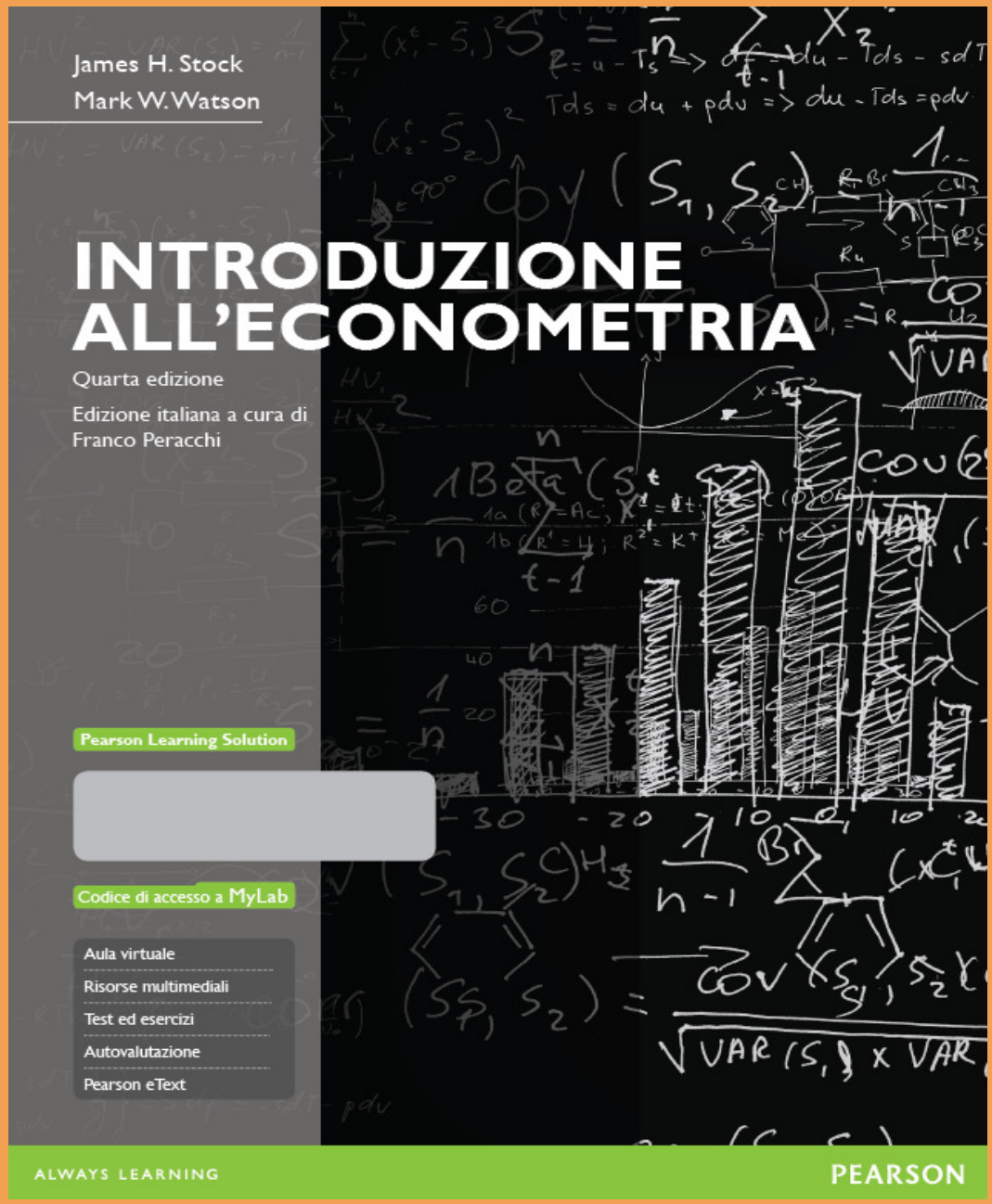
Introduzione all'econometria

Capitoli 1, 2 e 3

L'analisi statistica di dati economici (e correlati)



© 2016 Pearson Italia – Milano, Torino

Breve panoramica del corso

- L'economia suggerisce importanti relazioni, spesso con implicazioni politiche, ma praticamente mai fornisce dimensioni quantitative di effetti causali.
  - Qual è l'effetto *quantitativo* della riduzione delle dimensioni delle classi sui risultati degli studenti?
  - In che modo un anno in più di istruzione può influire sul reddito?
  - Qual è l'elasticità al prezzo delle sigarette?
  - Qual è l'effetto sulla crescita del prodotto interno lordo di un aumento di 1 punto percentuale nei tassi di interesse stabilito dalla Fed?
  - Qual è l'effetto sui prezzi delle case dei miglioramenti di tipo ambientale?

Questo corso tratta dell'uso dei dati per misurare effetti causali

- Idealmente vorremmo un esperimento
  - Quale sarebbe un esperimento per stimare l'effetto della dimensione delle classi sui punteggi nei test standardizzati?
- Ma quasi sempre abbiamo a disposizione soltanto dati osservazionali (non sperimentali).
  - rendimenti dell'istruzione
  - prezzi delle sigarette
  - politica monetaria
- La maggior parte del corso affronta le difficoltà che derivano dall'uso di dati non sperimentali per stimare effetti causali
  - effetti perturbativi (fattori omessi)
  - causalità simultanea
  - "la correlazione non implica causalità"

In questo corso:

- apprenderete metodi per stimare effetti causali usando dati non sperimentali;
- apprenderete l'uso di alcuni strumenti che possono essere impiegati per altri scopi, per esempio la previsione utilizzando serie di dati temporali;
- vi focalizzerete sulle applicazioni – si ricorre alla teoria solo ove necessario per comprendere i motivi alla base dei metodi;
- imparerete a valutare l'analisi di regressione effettuata da altri – questo significa che sarete in grado di leggere e comprendere articoli economici di carattere empirico in altri corsi di tipo economico;
- farete un po' di esperienza pratica con l'analisi di regressione nelle serie di esercizi.

Richiami di probabilità e statistica (Capitoli 2, 3)

- **Problema empirico:** Dimensione della classe e risultato dell'istruzione
  - Domanda: qual è l'effetto sui punteggi nei test (o su un'altra misura di risultato) della riduzione della dimensione delle classi di uno studente per classe? E di 8 studenti per classe?
  - Dobbiamo utilizzare i dati per rispondere (esiste un modo per rispondere a questa domanda *senza* dati?)

I dati dei punteggi nei test della California

Tutti i distretti scolastici K-6 e K-8 della California (n = 420)

Variabili:

- Punteggi nei test del quinto anno (*Stanford-9 achievement test*, combinazione di matematica e lettura), media del distretto
- Rapporto studenti/insegnanti (STR) = numero di studenti nel distretto diviso per numero di insegnanti a tempo pieno equivalente

Primo sguardo ai dati:

(dovreste già sapere come interpretare questa tabella)

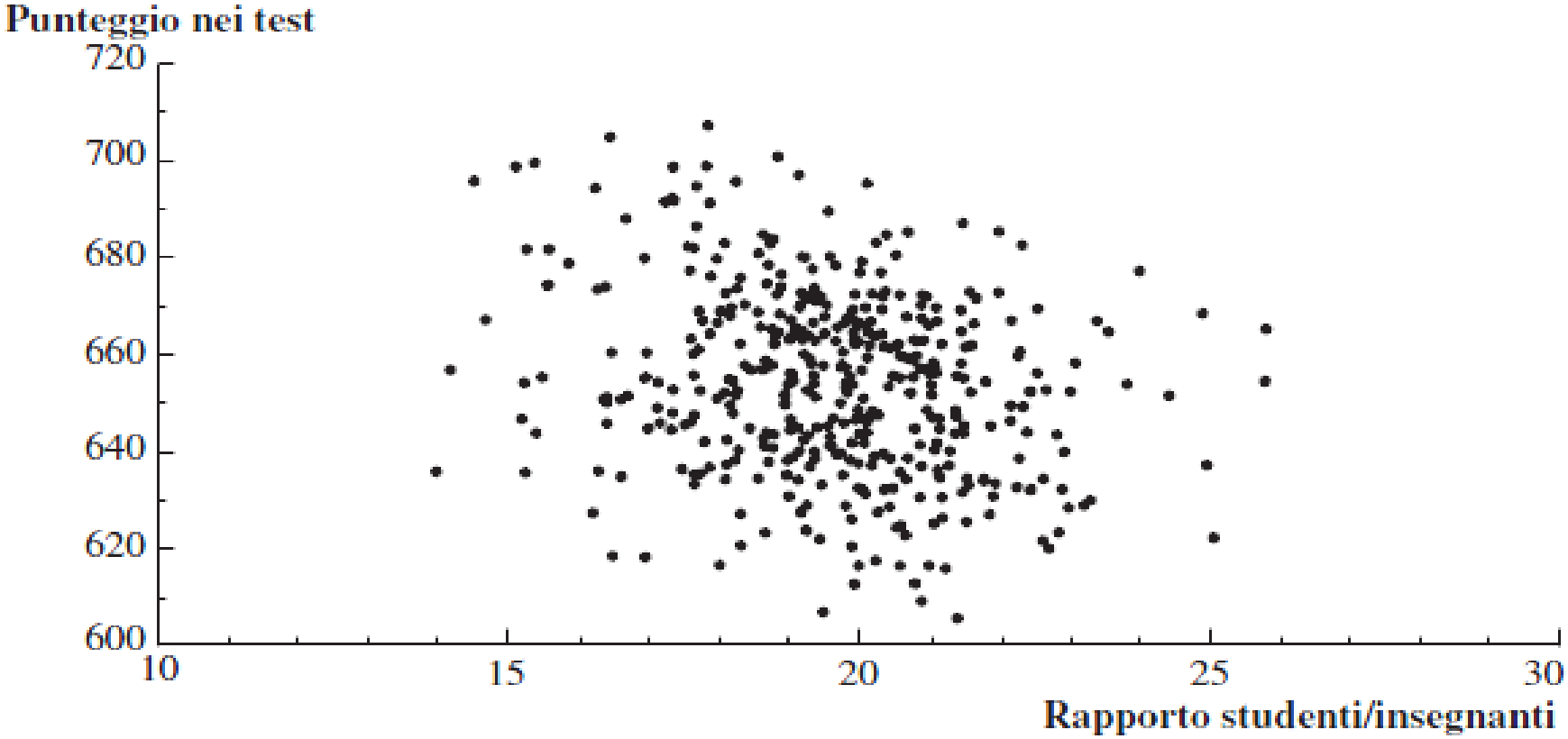
Tabella 4.1 Sintesi della distribuzione del rapporto studenti/insegnanti e del punteggio nei test relativa al quinto grado d'istruzione (quinta elementare) per 420 distretti K-8 in California nel 1998.

			Percentile						
	Media	**Deviazione standard**	10%	25%	40%	50% (mediana)	60%	75%	90%
Rapporto studenti/insegnanti	19,6	1,9	17,3	18,6	19,3	19,7	20,1	20,9	21,9
Punteggio nei test	654,2	19,1	630,4	640,0	649,1	654,5	659,4	666,7	679,1

Questa tabella non ci dice nulla sulla relazione tra punteggi nei test e *STR*.

I distretti con classi più piccole ottengono punteggi più elevati nei test?

Diagramma a nuvola di punteggio nei test e STR



Che cosa mostra questa figura?

Dobbiamo ottenere evidenza numerica che indichi se i distretti con basso STR hanno punteggi nei test più alti – ma come?

1. Confrontare i punteggi nei test nei distretti con basso STR a quelli con alto STR ("***stima***")
2. Sottoporre a verifica l'ipotesi "nulla" che i punteggi medi nei test nei due tipi di distretti siano gli stessi, contro l'ipotesi "alternativa" che siano diversi ("***test di ipotesi***")
3. Stimare un intervallo per la differenza nei punteggi medi nei test, nei distretti con alto vs basso STR ("***intervallo di confidenza***")

Analisi dei dati iniziali: confrontare i distretti con dimensioni delle classi "piccole" (STR < 20) e "grandi" (STR ≥ 20) :

Dimensione classe	Punteggio medio($\overline{Y}$)	Deviazione standard (s_Y)	*n*
Piccola	657,4	19,4	238
Grande	650,0	17,9	182

1. ***Stima*** di Δ = differenza tra medie dei gruppi
2. ***Verifica dell'ipotesi che*** Δ = 0
3. Costruire un ***intervallo di confidenza*** per Δ

1. Stima

$$\bar{Y}_{\text{piccola}} - \bar{Y}_{\text{grande}} = \frac{1}{n_{\text{piccola}}} \sum_{i=1}^{n_{piccola}} Y_i - \frac{1}{n_{\text{grande}}} \sum_{j=1}^{n_{\text{grande}}} Y_j$$

$$= 657{,}4 - 650{,}0$$

$$= 7{,}4$$

È una differenza da considerare grande nel mondo reale?

- Deviazione standard tra i distretti = 19,1
- Differenza tra 60-esimo and 75-esimo percentili della distribuzione dei punteggi nei test = 667,6 – 659,4 = 8,2
- È una differenza sufficientemente grande da risultare importante per discussioni sulla riforma della scuola, per i genitori o per un comitato scolastico?

2. Verifica di ipotesi

Test di differenza tra medie: calcolare la *statistica-t*,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$ (ricordate?)

- dove *SE*($\bar{Y}_s - \bar{Y}_l$) è l'"errore standard "di ($\bar{Y}_s - \bar{Y}_l$), i pedici ***s*** e ***l*** indicano distretti con STR "small" (piccolo) e "large" (grande), e

$$s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2$$

(ecc.)

Calcolare la statistica-*t* per la differenza tra medie

Dim	$\bar{Y}$	s_Y	*n*
piccola	657,4	19,4	238
grande	650,0	17,9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657,4 - 650,0}{\sqrt{\frac{19,4^2}{238} + \frac{17,9^2}{182}}} = \frac{7,4}{1,83} = 4,05$$

|t-act| > 1,96, perciò si rifiuta (al livello di significatività del 5%) l'ipotesi nulla che le due medie coincidano.

3. Intervallo di confidenza

Un intervallo di confidenza al 95% per la differenza tra medie $\Delta = \mu_s - \mu_l$ è

$$(\bar{Y}_s - \bar{Y}_l) \pm 1{,}96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7{,}4 \pm 1{,}96 \times 1{,}83 = (3{,}8,\ 11{,}0)$$

Due affermazioni equivalenti:

1. L'intervallo di confidenza al 95% per Δ non include 0;
2. L'ipotesi nulla H0: $\Delta = 0$, è rifiutata al livello di significatività del 5%.

E ora…

- I meccanismi di stima, verifica di ipotesi e intervalli di confidenza dovrebbero risultare familiari
- Questi concetti si estendono direttamente a regressione e relative varianti
- Prima di passare alla regressione, tuttavia, rivedremo alcuni elementi della teoria alla base di stima, verifica di ipotesi e intervalli di confidenza:
  - Perché queste procedure funzionano, e perché utilizzare proprio queste invece di altre?
  - Rivedremo i fondamenti teorici di statistica ed econometria

Richiami di teoria statistica

1. **Quadro di riferimento probabilistico per l'inferenza statistica**
2. Stima
3. Verifica
4. Intervalli di confidenza

Quadro di riferimento probabilistico per l'inferenza statistica

a) Popolazione, variabile casuale e distribuzione
b) Momenti di una distribuzione (media, varianza, deviazione standard, covarianza, correlazione)
c) Distribuzione condizionata e media condizionata
d) Distributione di un campione di dati estratto a caso da una popolazione: Y_1, …, Y_n

(a) Popolazione, variabile casuale e distribuzione

Popolazione

- Il gruppo o l'insieme di tutte le possibili unità di interesse (distretti scolastici)
- Considereremo le popolazioni infinitamente grandi (∞ è un'approssimazione di "molto grande")

Variabile casuale Y

- Rappresentazione numerica di un risultato casuale (punteggio medio nei test del distretto, STR del distretto)

Distribuzione di Y

- Le probabilità di diversi valori di *Y* che si verificano nella popolazione, per esempio
  Pr[*Y* = 650] (quando *Y* è discreta)
- oppure: le probabilità di insiemi di questi valori, per esempio Pr[640 ≤ *Y* ≤ 660] (quando *Y* è continua).
- In particolare per v.c. discrete e continue è detta **funzione di ripartizione o di probabilità cumulata** la funzione:

$$F(y) = prob\{Y \leq y\}, y \in R$$

(b) Momenti di una distribuzione: media, varianza, deviazione standard, covarianza, correlazione

media = valore atteso (aspettativa) di Y

$= E(Y)$

$= \mu_Y$

varianza $= E(Y - \mu_Y)^2$

$= \sigma_Y^2$

= misura della dispersione quadratica della distribuzione

deviazione standard $= \sqrt{\text{varianza}} = \sigma_Y$

=misura della dispersione della distribuzione nell'unità di misura della v.c. Y

Momenti (continua)

asimmetria = $\dfrac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$

= misura di asimmetria di una distribuzione

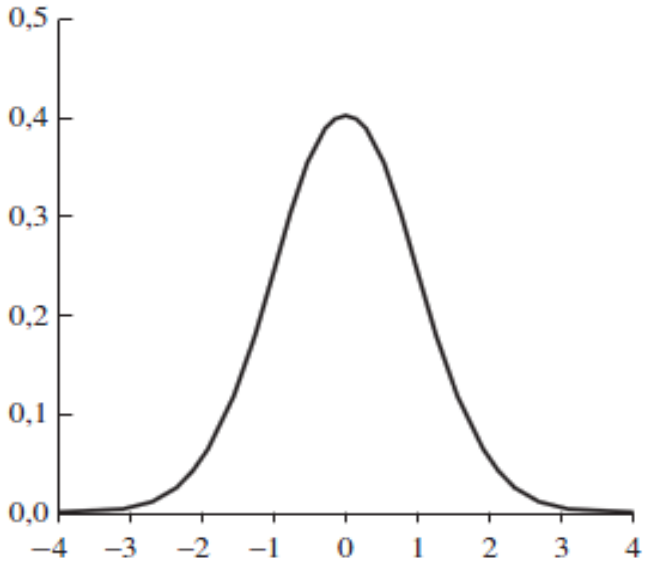
- *asimmetria* = 0: la distribuzione è simmetrica
- *asimmetria* > (<) 0: la distribuzione ha una coda lunga destra (sinistra)

curtosi = $\dfrac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$

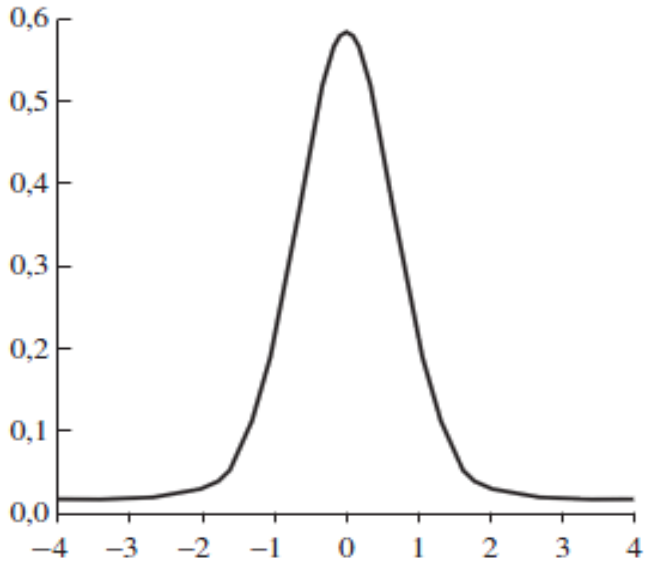
= misura di massa nelle code

= misura di probabilità di valori grandi

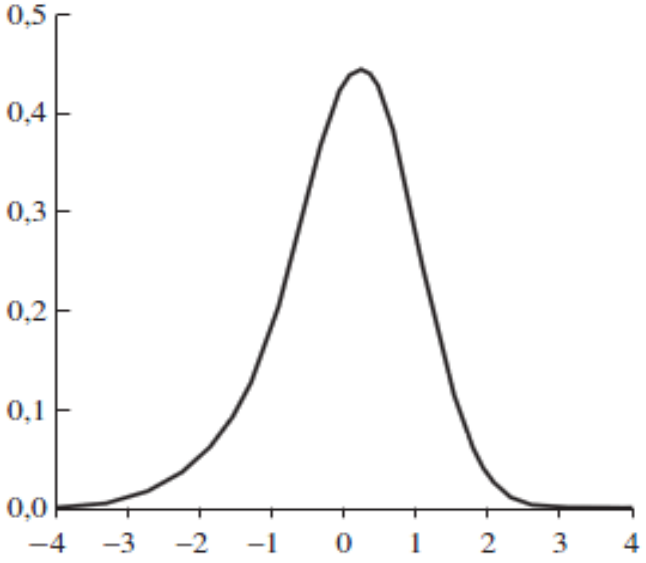
- *curtosi* = 3: distribuzione normale
- *asimmetria* > 3: code pesanti ("***leptocurtica***")



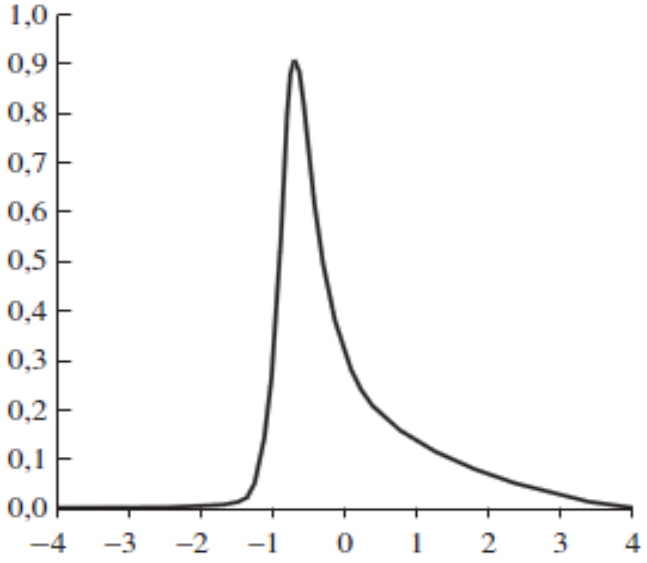
(a) Asimmetria = 0, Curtosi = 3



(b) Asimmetria = 0, Curtosi = 20



(c) Asimmetria = –0,1, Curtosi = 5



(d) Asimmetria = 0,6, Curtosi = 5

2 variabili casuali: distribuzioni congiunte e covarianza

- Le variabili casuali X e Z hanno una ***distribuzione congiunta***
- La ***covarianza*** tra X e Z è

  $\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$
- La covarianza è una misura dell'associazione lineare tra X e Z; le sue unità sono unità di X × unità di Z
- $\text{cov}(X,Z) > 0$ significa una relazione positiva tra X e Z
- Se X e Z sono indipendentemente distribuite, allora $\text{cov}(X,Z) = 0$ (ma non vale il vice versa!!)
- La covarianza di una variabile casuale con se stessa è la sua varianza:

  $\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$

La covarianza tra *Punteggio nei test* e *Rapporto studenti/insegnanti* è negativa:

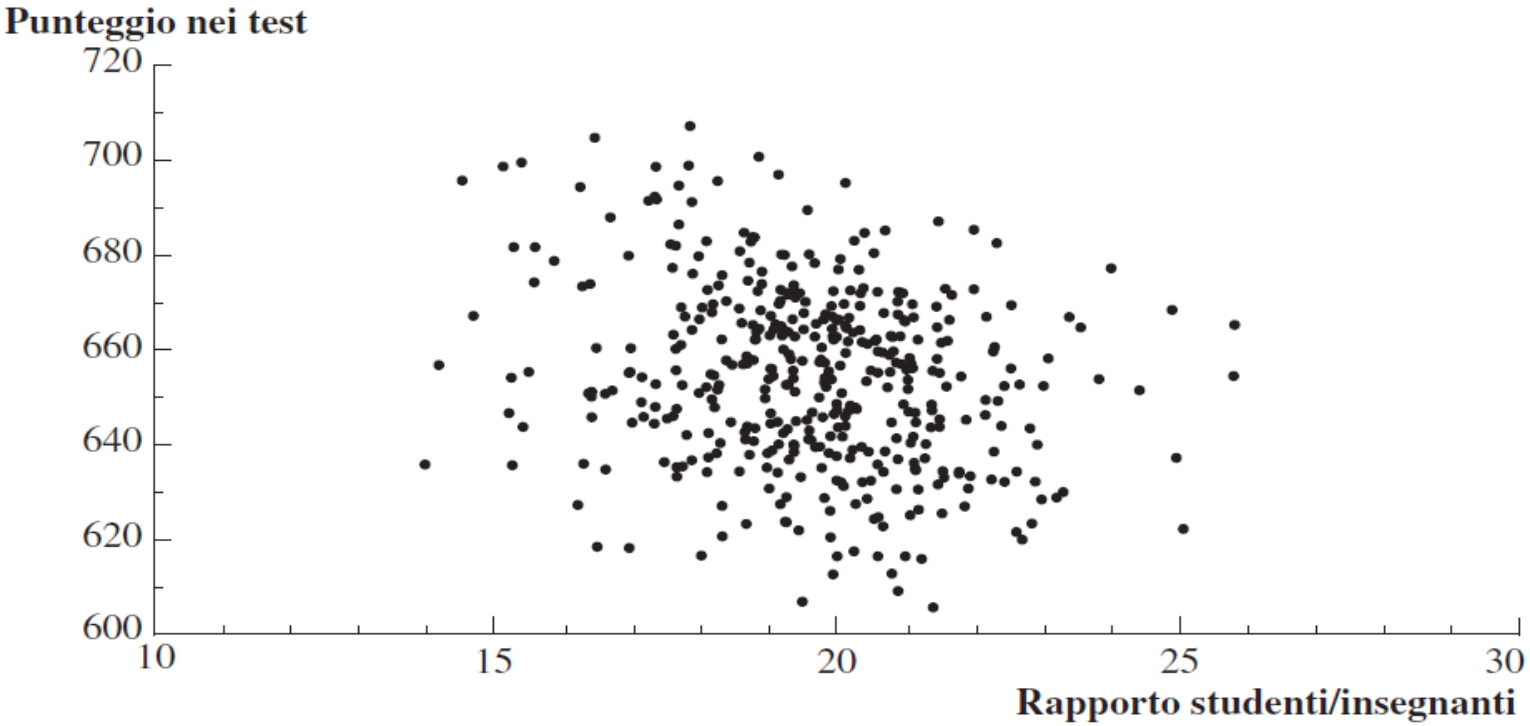


Figura 4.2

Diagramma a nuvola del punteggio nei test e del rapporto studenti/insegnanti (dati relativi ai distretti scolastici della California).

Dati per i 420 distretti scolastici della California. C'è una debole relazione negativa tra il rapporto studenti/insegnanti e il punteggio nei test: la correlazione campionaria è pari a –0,23.

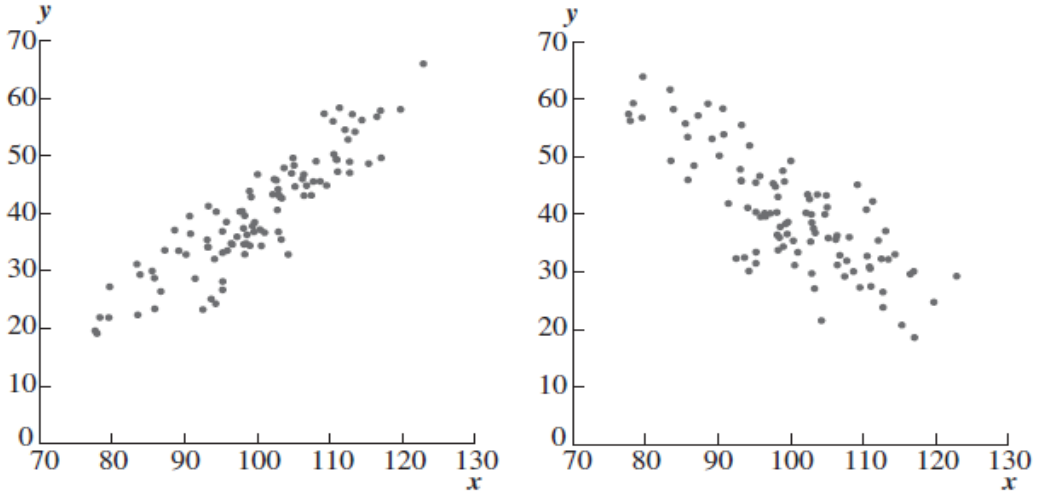
E così la ***correlazione***…

Il ***coefficiente di correlazione*** è definito in termini di covarianza:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}$$

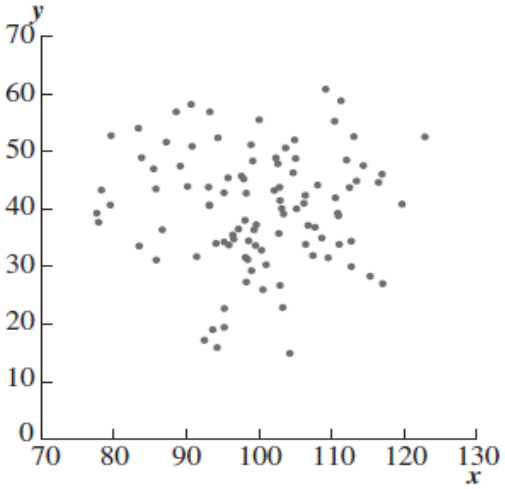
- $-1 \leq \text{corr}(X,Z) \leq 1$
- $\text{corr}(X,Z) = 1$ significa associazione lineare positiva perfetta
- $\text{corr}(X,Z) = -1$ significa associazione lineare negativa perfetta
- $\text{corr}(X,Z) = 0$ significa che non c'è associazione lineare

Il coefficiente di correlazione misura l'associazione lineare

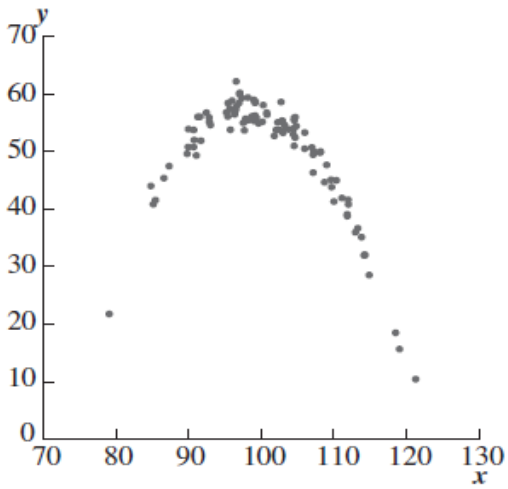


(a) Correlazione = +0,9

(b) Correlatzone = –0,8



(c) Correlazione = 0,0



(d) Correlazione = 0,0 (quadratica)

(c) Distribuzione condizionata e media condizionata

Distribuzione condizionata

- La distribuzione di *Y* dato il valore (o i valori) di un'altra variabile casuale *X*
- Es: la distribuzione dei punteggi nei test dato STR < 20

Valore atteso condizionato e momento condizionato

- *media condizionata* = media della distribuzione condizionata
  = $E(Y|X = x)$ (***concetto e notazione importanti***)
- *varianza condizionata* = varianza della distribuzione condizionata
- *Esempio*: $E(\textit{Punteggio test}|\textit{STR} < 20)$ = media dei punteggi nei test tra i distretti con dimensioni delle classi piccole

La differenza in media è la differenza tra le medie di due distribuzioni condizionate:

Media condizionata (continua)

$\Delta = E(Punteggio\ test|STR < 20) - E(Punteggio\ test|STR \geq 20)$

Altri esempi di media condizionata:

- Salari di tutte le lavoratrici femmine (Y = salari, X = genere)
- Tasso di mortalità di pazienti che ricevono una cura sperimentale (Y = vivo/morto; X = trattato/non trattato)
- Se $E(X|Z)$ = costante, allora corr(X,Z) = 0 (tuttavia non vale necessariamente il vice versa)

La media condizionata è un termine (forse nuovo) utilizzato per il concetto familiare di media di gruppo

(d) Distribuzione di un campione di dati estratto a caso da una popolazione: $Y_1,\ldots, Y_n$

Assumeremo un campionamento casuale semplice

- Scegliere a caso un individuo (distretto, unità) dalla popolazione

Casualità e dati

- Prima della selezione del campione, il valore di Y è casuale perché l'individuo selezionato è casuale
- Una volta selezionato l'individuo e osservato il valore di Y, Y è soltanto un numero – non casuale
- Il data set è (Y_1, Y_2,…, Y_n), dove Y_i = valore di Y per l'i-esimo individuo (distrettto, unità) del campione

Distribuzione di $Y_1,\ldots,Y_n$ sotto campionamento casuale semplice

- Poiché gli individui n. 1 e 2 sono selezionati a caso, il valore di Y_1 non contiene informazioni per Y_2. Quindi:
  - Y_1 e Y_2 sono ***indipendentemente distribuiti***
  - Y_1 e Y_2 provengono dalla stessa distribuzione, cioè Y_1, Y_2 sono ***identicamente distribuiti***
  - Ovvero, sotto campionamento casuale semplice, Y_1 e Y_2 sono indipendentemente e identicamente distribuiti (***i.i.d.***).
  - Più in generale, sotto campionamento casuale semplice, $\{Y_i\}$, $i = 1,\ldots, n$, sono i.i.d.

Questo quadro consente rigorose inferenze statistiche sui momenti di distribuzione utilizzando un campione di dati tratto dalla popolazione stessa…

1. Quadro probabilistico per inferenza statistica
2. **Stima**
3. Verifica
4. Intervalli di confidenza

Stima

$\overline{Y}$ è lo stimatore naturale della media $E(Y)=\mu_Y$. Ma:

a) quali sono le proprietà di $\overline{Y}$?

b) Perché dovremmo usare $\overline{Y}$ anziché un altro stimatore?

- Y_1 (prima osservazione)
- forse pesi non uniformi – non media semplice
- mediana(Y_1,…, Y_n)

Il punto di partenza è la distribuzione campionaria di $\overline{Y}$ …

(a) La distribuzione campionaria di $\overline{Y}$

$\overline{Y}$ è una variabile casuale e le sue proprietà sono determinate dalla ***distribuzione campionaria*** di $\overline{Y}$

- Gli individui nel campione sono estratti a caso.
- Quindi i valori di $(Y_1,\ldots,Y_n)$ sono casuali
- Quindi funzioni di $(Y_1,\ldots,Y_n)$, come $\overline{Y}$, sono casuali: se si fosse estratto un campione diverso, avrebbero assunto valori differenti
- La distribuzione di $\overline{Y}$ su diversi possibili campioni di dimensione *n* si chiama ***distribuzione campionaria*** di $\overline{Y}$.
- La media e la varianza di $\overline{Y}$ sono la media e la varianza della sua distribuzione campionaria, $E(\overline{Y})$ e var($\overline{Y}$).
- Il concetto di distribuzione campionaria è alla base di tutta l'econometria.

La distribuzione campionaria di $\overline{Y}$ (continua)

Esempio: Si supponga che *Y* assuma il valore 0 o 1 (variabile casuale di ***Bernoulli***) con la distribuzione di probabilità

Pr[*Y* = 0] = 0,22, Pr(*Y* =1) = 0,78

Allora

$E(Y) = p \times 1 + (1 - p) \times 0 = p = 0{,}78$

$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1 - p)$ *[ricordate?]*

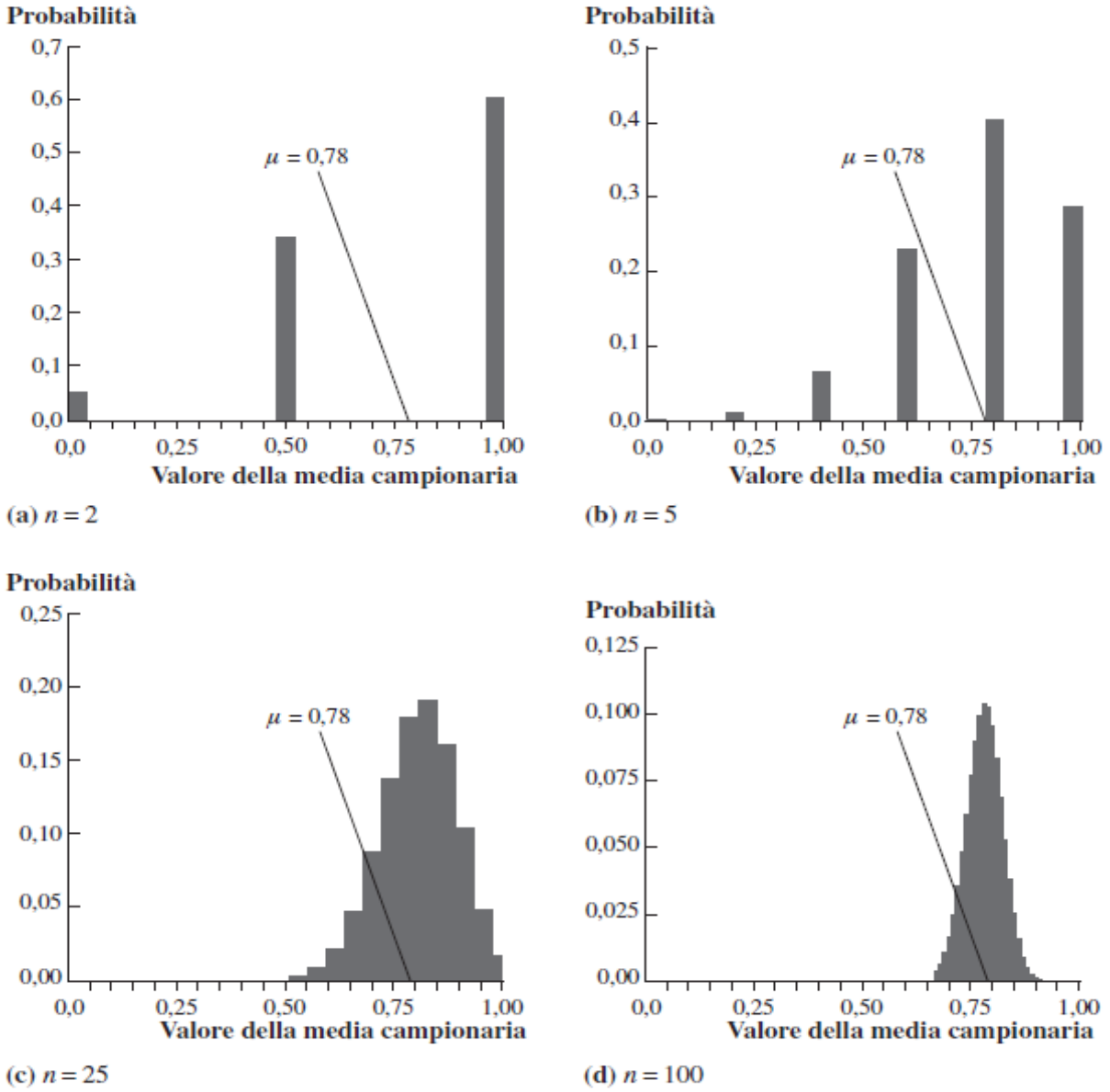
$= 0{,}78 \times (1 - 0{,}78) = 0{,}1716$

La distribuzione campionaria di $\overline{Y}$ dipende da *n*.

Si consideri *n* = 2. La distribuzione campionaria di $\overline{Y}$ è

- Pr($\overline{Y}$ = 0) = $0{,}22^2$ = 0,0484
- Pr($\overline{Y}$ = ½) = 2 × 0,22 × 0,78 = 0,3432
- Pr($\overline{Y}$ = 1) = $0{,}78^2$ = 0,6084

Distribuzione campionaria di $\overline{Y}$ quando Y è di Bernoulli (p = 0,78):



Cose che vogliamo sapere sulla distribuzione campionaria:

- Qual è la media di $\overline{Y}$?
  - Se $E(\overline{Y}) = \mu = 0{,}78$, allora $\overline{Y}$ è uno stimatore ***non distorto*** di μ
- Qual è la varianza di $\overline{Y}$?
  - In che modo var($\overline{Y}$) dipende da n (famosa formula $1/n$)
- Si avvicina a μ quando n è grande?
  - Legge dei grandi numeri: $\overline{Y}$ è uno stimatore ***consistente*** di μ
- $\overline{Y} - \mu$ assume forma a campana per n grande... questo è vero in generale?
  - In effetti, $\overline{Y} - \mu$ è approssimato da una distribuzione normale per n grande (teorema limite centrale)

Media e varianza della distribuzione campionaria di $\bar{Y}$

- Caso generale – cioè, per Y_i i.i.d. da qualsiasi distribuzione, non solo di Bernoulli:
- media: $E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$
- Varianza: $\text{var}(\bar{Y}) = E[\bar{Y} - E(\bar{Y})]^2$

$$= E[\bar{Y} - \mu_Y]^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_Y)\right]^2$$

quindi $\text{var}(\bar{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i - \mu_Y)(Y_j - \mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(Y_i, Y_j)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

Media e varianza della distribuzione campionaria di $\bar{Y}$ *(continua)*

$$E(\bar{Y}) = \mu_Y$$

$$\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Implicazioni:

1. $\bar{Y}$ è uno stimatore non distorto di μ_Y (cioè $E(\bar{Y}) = \mu_Y$)
2. var($\bar{Y}$) è inversamente proporzionale a *n*
    1. la dispersione della distribuzione campionaria è proporzionale a $1/\sqrt{n}$
    2. Quindi l'incertezza campionaria associata con $\bar{Y}$ è proporzionale a $1/\sqrt{n}$ (grandi campioni, meno incertezza, ma legge con radice quadrata)

Distribuzione campionaria di $\overline{Y}$ quando *n* è grande

Per piccoli campioni, la distribuzione di $\overline{Y}$ è complicata, ma se *n* è grande, la distribuzione campionaria è semplice!

1. All'aumentare di *n*, la distribuzione di $\overline{Y}$ diventa più strettamente centrata su μ_Y (*legge dei grandi numeri*)

1. Inoltre, la distribuzione di $\overline{Y} - \mu_Y$ diventa normale (*teorema limite centrale*)

***Legge dei grandi numeri*:**

Uno stimatore è ***consistente*** se la probabilità che ricada entro un intervallo del vero valore della popolazione tende a uno all'aumentare della dimensione del campione.

Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $\sigma_Y^2 < \infty$, allora $\overline{Y}$ è uno stimatore consistente di μ_Y, cioè

$$\Pr[|\overline{Y} - \mu_Y| < \mu] \rightarrow 1 \text{ per } n \rightarrow \infty$$

che si può scrivere $\overline{Y} \xrightarrow{p} \mu_Y$

("$\overline{Y} \xrightarrow{p} \mu_Y$" significa "$\overline{Y}$ converge in probabilità a μ_Y").

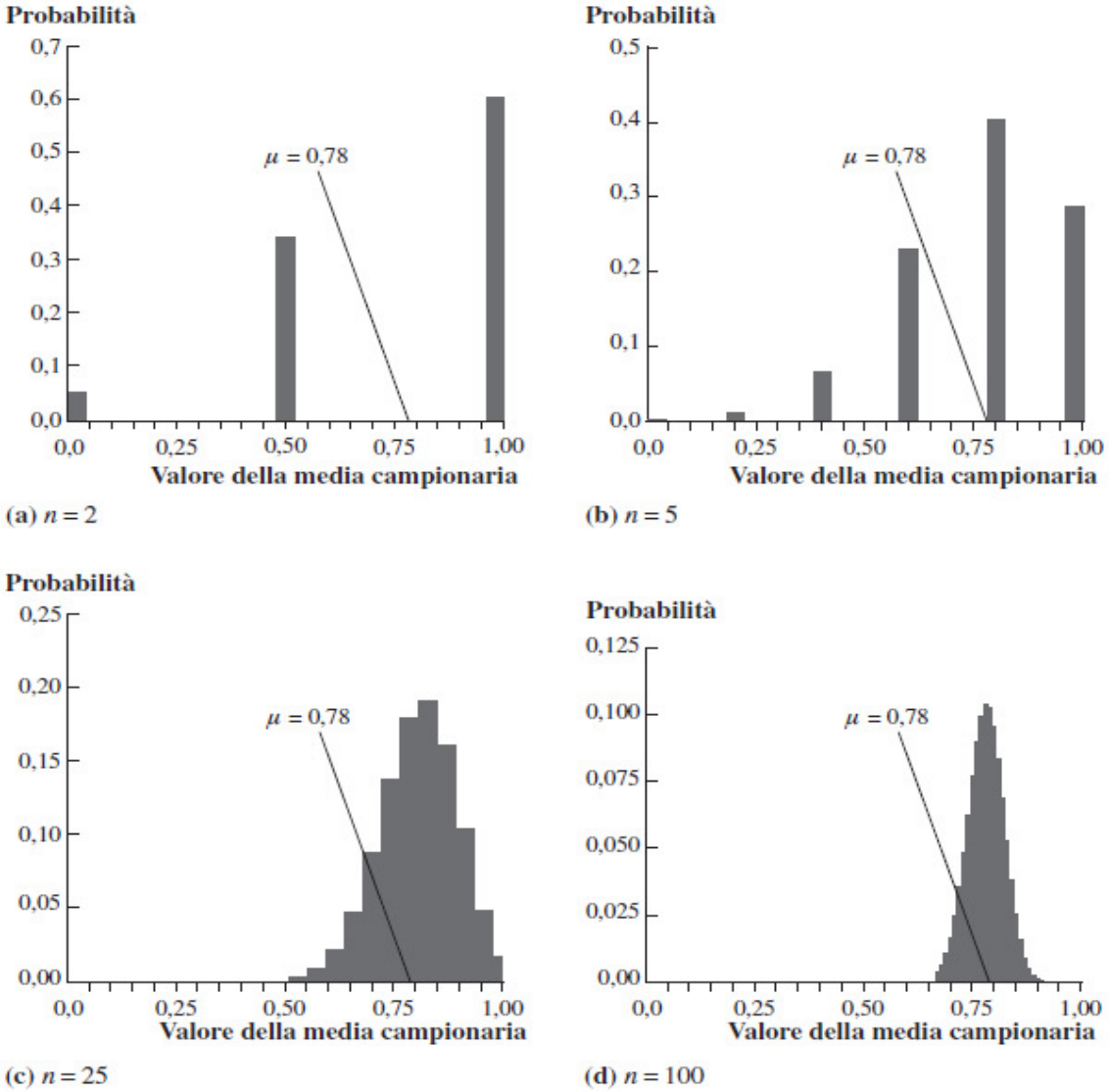
(*matematica*: per $n \rightarrow \infty$, var($\overline{Y}$) = $\frac{\sigma_Y^2}{n} \rightarrow 0$, il che implica che $\Pr[|\overline{Y} - \mu_Y| < \varepsilon] \rightarrow 1$.)

Teorema limite centrale (TLC):

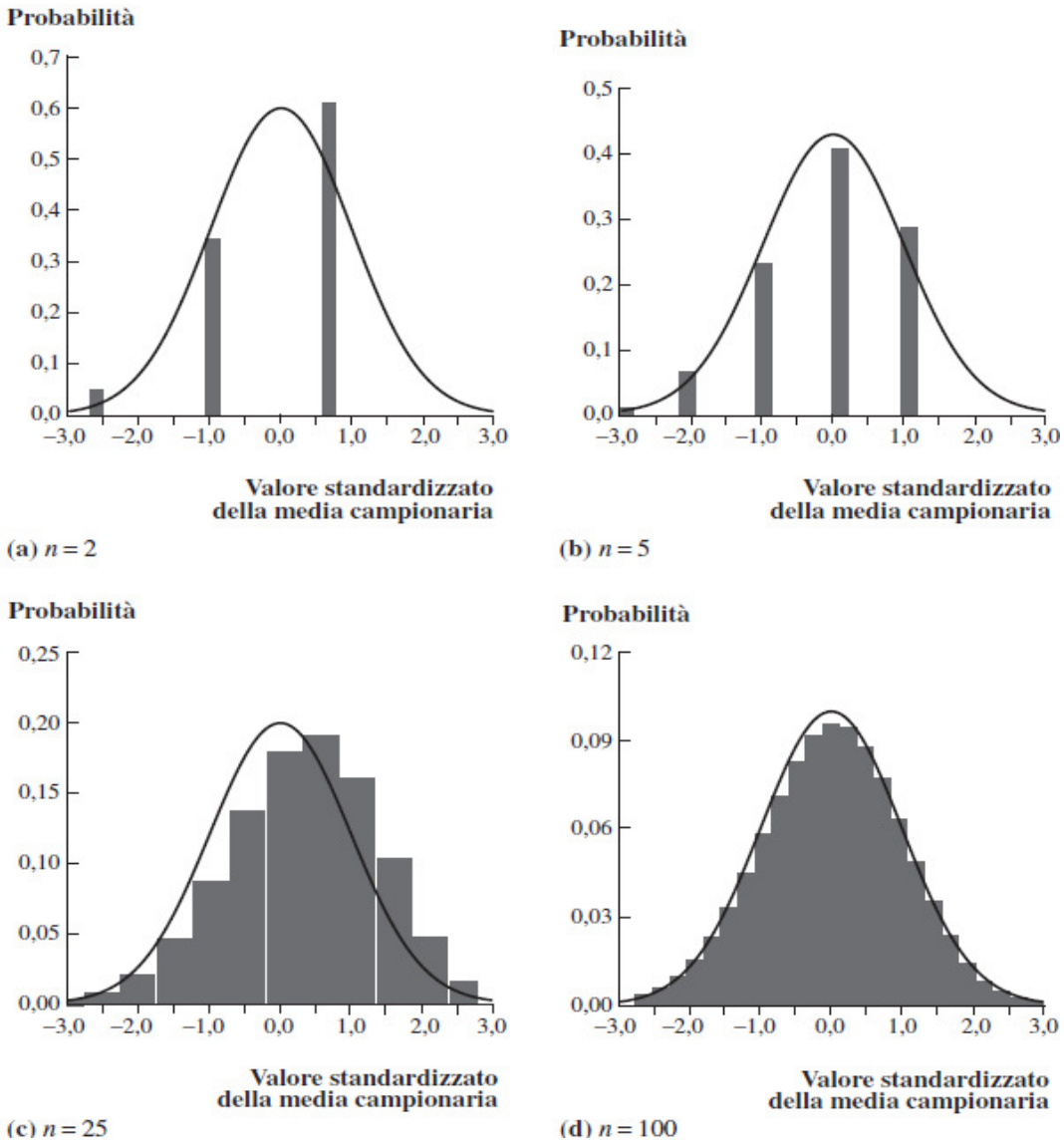
Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $0 < \sigma_Y^2 < \infty$, allora quando *n* è grande la distribuzione di $\overline{Y}$ è bene approssimata da una distribuzione normale.

- $\overline{Y}$ è distribuita approssimativamente come $N(\mu_Y, \frac{\sigma_Y^2}{n})$ ("distribuzione normale con media μ_Y e varianza σ_Y^2/n")
- $\sqrt{n}\,(\overline{Y} - \mu_Y)/\sigma_Y$ è distribuita approssimativamente come $N(0,1)$ (normale standard)
- **Cioè, $\overline{Y}$ "standardizzata" = $\dfrac{\overline{Y} - E(\overline{Y})}{\sqrt{\mathrm{var}(\overline{Y})}}$ = $\dfrac{\overline{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$ è distribuita approssimativamente come *N*(0,1)**
- **Più grande è *n*, migliore è l'approssimazione.**

Distribuzione campionaria di $\overline{Y}$ quando Y è di Bernoulli, $p = 0{,}78$:



Stesso esempio: distribuzione campionaria di $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$:



Riepilogo: distribuzione campionaria di $\bar{Y}$

Per $Y_1,\ldots,Y_n$ i.i.d. con $0 < \sigma_Y^2 < \infty$,

- La distribuzione campionaria esatta (campione finito) di $\bar{Y}$ ha media μ_Y ("$\bar{Y}$ è uno stimatore non distorto di μ_Y") e varianza σ_Y^2/n
- Al di là di media e varianza, la distribuzione esatta di $\bar{Y}$ è complessa e dipende dalla distribuzione di Y (la distribuzione della popolazione)
- Quando n è grande, la distribuzione campionaria si semplifica:
  - $\bar{Y} \xrightarrow{p} \mu Y$ (Legge dei grandi numeri)
  - $\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$ è approssimata da $N(0,1)$ (TLC)

(b) Perché usare $\bar{Y}$ per stimare μ_Y?

- $\bar{Y}$ è non distorto: $E(\bar{Y}) = \mu Y$
- $\bar{Y}$ è consistente: $\bar{Y} \xrightarrow{p} \mu Y$
- $\bar{Y}$ è lo stimatore "dei minimi quadrati" di μY; $\bar{Y}$ risolve

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

quindi $\bar{Y}$ minimizza la somma dei quadrati dei "residui": *derivazione facoltativa (cfr. anche Appendice 3.2)*

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = 2\sum_{i=1}^{n}(Y_i - m)$$

Si pone la derivata a zero e si denota il valore ottimale di m con : $\hat{m}$

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n} \hat{m} = n\hat{m} \quad \text{o} \quad \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Perché usare $\overline{Y}$ per stimare μ_Y *(continua)*

- $\overline{Y}$ ha una varianza minore di tutti gli altri *stimatori lineari non distorti*: si consideri lo stimatore $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, dove gli $\{a_i\}$ sono tali che $\hat{\mu}_Y$ è non distorto; allora var($\overline{Y}$) ≤ var($\hat{\mu}_Y$)

  (dimostrazione: Capitolo 17 del volume stampato)
- $\overline{Y}$ non è l'unico stimatore di μ_Y – vi viene in mente un caso in cui potrebbe essere preferibile utilizzare la mediana?

1. Quadro di riferimento probabilistico per l'inferenza statistica
2. Stima
3. **Verifica di ipotesi**
4. Intervalli di confidenza

Verifica di ipotesi

Il problema della ***verifica di ipotesi*** (per la media): prendere una decisione preliminare in base all'evidenza disponibile che un'ipotesi nulla è vera, o che è vera, invece, un'ipotesi alternativa. Cioè verificare

- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (monodirezionale, $>$)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (monodirezionale, $<$)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (bidirezionale)

Terminologia per la verifica di ipotesi statistiche

valore-p = probabilità di ricavare una statistica (per es. $\overline{Y}$) sfavorevole all'ipotesi nulla almeno quanto il valore effettivamente calcolato con i dati, supponendo che l'ipotesi nulla sia corretta.

Il ***livello di significatività*** di un test è una probabilità predeterminata di rifiutare in modo errato l'ipotesi nulla, quando invece è corretta.

Calcolo del valore-p in base a $\overline{Y}$:

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

Dove $\overline{Y}^{act}$ è il valore di $\overline{Y}$ effettivamente osservato (non casuale)

Calcolo del valore-p (continua)

- Per calcolare il valore-*p* è necessario conoscere la distribuzione campionaria di $\overline{Y}$, che è complessa se *n* è piccolo.
- Se *n* è grande, si può usare l'approssimazione normale (TLC):

valore-*p* = $\Pr_{H_0}[|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|]$

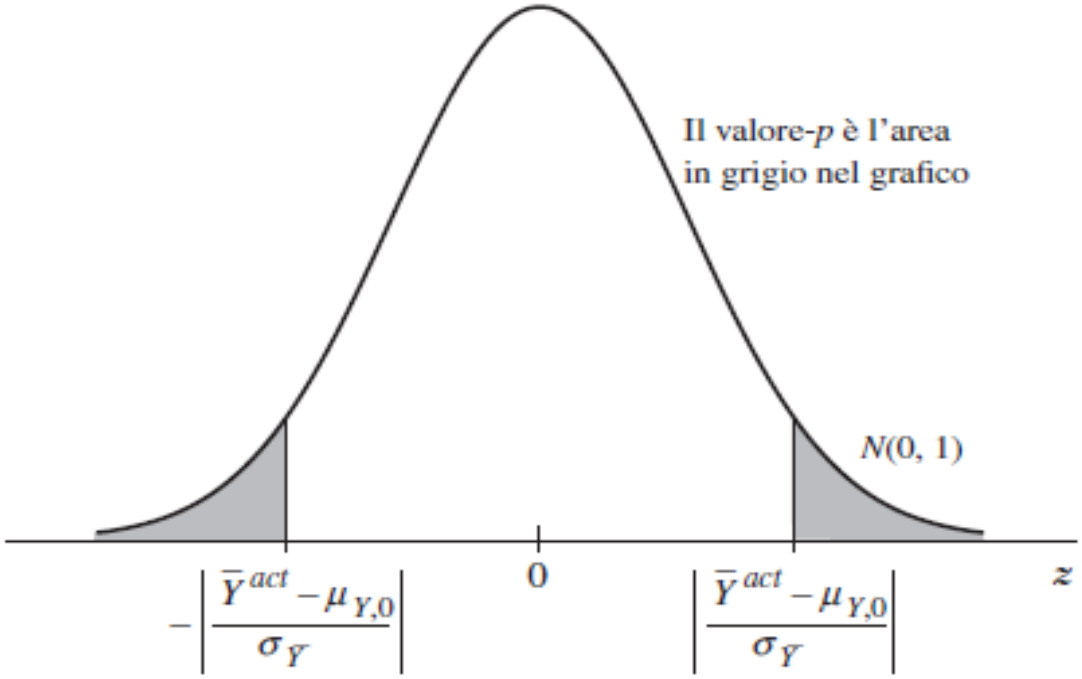
= $\Pr_{H_0}[|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|]$

= $\Pr_{H_0}[|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}}| > |\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}|]$

$\cong$ probabilità sotto code *N*(0,1) sin+destra

dove $\sigma_{\overline{Y}}$ = dev. std della distribuzione di $\overline{Y} = \sigma_Y / \sqrt{n}$.

Calcolo del valore-p con σ_Y nota:



- Per *n* grande, valore-*p* = probabilità che una variabile casuale *N*(0,1) ricada al di fuori di $|(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}|$
- In pratica, $\sigma_{\bar{Y}}$ è ignota – deve essere stimata

Stimatore della varianza di Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“varianza campionaria di } Y\text{”}$$

Se $(Y_1,\ldots,Y_n)$ sono i.i.d. e $E(Y^4) < \infty$, allora

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

Perché si applica la legge dei grandi numeri?

- Perché s_Y^2 è una media campionaria; cfr. Appendice 3.3
- Nota tecnica: si assume $E(Y^4) < \infty$ perché la media non è di Y_i, ma del suo quadrato; cfr. Appendice 3.3.

Calcolo del valore-p con σ_Y^2 *stimato*

$$\textit{valore-p} = \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}\right|\right]$$

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

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y / \sqrt{n}}\right|\right] \text{ (}n\text{ grande)}$$

quindi

$$\text{valore-}p = \Pr_{H_0}\left[|t| > |t^{act}|\right] \quad \textbf{(} \sigma_Y^2 \textbf{ stimato)}$$

$\cong$ probabilità sotto code normali al di fuori di $|t^{act}|$

dove $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$ (la consueta statistica t)

Che collegamento c'è tra il valore-*p* e il livello di significatività?

- Il livello di significatività è specificato in anticipo. Per esempio, se tale livello è del 5%,
  - si rifiuta l'ipotesi nulla se $|t| \geq 1{,}96$.
  - in modo equivalente, la si rifiuta se $p \leq 0{,}05$.
  - Il valore-*p* è detto talvolta ***livello di significatività marginale***.
  - Spesso è meglio comunicare il valore-*p* che limitarsi a indicare se un test rifiuta o no – il valore-*p* contiene più informazioni di un semplice risultato "sì/no" in riferimento a un test.

A questo punto potreste chiedervi...
Che ne è della tabella-*t* e dei gradi di libertà?

Digressione: la distribuzione *t* di Student

Se Y_i, $i = 1,\ldots,n$ sono i.i.d. $N(\mu_Y, \sigma_Y^2)$, allora la statistica t ha la distribuzione t di Student *con* $n - 1$ gradi di libertà.

I valori critici della distribuzione t di Student sono elencati in tutti i libri di statistica. Ricordate la procedura?

1. Calcolare la statistica t
2. Calcolare i gradi di libertà, $n - 1$
3. Cercare il valore critico al 5%
4. Se la statistica t supera (in valore assoluto) questo valore critico, rifiutare l'ipotesi nulla.

Commenti su questa procedura e sulla distribuzione *t* di Student

1. La teoria della distribuzione *t* è stata uno dei primi trionfi della statistica matematica. È davvero sorprendente: se *Y* è i.i.d. e normale, allora è possibile conoscere la distribuzione *esatta*, *a campione finito* della statistica *t* – è la *t* di Student. Perciò si possono costruire intervalli di confidenza (usando il valore critico *t* di Student) che hanno *esattamente* lo stesso tasso di copertura, indipendentemente dalla dimensione del campione. Questo risultato è stato molto utile in tempi in cui "calcolatore" era una posizione lavorativa, la raccolta di dati era costosa e il numero di osservazioni si aggirava attorno alla decina. È anche un risultato concettualmente splendido, e anche la matematica è molto elegante – il che probabilmente spiega perché i docenti amano insegnare la distribuzione *t*. Tuttavia...

Commenti sulla distribuzione *t* di Student *(continua)*

2. Se la dimensione del campione è moderata (varie dozzine) o grande (centinaia o più), la differenza tra la distribuzione *t* e i valori critici N(0,1) è trascurabile. Riportiamo di seguito alcuni valori critici al 5% per test bidirezionali:

gradi di libertà (n – 1)	valore critico distribuzione *t* al 5%
10	2,23
20	2,09
30	2,04
60	2,00
∞	1,96

Commenti sulla distribuzione *t* di Student *(continua)*

3. Perciò la distribuzione *t* di Student è di interesse soltanto quando la dimensione del campione è molto piccola; ma in quel caso, affinché sia corretta, è necessario assicurarsi che la distribuzione di *Y* sia normale. Per dati economici, l'assunzione di normalità è raramente credibile. Ecco le distribuzioni di alcuni dati economici.
   - Pensate che i guadagni abbiano distribuzione normale?
   - Supponete di avere un campione di $n = 10$ osservazioni da una di queste distribuzioni – vi sentireste di usare la distribuzione *t* di Student?

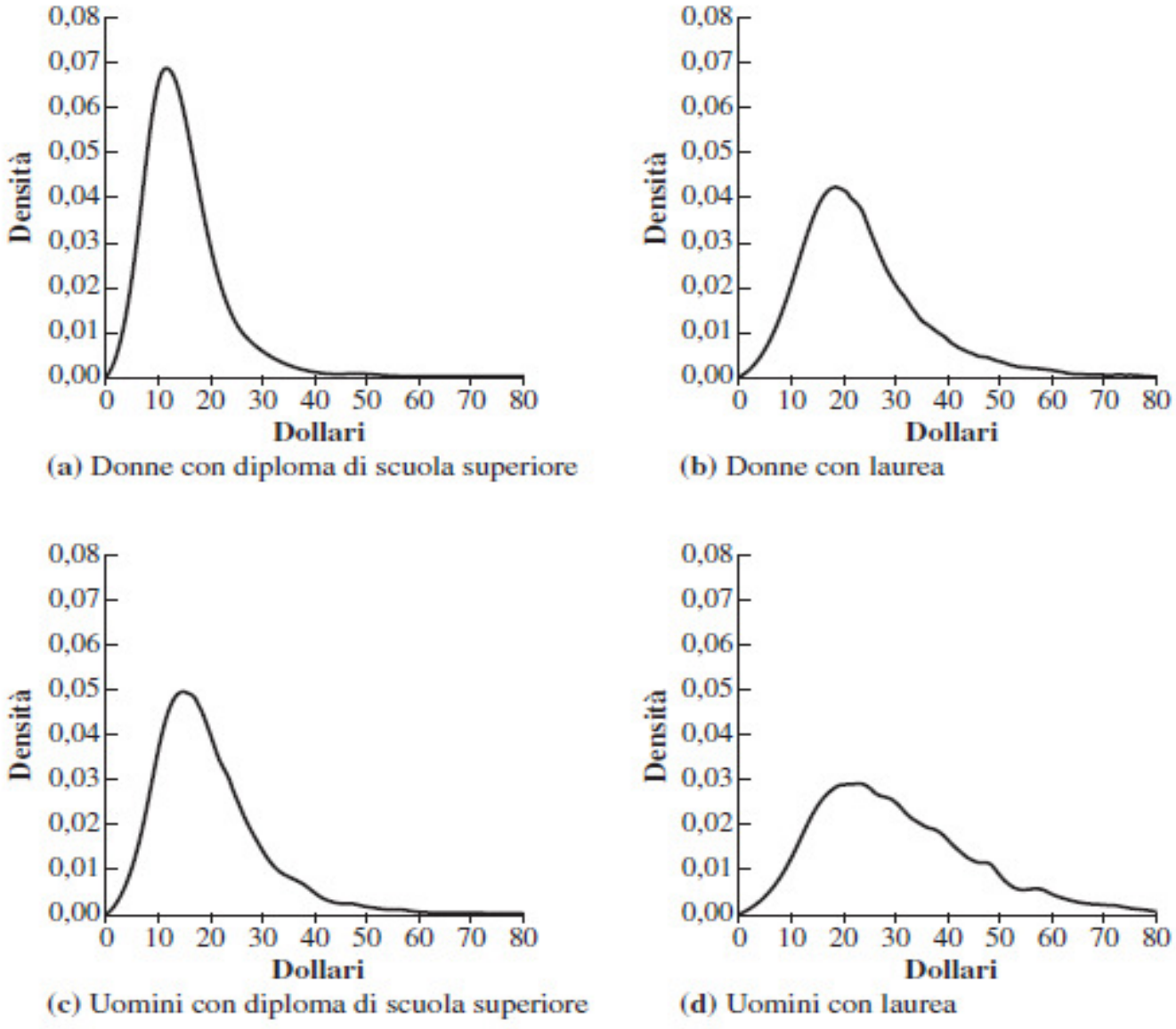


Figura 2.4

Distribuzione condizionata delle retribuzioni orarie medie dei lavoratori statunitensi a tempo pieno nel 2004, dati il livello d'istruzione e il sesso.

Le quattro distribuzioni delle retribuzioni sono per uomini e donne, per coloro che hanno solo un diploma di scuola superiore (a e c) e coloro che hanno una laurea (b e d).

Commenti sulla distribuzione *t* di Student *(continua)*

4. Forse non lo sapete. Considerate la statistica *t* che verifica l'ipotesi che due medie (gruppi *s*, *l*) siano uguali:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

Anche se la distribuzione di *Y* nei due gruppi è normale, questa statistica non ha una distribuzione *t* di Student!

Esiste una statistica che verifica questa ipotesi e ha distribuzione normale, la statistica *t* "a varianza aggregata" – cfr. il Paragrafo 3.6 del volume stampato – tuttavia essa è valida soltanto se le varianze delle distribuzioni normali sono le stesse nei due gruppi. Pensate che questo sia vero, per esempio, per i salari di uomini vs donne?

La distribuzione *t* di Student – Riepilogo

- L'ipotesi che Y abbia distribuzione $N(\mu_Y, \sigma_Y^2)$ è raramente plausibile nella pratica (reddito? numero di figli?)
- per $n > 30$, la distribuzione t e $N(0,1)$ sono molto vicine (al crescere di n, la distribuzione t_{n-1} converge a $N(0,1)$)
- La distribuzione t è un artefatto che risale ai tempi in cui le dimensioni dei campioni erano piccole e i "calcolatori" erano persone
- Per motivi storici, il software statistico utilizza generalmente la distribuzione t per calcolare valori-p ma questo è irrilevante quando la dimensione del campione è moderata o grande.
- Per questi motivi, in questo corso ci concentreremo sull'approssimazione con n grande data dal TLC
    1. Quadro probabilistico per l'inferenza statistica
    2. Stima
    3. Verifica
    4. **Intervalli di confidenza**

Intervalli di confidenza

- Un ***intervallo di confidenza al 95%*** per μ_Y è un intervallo che contiene il valore vero di μ_Y nel 95% dei campioni ripetuti.
- *Digressione*: qual è la casualità qui? I valori di $Y_1,...,Y_n$ e quindi qualsiasi funzione degli stessi – incluso l'intervallo di confidenza, che differirà da un campione all'altro. Il parametro della popolazione, μ_Y, non è casuale; semplicemente, non lo conosciamo.

Intervalli di confidenza (continua)

Un intervallo di confidenza al 95% può sempre essere costruito come insieme di valori dei μ_Y non rifiutati da un test di ipotesi con un livello di significatività del 5%.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1{,}96\} = \{\mu_Y: -1{,}96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1{,}96\}$$

$$= \{\mu_Y: -1{,}96 \leq -\mu_Y \leq 1{,}96 \frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \varepsilon \left(\bar{Y} - 1{,}96 \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1{,}96 \frac{s_Y}{\sqrt{n}}\right)\}$$

Questo intervallo di confidenza si basa sugli n-grande risultati che $\bar{Y}$ *è approssimata da una distribuzione normale e* $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Riepilogo:

Dalle due ipotesi di:

1. campionamento casuale semplice di una popolazione, cioè $\{Y_i, i = 1,\ldots,n\}$ sono i.i.d.
2. $0 < E(Y^4) < \infty$

abbiamo sviluppato, per grandi campioni (*n* grande):

- Teoria della stima (distribuzione campionaria di $\overline{Y}$)
- Teoria della verifica di ipotesi (distribuzione con *n* grande della statistica *t* e calcolo del valore-*p*)
- Teoria degli intervalli di confidenza (costruita invertendo la statistica test)

Le ipotesi (1) e (2) sono plausibili nella pratica? **Sì**

Torniamo alla domanda politica di partenza:

Qual è l'effetto sui punteggi nei test della riduzione della dimensione delle classi di uno studente per classe?

Abbiamo risposto a questa domanda?

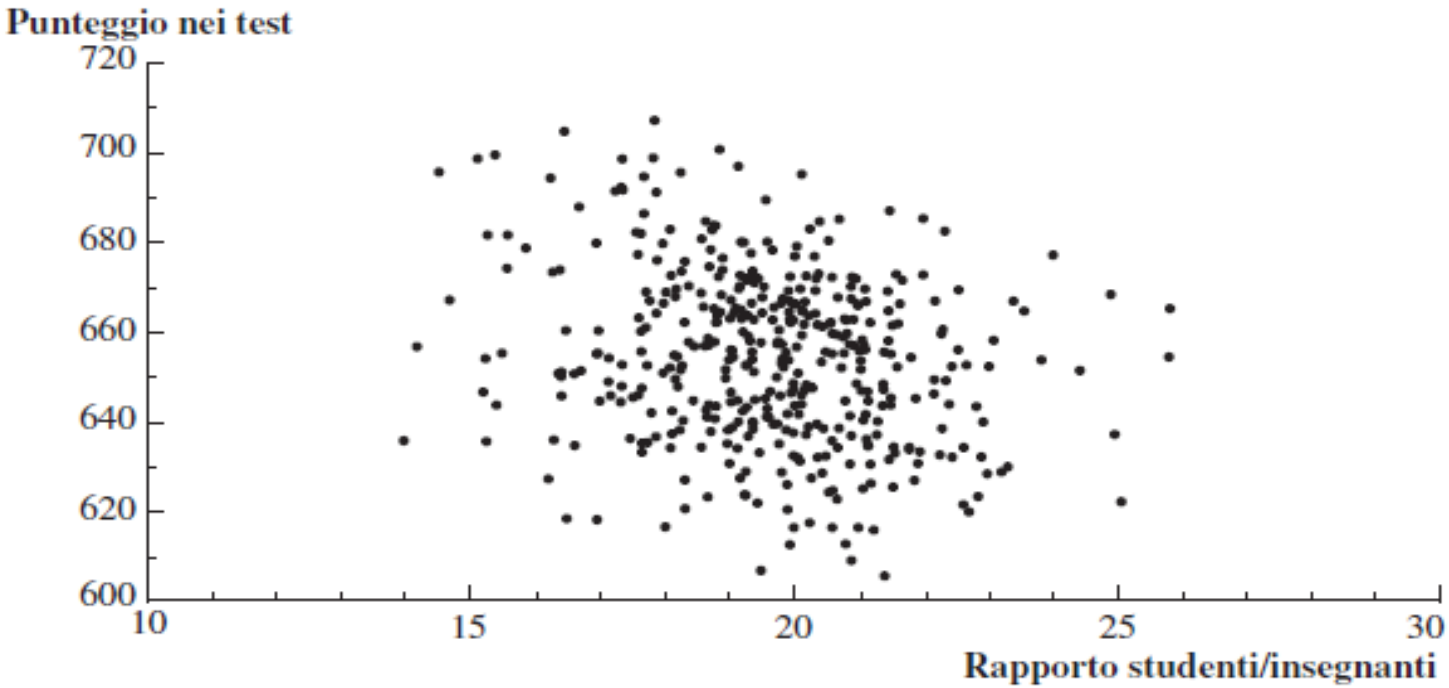


Figura 4.2

Diagramma a nuvola del punteggio nei test e del rapporto studenti/insegnanti (dati relativi ai distretti scolastici della California).

Dati per i 420 distretti scolastici della California. C'è una debole relazione negativa tra il rapporto studenti/insegnanti e il punteggio nei test: la correlazione campionaria è pari a –0,23.