Econometria | 2022/2023

Lezione 1: Introduzione all'econometria

Lezione 2: Richiami di probabilità e statistica I

Lezione 3: Richiami di probabilità e statistica II

Giuseppe Ragusa

https://gragusa.org

Roma, 20 febbraio 2023

Cos’è l’econometria

Il significato letterale è misura dell’economia, ma il suo scopo è molto più ampio

Goldberger (1964)

> Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena

Theil (1971)

> Econometrics is concerned with the empirical determination of economic laws

Le due facce dell’econometria

Predizioni/previsioni

a. predire le performance degli studenti in base alla dimensione delle classi?
b. predire il reddito di individui in base al loro livello di istruzione?
c. predire il numero di sigarette vendute in base al loro prezzo?
d. prevedere il prezzo futuro di un’azione in base al valore odierno dei fondamentali
e. prevedere l’inflazione futura in base alla stance di politica monetaria della BCE

Causalità

a. Quanto la performance degli studenti è legata alla dimensioni delle classi?
b. In che modo un anno in più di istruzione influisce sul reddito?
c. Qual è l’elasticità al prezzo delle sigarette?
d. Quanto sono influenzati i prezzi delle azioni dal dividend-yield ratio?
e. Qual è l’effetto sulla crescita del PIL di un aumento di 1 punto percentuale nei tassi di interesse?

Le due facce dell'econometria

Predizioni/previsioni

a. predire le performance degli studenti in base alla dimensione delle classi?
b. predire il reddito di individui in base al loro livello di istruzione?
c. predire il numero di sigarette vendute in base al loro prezzo?
d. prevedere il prezzo futuro di un'azione in base al valore odierno dei fondamentali
e. prevedere l'inflazione futura in base alla stance di politica monetaria della BCE

Causalità

a. Quanto la performance degli studenti è legata alla dimensioni delle classi?
b. In che modo un anno in più di istruzione influisce sul reddito?
c. Qual è l'elasticità al prezzo delle sigarette?
d. Quanto sono influenzati i prezzi delle azioni dal dividend-yield ratio?
e. Qual è l'effetto sulla crescita del PIL di un aumento di 1 punto percentuale nei tassi di interesse?

Questo corso tratta soprattutto l’identificazione e la stima di effetti causali, ma parleremo spesso di predizioni e previsioni

Esperimento ideale

- Idealmente vorremmo un esperimento per dare risposte quantitativamente rilevanti alle domande causali
  - benefici
  - costi
  - problemi etici

Progetto STAR (Student-Teacher Achievement Ratio)

- Studio quadriennale del costo di 12 milioni di dollari (1985/1990)
- Studenti (K3) assegnati casualmente a tre gruppi:
  1. classe normale (22 – 25 studenti)
  2. classe normale + assistente
  3. classe piccola (13 – 17 studenti)
- studenti delle classi normali riassegnati casualmente dopo il primo anno a classi normali o normali con assistente

La sfida: causalità con dati non sperimentali

- Spesso a disposizione soltanto dati non sperimentali
  - performance/classi (420 distretti scolastici California 1998)
  - prezzi/quantità sigarette in diversi mercati
  - inflazione e tassi per diversi periodi
- L'uso di dati non sperimentali per stimare effetti causali pone enormi ostacoli:
  - effetti perturbativi (fattori omessi)
  - causalità simultanea
  - selezione del campione
  - errori nelle variabili

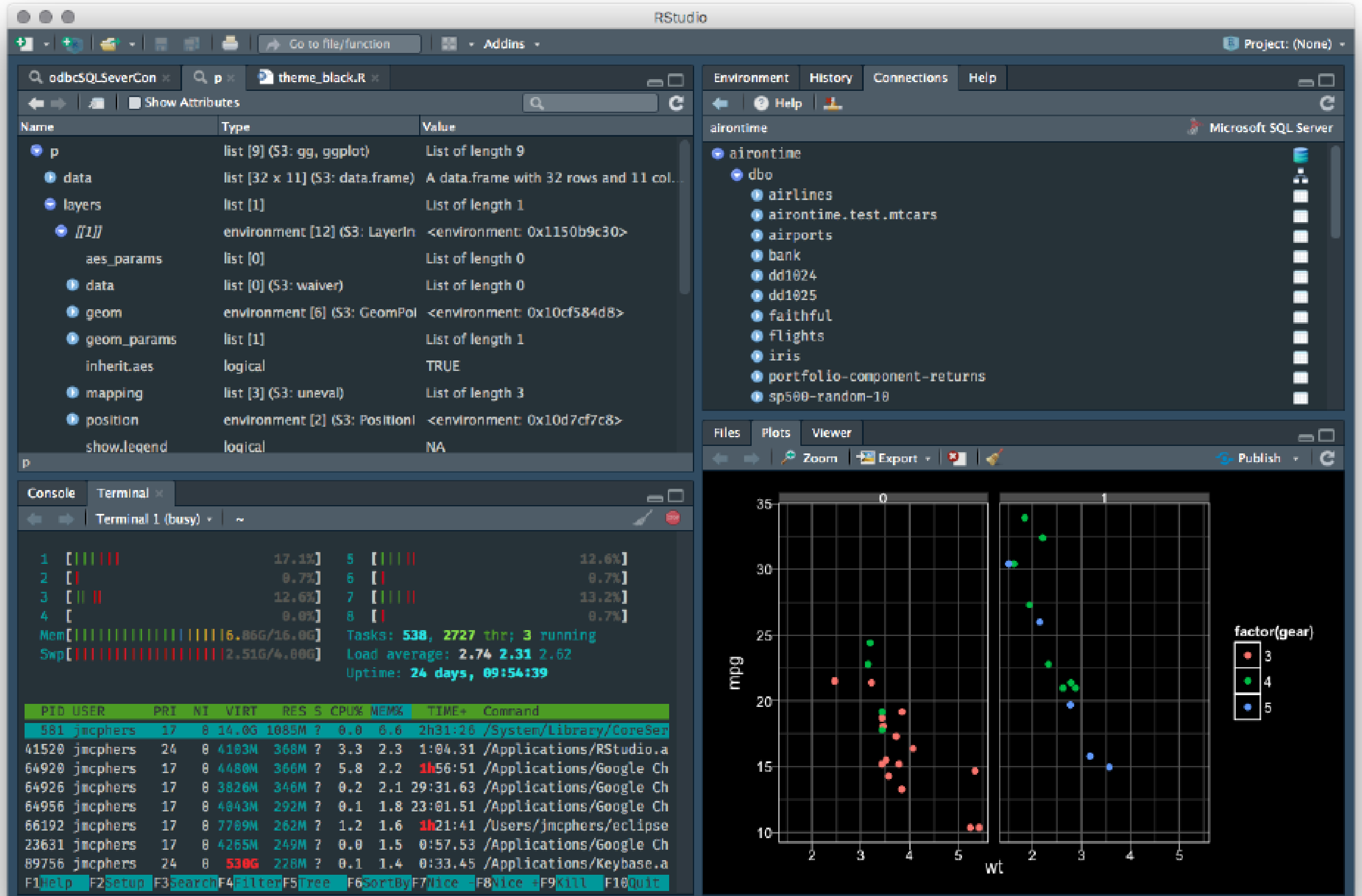
La correlazione $\neq$ causalità

1. i modelli di predizione/predizione non soffrono degli stessi problemi
2. un fenomenale modello di predizione/previsione (machine learning) potrebbe essere (e normalmente è) incapace di dirci qualcosa di “causale”
3. Spesso diciamo che la presenza di correlazione fra due o più variabili non implica che queste due variabili siano legate da un nesso di cause ed effetto

In questo corso:

- metodi per predizione/previsione (data science)
- metodi per stimare effetti causali da dati non sperimentali;
- molte applicazioni
- 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
- R

R


RStudio
Go to file/function
Addins
Project: (None)
odbcSQLSeverCon
theme_black.R
Show Attributes
Name
Type
Value
Environment
History
Connections
Help
airontime
Microsoft SQL Server
dbo
airlines
airontime.test.mtcars
airports
bank
dd1024
dd1025
faithful
flights
iris
portfolio-component-returns
sp500-random-10
Files
Plots
Viewer
Zoom
Export
Publish
Console
Terminal
Terminal 1 (busy)
factor(gear)
mpg
wt

Tipi di dati

- **Dati sezionali**

  riguardano diverse entità (scuole, lavoratori, consumatori, imprese, stati) osservate in un unico periodo

- Dati panel (longitudinali)

  riguardano più entità osservate in due o più peridi.

- Serie temporali

  riguardano una singola entità (persona, impresa, paese) osservata nel tempo

	distcod	mathscr	readscr	enrltot	teachers
1	75119	690.0	691.6	195	10.90
2	61499	661.9	660.5	240	11.15
3	61549	650.9	636.3	1550	82.90
4	61457	643.5	651.9	243	14.00
5	61523	639.9	641.8	1335	71.50
6	62042	605.4	605.7	137	6.40
7	68536	609.0	604.5	195	10.00
8	63834	612.5	605.5	888	42.50
9	62331	616.1	608.9	379	19.00
10	67306	613.4	611.9	2247	108.00
11	65722	618.7	612.8	446	21.00
12	62174	616.0	616.6	987	47.00
13	71795	619.8	612.8	103	5.00
14	72181	622.6	610.0	487	24.34
15	72298	621.0	611.9	649	36.00
16	72041	619.9	614.8	852	42.07
17	63594	624.4	611.7	491	28.92
18	63370	621.7	614.9	421	25.50
19	64709	620.5	619.1	6880	303.03
20	63560	619.3	621.3	2688	135.00
21	63230	625.4	615.6	440	24.00
22	72058	622.9	619.9	475	21.00
23	63842	620.6	622.9	2538	130.50
24	71811	623.4	620.7	476	19.00
25	65748	625.7	619.5	2357	114.00
26	72272	621.2	625.0	1588	85.00
27	65961	626.0	620.4	7306	319.80
28	63313	630.4	616.5	2601	135.00
29	72199	627.1	620.1	847	44.00
30	72215	620.4	627.9	452	22.00

Tipi di dati

- Dati sezionali

  riguardano diverse entità (scuole, lavoratori, consumatori, imprese, stati) osservate in un unico periodo

- Dati panel (longitudinali)

  riguardano più entità osservate in due o più peridi.

- Serie temporali

  riguardano una singola entità (persona, impresa, paese) osservata nel tempo

	airline	year	cost	output	pf	lf
1	1	1	1140640	0.952757	106650	0.534487
2	1	2	1215690	0.986757	110307	0.532328
3	1	3	1309570	1.091980	110574	0.547736
4	1	4	1511530	1.175780	121974	0.540846
5	1	5	1676730	1.160170	196606	0.591167
6	1	6	1823740	1.173760	265609	0.575417
7	1	7	2022890	1.290510	263451	0.594495
8	1	8	2314760	1.390670	316411	0.597409
9	1	9	2639160	1.612730	384110	0.638522
10	1	10	3247620	1.825440	569251	0.676287
11	1	11	3787750	1.546040	871636	0.605735
12	1	12	3867750	1.527900	997239	0.614360
13	1	13	3996020	1.660200	938002	0.633366
14	1	14	4282880	1.822310	859572	0.650117
15	1	15	4748320	1.936460	823411	0.625603
16	2	1	569292	0.520635	103795	0.490851
17	2	2	640614	0.534627	111477	0.473449
18	2	3	777655	0.655192	118664	0.503013
19	2	4	999294	0.791575	114797	0.512501
20	2	5	1203970	0.842945	215322	0.566782
21	2	6	1358100	0.852892	281704	0.558133
22	2	7	1501350	0.922843	304818	0.558799
23	2	8	1709270	1.000000	348609	0.572070
24	2	9	2025400	1.198450	374579	0.624763
25	2	10	2548370	1.340670	544109	0.628706
26	2	11	3137740	1.326240	853356	0.589150
27	2	12	3557700	1.248520	1003200	0.532612
28	2	13	3717740	1.254320	941977	0.526652
29	2	14	3962370	1.371770	856533	0.540163
30	2	15	4209390	1.389740	821361	0.528775

Tipi di dati

- Dati sezionali

  riguardano diverse entità (scuole, lavoratori, consumatori, imprese, stati) osservate in un unico periodo

- Dati panel (longitudinali)

  riguardano più entità osservate in due o più peridi.

- Serie temporali

  riguardano una singola entità (persona, impresa, paese) osservata nel tempo

	rfood	rdur	rcon	rmrf	rf	date
1	-4.59	0.87	-6.84	-6.99	0.33	1960-01-01
2	2.62	3.46	2.78	0.99	0.29	1960-02-01
3	-1.67	-2.28	-0.48	-1.46	0.35	1960-03-01
4	0.86	2.41	-2.02	-1.70	0.19	1960-04-01
5	7.34	6.33	3.69	3.08	0.27	1960-05-01
6	4.99	-1.26	2.05	2.09	0.24	1960-06-01
7	-1.52	-5.09	-3.79	-2.23	0.13	1960-07-01
8	3.96	4.38	-1.08	2.85	0.17	1960-08-01
9	-3.98	-4.23	-4.71	-6.00	0.16	1960-09-01
10	0.99	1.17	-1.44	-0.70	0.22	1960-10-01
11	9.22	10.58	6.53	4.72	0.13	1960-11-01
12	4.12	6.79	3.42	4.68	0.16	1960-12-01
13	4.75	0.26	6.08	6.23	0.19	1961-01-01
14	4.53	18.08	4.25	3.54	0.14	1961-02-01
15	4.43	3.68	2.08	2.86	0.20	1961-03-01
16	-1.14	-2.34	-4.23	0.39	0.17	1961-04-01
17	4.31	-1.27	2.74	2.40	0.18	1961-05-01
18	-2.23	-6.85	-3.24	-3.04	0.20	1961-06-01
19	2.57	-0.66	-0.30	2.81	0.18	1961-07-01
20	4.77	1.98	0.59	2.54	0.14	1961-08-01
21	-0.76	1.83	-2.87	-2.17	0.17	1961-09-01
22	3.45	-3.00	1.30	2.56	0.19	1961-10-01
23	5.22	1.91	5.93	4.39	0.15	1961-11-01
24	-3.32	-0.42	0.47	-0.10	0.19	1961-12-01
25	-6.42	-9.65	-5.18	-3.86	0.24	1962-01-01
26	-0.20	0.07	0.68	1.75	0.20	1962-02-01
27	0.87	-2.26	-1.20	-0.66	0.20	1962-03-01
28	-4.51	-8.04	-7.26	-6.56	0.22	1962-04-01
29	-11.05	-8.93	-8.64	-8.69	0.24	1962-05-01
30	-8.55	-11.07	-10.92	-8.46	0.20	1962-06-01

Richiami di probabilità e statistica

Problema empirico: Dimensione della classe e performance degli studenti

Domanda:

- qual è l'effetto sui punteggi nei test di una riduzione (o di un aumento) della dimensione delle classi di 1 unità?
- qual è l'effetto sui punteggi nei test di una riduzione (o di un aumento) della dimensione delle classi di 2 unità?

I dati della California

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

	distcod	mathscr	readscr	teachers	enrltot
1	75119	690.0	691.6	10.90	195
2	61499	661.9	660.5	11.15	240
3	61549	650.9	636.3	82.90	1550
4	61457	643.5	651.9	14.00	243
5	61523	639.9	641.8	71.50	1335
6	62042	605.4	605.7	6.40	137
7	68536	609.0	604.5	10.00	195
8	63834	612.5	605.5	42.50	888
9	62331	616.1	608.9	19.00	379
10	67306	613.4	611.9	108.00	2247
11	65722	618.7	612.8	21.00	446
12	62174	616.0	616.6	47.00	987
13	71795	619.8	612.8	5.00	103
14	72181	622.6	610.0	24.34	487
15	72298	621.0	611.9	36.00	649
16	72041	619.9	614.8	42.07	852
17	63594	624.4	611.7	28.92	491
18	63370	621.7	614.9	25.50	421
19	64709	620.5	619.1	303.03	6880
20	63560	619.3	621.3	135.00	2688
21	63230	625.4	615.6	24.00	440
22	72058	622.9	619.9	21.00	475
23	63842	620.6	622.9	130.50	2538
24	71811	623.4	620.7	19.00	476

Variabili:

- Punteggi nei test del quinto anno (Stanford-9 achievement test) media del distretto

$$testscr = \frac{\text{mathscr} + \text{readscr}}{2}$$

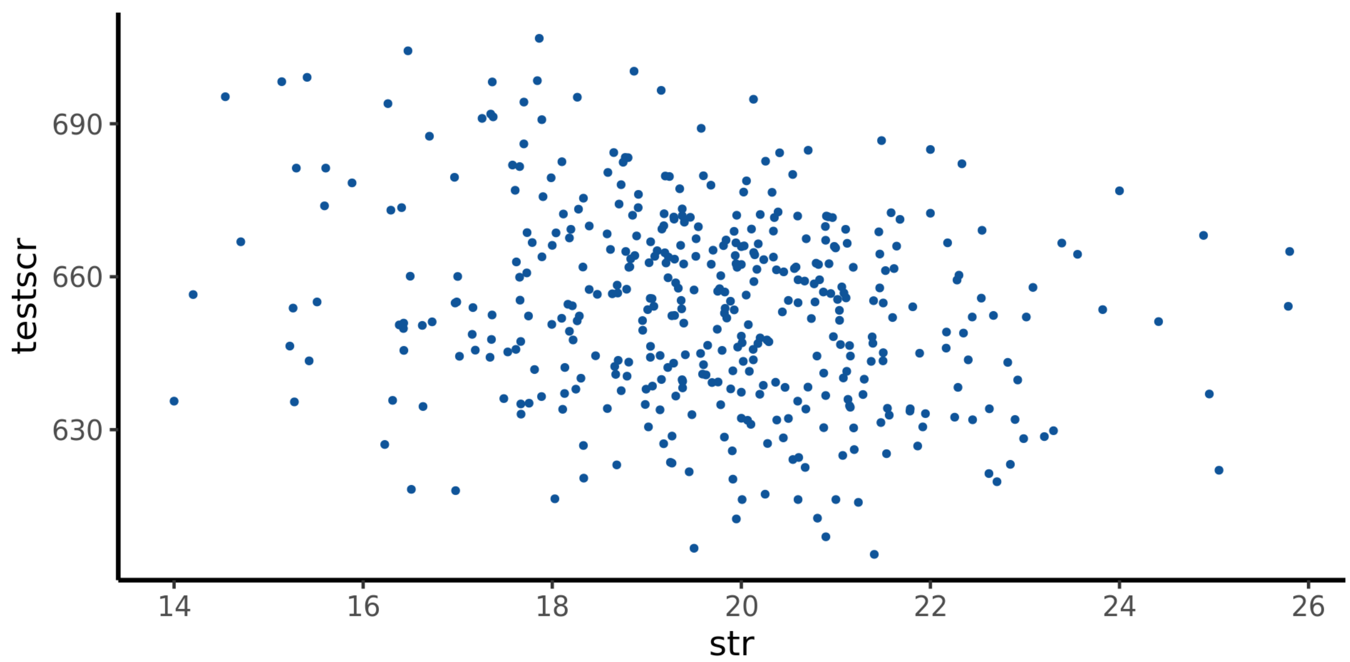
- Rapporto studenti/insegnanti

$$\begin{aligned} str &= \frac{\#\ \text{di studenti}}{\#\ \text{di insegnanti}} \\ &= \frac{\text{enrltot}}{\text{teachers}} \end{aligned}$$

Le scuole della California

```
   distcod testscr      str
1    75119  690.80 17.88991
2    61499  661.20 21.52466
3    61549  643.60 18.69723
4    61457  647.70 17.35714
5    61523  640.85 18.67133
6    62042  605.55 21.40625
7    68536  606.75 19.50000
8    63834  609.00 20.89412
9    62331  612.50 19.94737
10   67306  612.65 20.80556
11   65722  615.75 21.23809
12   62174  616.30 21.00000
13   71795  616.30 20.60000
14   72181  616.30 20.00822
15   72298  616.45 18.02778
16   72041  617.35 20.25196
17   63594  618.05 16.97787
18   63370  618.30 16.50980
19   64709  619.80 22.70402
20   63560  620.30 19.91111
21   63230  620.50 18.33333
22   72058  621.40 22.61905
23   63842  621.75 19.44828
24   71811  622.05 25.05263
25   65748  622.60 20.67544
26   72272  623.10 18.68235
27   65961  623.20 22.84553
28   63313  623.45 19.26667
```

```r
Caschool |>
  ggplot(aes(y=testscr,x=str)) +
  geom_point(color=blue) +
  geom_rug(fill="slategray", col='white') +
  theme_gragusa(base_size = 18)
```



Primo sguardo ai dati: (sapete già 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 punteggio test and str.

`str` piccoli associati a `testscr` elevati?

Evidenza numerica:

1. (stima) Confrontare i punteggi nei test nei distretti con basso `str` a quelli con alto `str`
2. (verifica di ipotesi) 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
3. (intervallo di confidenza) Costruire intervallo per la differenza nei punteggi medi nei test, nei distretti con alto vs basso `str`

Distretti con dimensioni delle classi “piccole” $(STR < 20)$ e “grandi” $(STR \geqslant 20)$

Dimensione classe	Punteggio medio	Deviazione standard	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 $\Delta = 0$
3. Costruire un intervallo di confidenza per Δ

Stima

$$\bar{Y}_s - \bar{Y}_l = \frac{1}{n_s}\sum_{i=1}^{n_s} Y_i - \frac{1}{n_l}\sum_{i=1}^{n_l} Y_i = 657,4 - 650,0 = 7,4$$

È una differenza da considerare grande nel mondo reale?

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

Note: i pedici s e l indicano distretti con `str` small (piccolo) e large (grande)

Verifica di ipotesi

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

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

dove:

- $SE(\bar{Y}_s - \bar{Y}_l)$ è l' errore standard di $\bar{Y}_s - \bar{Y}_l$
- $s_s^2 = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} (Y_i - \bar{Y}_s)^2$
- $s_l^2 = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} (Y_i - \bar{Y}_l)^2$

La statistica-t per la differenza tra medie

Dimensione classe	Punteggio medio	Deviazione standard	n
Piccola	657,4	19,4	238
Grande	650,0	17,9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} = \frac{7,4}{1,83} = 4,05$$

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

Intervallo di confidenza

Un intervallo di confidenza al 95% per la differenza tra medie è

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l) = 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)$$

$$\Delta = E(testscr|STR < 20) - E(testscr|STR \geqslant 20)$$

Due affermazioni equivalenti:

1. L'intervallo di confidenza al 95% per Δ non include 0;
2. L'ipotesi che $\Delta = 0$ è rifiutata al livello 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

Percorso:

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

Concetti:

- Popolazione, variabile casuale e distribuzione
- Momenti di una distribuzione (media, varianza, deviazione standard, covarianza, correlazione)
- Distribuzione condizionata e media condizionata
- Distribuzione di un campione di dati estratto a caso da una popolazione: $(Y_1, \ldots, Y_n)$

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:

- 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

  <u>Esempio:</u>

  - $Pr(Y = 650)$ (quando Y è discreta)
  - $\Pr(640 \leqslant Y \leqslant 660)$ (quando Y è continua).

Momenti

valore atteso di Y

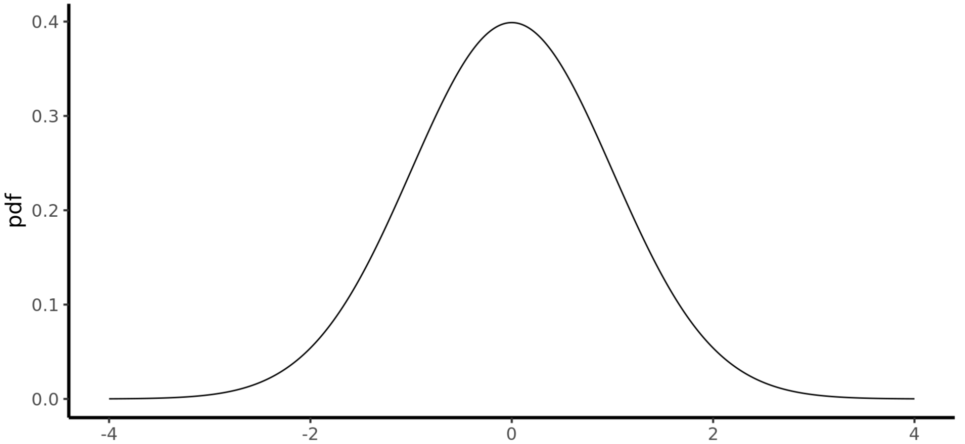
$$E(Y) = \mu_Y$$

varianza di Y

$$E[(Y - \mu_Y)^2] = \sigma_Y^2$$

deviazione standard di Y

$$\sqrt{E[(Y - \mu_Y)^2]} = \sigma_Y$$



Momenti, ctd.

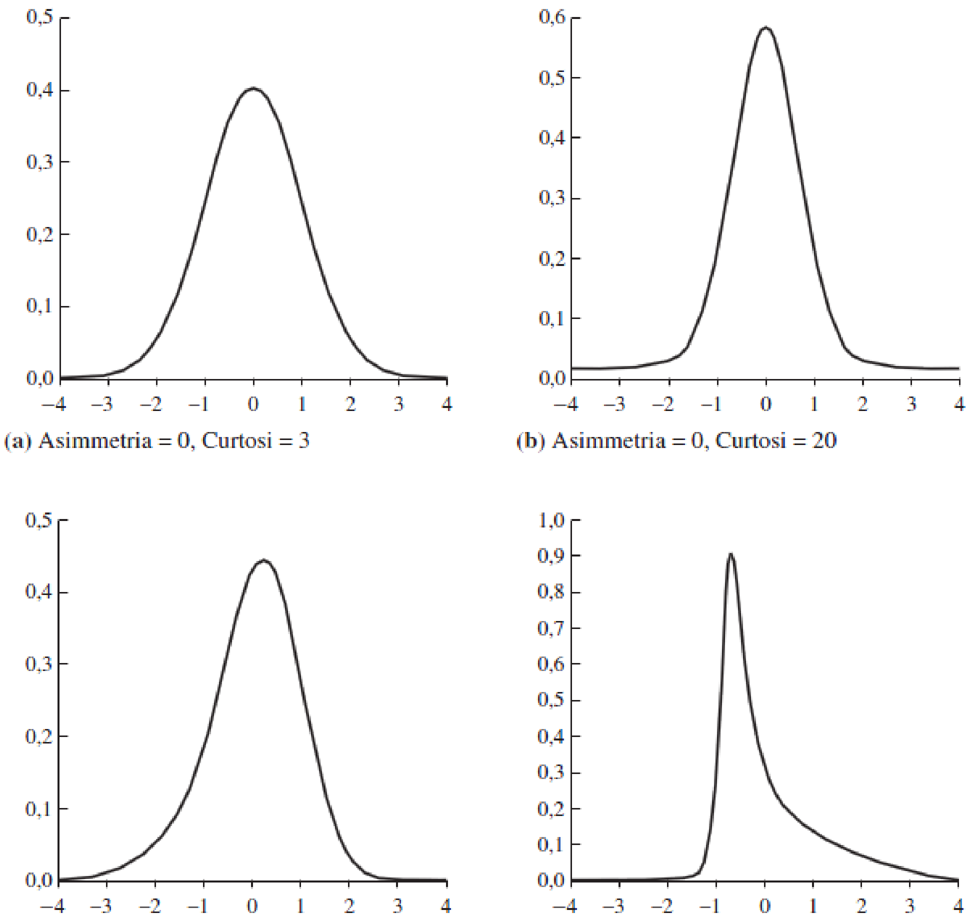
asimmetria

$$\frac{E[(Y-\mu_Y)^3]}{\sigma_Y^3}$$

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

curtosi

$$\frac{E[(Y-\mu_Y)^4]}{\sigma_Y^4}$$



(a) Asimmetria = 0, Curtosi = 3
(b) Asimmetria = 0, Curtosi = 20
(c) Asimmetria = –0,1, Curtosi = 5
(d) Asimmetria = 0,6, Curtosi = 5

La covarianza

La covarianza tra X e Z è

$$cov(X, Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

La covarianza è una misura dell'associazione lineare tra X e Z; le sue unità sono unità di X $\times$ unità di Z

- $cov(X, Z) > 0 \implies$ relazione positiva tra X e Z
- $cov(X, Z) < 0 \implies$ relazione negativa tra X e Z

Se X e Z sono indipendentemente distribuite, allora $cov(X, Z) = 0$ (ma non vale il vice versa!!)

La covarianza di una variabile casuale con se stessa è la sua varianza:

$$cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)2]$$

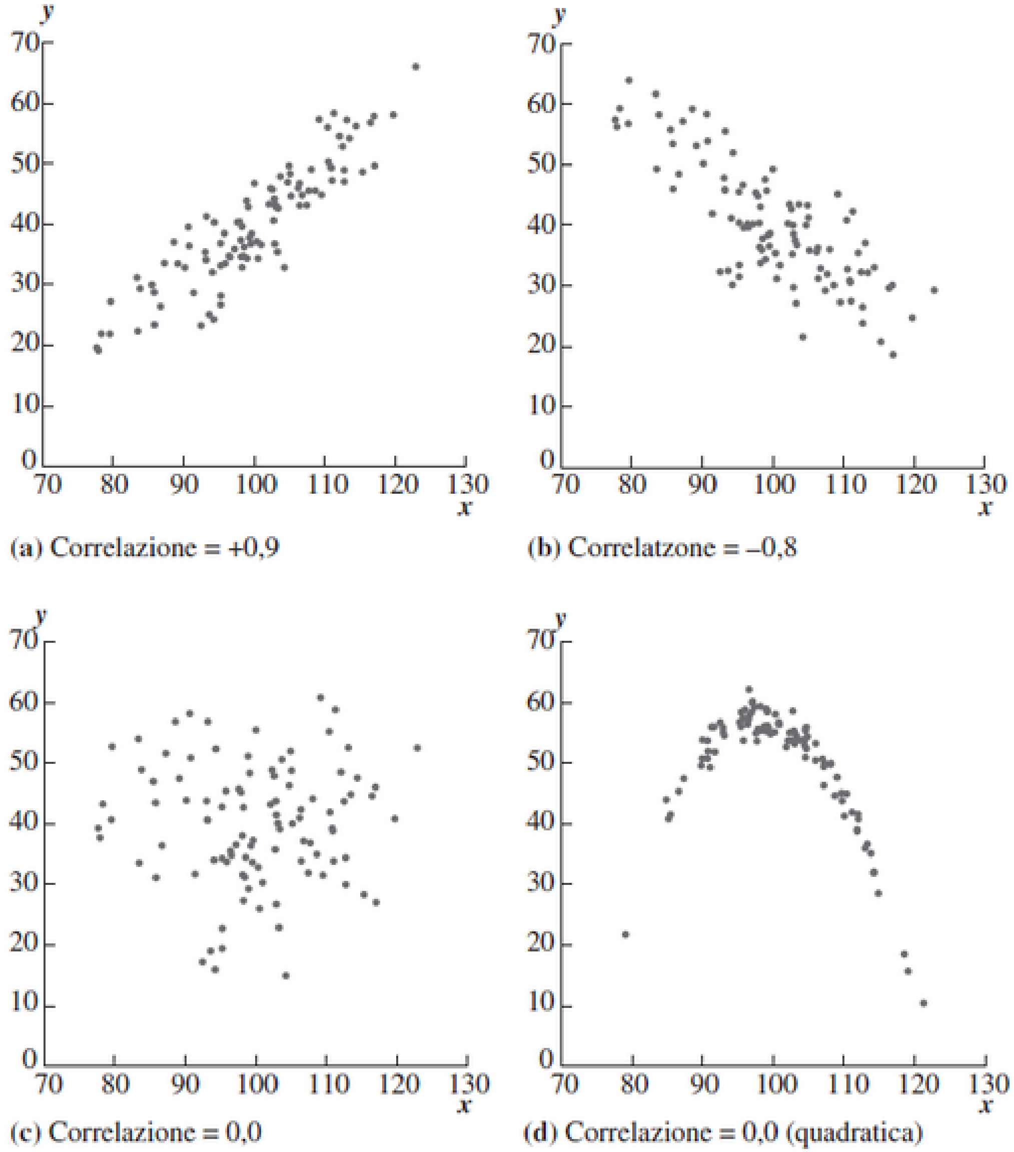
Il coefficiente di correlazione

Il coefficiente di correlazione è definito in termini di covarianza:

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

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

Il coefficiente di correlazione misura l'associazione lineare



corr

Momenti condizionato

- valore atteso condizionato

$$E(Y|X = x)$$

- varianza condizionata

$$Var(Y|X = x)$$

`testscr` e `str`

- Valore atteso dei punteggi nei test nei i distretti con classi piccole

$$E(testscr|str < 20)$$

- Valore atteso dei punteggi nei test nei distretti con classi grandi

$$E(testscr|str \geqslant 20)$$

`Altri esempi:`

- Valore atteso salario per lavoratori di genere femminile

$$E(Salario|genere = femminile)$$

- Tasso di mortalità di pazienti che ricevono una cura sperimentale (Y = vivo/morto; X = trattato/non trattato)

$$E(Y|X = trattato)$$

Media condizionata

Importante

Se

$$E(X|Z) = costante$$

allora

$$corr(X, Z) = 0$$

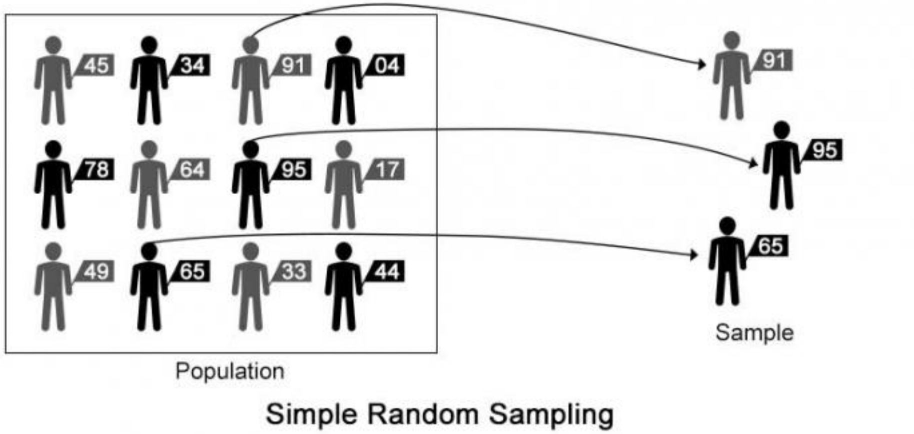
Tuttavia non vale necessariamente il contrario.

Campione

$$(Y_1, \ldots, Y_n)$$

Campionamento casuale semplice

- Individui scelti a caso dalla popolazione
- Il data set è $(Y_1, Y_2, \ldots, Y_n)$
- Prima della selezione, il valore di Y_i è casuale perché dipende dall'individuo selezionato
- Y_i è un numero dopo la selezione



Campione casuale

Poiché gli individui 1 e 2 sono selezionati a caso, il valore di Y_1 non contiene informazioni riguardo 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.

Percorso:

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

Concetti:

$\bar{Y}$ è lo stimatore naturale di $E(Y)$.

Ma:

- quali sono le proprietà di $\bar{Y}$?
- Perché dovremmo usare anziché un altro stimatore?

La distribuzione campionaria di $\bar{Y}$

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

- La distribuzione di su diversi possibili campioni di dimensione n si chiama distribuzione campionaria di $\bar{Y}$
- La media e la varianza di sono la media e la varianza della sua distribuzione campionaria, $E(\bar{Y})$ e $var(\bar{Y})$
- Il concetto di distribuzione campionaria è alla base di tutta l'econometria.

La distribuzione campionaria di $\bar{Y}$

Esempio

Y assume il valore 0 o 1 (variabile casuale di Bernoulli) con la distribuzione di probabilità

$$Y = \begin{cases} 0 & 0.22 \\ 1 & 0.78 \end{cases}$$

Valore atteso

$$E(Y) = p \times 1 + (1 - p) \times 0 = p = 0.78$$

Varianza

$$E[(Y - E(Y))^2] = p(1 - p) = 0.78 \times (1 - 0.78) = 0.1716$$

La distribuzione campionaria di $\bar{Y}$

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

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

- $\Pr(\bar{Y} = 0) = 0.22 \times 0.22 = 0.0484$
- $\Pr(\bar{Y} = 0.5) = 2 \times 0.22 \times 0.78 = 0.3432$
- $\Pr(\bar{Y} = 1) = 0.78 \times 0.78 = 0.6084$

Potremmo calcolare la distribuzione per $n = 5, n = 25, n = 100$, e cosi' via….

Distribuzione campionaria: Y è di Bernoulli ($p = 0.78$):

[]](bern.png){.center}

Vogliamo sapere:

- Qual è il valore atteso di $\bar{Y}$?
  - Se $E(\bar{Y}) = \mu = 0.78$, allora $\bar{Y}$ è uno stimatore non distorto di μ
- Qual è la varianza di $\bar{Y}$?
  - In che modo $var(\bar{Y})$ dipende da n?
- $\bar{Y}$ si avvicina a μ quando n è grande?
  - Legge dei grandi numeri: $\bar{Y}$ è uno stimatore consistente di μ
- Distribuzione di $\bar{Y}$
  - $\bar{Y} - \mu$ è approssimato da una distribuzione normale per n grande (teorema limite centrale)

Valore atteso e varianza di $\bar{Y}$

Caso generale - cioè, per $(Y_1, \ldots, Y_n)$ i.i.d. da qualsiasi distribuzione

- valore atteso

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{n\mu_Y}{n} = \mu_Y$$

- varianza

$$var(\bar{Y}) = var\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} var(Y_i) = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}$$

Valore atteso e varianza di $\bar{Y}$

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

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

$$sd(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

Implicazioni:

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

Distribuzione di $\bar{Y}$ quando $n \to \infty$

- Per piccoli campioni, la distribuzione di $\bar{Y}$ è complicata
- Se n è grande, derivare (almeno un'approssimazione) distribuzione campionaria diventa molto pià semplice:
  - legge dei grandi numeri All'aumentare di n, la distribuzione di diventa più strettamente centrata su μ_Y
  - teorema limite centrale Inoltre, la distribuzione di $\bar{Y} - \mu_Y$ può essere approssimata da una normale

Legge dei grandi numeri

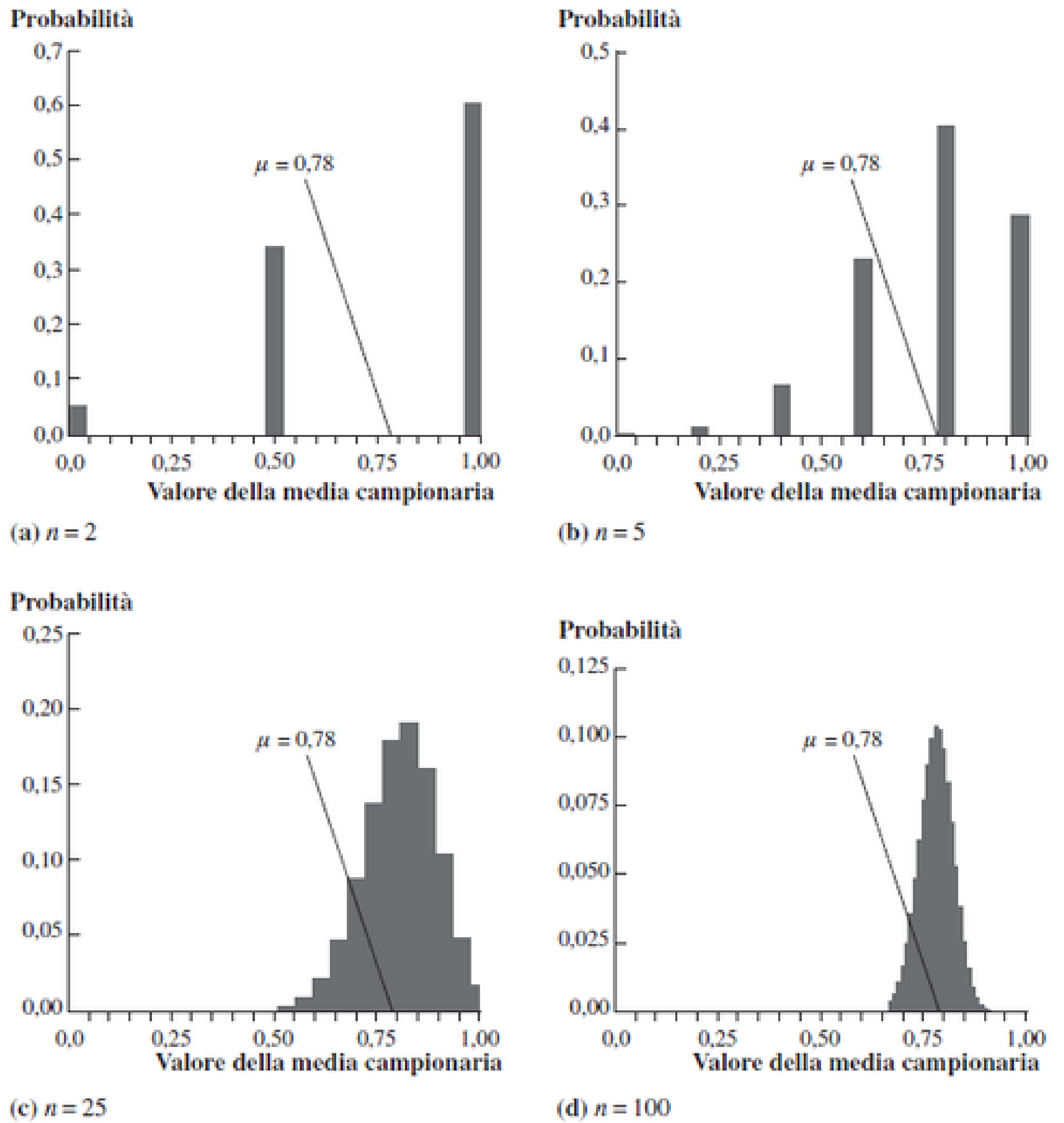
Se $(Y_1, \ldots, Y_n)$ sono i.i.d. e $\mu_Y < \infty$, allora

$$\lim_{n\to\infty} \Pr\left[|\bar{Y} - \mu_Y| < \epsilon\right] = 1,$$

per ogni $\epsilon > 0$.

Spesso scriviamo $\bar{Y} \overset{p}{\to} \mu_Y$, che significa che $\bar{Y}$ converge in probabilità a μ_Y.

Distribuzione campionaria di quando Y è di Bernoulli (p = 0.78):



bern

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 $\bar{Y}$ è bene approssimata da una distribuzione normale:

$$\bar{Y} \xrightarrow{d} N\left(\mu_Y, \frac{\sigma_Y}{n}\right)$$

o, equivalentemente,

$$\sqrt{n}\left(\frac{\bar{Y} - \mu_Y}{\sigma_Y}\right) \xrightarrow{d} N(0, 1)$$

Stimatore della varianza di Y

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

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \xrightarrow{p} \sigma_Y^2$$

Perché si applica la legge dei grandi numeri?

- Perché s_Y^2 è una media campionaria

$$s_Y^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_Y)^2 = \frac{1}{n-1}\sum_{i=1}^{n} Z_i$$

  e $E(Z_i) = var(Y_i) = \sigma_Y^2$.
- Nota tecnica: si assume $E(Y^4) < \infty$ perché la media non è di Y_i, ma del suo quadrato; cfr. Appendice 3.3.

Riepilogo: distribuzione 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_i (la distribuzione della popolazione)
- Quando n è grande, la distribuzione campionaria si semplifica:
  - Legge dei grandi numeri:

$$\bar{Y} \xrightarrow{p} \mu_Y$$

  - Teorema del limite centrale:

$$\frac{\sqrt{n}(\bar{Y} - \mu_Y)}{\sigma_Y^2} \xrightarrow{d} N(0,1)$$

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 μ_Y; $\bar{Y}$ è la soluzione di questo problema

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

Derivazione: Le condizioni del primo ordine sono:

$$\frac{\partial \sum_{i=1}^{n} (Y_i - m)^2}{\partial m} = 0 \implies -2\sum_{i=1}^{n} (Y_i - m) = 0 \implies m = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Perché usare $\bar{Y}$ per stimare μ_Y

- ha una varianza minore di tutti gli altri stimatori lineari non distorti
  - si consideri lo stimatore

$$\tilde{Y} = \frac{1}{n} \sum_{i=1}^{n} a_i Y_i$$

  dove gli a_i sono tali per cui $\tilde{Y}$ risulta non distorto allora

$$var(\bar{Y}) \leqslant var(\tilde{Y})$$

- $\bar{Y}$ non è l'unico stimatore di μ_Y – vi viene in mente un caso in cui potrebbe essere preferibile utilizzare la mediana?

Percorso

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

Il problema della verifica di ipotesi – prendere una decisione riguardo la veridicità di un’ipotesi su una quantità della popolazione in base all’evidenza disponibile

Verifica di ipotesi

Il problema della verifica di ipotesi – prendere una decisione riguardo la veridicità di un'ipotesi riguardo $E(Y)$ in base all'evidenza disponibile:

- l'ipotesi nulla: quella che si suppone essere vera

$$H_0 : E(Y) = \mu_{Y,0}$$

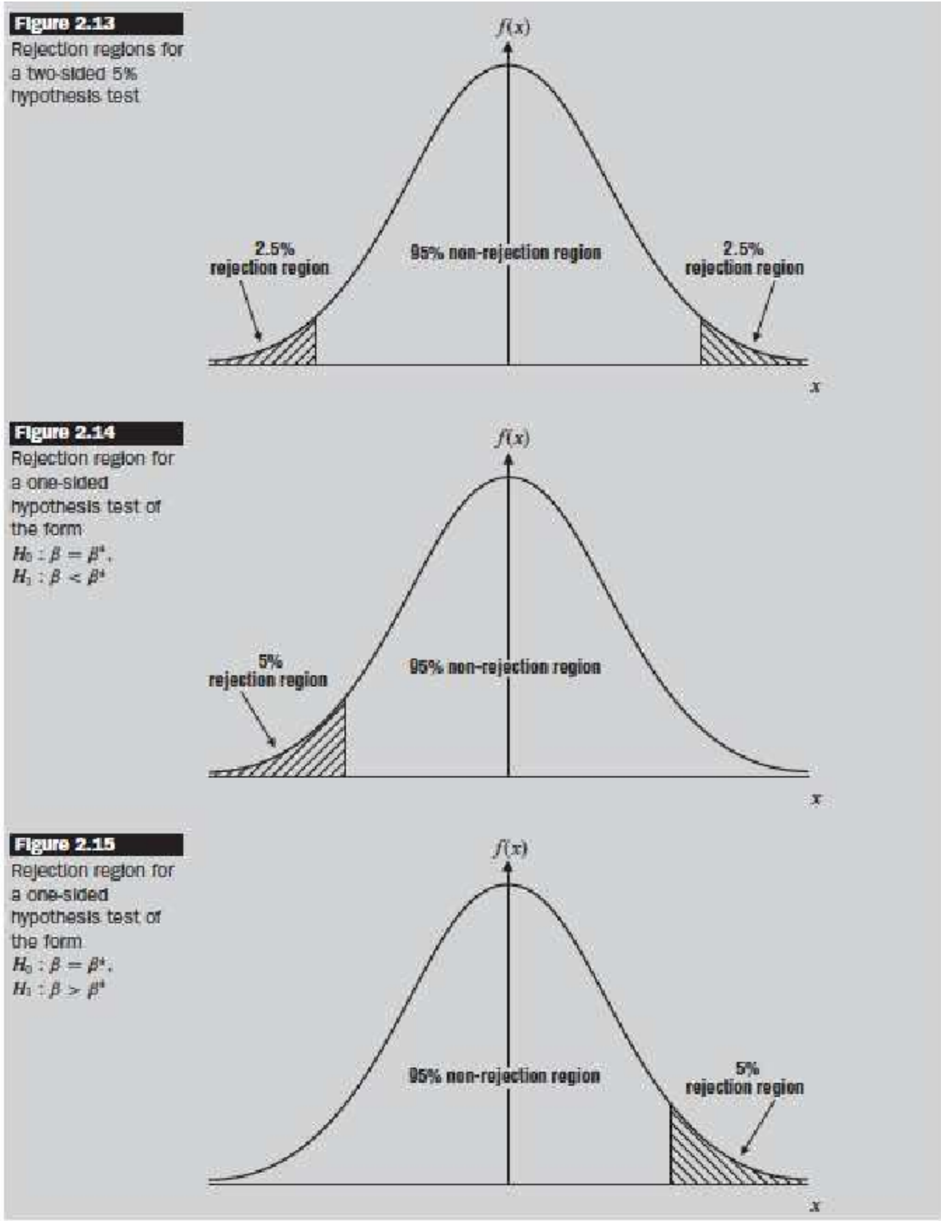
- l'ipotesi alternativa: quella che direttamente contraddice l'ipotesi nulla
  - bidirezionale: $H_1 : E(Y) \neq \mu_{Y,0}$
  - unidirezionale: $H_1 : E(Y) > \mu_{Y,0}$ o $H_1 : E(Y) < \mu_{Y,0}$

Regola decisionale: Errore I tipo

Basando la decisione di accettare o rifiutare l'ipotesi nulla in base all'evidenza empirica, si possono commettere due tipi di errore:

	H_0 vera	H_0 falsa
Accetto	OK	Errore II tipo
Rifiuto	Errore I tipo	OK

Regola decisionale



Terminologia per la verifica di ipotesi

- Il livello di significatività di un test è una probabilità predeterminata di rifiutare in modo errato l'ipotesi nulla quando invece è corretta, ovvero di commettere l'errore di I tipo.
- valore-p = il più piccolo livello di significatività per il quale non è possibile rifiutare l'ipotesi nulla

Calcolo del valore-p:

$$p - value = \Pr(|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|)$$

dove $\bar{Y}^{act}$ e' il valore effettivamente osservato di $\bar{Y}$

Se n è grande, si può usare l'approssimazione normale:

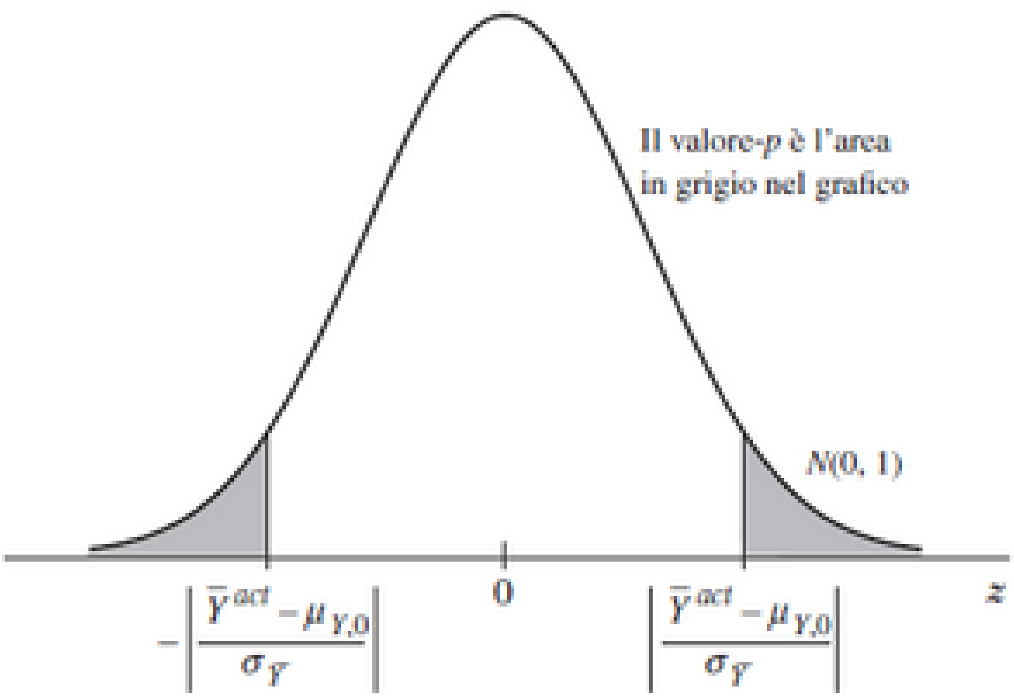
$$valore - p = \Pr_{H_0}\left[\frac{\sqrt{n}|\bar{Y} - \mu_{Y,0}|}{\sigma_Y} > \frac{\sqrt{n}|\bar{Y}^{act} - \mu_{Y,0}|}{\sigma_Y}\right]$$

che non e' altro che la probabilità sotto le code $N(0,1)$

Calcolo del valore-p con σ_Y^2 stimato

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

- Sostituire la varianze σ_Y^2 con una stima consistente non altera la validita’ del teorema del limite centrale. Pertanto, il p-value puo’ essere calcolato usando s_Y invece che σ_Y.



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 $|t| \geqslant 1.96$.
- in modo equivalente, la si rifiuta se $p \leqslant 0.05$
- il valore-p è detto talvolta livello di significatività marginale
- il valore-p e' chiamato in inglese p-value
- software statistico (come R) calcola il p-value per l'ipotesi nulla

Percorso:

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

Concetti:

Un intervallo di confidenza al $(1-\alpha)\%$ per una quantità della popolazione è un intervallo che contiene questa quantità nel 95% dei campioni su cui è ripetutamente calcolato.

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.
- 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%.

$$\begin{aligned}\{\mu_Y : |t| \leqslant 1.96\} &= \{\mu_Y : -1.96 \leqslant t \leqslant 1.96\} \\ &= \{\bar{Y} - 1.96 \times s_Y \sqrt{n}, \bar{Y} + 1.96 \times s_Y \sqrt{n}\}\end{aligned}$$

- Questo intervallo di confidenza si basa sui risultati asintotici - $n \to \infty$ - che ci permettono di approssimare la distribuzione di $\bar{Y}$ con quella di una normale.

Oltre la media

Covarianza

- Campione congiunto

$$\{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\}$$

- Obiettivo

$$\sigma_{Y,X} = E\left[(Y - \mu_Y)(X - \mu_X)\right]$$

- Stimatore

$$\hat{\sigma}_{Y,X} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$$

In campioni grandi ($n \to \infty$)

$$\begin{aligned}\hat{\sigma}_{Y,X} &\approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)(X_i - \mu_X) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i\end{aligned}$$

- $\hat{\sigma}_{Y,X}$ è (approssimativamente) la media campionaria di Z che è una funzione di variabili indipendenti e quindi a sua volta indipendente.
- la legge dei grandi numeri e il teorema del limite centrale si applicano

Covarianza

In campioni grandi ($n \to \infty$)

$$\hat{\sigma}_{Y,X} \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)(X_i - \mu_X)$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i$$

- $\hat{\sigma}_{Y,X}$ è (approssimativamente) la media campionaria di Z che è una funzione di variabili indipendenti e quindi a sua volta indipendente.
- la legge dei grandi numeri e il teorema del limite centrale si applicano

- Consistenza

$$\hat{\sigma}_{Y,X} \xrightarrow{p} \sigma_{Y,X}$$

- Normalità asintotica

$$\sqrt{n}(\hat{\sigma}_{Y,X} - \sigma_{Y,X}) \xrightarrow{d} N(0, V)$$

dove

$$V = E\left\{[Z_i - \sigma_{Y,X}]^2\right\}$$

- Stima della varianza

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\sigma}_{Y,X}]^2$$

Covarianza

In campioni grandi ($n \to \infty$)

$$\hat{\sigma}_{Y,X} \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)(X_i - \mu_X)$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i$$

- $\hat{\sigma}_{Y,X}$ è (approssimativamente) la media campionaria di Z che è una funzione di variabili indipendenti e quindi a sua volta indipendente.
- la legge dei grandi numeri e il teorema del limite centrale si applicano

```r
## Intervallo di confidenza per covarianza fra testscr e
n = 420
Y <- Caschool$testscr
X <- Caschool$str
Ybar <- mean(Y)
Xbar <- mean(X)
sigmahat_YX <- cov(Y, Y)
Z = (Y-Ybar)*(X-Xbar) - sigmahat_YX
V = var(Z)
## Intervallo di confidence per sigma_Y
c(sigmahat_YX - 1.96*sqrt(V)/sqrt(n),
  sigmahat_YX + 1.96*sqrt(V)/sqrt(n))
```

```
[1] 359.3692 366.6909
```