica t è grande e positivo invece che grande in valore assoluto.
- Il valore-p riporta l'area sottostante la distribuzione normale standard alla destra del valore osservato delle statistica t.

$$valore - p = \Pr_{H_0}(Z > t^{act}) = 1 - \Phi(t^{act})$$

Intervalli di confidenza per la media della popolazione

- Nonostante sia impossibile inferire la media della popolazione dai dati del campione, possiamo costruire un insieme di valori che contiene la media della popolazione con una certa probabilità prefissata
- Tale insieme è detto regione di confidenza e la probabilità prefissata è detta livello di confidenza.
- La regione di confidenza è quindi un intervallo detto intervallo di confidenza.

Intervalli di confidenza per la media della popolazione

- Richiamando la formula per la statistica t, sappiamo che un valore di prova pari a $\mu_{Y,0}$ è rifiutato al 5% se è lontano più di 1.96 errori standard da $\overline{Y}$.
- L'insieme dei valori di μ_Y che non sono rifiutati al livello del 5% è composto da quei valori compresi tra $\pm 1.96 * SE(\overline{Y})$
- Quindi un intervallo di confidenza di livello 95% per μ_Y è $\overline{Y} - 1.96 * SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.96 * SE(\overline{Y})$

Intervalli di confidenza per la media della popolazione - Riassunto

- Un intervallo di confidenza di livello 95% per μ_Y è costruito in modo da contenere il valore vero di μ_Y nel 95% dei casi.
- Quando la dimensione n del campione è elevata, gli intervalli di confidenza al 90%, 95% e 99% sono

$$\text{Intervallo di confidenza al } 90\% \ \mu_Y = \left\{\overline{Y} \pm 1.64 * SE(\overline{Y})\right\}$$

$$\text{Intervallo di confidenza al } 95\% \ \mu_Y = \left\{\overline{Y} \pm 1.96 * SE(\overline{Y})\right\}$$

$$\text{Intervallo di confidenza al } 99\% \ \mu_Y = \left\{\overline{Y} \pm 2.58 * SE(\overline{Y})\right\}$$

Esempio

- Costruiamo un intervallo di confidenza al 95% per la retribuzione media oraria dei neolaureati usando un campione ipotetico di 200 neolaureati in cui $\overline{Y}$=22.64$ e SE($\overline{Y}$) = 1.28.

- L'intervallo di confidenza al 95% è
  - 22.64 ± (1.96*1.28)=22.64 ± 2.51=(20.13$, 25.15$)

Test d'ipotesi per la differenza tra due medie

- Sia μ_w la retribuzione oraria media per la popolazione di donne neolaureate e sia μ_w la retribuzione oraria media per la popolazione di uomini neolaureati.
- L'ipotesi nulla che le retribuzioni di queste due popolazioni differiscano mediamente di un certo ammontare e l'ipotesi alternativa bilaterale sono,

$$H_0: \mu_m - \mu_w = d_0$$

$$H_1: \mu_m - \mu_w \neq d_0$$

- Quale è l'ipotesi nulla che non vi sia differenza tra uomini e donne?

Test d'ipotesi per la differenza tra due medie

- Le medie delle popolazioni sono ignote e quindi devono essere stimate
- Supponiamo di avere due campioni di n_m uomini e n_w donne estratti casualmente dalle rispettive popolazioni.
- Sia $\overline{Y}_m$ la media campionaria della retribuzione annuale per gli uomini e sia $\overline{Y}_w$ quella per le donne.
- Allora $\overline{Y}_m - \overline{Y}_w$ è uno stimatore di $\mu_m - \mu_w$

Test d'ipotesi per la differenza tra due medie

- Per verificare l'ipotesi nulla $\mu_m - \mu_w = d_0$ tramite $\overline{Y}_m - \overline{Y}_w$

dobbiamo conoscere la distribuzione di $\overline{Y}_m - \overline{Y}_w$

- Su che basi giustifichiamo l'affermazione che le due medie campionarie si distribuiscono approssimativamente come delle distribuzioni normali $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$, con $\sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n}$?

Test d'ipotesi per la differenza tra due medie

- Dato che le due medie campionarie sono ottenute da campioni casuali diversi, sono delle variabili casuali indipendenti.
- Quindi $\bar{Y}_m - \bar{Y}_w$ si distribuisce secondo una $N\left[\mu_m - \mu_w, \left(\frac{\sigma_m^2}{n_m}\right) + \left(\frac{\sigma_w^2}{n_w}\right)\right]$
- Solitamente le varianze della popolazione non sono note e devono essere stimate usando le varianze campionarie.
- L'errore standard di $\bar{Y}_m - \bar{Y}_w$ è

$$SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

Test d'ipotesi per la differenza tra due medie

- La statistica t per l'ipotesi nulla si ottiene analogamente al caso della media di una singola popolazione

$$t = \frac{\left(\bar{Y}_m - \bar{Y}_w\right) - d_0}{SE\left(\bar{Y}_m - \bar{Y}_w\right)}$$

- Se i campioni sono numerosi, la statistica t ha una distribuzione normale standard ed il valore-p si calcola analogamente al caso della media di una singola popolazione
- Il test con livello di significatività prefissato si conduce esattamente allo stesso modo.

Test d'ipotesi per la differenza tra due medie

- Gli intervalli di confidenza vengono costruiti secondo le modalità già viste
- Quindi l'intervallo di confidenza al 95% per $d = \mu_m - \mu_w$ è

$$\left(\overline{Y}_m - \overline{Y}_w\right) \pm 1.96 * SE\left(\overline{Y}_m - \overline{Y}_w\right)$$

Retribuzione oraria laureati USA, 25-34 anni

Uomini					Donne				Differenza uomini/ donne		
Anno	$\overline{Y}_m$	s_m	n_m		$\overline{Y}_w$	s_w	n_w		$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	Intervallo
1992	17.57	7.5	1591		15.22	5.97	1371		2.35**	0.25	1.87-2.84
1994	16.93	7.39	1598		15.01	6.41	1358		1.92**	0.25	1.42-2.42
1996	16.88	7.29	1374		14.42	6.07	1235		2.46**	0.26	1.94-2.97
1998	17.94	7.86	1393		15.49	6.80	1210		2.45**	0.29	1.89-3.02

- La statistica t per l'ipotesi nulla che il differenziale salariale sia nullo è
  - (2.45-0)/0.29=8.45

- Tale valore è maggiore di quello critico di un test bilaterale all'1% (2.58) e quindi il differenziale è significativo all'1%.
- L'intervallo di confidenza del 95% è
  - 2.45±(1.96*0.29)=(1.89, 3.02)

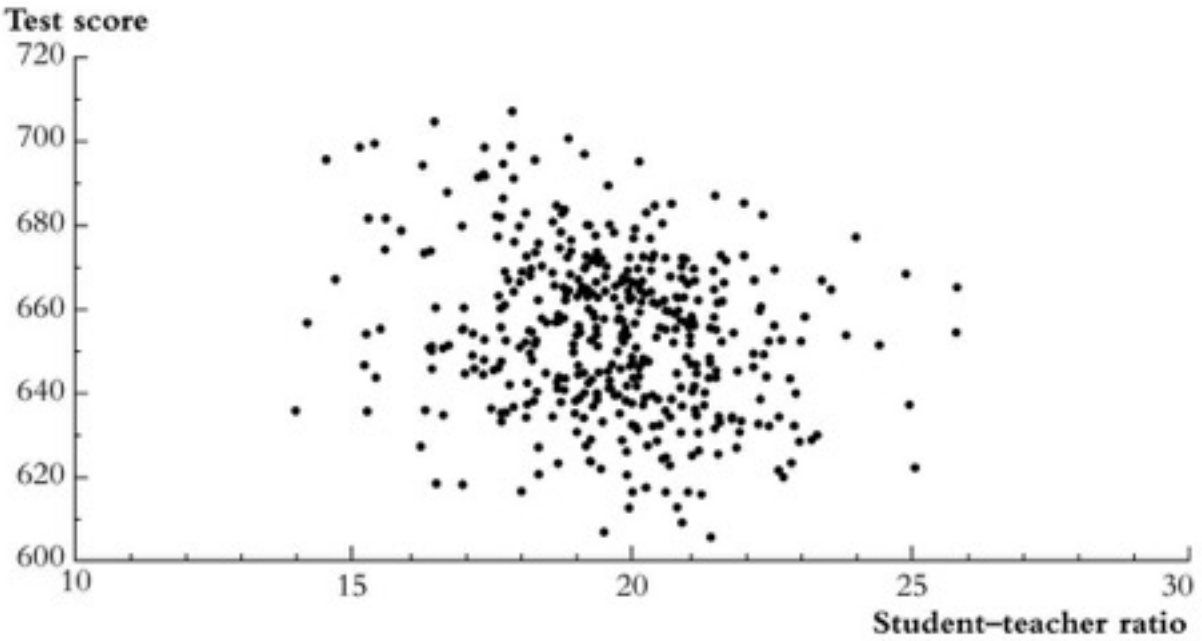
- Con probabilità pari a 0.95, il valore del differenziale stimato cade nel suddetto intervallo

Metodi sintetici per analizzare il legame tra 2 variabili

- Diagramma a nuvola di punti è un grafico delle n osservazioni su due variabili in cui ogni osservazioni è rappresentata dal punto (X_i, Y_i).

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



Covarianza e Correlazione

- La covarianza e la correlazione campionaria sono degli stimatori della covarianza e correlazione della popolazione.
- Come già fatto sostituiamo la media campionaria alla media della popolazione
- Nel caso della covarianza campionaria avremo

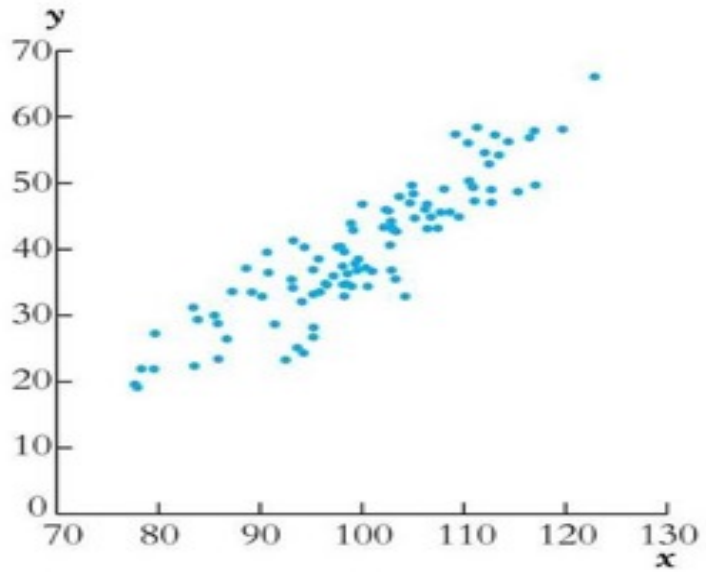
$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

- Nel caso della correlazione campionaria avremo

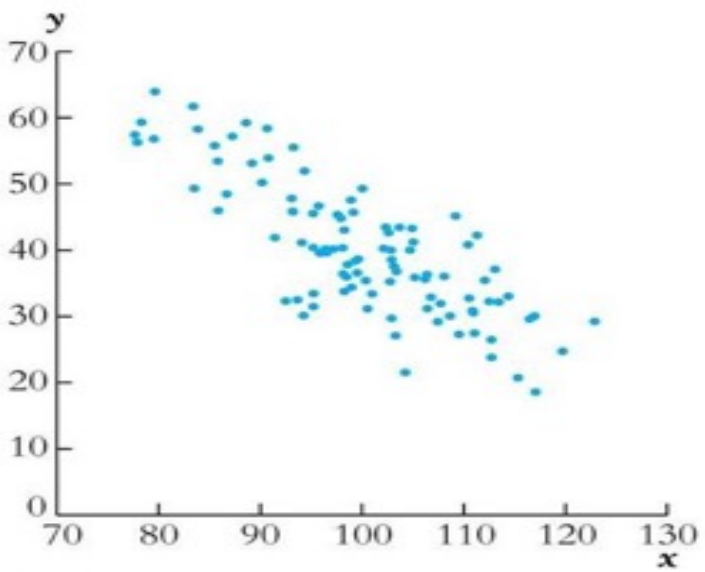
$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

- In generale, la correlazione è pari a ±1 se il diagramma a nuvola di punti è una linea retta

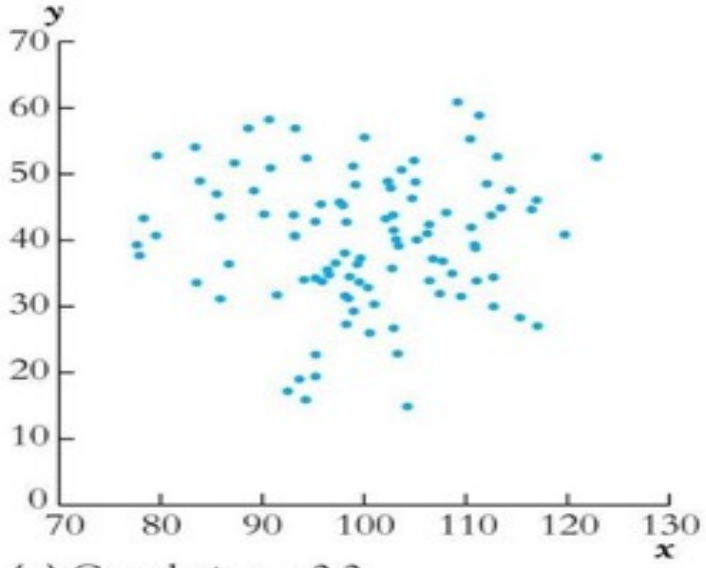
- Sia la covarianza sia varianza campionaria sono consistenti e, quindi, lo stesso vale per il coefficiente di correlazione



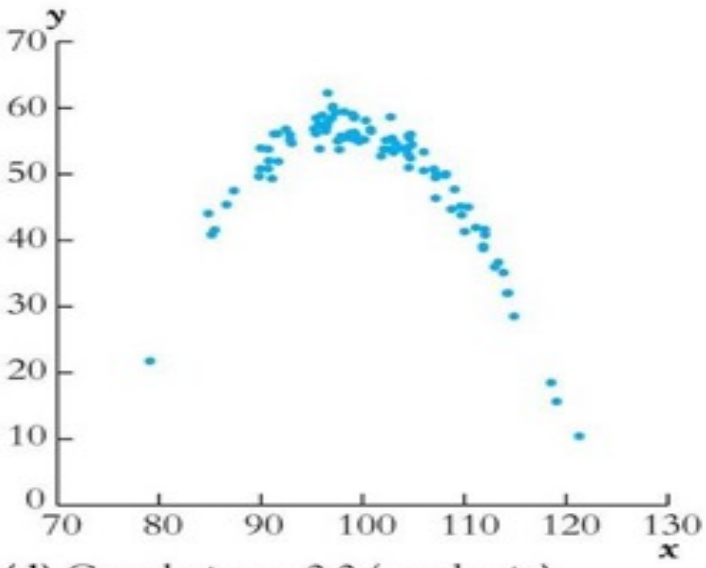
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Esempio

- Consideriamo i dati sull'età e la retribuzione dei lavoratori relativi ad un campione di 184 soggetti
- La deviazione standard campionaria dell'età è s_A=10.49 anni
- La deviazione standard campionaria della retribuzione è s_E=6.44$/h
- La covarianza campionaria è s_{AE}=24.29 ed coefficiente campionario di correlazione è r_{AE}=24.29/(10.49*6.44)=0.36
- Tale risultato mostra un relazione positiva tra le due variabili