Econometria

Lezione 1

Giuseppe Ragusa

giuseppe.ragusa@uniroma1.it

Sapienza, Università di Roma

https://gragusa.org/econometria

17 febbraio 2025

Cos’è l’Econometria?

Origini e Sviluppo

Disciplina nata all'inizio del XX secolo dall'esigenza di combinare:

- Teoria economica
- Metodi matematici
- Analisi statistica

per verificare empiricamente le teorie economiche

L'Econometric Society, fondata nel 1930 da:

- Irving Fisher
- Ragnar Frisch
- Charles Roos

con l'obiettivo di trasformare l'economia in una "scienza in senso stretto"

Sottodiscipline

Microeconometria

- Analisi delle relazioni fra individui, famiglie, imprese
- Utilizza principalmente dati sezionali

Macroeconometria

- Analisi di aggregati economici
- Studio delle politiche monetarie e fiscali

Financial Econometrics

- Focus su mercati e strumenti finanziari
- Studio di:
  - → Prezzi delle attività
  - → Rendimenti
  - → Rischio

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 in base alla 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 d'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 un punto percentuale nei tassi d'interesse?

Le due facce dell'econometria

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

! Importante

I modelli di predizione/predizione non sono adatti per analisi causale

Endogeneità

Un problema fondamentale

- L'endogeneità è il problema centrale dell'analisi econometrica
- Le variabili che studiamo rispondono esse stesse alle condizioni del sistema
- Rende difficile stabilire relazioni causali

Esempi di Endogeneità

Trasporto Pubblico

- Domanda: "Effetto della riduzione delle tariffe?"
- Problema: Le tariffe potrebbero essere ridotte in risposta ad un calo di utilizzo
- La correlazione osservata potrebbe mascherare l'effetto reale

Politica Monetaria

- Domanda: "Effetto della variazione dei tassi sull'inflazione?"
- Problema: I tassi vengono alzati anche in base alle previsioni dell'inflazione
- L'apparente inefficacia potrebbe mascherare successo della Bance Centrale

Endogeneità

Esperimenti

Idealmente vorremmo un esperimento per dare risposte quantitativamente rilevanti alle domande causali

- benefici
- costi
- problemi etici
- impossibile in molti contesti

Esperimenti

Progetto STAR

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

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

Sviluppi Metodologici

La “lotta” all’endogeneità con dati sperimentali ha stimolato lo sviluppo di tecniche econometriche:

- Microeconometria
  - → Variabili strumentali
  - → Tecniche difference-in-differences
  - → Analisi dei dati panel

- Macroeconometria
  - → Modelli VAR
  - → Modelli DSGE

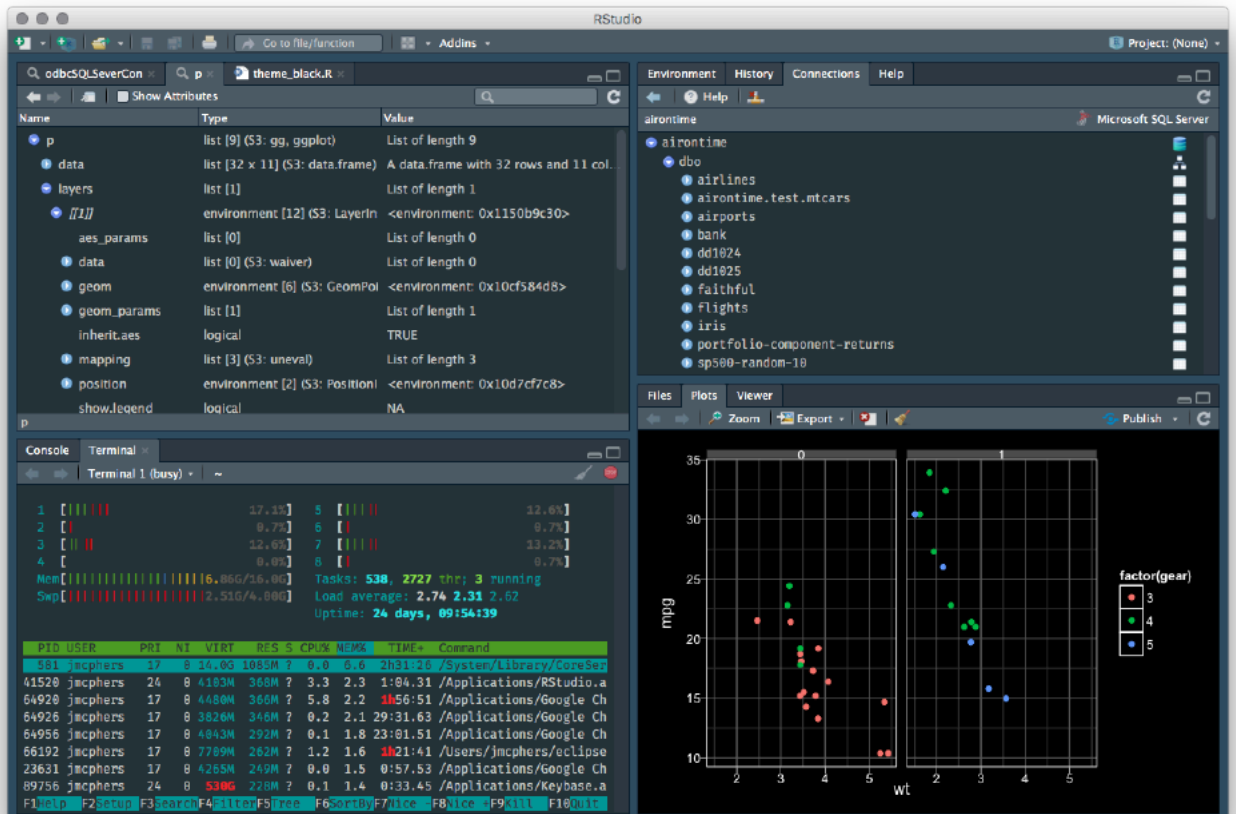
In questo corso:

- 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


RStudio
Go to file/function
Addins
Project: (None)
odbcSQLSeverCon
p
theme_black.R
Show Attributes
Name
Type
Value
p
list [9] (S3: gg, ggplot)
List of length 9
data
list [32 x 11] (S3: data.frame)
A data.frame with 32 rows and 11 col..
layers
list [1]
List of length 1
[[1]]
environment [12] (S3: LayerIn
<environment: 0x1150b9c30>
aes_params
list [0]
List of length 0
data
list [0] (S3: waiver)
List of length 0
geom
environment [6] (S3: GeomPoi
<environment: 0x10cf584d8>
geom_params
list [1]
List of length 1
inherit.aes
logical
TRUE
mapping
list [3] (S3: uneval)
List of length 3
position
environment [2] (S3: Position
<environment: 0x10d7cf7c8>
show.legend
logical
NA
Environment
History
Connections
Help
Help
airontime
Microsoft SQL Server
airontime
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)
Tasks: 538, 2727 thr; 3 running
Load average: 2.74 2.31 2.62
Uptime: 24 days, 09:54:39
PID USER PRI NI VIRT RES S CPU% MEM% TIME+ Command
F1Help F2Setup F3Search F4Filter F5Tree F6SortBy F7Nice F8Nice + F9Kill F10Quit
mpg
wt
factor(gear)
3
4
5

Tassonomia dei Dati

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

Tassonomia dei Dati

Dati sezionali (cross-section)

```
library(AER)
data(CASchools)
head(CASchools |> select(district, school, students, teachers, read, math), 10)
```

```
   district                          school students teachers  read  math
1     75119              Sunol Glen Unified      195    10.90 691.6 690.0
2     61499            Manzanita Elementary      240    11.15 660.5 661.9
3     61549     Thermalito Union Elementary     1550    82.90 636.3 650.9
4     61457 Golden Feather Union Elementary      243    14.00 651.9 643.5
5     61523        Palermo Union Elementary     1335    71.50 641.8 639.9
6     62042         Burrel Union Elementary      137     6.40 605.7 605.4
7     68536           Holt Union Elementary      195    10.00 604.5 609.0
8     63834             Vineland Elementary      888    42.50 605.5 612.5
9     62331        Orange Center Elementary      379    19.00 608.9 616.1
10    67306     Del Paso Heights Elementary     2247   108.00 611.9 613.4
```

Tassonomia dei Dati

Dati sezionali (cross-section)

```
# A tibble: 10 × 7
   RETRIC  ETAM EDULEV        REG                     C27 SG11    STACIM
    <dbl> <dbl> <chr>         <chr>                 <dbl> <chr>   <chr>
 1   1290    49 Diploma 4-5   Veneto                    1 Maschio Celibe/nubile
 2   1200    56 Diploma 2-3   Lombardia                 1 Maschio Coniugato/a
 3    600    38 Laurea        Sicilia                   2 Femmina Celibe/nubile
 4    720    62 Licenza media Piemonte                  2 Femmina Celibe/nubile
 5   1300    55 Licenza media Valle d'Aosta             1 Maschio Celibe/nubile
 6   3000    42 Laurea        Friuli Venezia Giulia     1 Maschio Celibe/nubile
 7   1450    48 Licenza media Lombardia                 1 Maschio Separato/a o …
 8    650    30 Laurea        Lombardia                 2 Femmina Coniugato/a
 9   3000    30 Diploma 4-5   Lombardia                 1 Maschio Coniugato/a
10   1030    58 Diploma 2-3   Lombardia                 2 Femmina Coniugato/a
```

Tassonomia dei Dati

Dati panel

```
data(Fatality)
head(Fatality, 11)
```

```
   state year   mrall   beertax  mlda jaild comserd   vmiles unrate   perinc
1      1 1982 2.12836 1.5393795 19.00    no      no 7.233887   14.4 10544.15
2      1 1983 2.34848 1.7889907 19.00    no      no 7.836348   13.7 10732.80
3      1 1984 2.33643 1.7142856 19.00    no      no 8.262990   11.1 11108.79
4      1 1985 2.19348 1.6525424 19.67    no      no 8.726917    8.9 11332.63
5      1 1986 2.66914 1.6099070 21.00    no      no 8.952854    9.8 11661.51
6      1 1987 2.71859 1.5599999 21.00    no      no 9.166302    7.8 11944.00
7      1 1988 2.49391 1.5014436 21.00    no      no 9.674323    7.2 12368.62
8      4 1982 2.49914 0.2147971 19.00   yes     yes 6.810157    9.9 12309.07
9      4 1983 2.26738 0.2064220 19.00   yes     yes 6.587495    9.1 12693.81
10     4 1984 2.82878 0.2967033 19.00   yes     yes 6.709970    5.0 13265.93
11     4 1985 2.80201 0.3813559 21.00   yes     yes 6.771263    6.5 13726.70
```

Tassonomia dei Dati

Dati temporali (time-series)

```
# A tibble: 12 × 8
   yyyymm      ret    AAA    BAA   d12      infl  `d/y`  `e/p`
    <dbl>    <dbl>  <dbl>  <dbl> <dbl>     <dbl>  <dbl>  <dbl>
 1 202301  0.0638  0.044  0.055   67.4  0.00800  0.0175 0.0426
 2 202302 -0.0243  0.0456 0.0559  67.8  0.00558  0.0166 0.0439
 3 202303  0.0375  0.046  0.0571  68.2  0.00331  0.0172 0.0426
 4 202304  0.0150  0.0447 0.0553  68.4  0.00506  0.0166 0.0425
 5 202305  0.00531 0.0467 0.0577  68.5  0.00252  0.0164 0.0428
 6 202306  0.0670  0.0465 0.0575  68.7  0.00323  0.0164 0.0407
 7 202307  0.0317  0.0466 0.0574  68.9  0.00191  0.0155 0.0397
 8 202308 -0.0154  0.0495 0.0602  69.1  0.00437  0.0151 0.0406
 9 202309 -0.0476  0.0513 0.0616  69.3  0.00249  0.0154 0.0430
10 202310 -0.0210  0.0561 0.0663  69.6 -0.000383 0.0162 0.0446
11 202311  0.0908  0.0528 0.0629  70.0 -0.00202  0.0167 0.0415
12 202312  0.0453  0.0474 0.0564  70.3 -0.000993 0.0154 0.0403
```

Variabili

Tipologie principali

1. **Continue**
   - Possono assumere qualsiasi valore in un intervallo
   - Esempio: salari, prezzi
2. **Discrete**
   - Assumono solo valori specifici
   - Esempio: numero di figli
3. **Categoriche**
   - Classificano in gruppi
   - Esempio: regione di residenza

Notazione

Singola Variabile

- x_i = i-esimo valore osservato
- $i = 1, ..., n$ indica l'osservazione

Esempio

Se x = salario:

- x_1 = salario primo lavoratore
- x_2 = salario secondo lavoratore

Campione:

$$(x_1, x_2, ..., x_n)$$

Notazione Multivariata

Multiple Caratteristiche

Per l'osservazione i:

- x_{1i} = prima caratteristica (es: salario)
- x_{2i} = seconda caratteristica (es: età)
- ⋮
- x_{ki} = k-esima caratteristica (es: istruzione)

Dataset Completo

$$\left\{ x_{1i}, x_{2i}, ..., x_{ki} \right\}_{i=1}^{n}$$

Convenzione Importanti

- Primo vs Secondo Pedice
  - → Primo pedice (x_{1i}): quale caratteristica
  - → Secondo pedice (x_{1i}): quale osservazione

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

Nota: Quando usiamo la parola effetto ci riferiamo alleffetto causale.

I dati della California

Dati sui Distretti scolastici K-6 e K-8 (n = 420)

```
head(CASchools |> select(district, students, teachers, read, math), 12)
```

	district	students	teachers	read	math
1	75119	195	10.90	691.6	690.0
2	61499	240	11.15	660.5	661.9
3	61549	1550	82.90	636.3	650.9
4	61457	243	14.00	651.9	643.5
5	61523	1335	71.50	641.8	639.9
6	62042	137	6.40	605.7	605.4
7	68536	195	10.00	604.5	609.0
8	63834	888	42.50	605.5	612.5
9	62331	379	19.00	608.9	616.1
10	67306	2247	108.00	611.9	613.4
11	65722	446	21.00	612.8	618.7
12	62174	987	47.00	616.6	616.0

Variabili:

- Media (Stanford-9 achievement test)

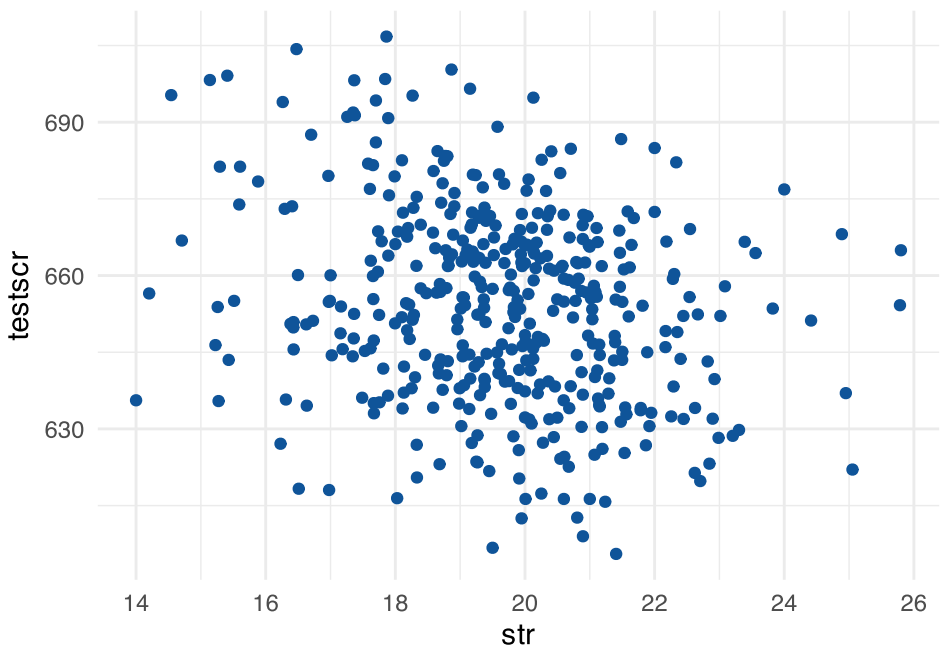
$$testscr = \frac{\text{mathscr} + \text{readscr}}{2}$$

- Rapporto studenti/insegnanti

$$str = \frac{\#\text{studenti}}{\#\text{insegnanti}} = \frac{\text{students}}{\text{teachers}}$$

Le scuole della California

```
CASchools <- CASchools |> mutate(testscr = (read+math)/2,  str = students/teachers)
library(ggplot2)
ggplot(CASchools, aes(y=testscr,x=str)) + geom_point(color=blue) + theme_minimal()
```



Primo sguardo ai dati:

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"* $(STR \geq 20)$

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 (nella popolazione)
2. Verifica dell'ipotesi che $\Delta = 0$
3. Costruire un *intervallo di confidenza* per Δ

Stima

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

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

È una differenza sufficientemente grande da risultare importante per discussioni sulla riforma della scuola, per i genitori o per un comitato scolastico?

Stima

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

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

È una differenza sufficientemente grande da risultare importante per discussioni sulla riforma della scuola, per i genitori o per un comitato scolastico?

- Deviazione standard tra i distretti $\Longrightarrow 19.1$
- Differenza tra 60-esimo and 75-esimo percentili della distribuzione dei punteggi nei test $\Longrightarrow 667.6 - 659.4 = 8.2$

Verifica di ipotesi

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

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

dove:

$SE\left(\overline{Y}_s - \overline{Y}_l\right)$ è l' errore standard di $\overline{Y}_s - \overline{Y}_l$

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

$s_l^2 = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \left(Y_i - \overline{Y}_l\right)^2$

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

$$t = \frac{\overline{Y}_s - \overline{Y}_l}{SE\left(\overline{Y}_s - \overline{Y}_l\right)} = \frac{7.4}{1.83} = 4.05$$

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

Intervallo di confidenza

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

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

Due affermazioni equivalenti:

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

E ora...

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

Percorso

Quadro di riferimento probabilistico per l'inferenza statistica

1. Stima
2. Verifica di ipotesi
3. Intervalli di confidenza

Concetti:

- Popolazione, variabile casuale e distribuzioni marginali e congiunte
- Momenti (media, varianza, deviazione standard, covarianza, correlazione)
- Distribuzione condizionata e media condizionata

Popolazione

- Il gruppo o l'insieme di tutte le possibili unità di interesse

Popolazione Finita

- Insieme definito di unità in un momento specifico
- Es: 17 milioni di lavoratori italiani nel 2018
- Limitata nel tempo e nello spazio

Superpopolazione

- Insieme infinito di possibili popolazioni
- Generate dagli stessi meccanismi sottostanti
- Trascende il contesto specifico

L’Esempio delle Scuole Californiane

`CASchool` ha dati su tutte le scuole della California nel 2008

Approccio Superpopolazione

- Scuole CA come realizzazione di un processo più ampio
- Interesse nelle relazioni generali (es: dimensione classi e apprendimento)
- Permette generalizzazione ad altri contesti

Incertezza campionaria

L'Analogia dell'Urna

- Popolazione = urna gigante con tanti (infiniti) biglietti
- Ogni "biglietto" = un'unità con le sue caratteristiche
- L'estrazione = processo di campionamento

Elementi di Incertezza

- Non sappiamo quale unità estrarremo
- Ogni estrazione è casuale
- Il risultato è aleatorio

Dai Valori Osservati alle Variabili Aleatorie

Valori Osservati (lettere minuscole)

- x_1 = primo valore estratto
- x_2 = secondo valore estratto
- x_3 = terzo valore estratto
- $\vdots$
- x_n = n-esimo valore estratto

Variabili Aleatorie (lettere maiuscole)

- X_1 = prima estrazione potenziale
- X_2 = seconda estrazione potenziale
- X_3 = terza estrazione potenziale
- $\vdots$
- X_n = n-esimo valore potenziale

I valori osservati sono realizzazioni specifiche delle variabili aleatorie

Definizione Formale

Variabile Aleatoria

Una funzione che associa a ogni possibile esito dell'esperimento (estrazione) un numero reale

Esempio: Salario dei lavoratori - X_i = variabile aleatoria del salario - Prima dell'estrazione: non conosciamo il valore - Dopo l'estrazione: osserviamo x_i

Distribuzione di Y:

- Le probabilità di diversi valori di Y che si verificano nella popolazione

  Esempio:

  → $Pr(Y = 650)$ (quando Y è discreta)

  → $\Pr(640 \leq Y \leq 660)$ (quando Y è continua).

Momenti

valore atteso di Y

Momenti

$$E(Y) = \mu_Y$$

varianza di Y

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

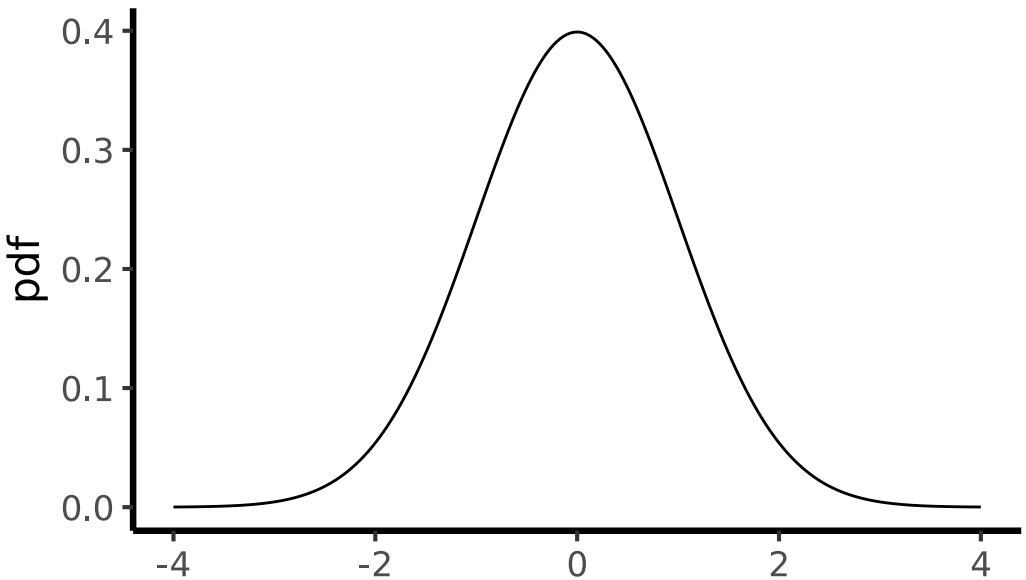
deviazione standard di Y

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

```
x <- seq(-4,4, by = 0.01)
y <- dnorm(x)
df <- data.frame(x=x, y=y)
ggplot(df, aes(x=x, y=y)) + geom_line() + theme_gragusa() + xlab("") + ylab("pdf")
```

```
Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
ℹ Please use the `linewidth` argument instead.
```

Momenti, ctd.



Momenti, ctd.

asimmetria

Momenti, ctd.

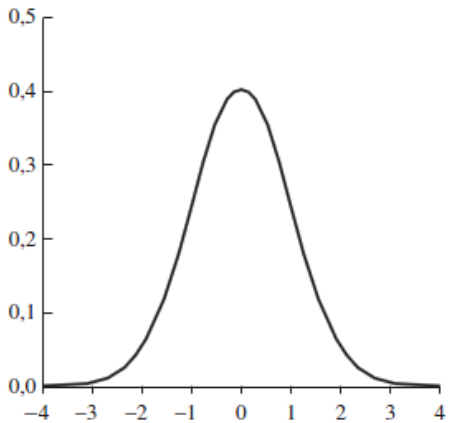
$$\frac{E\left[(Y-\mu_Y)^3\right]}{\sigma_Y^3}$$

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

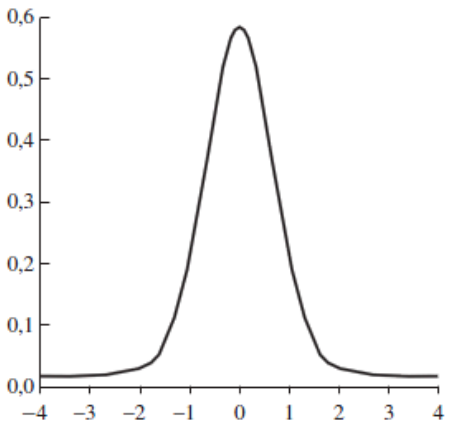
curtosi

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

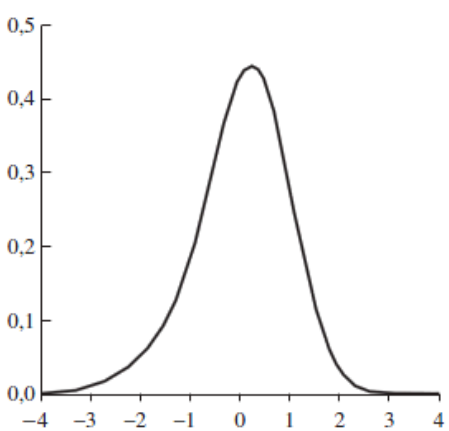
Momenti, ctd.



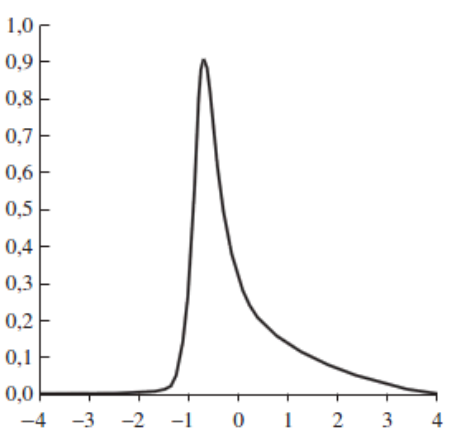
(a) Asimmetria = 0, Curtosi = 3



(b) Asimmetria = 0, Curtosi = 20



(c) Asimmetria = –0,1, Curtosi = 5



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

Il coefficiente di correlazione

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

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

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

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

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

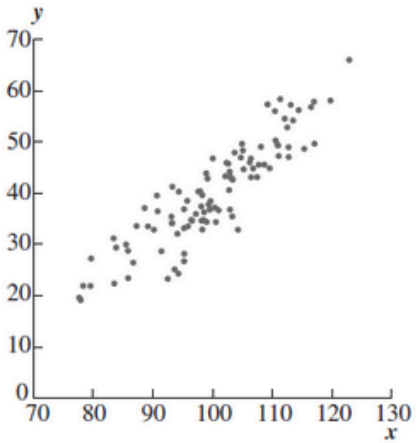
Il coefficiente di correlazione

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

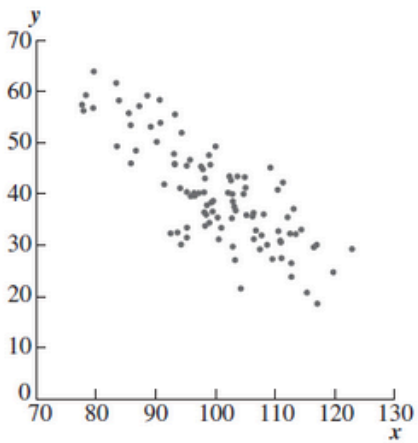
$$corr(X, Z) = \frac{cov(X, Z)}{\sqrt{var(X)var(Z)}} = \frac{cov(X, Z)}{sd(X)sd(Z)} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = \rho_{XZ}$$

- $-1 \leq corr(X, Z) \leq 1$
- $corr(X, Z) = 1$ 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

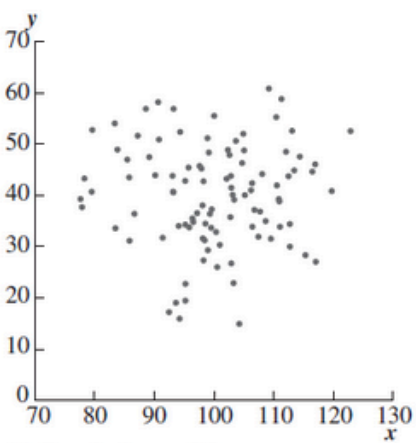
Il coefficiente di correlazione misura l'associazione lineare



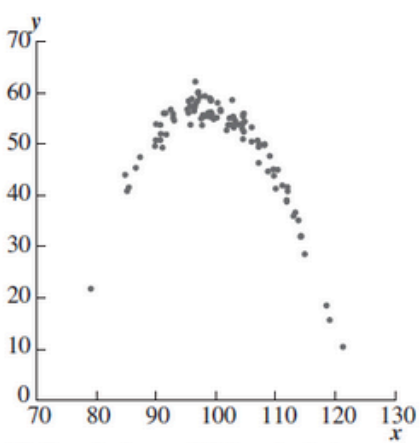
(a) Correlazione = +0,9



(b) Correlatzone = –0,8



(c) Correlazione = 0,0



(d) Correlazione = 0,0 (quadratica)

Momenti condizionati

- **Valore Atteso Condizionato**: Il valore atteso condizionato,

$$E(Y \mid X = x),$$

  rappresenta il valore atteso Y quando $X = x$. Questo concetto è fondamentale nell'econometria per comprendere come il valore atteso di una variabile risponda a cambiamenti in un'altra variabile.

- **Varianza Condizionata**: La varianza condizionata, denotata come

$$Var(Y \mid X = x),$$

  misura la variabilità di Y quando $X = x$. Ci dice quanto i valori di Y si discostano dal loro valore atteso condizionato $E(Y \mid X = x)$ quando X è noto.

`testscr` e `str`

Media condizionata

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

$$E(testscr \mid str < 20)$$

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

$$E(testscr \mid str \geq 20)$$

`Altri esempi:`

- Valore atteso salario per lavoratori di genere femminile

$$E(Salario \mid genere = femminile)$$

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

$$E(Y \mid X = trattato)$$

Media condizionata

Campione

! Importante

$$E(X \mid Z) = \text{costante} \implies corr(X, Z) = 0.$$

Tuttavia, non vale necessariamente il contrario.

Se $E(X \mid Z) = \text{costante}$ il valore atteso di X è lo stesso indipendentemente da Z: conoscere (Z) non fornisce alcuna informazione aggiuntiva su (X) rispetto a quella già nota dalla distribuzione marginale di (X). Ciò implica che non c'è una relazione lineare tra X e Z.

Campione

$$(Y_1, ..., Y_n)$$

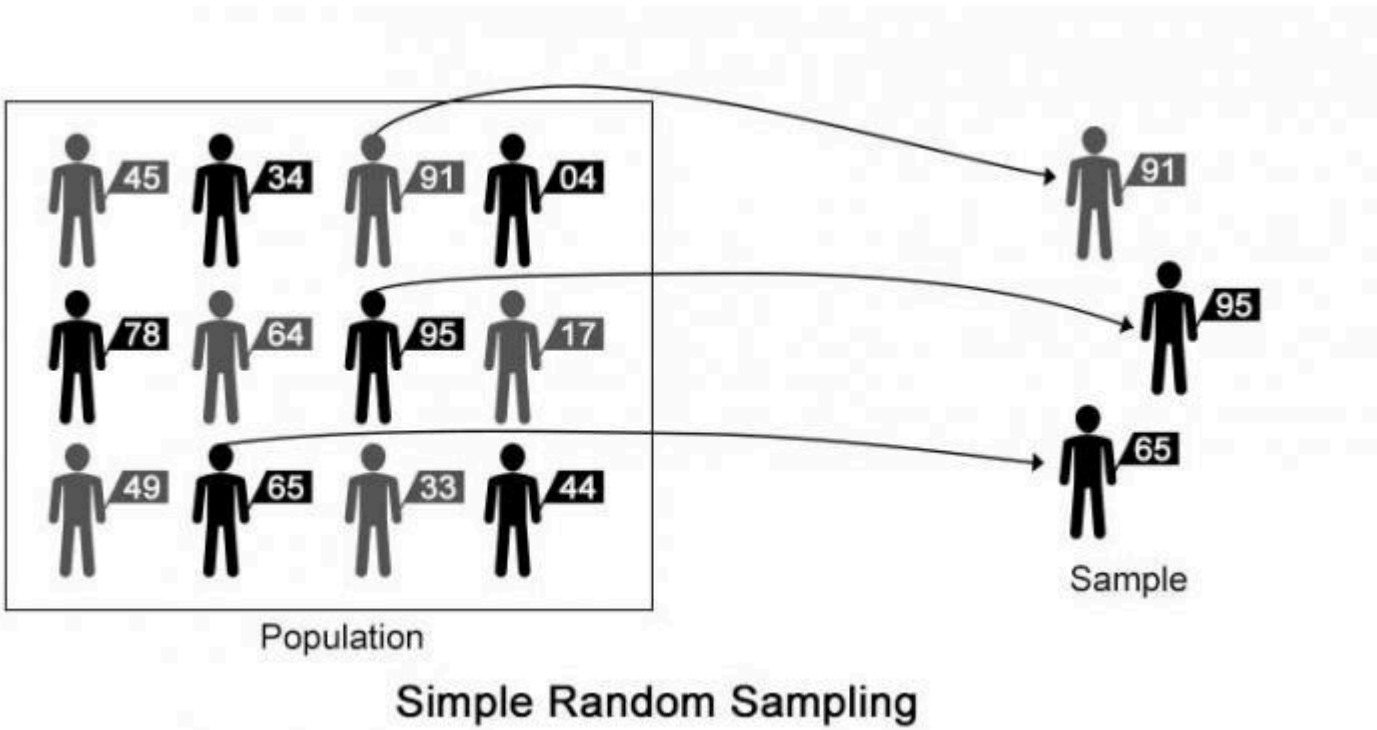
Campionamento casuale semplice

- Individui scelti a caso dalla popolazione

Campione

- Il data set è $(Y_1, Y_2, ..., Y_n)$
- 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



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, $\{Y_i\}$, $i = 1, ..., n$, sono ***i.i.d.***

Percorso:

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

Concetti:

La distribuzione campionaria di $\overline{Y}$

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

Ma:

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

La distribuzione campionaria di $\overline{Y}$

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

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

La distribuzione campionaria di $\overline{Y}$

La distribuzione campionaria di $\overline{Y}$

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

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

La distribuzione campionaria di $\overline{Y}$

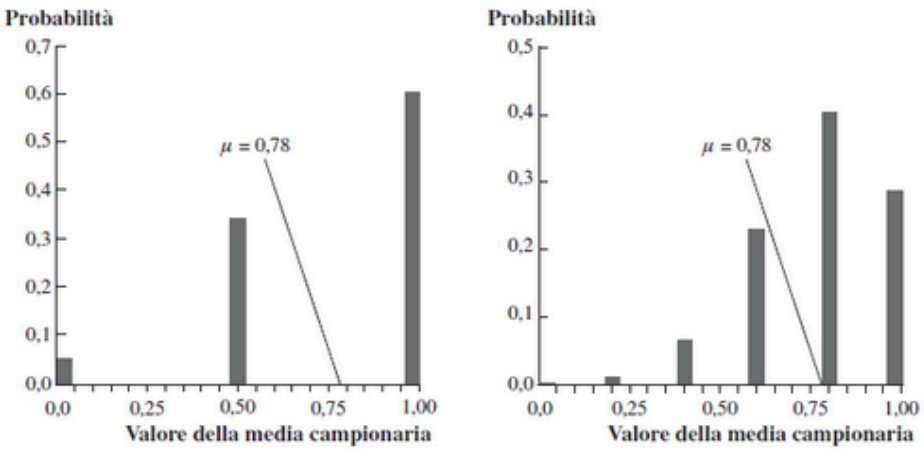
La distribuzione campionaria di $\overline{Y}$ dipende da n.

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

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

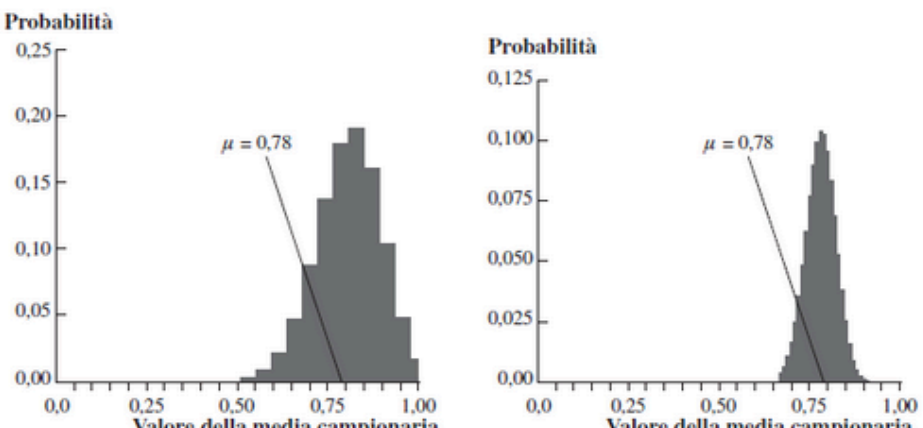
Potremmo calcolare la distribuzione per $n = 5$, $n = 25$, $n = 100$, e cosi' via....

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



(a) $n = 2$

(b) $n = 5$



(c) $n = 25$

(d) $n = 100$

Vogliamo sapere:

Vogliamo sapere:

- Qual è il valore atteso di $\overline{Y}$?

  $\rightarrow$ Se $E\left(\overline{Y}\right) = \mu = 0.78$, allora $\overline{Y}$ è uno stimatore ***non distorto*** di μ

- Qual è la varianza di $\overline{Y}$?

  $\rightarrow$ In che modo $var\left(\overline{Y}\right)$ dipende da ***n***?

- $\overline{Y}$ si avvicina a μ quando ***n*** è grande?

  $\rightarrow$ Legge dei grandi numeri: $\overline{Y}$ è uno stimatore ***consistente*** di μ

- Distribuzione di $\overline{Y}$

  $\rightarrow$ $\overline{Y} - \mu$ è approssimato da una distribuzione normale per ***n*** grande (teorema limite centrale)

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

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

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

- valore atteso

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

- varianza

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

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

$$E\left(\overline{Y}\right) = \mu_Y$$

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

$$var\left(\overline{Y}\right) = \frac{\sigma_Y^2}{n}$$

$$sd\left(\overline{Y}\right) = \frac{\sigma_Y}{\sqrt{n}}$$

Implicazioni:

- $\overline{Y}$ è uno stimatore non distorto di μ_Y
- $var\left(\overline{Y}\right)$ 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 $\overline{Y}$ quando $n \to \infty$

- Per piccoli campioni, la distribuzione di $\overline{Y}$ è complicata

Legge dei grandi numeri

- 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 $\overline{Y} - \mu_Y$ può essere approssimata da una normale

Legge dei grandi numeri

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

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

per ogni $\epsilon > 0$.

Spesso scriviamo $\overline{Y} \xrightarrow{p} \mu_Y$, che significa che $\overline{Y}$ converge in probabilità a μ_Y.

Stimatore della varianza di Y

Teorema limite centrale (TLC)

Se $(Y_1, ..., Y_n)$ 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:

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

o, equivalentemente,

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

Stimatore della varianza di Y

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

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

Riepilogo: distribuzione di $\overline{Y}$

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 $\overline{Y}$

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

- La distribuzione campionaria esatta (campione finito) di $\overline{Y}$ ha media μ_Y ($\overline{Y}$ è uno stimatore non distorto di μ_Y) e varianza σ_Y^2/n

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

- Al di là di media e varianza, la distribuzione esatta di $\overline{Y}$ 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:

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

  - → Teorema del limite centrale:

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

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

- $\overline{Y}$ è **non distorto**: $E\left(\overline{Y}\right) = \mu_Y$
- $\overline{Y}$ è **consistente**: $\overline{Y} \xrightarrow{p} \mu_Y$

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

- $\overline{Y}$ è lo stimatore ***dei minimi quadrati*** di μ_Y; $\overline{Y}$ è la soluzione di questo problema

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

Derivazione: Le condizioni del primo ordine sono:

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

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

- ha una varianza minore di tutti gli altri ***stimatori lineari non distorti***

  $\longrightarrow$ 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

$$var\left(\overline{Y}\right) \leq var\left(\tilde{Y}\right)$$

- $\overline{Y}$ non è l'unico stimatore di μ_Y

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

Regola decisionale: Errore I tipo

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: Errore I tipo

Regola decisionale

Regola decisionale

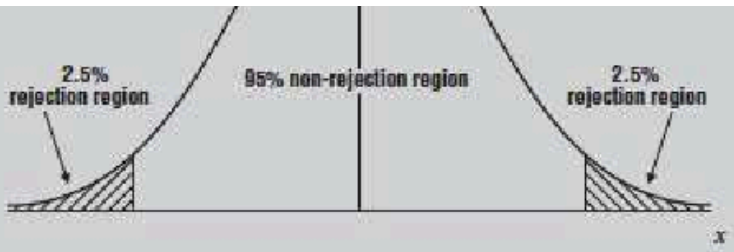


Figure 2.14
Rejection region for a one-sided hypothesis test of the form
$H_0 : \beta = \beta^*$,
$H_1 : \beta < \beta^*$

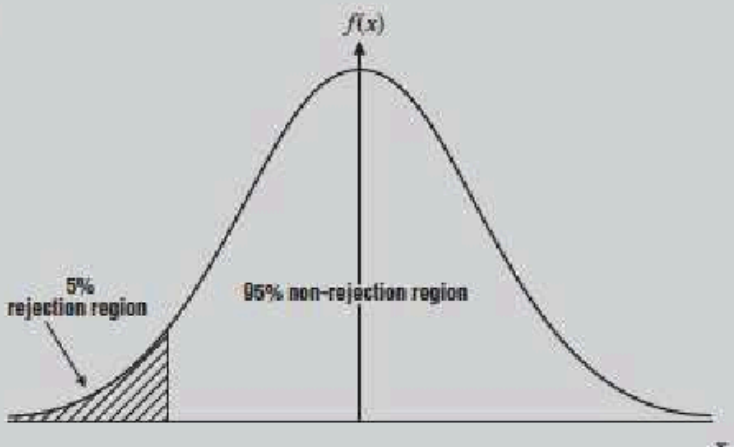
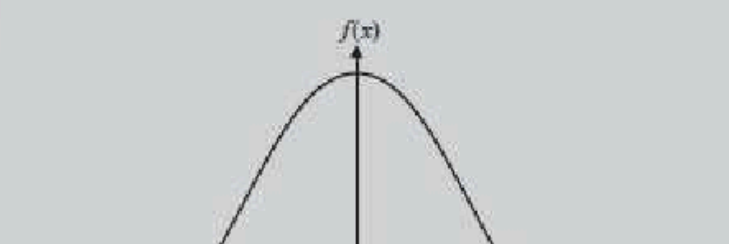


Figure 2.15
Rejection region for a one-sided hypothesis test of the form
$H_0 : \beta = \beta^*$,
$H_1 : \beta > \beta^*$



Terminologia per la verifica di ipotesi

- Il *livello di significatività* (α) di un test è la probabilità 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 α per il quale non è possibile rifiutare l'ipotesi nulla

Calcolo del valore-p:

$$\text{valore} - \text{p} = \Pr\left(\left|\overline{Y} - \mu_{Y,0}\right| > \left|\overline{Y}^{act} - \mu_{Y,0}\right|\right)$$

dove $\overline{Y}^{act}$ è il valore effettivamente osservato di $\overline{Y}$

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

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

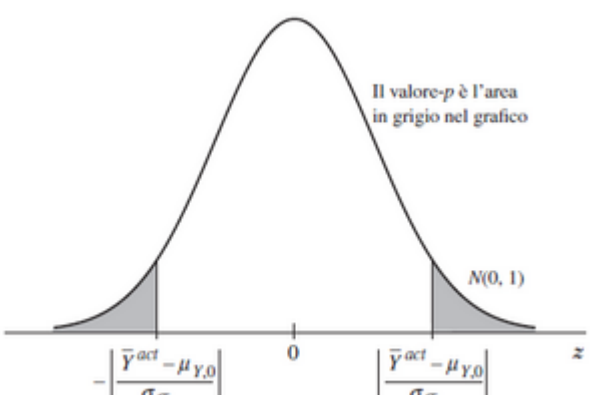
Calcolo del valore-p con σ_Y^2 stimato

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

Calcolo del valore-p con σ_Y^2 stimato

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

- 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 $|t| \geq 1.96$.
- in modo equivalente, la si rifiuta se $p \leq 0.05$
- 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**

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%.

$$\begin{aligned}\{\mu_Y : |t| \leq 1.96\} &= \{\mu_Y : -1.96 \leq t \leq 1.96\} \\ &= \left\{\overline{Y} - 1.96 \times \frac{s_Y}{\sqrt{n}}, \overline{Y} + 1.96 \times \frac{s_Y}{\sqrt{n}}\right\}\end{aligned}$$

Covarianza

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

Covarianza

- Campione congiunto

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

- Obiettivo

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

- Stimatore

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

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

Covarianza

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

- $\hat{\sigma}_{Y,X}$ 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

Covarianza

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

$$\hat{\sigma}_{Y,X} \approx \frac{1}{n}\sum_{i=1}^{n} Z_i$$

dove $Z_i = (Y_i - \mu_Y)(X_i - \mu_X)$

Covarianza

- $\hat{\sigma}_{Y,X}$ è ($\approx$) la media campionaria di Z_i – una funzione di variabili indipendenti e quindi 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

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

dove

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

- Stima della varianza

Covarianza

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

Covarianza

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

$$\hat{\sigma}_{Y,X} \approx \frac{1}{n}\sum_{i=1}^{n} Z_i$$

dove $Z_i = (Y_i - \mu_Y)(X_i - \mu_X)$

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

Covarianza

```
## Intervallo di confidenza per covarianza fra testscr e str
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
```