



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

**DISIA**

Dipartimento di Statistica,  
Informatica, Applicazioni  
"Giuseppe Parenti"

# Note di Statistica

ultimo aggiornamento: 9 dicembre 2019

insegnamento di **Statistica (L-Z)**  
*CdS in Scienze e Tecniche Psicologiche*

a cura di **Bruno Bertaccini**

Materiale didattico a disposizione degli studenti,  
scaricabile all'indirizzo <http://local.disia.unifi.it/bertaccini>

È vietata la riproduzione non autorizzata a fini commerciali.



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

**DISIA**

Dipartimento di Statistica,  
Informatica, Applicazioni  
"Giuseppe Parenti"

## **Bruno Bertaccini**

Dipartimento di Statistica, Informatica, Applicazioni «G. Parenti»  
v.le Morgagni, 59 - Firenze

[bruno.bertaccini@unifi.it](mailto:bruno.bertaccini@unifi.it)

ricevimento: **su appuntamento (fissando data e luogo)**

orario delle lezioni / esercitazioni:

**Lunedì, Martedì e Giovedì**      **dalle 9:15 alle 10:45**  
**dal 30 settembre al 15 dicembre 2019**

È vietata la riproduzione non autorizzata a fini commerciali.

## Testo di riferimento e altro materiale didattico

- ❑ Agresti Alan e Finlay Barbara (2012)  
"Metodi statistici di base e avanzati per le scienze sociali". Pearson, Prentice Hall.
  
- ❑ queste dispense  
predisposte con l'obiettivo d'essere d'ausilio alla studio  
delle parti del testo da studiare  
(scaricabili su internet all'indirizzo <http://local.disia.unifi.it/bertaccini>)

## Modalità d'esame

Test scritto con domande di varia natura (vero/falso; risposta multipla, esercizi brevi).

Eventuale discussione orale dell'esito dello scritto.

## Appelli

Gli esami di profitto si svolgono in tre diverse sessioni, per complessivi otto appelli:

- ❑ sessione Invernale (tre appelli)
- ❑ sessione Estiva (tre appelli)
- ❑ sessione Autunnale (due appelli)

## INDICE (programma del corso)

- Introduzione alla statistica
- I Principi della Probabilità
- I Principi dell'Inferenza
- Note di Campionamento statistico
- Note di Inferenza parametrica (stima puntuale e per intervallo)
- Note di Inferenza parametrica (verifica d'ipotesi)
- Analisi dell'associazione tra variabili categoriali
- Analisi dell'associazione tra variabili quantitative
- Regressione lineare semplice
- Relazioni multivariate
- Regressione lineare multipla

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 1

# Introduzione

È vietata la riproduzione non autorizzata a fini commerciali.

## la Statistica...

è la disciplina che si occupa dell'elaborazione dei risultati dell'osservazione di uno o più caratteri posseduti dagli elementi di un insieme determinato, con l'intento di

- esprimere un giudizio e/o
- prendere una decisione

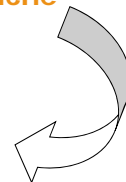
in merito ad alcuni aspetti di una realtà di interesse che, in quanto riferita ad un insieme e non ai singoli elementi che lo compongono, viene chiamata **fenomeno collettivo**.

## L'essenza della Statistica

La ragion d'essere della statistica è la presenza di un certo livello di variabilità nei dati (altrimenti, paradossalmente, la comprensione del fenomeno sarebbe possibile ricorrendo ad una sola osservazione)

nello studio dei fenomeni collettivi si è consapevoli che **al variare dell'unità statistica  $u$  entro una certa popolazione  $P = \{ u \}$  variano certe caratteristiche misurate su  $u$**

in altre parole,  
per lo studio di un fenomeno caratterizzato da assenza di variabilità **non serve scomodare uno statistico**



## L'essenza della Statistica

### Alcuni semplici esempi:

- altezza e peso degli studenti di una classe
- reddito dei parlamentari
- votazioni riportate all'esame di Statistica dagli studenti di un certo corso di studi universitario
- valutazione dell'efficacia dei titoli di studio universitari
- durata delle lampadine ad alto risparmio energetico
- soddisfazione nei confronti del trasporto pubblico locale
- ...

È vietata la riproduzione non autorizzata a fini commerciali.

## la Statistica...

... è quindi il  
**fondamento logico e metodologico  
per la risoluzione  
dei problemi decisionali  
in condizioni di incertezza**

È vietata la riproduzione non autorizzata a fini commerciali.

## Le branche della Statistica (1)

### □ Statistica Descrittiva

In questo settore rientrano i metodi per **sintetizzare** con opportune grandezze le caratteristiche salienti dei fenomeni collettivi.

La **descrizione** passa attraverso le fasi di formazione del dato statistico e del trattamento matematico dello stesso.

Per formazione del dato statistico si può intendere:

- l'elaborazione di dati preesistenti in natura (dati anagrafici, indici aziendali di bilancio, dati di produzione industriale ecc.)
- la necessità di procedere **all'effettiva rilevazione delle informazioni** necessarie alla comprensione del fenomeno di interesse

## Il processo di rilevazione delle informazioni

Il **processo di rilevazione delle informazioni** è generalmente distinto nelle fasi di:

- **definizione del piano di rilevazione,**
- **raccolta delle informazioni,**
- **spoglio e classificazione.**

La fase più delicata è senza dubbio la prima, soprattutto in relazione al tipo di fenomeno collettivo che si vuole indagare:

- altezza -> **metro**; peso -> **bilancia**
- reddito dei parlamentari -> **modello 730**
- performance esami di profitto -> **voto conseguito**
- durata lampadine -> **cronometro**
- soddisfazione nei confronti del trasporto pubblico locale -> **???**

## La POPOLAZIONE

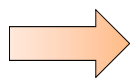
**Popolazione ( $P$ ):** insieme finito o infinito di unità che non interessano prese singolarmente ma per il contributo che danno allo studio del fenomeno collettivo d'interesse (**carattere**)  $F$ .

Se si è interessati alla conoscenza di un certo fenomeno  $F$  si possono rappresentare le sue possibili manifestazioni (**modalità del carattere**) come punti dell'insieme  $P$ .  
 Ovviamente non tutti i punti avranno lo stesso peso, perché può accadere che una determinata manifestazione si realizzi più frequentemente di un'altra.

**$N \rightarrow$  dimensione della Popolazione**

## Le rilevazioni campionarie

Stabilito con quale strumento misurare  $F$ :



- **Rilevazioni complete (censuarie)**
- **Rilevazioni campionarie**

**NB: la rilevazione completa è teoricamente semplice**

In realtà, motivazioni legate:

- alla **numerosità della Popolazione** (sovente non finita),
- ai **costi e/o ai tempi d'indagine**

inducono a optare per la **strategia campionaria**.

## Le rilevazioni campionarie

Fondamentale diviene quindi in statistica il ruolo dell'esperimento campionario.

**Campione:** un qualsiasi aggregato di unità statistiche appartenenti ad una certa popolazione e selezionate mediante una certa procedura.

$n$  → dimensione del campione

**NB1:** la strategia campionaria è la sola possibile quando:

- la popolazione è virtualmente infinita;
- l'osservazione è distruttiva.

**NB2:** la popolazione da cui si estrae il campione, detta popolazione campionata, non sempre coincide con la popolazione obiettivo.

## Le rilevazioni campionarie

**Importante distinzione:**

**Campioni probabilistici:**

- è possibile definire l'**insieme (Universo) di tutti i possibili campioni** che potrebbero formarsi seguendo una determinata procedura di estrazione di tipo randomizzato;
- è possibile associare a ciascun campione una probabilità di selezione nota;
- è possibile attribuire ad ogni unità componente la popolazione una probabilità strettamente positiva di essere estratta.

**Campioni non probabilistici:** tutti gli altri...



## Le rilevazioni campionarie

### I principali vantaggi derivanti dall'adozione di una strategia di campionamento

- contenere i costi dell'indagine entro limiti accettabili;
- svolgere l'indagine in tempi relativamente brevi;
- raccogliere per ogni unità inclusa nell'indagine un maggior numero di informazioni;
- raccogliere le informazioni con maggior accuratezza grazie all'utilizzazione di personale qualificato e/o di tecniche specialistiche.

È vietata la riproduzione non autorizzata a fini commerciali.

## Le rilevazioni campionarie

### ... però, distorsione indotta dal campionamento:

**in generale, un campione non costituisce quasi mai una riproduzione fedele della popolazione su piccola scala**

#### Inoltre :

- Distorsioni dovute alla risposte:** a causa di risposte non corrette o quesiti mal posti;
- Distorsioni dovute alle non-risposte:** a causa di soggetti campionati che rifiutano di partecipare o rispondere ad alcune domande del questionario.

È vietata la riproduzione non autorizzata a fini commerciali.

## Le rilevazioni campionarie

... quindi

(dato che molto spesso non possiamo fare a meno di condurre un'indagine campionaria):

- ❑ come estrarre il campione (secondo quale tecnica)?
- ❑ come estendere i risultati campionari all'intera popolazione?

## Le rilevazioni campionarie

Le fasi relative alla selezione del campione costituiscono il cosiddetto **disegno di campionamento**.

### Disegno di indagine

- definizione della popolazione obiettivo;
- scelta dei caratteri da studiare e dello strumento per misurarli;
- scelta dei domini spazio-temporali dell'indagine;
- definizione del **disegno di campionamento**;
- definizione dei metodi di raccolta, codifica ed elaborazione dati;
- definizione dei costi e dei livelli di precisione desiderati;
- definizione dei metodi di stima e di calcolo degli errori campionari;
- definizione dei metodi di controllo degli errori non campionari;
- analisi e presentazione dei risultati.



## Le branche della Statistica (2)

### □ Statistica Inferenziale

Se l'estrazione del campione è casuale, i dati possono fornire informazioni sulla variabilità della popolazione e sulla fiducia da accordare a tali informazioni. Questi problemi sono oggetto della Statistica Inferenziale o Induttiva.

**Il termine inferenza deriva dal latino e letteralmente significa: argomentare, desumere.**

**Si effettua inferenza quando si generalizza l'esperimento, operando una sorta di estensione dal particolare al generale;  
le generalizzazioni però non sono certe.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza statistica e probabilità

**L'inferenza è quindi un processo d'azzardo e l'incertezza viene misurata in termini probabilistici.**



**La PROBABILITÀ è il fondamento logico per fare inferenza sulla Popolazione oggetto d'indagine.**

Ma ...

**che cos'è la PROBABILITÀ?**

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 2

# I Principi della Probabilità

È vietata la riproduzione non autorizzata a fini commerciali.

## La Probabilità...

... è un concetto primitivo.

Per definirla occorre introdurre alcuni ingredienti:

❑ **esperimento casuale**  $\longrightarrow$   $\Omega$  spazio dei possibili  
risultati dell' esperimento  
(es: lancio del dado o di una moneta)

❑ **evento**

❑ **spazio degli eventi  $\mathcal{B}$**

La probabilità è una funzione matematica su  $\mathcal{B}$   
con certe proprietà

È vietata la riproduzione non autorizzata a fini commerciali.

## La rappresentazione degli eventi

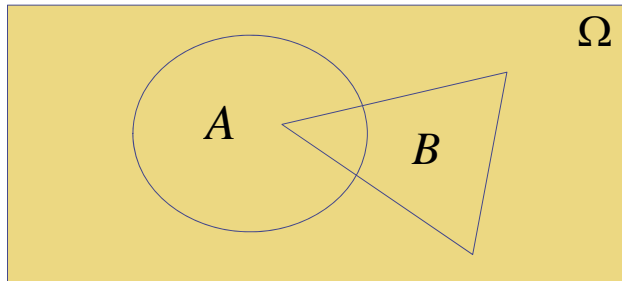


Diagramma di Venn

È vietata la riproduzione non autorizzata a fini commerciali.

## La probabilità

Approcci alla probabilità (*in ordine cronologico*):

- **impostazione Classica;**  
es: moneta
- **impostazione Frequentista;**  
es: moneta truccata
- **impostazione Soggettiva;**  
es: uomo su Marte
  
- **impostazione Assiomatica**

È vietata la riproduzione non autorizzata a fini commerciali.

### L'impostazione assiomatica delle Probabilità (Kolmogorov)

- 1)  $P(A) \geq 0$
- 2)  $P(\Omega) = 1$
- 3)  $P(A \cup B) = P(A) + P(B)$   
*se*  $A \cap B = \emptyset$

$$4) P(A / B) = \frac{P(A \cap B)}{P(B)} \quad \text{Principio delle Probabilità condizionate}$$

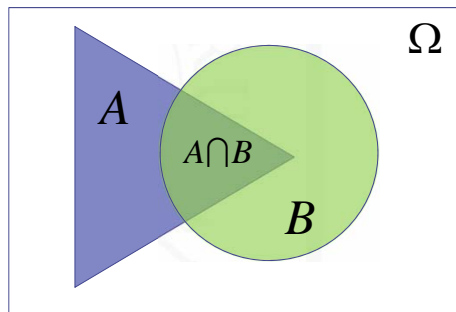
È una formalizzazione matematica di concetti intuitivi

È vietata la riproduzione non autorizzata a fini commerciali.

### L'impostazione assiomatica delle Probabilità (Kolmogorov)

In generale ( **Principio delle Probabilità Totali** ):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



NB: *se*  $A \cap B = \emptyset \Rightarrow$  assioma 3

È vietata la riproduzione non autorizzata a fini commerciali.

### Eventi incompatibili e indipendenti

se  $A \cap B = \emptyset \Rightarrow$  A e B sono eventi **incompatibili**

se  $P(A / B) = P(A)$

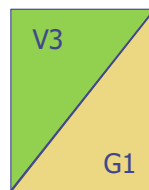
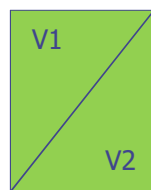
$\Rightarrow$  A e B sono eventi **indipendenti**

ovvero il verificarsi di B non incide sulla probabilità di A

se 
$$P(A / B) = \frac{P(A \cap B)}{P(B)} = P(A)$$

$\Rightarrow$  
$$P(A \cap B) = P(A) \cdot P(B)$$

### Il gioco delle tre cartine colorate

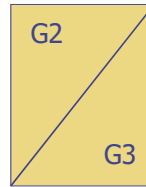
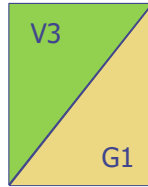
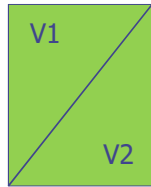


Si mescolano le carte e le facce (che, al di là del colore, sono indistinguibili), per cui mentre si mescola si possono anche ruotare le carte.

Quindi si estrae una carta e la si pone su un tavolo. Il colore che la carta mostra è il VERDE.

Ci si chiede quale sia la probabilità che quella carta mostri lo stesso colore anche sull'altra faccia.

## Il gioco delle tre cartine colorate



$$P(\text{osservare VERDE}) = P(V_1 \cup V_2 \cup V_3) = \frac{3}{6}$$

$$P(V_1 \cup V_2 / \text{VERDE}) = \frac{P((V_1 \cup V_2) \cap \text{VERDE})}{P(\text{VERDE})}$$

$$= \frac{P(V_1 \cup V_2)}{P(\text{VERDE})} = \frac{1/3}{3/6} = \frac{2}{3}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## dal Principio delle Probabilità condizionate...

$$P(A / B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B / A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = \begin{cases} P(A / B) \cdot P(B) \\ P(B / A) \cdot P(A) \end{cases}$$

e, in caso di indipendenza tra A e B:

$$\Rightarrow P(A \cap B) = P(A) \cdot P(B)$$

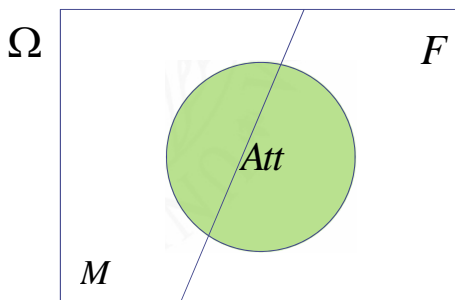
È vietata la riproduzione non autorizzata a fini commerciali.



### Teorema di Bayes

$$P(B / A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A / B) \cdot P(B)}{P(A)}$$

esempio:

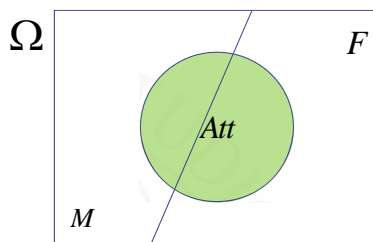


su una certa popolazione studentesca (Femmine e Maschi), è stato condotto un test psicometrico per la valutazione di un certo atteggiamento.

Punteggi elevati del test sono suscettibili di attenzione ( $Att$ )

È vietata la riproduzione non autorizzata a fini commerciali.

### Teorema di Bayes



Sono note le quote  $F$  e  $M$  per cui

$$P(M) = .45 \quad P(F) = 1 - P(M) = .55$$

Sulle due popolazioni il test ci dice che:

$$P(Att / M) = .60 \quad P(Att / F) = .25$$

Ma dato un punteggio del test selezionato casualmente, qual è la probabilità che questo sia stato prodotto da un maschio?

$$\begin{aligned} P(M / Att) &= \frac{P(M \cap Att)}{P(Att)} = \frac{P(Att / M) \cdot P(M)}{P(Att)} \\ &= \frac{P(Att / M) \cdot P(M)}{P((Att \cap M) \cup (Att \cap F))} \\ &= \frac{P(Att / M) \cdot P(M)}{P(Att / M) \cdot P(M) + P(Att / F) \cdot P(F)} = \frac{0.6 \cdot 0.45}{0.6 \cdot 0.45 + 0.25 \cdot 0.55} \cong 0.663 \end{aligned}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## I test diagnostici nelle scienze mediche

### Tipica applicazione del Teorema di Bayes:

- **D** = soggetto in presenza di una certa condizione patologica (*disease*) o fisiologica (es: gravidanza)
- **D'** = soggetto non in quella condizione
- **T+** = Test positivo (segnala la presenza della condizione)
- **T-** = Test negativo

Fino a che non si conosce l'esito del test, il generico soggetto ha una probabilità  $P(D)$  di essere nello status in questione.

Tale probabilità viene stimata tramite la cosiddetta *prevalenza* nella popolazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## I test diagnostici nelle scienze mediche

Il test viene però, in genere, messo a punto da un soggetto terzo (ente / azienda) che cerca di massimizzare la probabilità di identificare correttamente la presenza dello status partendo da due distinte popolazioni:

- quella dei soggetti per i quali la condizione è **certamente presente**:

$$P(T + / D) = \text{sensibilità del test}$$

- quella dei soggetti per i quali la condizione è **certamente assente**:

$$P(T - / D') = \text{specificità del test}$$

È vietata la riproduzione non autorizzata a fini commerciali.

**NB:**

- la **sensibilità** è quindi la capacità del test di **individuare correttamente la presenza della condizione**

e:

$$1 - P(T + / D) = P(T - / D) = \text{falso negativo}$$

- la **specificità** è quindi la capacità del test di **individuare correttamente l'assenza della condizione**

e:

$$1 - P(T - / D') = P(T + / D') = \text{falso positivo}$$

**Tipi di errore:**

Esito (Test) Status (natura delle cose)	<i>Negativo</i> (T-)	<i>Positivo</i> (T+)
<i>ASSENTE</i> (D')	<b>OK</b> P(T- D')	<i>falso positivo</i> P(T+ D')
<i>PRESENTE</i> (D)	<i>falso negativo</i> P(T- D)	<b>OK</b> P(T+ D)

I due tipi di errore sono concettualmente diversi.

**E tra loro ce n'è uno più grave dell'altro...**

Ma chi si sta accingendo a sottoporsi ad un test diagnostico, in realtà riterrebbe importante sapere con quale probabilità sarà interessato dalla condizione patologica o fisiologica, nell'eventuale presenza di un riscontro positivo.

In altre parole:

note:  $P(D)$

$$P(D') = 1 - P(D)$$

$$P(T + / D)$$

$$P(T - / D')$$

ci chiediamo quale sia:  $P(D / T +)$

Cosa accade in presenza di patologie molto rare?

Supponiamo:

$$P(D) = 3 / 10000 \quad \Rightarrow \quad P(D') = 9997 / 10000$$

e che il test abbia una sensibilità e una specificità molto elevate:

$$P(T + / D) = 0.95 \quad \Rightarrow \quad P(T - / D) = 1 - 0.95 = 0.05$$

$$P(T - / D') = 0.90 \quad \Rightarrow \quad P(T + / D') = 1 - 0.90 = 0.10$$

$$\begin{aligned}
 P(D/T+) &= \frac{P(T+ \cap D)}{P(T+)} = \frac{P(T+ \cap D)}{P((T+ \cap D) \cup (T+ \cap D'))} \\
 &= \frac{P(T+/D)P(D)}{P(T+/D)P(D) + P(T+/D')P(D')} \\
 &= \frac{0.95 \cdot 0.0003}{0.95 \cdot 0.0003 + 0.10 \cdot 0.9997} \\
 &= \frac{0.000285}{0.000285 + 0.09997} \cong 0.0028
 \end{aligned}$$

che è più grande di:  $P(D) = 3/10000$

ma molto più piccola di:  $P(T+/D) = 0.95$

È vietata la riproduzione non autorizzata a fini commerciali.

Talvolta i medici confondono  $P(D/T+)$  con  $P(T+/D)$  affermando che un soggetto positivo al test ha una probabilità  $P(T+/D)$ , nell'esempio precedente pari a 0.95, di presentare la condizione patologica o fisiologica in questione.

Nel caso di una patologia rara, se si riuscisse a sviluppare un test diagnostico con elevati valori di sensibilità e specificità, questo comunque produrrebbe **una quantità di falsi positivi che in proporzione sarebbe molto più elevati dei positivi reali.**

Questo fa capire perché gli screening di massa siano spesso problematici.

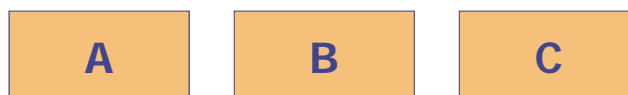
È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago:

- **Gioco del lotto:** prob. che esca 23 al secondo estratto
- **Mazzo di 40 carte:** prob. che esca un K alla seconda estraz.
- **Le 3 buste**
- **I 3 prigionieri**
- **35 studenti su uno scuolabus:** prob. che almeno 2 abbiano stessa data di nascita (gg/mm)
- **Il valore atteso ed il Paradosso di San Pietroburgo**
- **Come misurare l'area di un lago**

È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago: le 3 buste



Solo una contiene un grosso premio; le altre due sono vuote.  
 Giochiamo con Gino e proponiamo a Gino di scegliere una busta.

Gino sceglie la busta **A**.

Una volta scelta, facciamo vedere a Gino, aprendola, che una tra le buste **B** e **C** è vuota.

Offriamo a Gino la possibilità di poter cambiare la busta **A** con la busta chiusa rimasta sul tavolo.

**Il dubbio di Gino: cosa conviene fare?**

È vietata la riproduzione non autorizzata a fini commerciali.

### Un po' di svago: le 3 buste

Inizialmente:  $P(A_{vince}) = P(B_{vince}) = P(C_{vince}) = 1/3$

Supponiamo a Gino venga mostrato che B è vuota;  
 Gino lo considera un evento e condiziona la sua decisione a questo.

$$\begin{aligned}
 P(A_{vince}/B_{vuota}) &= \frac{P(A_{vince} \cap B_{vuota})}{P(B_{vuota})} \\
 &= \frac{P(B_{vuota}/A_{vince})P(A_{vince})}{P(B_{vuota})} \\
 &= \frac{P(B_{vuota}/A_{vince})P(A_{vince})}{P(B_{vuota}/A_{vince})P(A_{vince}) + P(B_{vuota}/C_{vince})P(C_{vince})} \\
 &= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{2}
 \end{aligned}$$

ovvero sapendo che "B è vuota" è indifferente conservare A o operare lo scambio

È vietata la riproduzione non autorizzata a fini commerciali.

### Un po' di svago: le 3 buste

Proviamo a consideriamo invece l'evento "mostriamo B vuota" e condiziamo la decisione di Gino a questo:

$$\begin{aligned}
 P(A_{vince}/mB_{vuota}) &= \frac{P(A_{vince} \cap mB_{vuota})}{P(mB_{vuota})} \\
 &= \frac{P(mB_{vuota}/A_{vince})P(A_{vince})}{P(mB_{vuota})} \\
 &= \frac{P(mB_{vuota}/A_{vince})P(A_{vince})}{P(mB_{vuota}/A_{vince})P(A_{vince}) + P(mB_{vuota}/C_{vince})P(C_{vince})} \\
 &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}
 \end{aligned}$$

si è indifferenti tra mostrare a Gino la busta B o la C, se A è la vincente.

ovvero il vero evento non è quello che Gino vede, ma l'azione che noi facciamo a seguito della scelta iniziale di Gino.  
**È certamente conveniente operare lo scambio.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago: i tre prigionieri

Tre prigionieri **A**, **B** e **C** l'indomani saranno condannati a morte.

Il Governatore decide di graziarne uno e comunica la sua scelta al secondino **S**, obbligandolo al silenzio sulla scelta fatta.

**A** chiede ad **S** di rivelargli il nome di chi si salverà.  
**S** non può parlare pena la sua esecuzione.

In alternativa, **A** chiede ad **S** di comunicargli il nome di chi degli altri due verrà sicuramente condannato.  
**S** accetta ritenendo di non contravvenire agli ordini ricevuti.

**A** adesso ritiene che la sua probabilità di salvarsi sia pari a  $\frac{1}{2}$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago: i tre prigionieri

Dopo la grazia del Governatore:  $P(A) = P(B) = P(C) = 1/3$

$$\begin{aligned}
 P(A/S_{diceB}) &= \frac{P(A \cap S_{diceB})}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB}/A)P(A) + P(S_{diceB}/C)P(C)} \\
 &= ?
 \end{aligned}$$

Occorre fare delle ipotesi.

**In primis, assumiamo che S non dica bugie ...**

È vietata la riproduzione non autorizzata a fini commerciali.



## Un po' di svago: i tre prigionieri

S è indifferente tra B e C:

$$\begin{aligned}
 P(A/S_{diceB}) &= \frac{P(A \cap S_{diceB})}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB}/A)P(A) + P(S_{diceB}/C)P(C)} \\
 &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}
 \end{aligned}$$

La probabilità di A non cambia (risultato analogo al gioco delle tre buste)

È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago: i tre prigionieri

Ad S è estremamente antipatico B, per cui se può fa il suo nome:

$$\begin{aligned}
 P(A/S_{diceB}) &= \frac{P(A \cap S_{diceB})}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB}/A)P(A) + P(S_{diceB}/C)P(C)} \\
 &= \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{2}
 \end{aligned}$$

S è costretto a fare il nome di B perché è C a salvarsi

S se può fa il nome di B

La probabilità di A sale a 1/2.

È vietata la riproduzione non autorizzata a fini commerciali.

## Un po' di svago: i tre prigionieri

Ad S è estremamente antipatico C, per cui se può fa il suo nome:

$$\begin{aligned}
 P(A/S_{diceB}) &= \frac{P(A \cap S_{diceB})}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB})} \\
 &= \frac{P(S_{diceB}/A)P(A)}{P(S_{diceB}/A)P(A) + P(S_{diceB}/C)P(C)}
 \end{aligned}$$

Se è A a salvarsi, S farebbe il nome di C per cui, in tal caso, la probabilità che dica B è zero

$$= \frac{0 \cdot \frac{1}{3}}{0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = 0$$

A è sicuramente condannato perché se S dice B, è certo che è stato costretto a dirlo visto che C si salverà

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 3

# I Principi dell'Inferenza

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza...

- ❑ **Deduttiva:** è un metodo per derivare informazioni da fatti accertati; le conclusioni cui si arriva con l'inferenza deduttiva sono definitive.

È l'inferenza che in matematica si usa per dimostrare i teoremi.

es: **SE** un triangolo rettangolo ha un angolo di  $90^\circ$  e il triangolo A è rettangolo **ALLORA** il triangolo A ha un angolo di  $90^\circ$

- ❑ **Induttiva:** si generalizza l'esperimento singolo alla classe di tutti gli esperimenti simili operando una sorta di estensione dal particolare al generale. Le generalizzazioni però non sono certe.

**L'inferenza induttiva è un processo d'azzardo e l'incertezza viene misurata in termini probabilistici.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza induttiva

Uno dei **compiti della statistica** è quello di fornire metodi per fare delle inferenze induttive e misurarne il grado di incertezza.

- ❑ **Inferenza Statistica Parametrica**  
si presuppone di conoscere il modello probabilistico caratterizzante il fenomeno oggetto di studio, **ma non si conoscono i suoi parametri.**
- ❑ **Inferenza Statistica Non Parametrica**  
non si conosce neanche il modello probabilistico caratterizzante il fenomeno oggetto di studio.

È vietata la riproduzione non autorizzata a fini commerciali.

## Scopo dell' Inferenza Statistica Parametrica...

... è utilizzare i risultati dell'esperienza campionario per giungere alla conoscenza (dei **parametri**) della **Popolazione** che ha generato quei risultati

*dai dati osservati per un campione*



*ad affermazioni che riguardano la popolazione*

È vietata la riproduzione non autorizzata a fini commerciali.

## La Popolazione e i suoi parametri

### Popolazioni finite

Una **Popolazione finita** è un insieme di unità su cui si può osservare un certo carattere. (es: gli investimenti annui di tutte le aziende di un paese; il numero di figli di ogni famiglia italiana)

I **parametri della popolazione** sono delle costanti che descrivono aspetti caratteristici della distribuzione del carattere nella popolazione stessa.

Es:  
 media e varianza della popolazione

### Popolazioni infinite

Una **Popolazione infinita** è composta da tutte le unità potenzialmente osservabili e non necessariamente già esistenti fisicamente.

Il carattere d'interesse può essere rappresentato da una variabile casuale con una certa distribuzione di probabilità. In questo caso si indicherà con "popolazione  $Y$ " la v.c.  $Y$ .

I **parametri della popolazione** sono le costanti caratteristiche della distribuzione di probabilità della v.c.  $Y$

È vietata la riproduzione non autorizzata a fini commerciali.

## La Popolazione e i suoi parametri

Parametri (costanti) di maggior interesse:

- **Totali** (occupati, forza lavoro, ...)
- **Medie** (reddito pro-capite, ...)
- **Proporzioni** (% di laureati, % di soddisfatti, ...)
- **Rapporti** (tra totali, tra medie, ecc.)

## Teoria della Stima

Attraverso l'osservazione di un campione si cerca di valutare un **parametro** (una costante) della Popolazione.

- Stima Puntuale**
- Stima per Intervallo**
  
- Verifica o Test di Ipotesi**

**NB:** tutte le affermazioni della statistica inferenziale sono incerte, **ma certe probabilisticamente**

## Teoria della Stima puntuale

- (domani) si estrae un campione casuale  $(Y_1, Y_2, \dots, Y_n) \in R^n$
- oggi i valori estratti non sono noti per cui  $(Y_1, Y_2, \dots, Y_n)$  è una v.c.
- si utilizza un'opportuna funzione di riduzione dei dati  $T_n$

$$T_n : R^n \rightarrow R$$

- $T_n$  è detta **statistica campionaria** se **NON dipende da altre quantità incognite**
- La statistica campionaria  $T_n$  è una v.c., in quanto è funzione delle v.c.  $(Y_1, Y_2, \dots, Y_n)$ .

$T_n$  assume valori nell'universo dei campioni per cui la sua distribuzione di probabilità è detta **distribuzione campionaria**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Un esempio di statistica (campionaria): la media campionaria

Media campionaria: **oggi è una v.c.** **domani è un numero**

$$T_n(Y_1, \dots, Y_n) = \bar{y} = \sum_{i=1}^n \frac{Y_i}{n} \longrightarrow \bar{y} = \sum_{i=1}^n \frac{y_i}{n}$$

- i valori che  $\bar{y}$  potrà assumere saranno in numero uguale al numero dei campioni e varieranno in funzione di tali campioni
- la distribuzione di  $\bar{y}$  dipenderà dalla distribuzione della Popolazione  $Y$  e sarà caratterizzata, come tutte le distribuzioni di probabilità, da una sua media, una sua varianza, ...



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

$$Var(\bar{y}) = Var\left(\sum_{i=1}^n \frac{Y_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^n Y_i\right) = ? \longrightarrow \begin{array}{l} \text{dipende} \\ \text{se le estrazioni} \\ \text{sono indipendenti} \\ \text{o meno ...} \end{array}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Stima puntuale

**Stimatore:** è una statistica (ovvero una funzione di v.c. che è essa stessa v.c.) **utilizzata per stimare il parametro incognito**  $\theta \in \Theta$

per esempio:  $\mu \in \mathbb{R}$   
 $\sigma^2 \in \mathbb{R}^+ \cup \{0\}$

$T(Y_1, Y_2, \dots, Y_n)$  oggi è una v.c.



$T(y_1, y_2, \dots, y_n) = t$  domani è un numero  
 ovvero **una stima di  $\theta$**

Supponendo di voler stimare  $\theta$

**qual è il miglior stimatore che possiamo utilizzare?**

È vietata la riproduzione non autorizzata a fini commerciali.

## Stimatori e stime

**Idealmente** vorremmo che, domani, **la stima**

$$T(y_1, y_2, \dots, y_n) = \theta \longrightarrow T \text{ stimatore ottimale}$$

qualunque sia il campione che estrarremo e qualunque sia il valore di  $\theta$ .



**NB: non esiste alcun metodo di stima che garantisca stimatori ottimali in tutte le situazioni**

$$d = t - \theta \longrightarrow \text{Errore campionario o errore di stima}$$

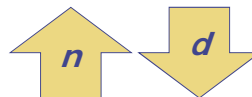
È vietata la riproduzione non autorizzata a fini commerciali.

## Stimatori e stime

L'errore di stima  $d$  non può in generale essere azzerato nell'indagine campionaria;  $d = 0$  solo nei censimenti (in assenza di non riposte).

Come cercare di ridurre  $d$  nell'indagine campionaria?

➤ dimensione  $n$  del campione



➤ Piano di campionamento

**NB:** per quanto detto in precedenza  $n$  non può essere aumentato a piacere.

È vietata la riproduzione non autorizzata a fini commerciali.

## Una proprietà degli estimatori

Uno stimatore  $T$  è **NON DISTORTO** sse

$$E(T) = \theta \quad \forall \theta \in \Theta$$

La non distorsione è da considerarsi più come un vincolo che come una proprietà auspicabile

È vietata la riproduzione non autorizzata a fini commerciali.



## Lezione 4

# Note di Campionamento Statistico

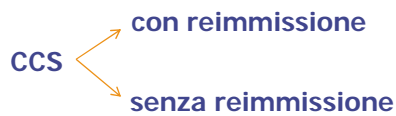
È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Casuale Semplice (CCS)

- ❑ È lo schema di campionamento più semplice: **corrisponde all'estrazione da un'urna** (tipo numeri della tombola).
- ❑ Le unità vengono scelte **CASUALMENTE** dalla lista e **ogni unità ha la stessa probabilità di entrare a far parte del campione**.
- ❑ **CASUALMENTE** però non vuol dire A CASACCIO.  
Il termine "CASUALE" è infatti strettamente connesso con quello di probabilità.
- ❑ Ci sono vari modi per fare un'estrazione casuale, tutti riferibili allo schema di estrazione da un'urna:
  - Tavola dei numeri casuali
  - Generazione di numeri casuali e estrazione con il calcolatore

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Casuale Semplice (CCS)



### Quali sono le differenze?

Supponiamo che ci sia una Popolazione di 4 studenti ( $N = 4$ )

$Y$  è la v.c. età degli individui (in anni)  
 Valori assunti da  $Y$ : 18, 20, 22, 25



Supponiamo (**domani**) di estrarre un campione di  $n = 2$  studenti: ( $Y_1, Y_2$ )

Oggi, ci chiediamo qual è la distribuzione di  $Y_1$  ? e quella di  $Y_2$  ?

È vietata la riproduzione non autorizzata a fini commerciali.

### CCS con reimmissione

$Y_1$ : 1° estratto;  $Y_2$ : 2° estratto ...

**NB:** le  $Y_i$  oggi sono v.c., domani saranno numeri

**NB<sub>2</sub>:** le  $Y_i$  saranno  $n$  v.c. **indipendenti** (perché l'estrazione è con rimessa) ciascuna delle quali:

- potrà assumere gli stessi valori della variabile  $Y$ ;
- avrà una distribuzione esattamente identica a quella della variabile  $Y$ .

➔  $Y_1 \sim Y_2 \sim \dots \sim Y_n \sim Y$

➔ le  $Y_i$  sono v.c. **I.I.D.** (indip. identicam. distribuite)

È vietata la riproduzione non autorizzata a fini commerciali.

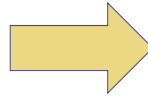
### CCS con reimmissione

Y	prob
18	1/4
20	1/4
22	1/4
25	1/4

Verifichiamola ricorrendo all'esempio dei 4 studenti:

Osservando l'insieme dei possibili risultati dell'estrazione:

(Y1, Y2)	Prob
(18 ; 18)	1/4*1/4 = 1/16
(18 ; 20)	1/4*1/4 = 1/16
(18 ; 22)	1/4*1/4 = 1/16
(18 ; 25)	1/4*1/4 = 1/16
(20 ; 18)	1/4*1/4 = 1/16
(20 ; 20)	1/4*1/4 = 1/16
(20 ; 22)	1/4*1/4 = 1/16
(20 ; 25)	1/4*1/4 = 1/16
(22 ; 18)	1/4*1/4 = 1/16
(22 ; 20)	1/4*1/4 = 1/16
(22 ; 22)	1/4*1/4 = 1/16
(22 ; 25)	1/4*1/4 = 1/16
(25 ; 18)	1/4*1/4 = 1/16
(25 ; 20)	1/4*1/4 = 1/16
(25 ; 22)	1/4*1/4 = 1/16
(25 ; 25)	1/4*1/4 = 1/16



Y1	prob
18	4/16 = 1/4
20	4/16 = 1/4
22	4/16 = 1/4
25	4/16 = 1/4

Y2	prob
18	4/16 = 1/4
20	4/16 = 1/4
22	4/16 = 1/4
25	4/16 = 1/4

È vietata la riproduzione non autorizzata a fini commerciali.

### CCS senza reimmissione

$Y_1$ : 1° estratto;  $Y_2$ : 2° estratto ...

**NB:** anche in questo caso le  $Y_i$  oggi sono v.c., domani saranno numeri

**NB<sub>2</sub>:** le  $Y_i$  saranno  $n$  v.c. **dependenti** (perché l'estrazione è ora senza rimessa) ciascuna delle quali:

- potrà assumere gli stessi valori della variabile  $Y$ ;
- avrà una distribuzione esattamente identica a quella della variabile  $Y$ .

➡  $Y_1 \sim Y_2 \sim \dots \sim Y_n \sim Y$  le  $Y_i$  sono v.c. **I.D.** (identicam. distribuite)

**NB:** dal momento che il campionamento è senza rimessa, la seconda proprietà può apparire di non così immediata comprensione.

È vietata la riproduzione non autorizzata a fini commerciali.

### CCS senza reimmissione

$Y$	$prob$
18	1/4
20	1/4
22	1/4
25	1/4

Verifichiamola ricorrendo all'esempio dei 4 studenti:

Osservando l'insieme dei possibili risultati dell'estrazione:

$(Y_1, Y_2)$	$Prob$
(18 ; 20)	$1/4 * 1/3 = 1/12$
(18 ; 22)	$1/4 * 1/3 = 1/12$
(18 ; 25)	$1/4 * 1/3 = 1/12$
(20 ; 18)	$1/4 * 1/3 = 1/12$
(20 ; 22)	$1/4 * 1/3 = 1/12$
(20 ; 25)	$1/4 * 1/3 = 1/12$
(22 ; 18)	$1/4 * 1/3 = 1/12$
(22 ; 20)	$1/4 * 1/3 = 1/12$
(22 ; 25)	$1/4 * 1/3 = 1/12$
(25 ; 18)	$1/4 * 1/3 = 1/12$
(25 ; 20)	$1/4 * 1/3 = 1/12$
(25 ; 22)	$1/4 * 1/3 = 1/12$

$Y_1$	$prob$
18	$3/12 = 1/4$
20	$3/12 = 1/4$
22	$3/12 = 1/4$
25	$3/12 = 1/4$

$Y_2$	$prob$
18	$3/12 = 1/4$
20	$3/12 = 1/4$
22	$3/12 = 1/4$
25	$3/12 = 1/4$

È vietata la riproduzione non autorizzata a fini commerciali.

### CCS con e senza reimmissione: riepilogo

**NB:** le  $Y_i$  oggi sono v.c., domani saranno numeri

**NB<sub>2</sub>:** quando l'estrazione è con rimessa, le  $Y_i$  sono  $n$  v.c. **indipendenti** ciascuna delle quali avrà una distribuzione esattamente identica a quella della variabile  $Y$  (Popolazione).

**NB<sub>3</sub>:** quando l'estrazione è senza rimessa, le  $Y_i$  sono  $n$  v.c. **dipendenti** ciascuna delle quali avrà una distribuzione esattamente identica a quella della variabile  $Y$  (Popolazione). In altre parole, la **distribuzione marginale** di  $Y_2$  (cioè quella senza alcun condizionamento ai possibili valori assunti dalla v.c.  $Y_1$ ) **non cambia**.

Quello che cambia è la distribuzione di  $Y_2$  condizionata ad  $Y_1$ , perché, ad ogni estrazione, la popolazione subisce un cambiamento in termini di frequenze relative.

È vietata la riproduzione non autorizzata a fini commerciali.

### L'esempio dei 4 studenti: CCS senza reimmissione

NB: l'esempio è puramente didattico.

In realtà i campioni si distinguono per la natura e non per l'ordine per cui, nel caso del CCS senza reimmissione, l'universo dei campioni  $\{s\}$  è di fatto formato da soli 6 campioni:

$$(Y_1, Y_2)_i; (Y_1, Y_3)_i; (Y_1, Y_4)_i; (Y_2, Y_3)_i; (Y_2, Y_4)_i; (Y_3, Y_4)_i$$

tutti con la stessa probabilità di essere estratti.

CCS senza reimmissione

➔  $\text{Prob}(s) = 2 \cdot 1/12 = 1/6$

### L'esempio dei 4 studenti: CCS con reimmissione

NB: poiché i campioni si distinguono per la natura e non per l'ordine, in caso di reimmissione invece l'universo dei campioni  $\{s\}$  è di fatto formato da 10 campioni:

$$(Y_1, Y_1)_i; (Y_1, Y_2)_i; (Y_1, Y_3)_i; (Y_1, Y_4)_i; (Y_2, Y_2)_i; (Y_2, Y_3)_i; (Y_2, Y_4)_i; (Y_3, Y_3)_i; (Y_3, Y_4)_i; (Y_4, Y_4)_i$$

Attenzione: NON tutti con la stessa probabilità di essere estratti.

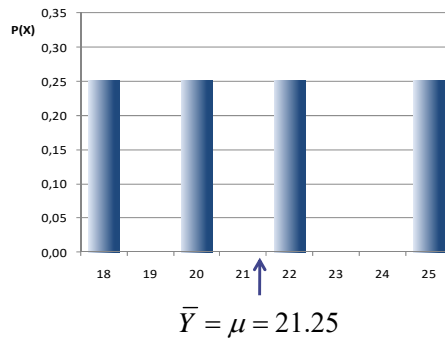
➔  $\text{Pr}((Y_i, Y_j)) = \begin{cases} 1/16 & \text{se } i = j \\ 1/8 & \text{se } i \neq j \end{cases}$

### Stima della media da CCS

esempio sulla Popolazione dei 4 studenti

$Y$	$prob$
18	1/4
20	1/4
22	1/4
25	1/4

$\bar{Y} = \mu = \frac{\sum_{i=1}^N Y_i}{N}$  media da stimare



È vietata la riproduzione non autorizzata a fini commerciali.

### Stima della media da CCS (senza reimmissione)

**stimatore**  $\rightarrow \bar{y} = \frac{\sum_{i=1}^n Y_i}{n}$  oggi è una v.c.

$s$	$\bar{y}$
(18 ; 20)	19
(18 ; 22)	20
(18 ; 25)	21,5
(20 ; 22)	21
(20 ; 25)	22,5
(22 ; 25)	23,5

domani è un numero  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$

**NB:** nessuna delle possibili medie campionarie che si possono verificare assume un valore identico alla media della popolazione

È vietata la riproduzione non autorizzata a fini commerciali.

## Stima della media da CCS (con reimmissione)

stimatore



$$\bar{y} = \frac{\sum_{i=1}^n Y_i}{n} \quad \text{oggi è una v.c.}$$



domani è un numero  $\bar{y} = \sum_{i=1}^n \frac{y_i}{n}$

s	P(s)	$\bar{y}$
(18 ; 18)	1/16	18
(18 ; 20)	1/8	19
(18 ; 22)	1/8	20
(18 ; 25)	1/8	21,5
(20 ; 20)	1/16	20
(20 ; 22)	1/8	21
(20 ; 25)	1/8	22,5
(22 ; 22)	1/16	22
(22 ; 25)	1/8	23,5
(25 ; 25)	1/16	25



- NB:** anche in questo caso nessuna delle possibili medie campionarie che si possono verificare assume un valore identico alla media della popolazione
- NB2:** le possibili medie campionarie hanno, ovviamente, probabilità di verificarsi equivalente a quella del relativo campione

È vietata la riproduzione non autorizzata a fini commerciali.

## La varianza della media campionaria

### CCS con reimmissione

$$\begin{aligned} \text{Var}(\bar{y}) &= \text{Var}\left(\sum_{i=1}^n \frac{Y_i}{n}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n Y_i\right) = \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

le variabili sono indipendenti per cui tutte le possibili covarianze sono nulle

- ❑ la varianza di tutte le medie di tutti i possibili campioni di dimensione  $n$  che potremmo estrarre è uguale alla varianza della Popolazione fratto  $n$
- ❑ la distribuzione della media campionaria è più concentrata della distribuzione della Popolazione, perché

$$\text{Var}(\bar{y}) = \frac{\sigma^2}{n} < \sigma^2 = \text{Var}(Y)$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La varianza della media campionaria

### CCS senza reimmissione

$$\begin{aligned} \text{Var}(\bar{y}) &= \text{Var}\left(\sum_{i=1}^n \frac{Y_i}{n}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n Y_i\right) = \\ &= \frac{1}{n^2} \left[ \sum_{i=1}^n \text{Var}(Y_i) + \sum_{i \neq j} \sum_{j=1}^n \text{Cov}(Y_i, Y_j) \right] = \dots \\ &= \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \end{aligned}$$

**Fattore di correzione per Popolazioni finite**

**NB:** in generale:  $\frac{N-n}{N-1} \leq 1$  per cui  $\text{Var}(\bar{y}) \leq \text{Var}(\bar{y})$   
senza rimessa                      con rimessa

- a) l'estrazione con rimessa coincide con quella senza rimessa quando  $n = 1$  o  $N \rightarrow \infty$
- b) il fattore di correzione tende a 1 quando  $N$  è molto grande rispetto a  $n$

## La varianza della media campionaria

(notazione alternativa)

### CCS senza reimmissione

Se indichiamo con:

$$S_{POP}^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1}$$

**la varianza elementare della Popolazione**

allora la varianza dello stimatore della media è:

$$\begin{aligned} \text{Var}(\bar{y}) &= \frac{S^2}{n} \left(1 - \frac{n}{N}\right) \\ &= \frac{S^2}{n} \left(\frac{N-n}{N}\right) \\ &= \frac{\sum (Y_i - \bar{Y})^2}{N-1} \frac{1}{n} \left(\frac{N-n}{N}\right) = \\ &= \frac{\sum (Y_i - \bar{Y})^2}{N} \frac{1}{n} \left(\frac{N-n}{N-1}\right) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1}\right) \end{aligned}$$

ovvero è la varianza dello stimatore della media nel caso con reimmissione, moltiplicata per il Fattore di correzione per Popolazioni finite



## La varianza della media campionaria: un esempio

### CCS con reimmissione

$\Omega$	$P(s)$	$\bar{y}$
(18, 18)	0,0625	18
(18, 20)	0,125	19
(18, 22)	0,125	20
(18, 25)	0,125	21,5
(20, 20)	0,0625	20
(20, 22)	0,125	21
(20, 25)	0,125	22,5
(22, 22)	0,0625	22
(22, 25)	0,125	23,5
(25, 25)	0,0625	25
<b>media</b>	<b>21,25</b>	
<b>varianza</b>	<b>3,34375</b>	

### CCS senza reimmissione

$\Omega$	$P(s)$	$\bar{y}$
(18, 20)	0,166667	19
(18, 22)	0,166667	20
(18, 25)	0,166667	21,5
(20, 22)	0,166667	21
(20, 25)	0,166667	22,5
(22, 25)	0,166667	23,5
<b>media</b>	<b>21,25</b>	
<b>varianza</b>	<b>2,229167</b>	

$$3.34 \cdot \frac{4-2}{4-1} = 2.23$$

fattore di correzione per Popolazioni finite

Entrambe le due distribuzioni non sono più uniformi a differenza della Popolazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## CCS senza reimmissione: codice R di simulazione

```
# installare preventivamente la libreria combinat
library(combinat)

# numerosità della POPOLAZIONE da variare a piacere (suggerirei minore di 20)
N=10
POP=sample(1:99,N,rep=F)
POP=sort(POP)

# numerosità del CAMPIONE da variare a piacere (suggerirei minore di 10)
n=5

# cardinalità dell'universo dei possibili campioni
choose(N,n)

# universo dei possibili campioni
# estrazioni senza ripetizione: ogni campione ha la stessa probabilità
# di essere estratto
U=combn(POP,n)

# la funzione t() traspone la matrice U (inverte righe con colonne)
U=t(U)

# la funzione paste() incolla alla stringa "X" il numero dell'estrazione
colnames(U)=paste("X",1:n,sep="")
```

È vietata la riproduzione non autorizzata a fini commerciali.

## CCS senza reimmissione: codice R di simulazione

```
# la funzione apply() applica ad ogni riga (ovvero ogni campione) di U
# la media e la varianza campionaria corretta
mu=apply(U,1,mean)
s2=apply(U,1,var)

# la funzione cbind() concatena per colonna ad U i vettori
# delle medie e delle varianze
U=cbind(U,mu,s2)

# calcolo dei valori attesi (media) di mu e s2 nell'Universo dei campioni
exp.mu=mean(mu)
exp.mu
exp.s2=mean(s2)
exp.s2

# calcolo della media e della varianza sigma^2 della POP
# che devono essere confrontati con i valori attesi calcolati su U
mu.POP=mean(POP)
mu.POP
s2.POP=var(POP)
s2.POP
```

È vietata la riproduzione non autorizzata a fini commerciali.

## Altre tecniche di campionamento (probabilistico)

- Campionamento Casuale Stratificato**
  - proporzionale
  - non proporzionale
- Campionamento Sistemático**
- Campionamento a Grappoli o a Stadi**

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento casuale Stratificato

### Metodo:

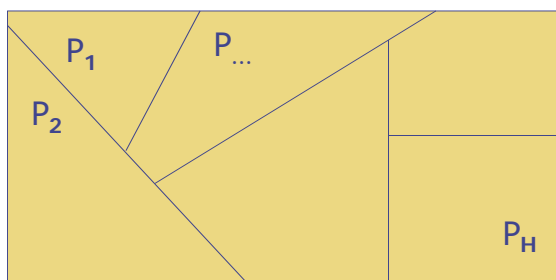
1. suddivisione della Popolazione in **STRATI** (**partizione** della Popolazione in sottoinsiemi esaustivi e mutualmente escludentesi);
2. selezione di campioni indipendenti da ciascuno strato.

### Obiettivi:

1. ottenere **stimatori più precisi** rispetto al CCS;
2. Garantire la partecipazione all'indagine di unità appartenenti a tutti i **domini di studio**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento casuale Stratificato



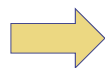
Partizione della Popolazione in **H Strati**

### Notazione:

$N_h$

$n_h$

$W_h = \frac{N_h}{N}$  proporzione di popolazione nello strato  $h$



$$\left\{ \begin{array}{l} \sum_h N_h = N \\ \sum_h n_h = n \\ \sum_h W_h = 1 \end{array} \right.$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento casuale Stratificato

Parametro da stimare:  $\bar{Y} = \sum_h W_h \bar{Y}_h$

Stimatore:  $\bar{y}_{str} = \sum_h W_h \bar{y}_h$

Varianza dello stimatore:  $Var(\bar{y}_{str}) = \sum_h W_h^2 Var(\bar{y}_h)$

Le covarianza sono zero perché i campioni sono estratti in maniera indipendente da uno strato all'altro

**NB:** la varianza dello stimatore è quindi funzione di quella elementare interna ai vari strati.

La possibilità di ridurre la varianza dello stimatore è quindi legata a quella di ottenere **strati** che risultino (rispetto alla variabile d'indagine) **più omogenei della Popolazione presa nel suo complesso.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Stratificato Proporzionale

È caratterizzato da **frazione di campionamento costante**:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

### Vantaggi:

La stratificazione proporzionale è molto diffusa e dà luogo a stimatori molto semplici e di precisione non inferiore a quella che si otterrebbe con il CCS:

$$\frac{n_h}{N_h} = \frac{n}{N} \Rightarrow W_h = \frac{n_h}{n} \longrightarrow \bar{y}_{st.pr} = \sum_h W_h \bar{y}_h = \sum_h \frac{n_h}{n} \bar{y}_h$$

$$= \frac{1}{n} \sum_h n_h \bar{y}_h = \frac{1}{n} \sum_{h=1}^H \sum_{i \in S_h} y_{hi}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Stratificato NON Proporzionale

Spesso il principale obiettivo che si persegue con la stratificazione è quello di ottenere stime di adeguata precisione per particolari sottopopolazioni, dette **domini di studio**, che vengono fatte coincidere con gli strati.

Se un dominio è rappresentato da uno strato molto più piccolo rispetto agli altri è probabile che una stratificazione proporzionale non risulti adeguata a garantire al suo interno una sufficiente precisione degli stimatori.

La soluzione consiste nell'applicare in quello strato una frazione di campionamento diversa (maggiore) dalle altre.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Stratificato NON Proporzionale

### Ripartizione OTTIMALE:

volendo massimizzare la precisione delle stime, tenuto conto delle risorse economiche disponibili, la frazione di campionamento negli strati dovrà tener conto:

$$f_h \propto \frac{S_h}{\sqrt{c_h}}$$



- **variabilità** (dev.standard) **elementare degli strati** (in proporzione diretta);
- radice quadrata del **costo di rilevazione di un'unità negli strati** (in proporzione inversa).

Negli strati più eterogenei occorre applicare una  $f_h$  maggiore rispetto a quella per gli strati più omogenei, tenendo conto delle eventuale differenziale del costo di rilevazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Stratificato NON Proporzionale

Se il costo unitario di rilevazione non varia da strato a strato:



$$f_h \propto S_h$$

Se si è interessati a **confrontare tra loro le stime dei vari strati** (piuttosto che a «fonderle» in un unico stimatore) e se varianze e costi di rilevazione possono essere ipotizzati approssimativamente uguali negli strati

$$f_1 \cong f_2 \cong \dots \cong f_H$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Stratificato NON Proporzionale

### Svantaggi della ripartizione ottimale:

- ❑ all'atto pratico la ripartizione ottimale presuppone una qualche conoscenza di  $S_h^2$ . Approssimazioni grossolane di tali valori possono vanificare gli effetti della stratificazione, fino a condurre a perdite di precisione rispetto al CCS;
- ❑ dato che le variabili d'indagine sono generalmente numerose, non è detto che la ripartizione ottimale per una o alcune lo sia per tutte le altre.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Sistemático

### Metodo:

Il campione è formato prendendo una unità ogni  $k$  presenti nella lista della Popolazione, a partire dalla prima estratta, con  $k$  pari al reciproco della frazione di campionamento:  $k = \frac{N}{n}$

es:

$N = 1500, n = 100 \Rightarrow k = 15$

quindi si estrae un numero casuale  $r$  tra 1 e 15

e si procede con passo  $k$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento Sistemático

### NB:

- ❑ nel campionamento sistemático, come nel CCS, ogni unità della popolazione ha la stessa probabilità di entrare a far parte del campione;
- ❑ diversamente dal CCS, non tutte le  $n$ -uple hanno la stessa probabilità di essere estratte.

In altre parole, sono solo  $k$  i possibili campioni selezionabili a partire da tutte le possibili  $n$ -uple;

- ❑ il campionamento sistemático può essere ricondotto a una selezione equivalente al CCS, se si opera un preliminare disordinamento casuale della lista della Popolazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il Campionamento a Grappoli e a più Stadi

In gran parte delle popolazioni oggetto di indagine le unità di studio sono raggruppate in sottopopolazioni di varia natura.

Esempi:

- La popolazione presente sul territorio italiano è la somma delle sottopopolazioni presenti sui territori regionali. All'interno di ciascuna regione, la popolazione è distribuita in province, quindi in comuni; nei comuni, infine, la popolazione è aggregata in famiglie.
- Gli studenti di un Ateneo sono classificati in facoltà, quelli di una scuola, in classi, ecc.

Questi raggruppamenti di unità possono essere utilizzati come **strati** al cui interno estrarre unità.

Alternativamente, possono essere utilizzati come **vere e proprie unità di selezione** e in questo caso sono denominati **grappoli**.

## Il Campionamento a Grappoli e a più Stadi

Metodo 1:

L'elenco dei grappoli forma la lista da cui viene estratto il campione. Per cui il campione è formato da tutte le unità appartenenti ai grappoli estratti.



**campionamento a grappoli**

Metodo 2:

Nel campione vengono incluse solo alcune unità selezionate da ciascuno dei grappoli estratti.



**campionamento a due o più stadi**



## Il Campionamento a Grappoli e a più Stadi

NB:

il **numero degli stadi** dipende da quello dei livelli gerarchici di aggregazione delle unità che vengono individuati per effettuare la selezione.

es:

un campione di italiani potrebbe essere estratto selezionando inizialmente alcune regioni, da ognuna di queste alcune province, da ciascuna provincia dei comuni, da questi delle famiglie e, infine, dalle famiglie, le persone che sono oggetto di studio.

## Strati VS Grappoli

Gli stessi aggregati di popolazione possano essere utilizzati come **strati** e come **grappoli**. Però gli scopi che si perseguono con la stratificazione sono profondamente diversi da quelli che si perseguono con la stadificazione.

Gli **strati** dovrebbero essere **omogenei il più possibile al loro interno e il più eterogenei possibile tra loro**, in quanto ognuno di essi è rappresentato nel campione.

## Strati VS Grappoli

Al contrario, solo alcuni dei grappoli vengono selezionati, e questi devono rappresentare anche quelli esclusi dalla selezione.

L'ideale sarebbe quindi che tutti i **grappoli** fossero più **eterogenei possibile al loro interno e, conseguentemente, più simili possibile tra loro.**

Ipotesi estrema: se i grappoli fossero tutti uguali, ciascuno sarebbe una copia ridotta della Popolazione; sarebbe quindi sufficiente selezionarne solo uno per avere la stessa informazione che si otterrebbe da un'indagine completa.

## Strati VS Grappoli

Purtroppo, spesso, i grappoli non vengono formati da chi estrae il campione, ma sono **aggregazioni preesistenti nella popolazione** (si pensi agli esempi fatti in precedenza), caratterizzate da una certa omogeneità interna che risulta generalmente tanto più marcata quanto minore è la loro dimensione.

Ma l'**omogeneità**, che nella stratificazione è sinonimo di precisione degli stimatori, **nel campionamento a grappoli produce normalmente una perdita in precisione rispetto al CCS.** Quindi, in generale, **nel campionamento a grappoli, per ottenere stimatori caratterizzati dalla stessa precisione che hanno quelli di un CCS di dimensione  $n$ , occorre un campione di dimensione maggiore di  $n$ .**

## Strati VS Grappoli

Quindi il ricorso ad un campionamento a grappoli o a più stadi è legato agli aspetti pratici ed economici ad esso collegati:

- ❑ risulta spesso impossibile (economicamente o materialmente) formare una lista delle unità di studio, mentre può essere disponibile una lista di grappoli della popolazione;
- ❑ per una prestabilita dimensione campionaria, il campionamento a grappoli comporta costi generalmente molto inferiori a quelli del CCS, in massima parte per la minore dispersione delle unità del campione.

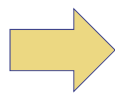
## Lezione 5

# Note di Inferenza parametrica (stima puntuale e per intervallo)

## Prima ipotesi di lavoro

### Ipotesi sulla Popolazione:

se non diversamente indicato,  
 nel proseguo della trattazione **supporremo**  
 che quella di riferimento  
 sia una **Popolazione infinita**  $Y \sim ?(\mu, \sigma^2)$



- nel **CCS**, i due schemi di campionamento (con e senza reimmissione)  $\frac{N-n}{N-1} \rightarrow 1$  sono di fatto coincidenti;
- ad ogni estrazione, la popolazione **NON** subisce un cambiamento in termini di frequenze relative per cui le  $Y_i$  sono v.c. **I.I.D.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Stima puntuale: stimatori di uso frequente nel caso di variabili I.I.D.

- **Stimatore per la media**  $\mu$

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

- è **non distorto**:

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

- **ha varianza**:

$$VAR(\bar{Y}) = VAR\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} VAR\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \sum_{i=1}^n VAR(Y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Stima puntuale: stimatori di uso frequente nel caso di variabili I.I.D.

- **Stimatore per la varianza  $\sigma^2$**

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- **è distorto!!!!**

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2\right) = \dots = \frac{n-1}{n} \sigma^2 < \sigma^2$$



- **Stimatore non distorto della varianza  $\sigma^2$**

$$S^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Alcuni teoremi utili

### Teorema 1:

Se  $Y \sim N(\mu_Y, \sigma_Y^2)$  allora  $W = a + bY \sim N(a + b\mu_Y, b^2\sigma_Y^2)$

Una trasformazione lineare di una normale è ancora una Normale

### Teorema 2:

Se  $Y_i \sim N(\mu_i, \sigma_i^2)$  sono  $n$  v.c. **indipendenti** allora

$$W = \sum_{i=1}^n Y_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

La somma di  $n$  v.c. Normali indipendenti è ancora una distribuzione Normale

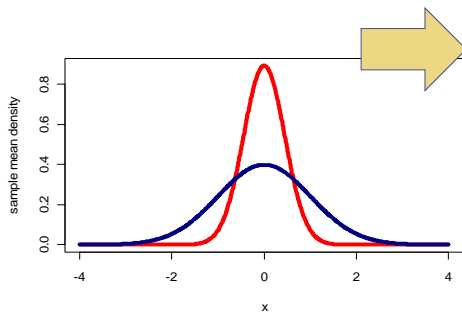
È vietata la riproduzione non autorizzata a fini commerciali.

## Seconda ipotesi di lavoro

### Ipotesi sulla distribuzione della Popolazione:

se non diversamente indicato, nel proseguo della trattazione **supporremo** che la Popolazione si distribuisca secondo una **Normale**

per i teoremi precedenti:



$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z_{\bar{Y}} = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

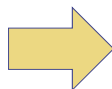
È vietata la riproduzione non autorizzata a fini commerciali.

## La distribuzione *t* di Student

Se sostituiamo il parametro  $\sigma$  con una sua stima  $S$  ottenuta mediante:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

( **NB:**  $S^2$  è ora la varianza elementare del campione )

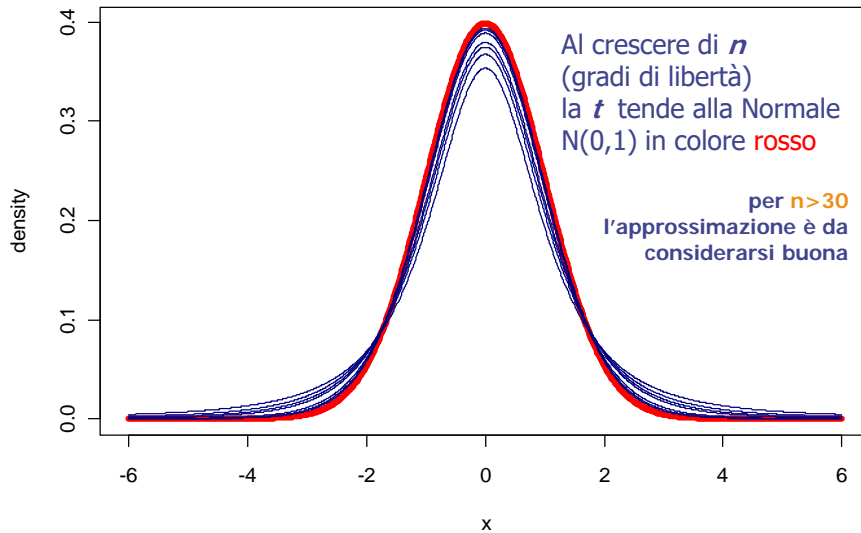


$$T = \frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

***t* di Student**  
 con  $n - 1$  gradi di libertà

È vietata la riproduzione non autorizzata a fini commerciali.

## La distribuzione $t$ di Student



È vietata la riproduzione non autorizzata a fini commerciali.

## Stime per intervallo

Valgono le ipotesi distributive sulla Popolazione fatte in precedenza.

Supponiamo di voler costruire una **stima per intervallo** per il parametro  $\theta$  della Popolazione

cioè supponiamo di voler costruire un **intervallo di confidenza** per  $\theta$

Il livello di confidenza è **la probabilità che  $\theta$  cada in tale intervallo.**

**Confidenza  $\equiv$  Fiducia**

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza

In generale, l'intervallo di confidenza per  $\theta$  risulta definito da

$$\Pr \{l \leq \theta \leq L\} = 1 - \alpha$$

con:

$l = f(Y_1, Y_2, \dots, Y_n)$  limite inferiore (è una v.c.)

$L = g(Y_1, Y_2, \dots, Y_n)$  limite superiore (è una v.c.)

$1 - \alpha$  **Livello di confidenza**

$\alpha$  Probabilità di sbagliare

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza

### Livello di confidenza:

la probabilità che l'intervallo casuale  $[l(Y_1, \dots, Y_n), L(Y_1, \dots, Y_n)]$  contenga al suo interno il parametro  $\theta$  è pari a  $1 - \alpha$

### Informatività dell'intervallo:

sarà tanto più alta quanto più è stretto l'intervallo

**Situazione ottimale:**  Intervallo stretto  
 Livello di confidenza elevato



Se aumenta il livello di confidenza, aumenta l'ampiezza dell'intervallo **MA** diminuisce l'informatività dello stesso, **a meno che non si aumenti la dimensione del campione**

È vietata la riproduzione non autorizzata a fini commerciali.



## Intervalli di confidenza

Per determinare l'intervallo di confidenza per un generico parametro, si cerca una espressione (**quantità pivotale**) in cui:

- ❑ deve comparire solo il parametro da stimare e non altri parametri incogniti (o di disturbo);
- ❑ la cui distribuzione è perfettamente nota.

Una volta individuata questa espressione si può, **isolando il parametro**, costruire l'intervallo di confidenza (questo metodo è detto **metodo del pivot**).

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ nota)

Se la varianza della Popolazione è nota:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \sim N\left(\mu, \frac{\sigma^2}{n}\right) \leftarrow \text{Non è quantità pivotale perché solo la forma della distribuzione è nota ma non la distribuzione esatta}$$

Standardizziamo  $\bar{Y}$  :

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Questa è **quantità pivotale**, perché la distribuzione è perfettamente nota (tabulata) e l'espressione contiene un unico parametro incognito

➡ possiamo applicare il **Metodo del Pivot**

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ nota)

Partiamo da un'affermazione **probabilisticamente vera** relativa alla quantità pivotale:

$$\Pr \left\{ -z_{\alpha/2} \leq \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2} \right\} = 1 - \alpha$$

**NB:** date le proprietà della distribuzione, questo è il più piccolo intervallo ottenibile al livello di probabilità desiderato

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ nota)

Pivotiamo rispetto al parametro incognito  $\mu$  :

$$\Pr \left\{ -z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{Y} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

$$\Pr \left\{ -\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

$$\Pr \left\{ \bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha$$

**Intervallo di confidenza per la media di una Popolazione Normale con varianza nota**

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ NON nota)

Se la varianza della Popolazione **NON** è nota:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

← Non è quantità pivotale perché solo la forma della distribuzione è nota ma non la distribuzione esatta

Standardizziamo  $\bar{Y}$  :

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

**Neanche questa è quantità pivotale**, perché la distribuzione non è nota in quanto l'espressione contiene il parametro incognito ed un parametro di disturbo.

➔ **sostituiamo il parametro di disturbo (incognito) con una sua stima**

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ NON nota)

$$\frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Questa è **quantità pivotale**, perché la distribuzione è perfettamente nota (tabulata) e l'espressione contiene un unico parametro incognito

➔ possiamo applicare il **Metodo del Pivot**

Partiamo da un'affermazione **probabilisticamente vera** relativa alla quantità pivotale:

$$\Pr \left\{ -t_{\frac{\alpha}{2}, n-1} \leq \frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}, n-1} \right\} = 1 - \alpha$$

La  $t$  di Student ha le stesse proprietà della Normale, per cui **questo è il più piccolo intervallo ottenibile al livello di probabilità desiderato**

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $\mu$ ( $\sigma^2$ NON nota)

Pivotiamo rispetto al parametro incognito  $\mu$  :

$$\Pr \left\{ -t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq \bar{Y} - \mu \leq t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right\} = 1 - \alpha$$

$$\Pr \left\{ -\bar{Y} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq -\mu \leq -\bar{Y} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right\} = 1 - \alpha$$

$$\Pr \left\{ \bar{Y} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right\} = 1 - \alpha$$

**Intervallo di confidenza per la media di una  
Popolazione Normale con varianza NON nota**

È vietata la riproduzione non autorizzata a fini commerciali.

## Teorema Limite Centrale

Supponiamo ora che NON valga più l'ipotesi  
sulla Normalità della Popolazione.

Se  $Y_1, \dots, Y_n$  sono  $n$  v.c. I.I.D.  $\sim ?(\mu, \sigma^2)$   
con parametri finiti, allora

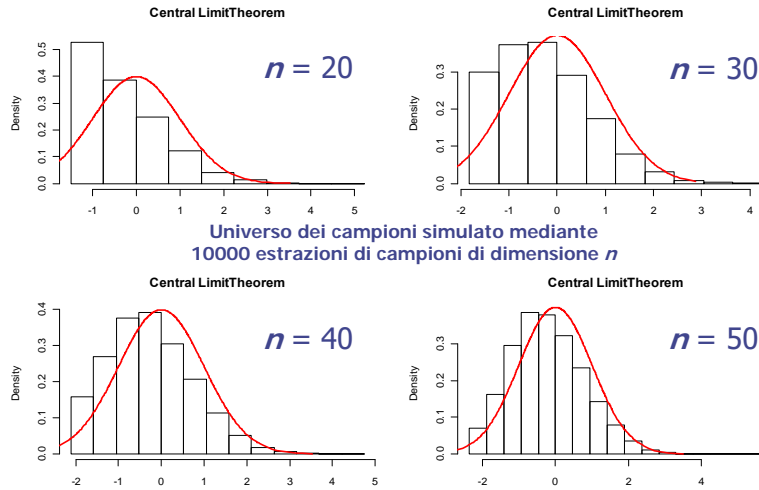
$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{n \rightarrow \infty} N(0, 1)$  ovvero la standardizzazione della  
media campionaria tende,  
al crescere di  $n$ ,  
a distribuirsi come una **Normale  
standard**

**Corollario al TLC:**  $\bar{Y} \xrightarrow{n \rightarrow \infty} N\left(\mu, \frac{\sigma^2}{n}\right)$

È vietata la riproduzione non autorizzata a fini commerciali.

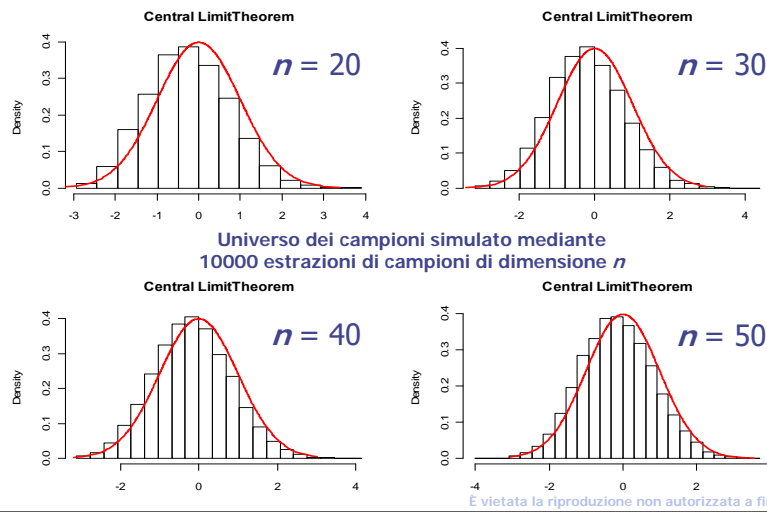
## Teorema Limite Centrale

Esempio: Popolazione di tipo Bernoulliano ( $p = 0.1$ )



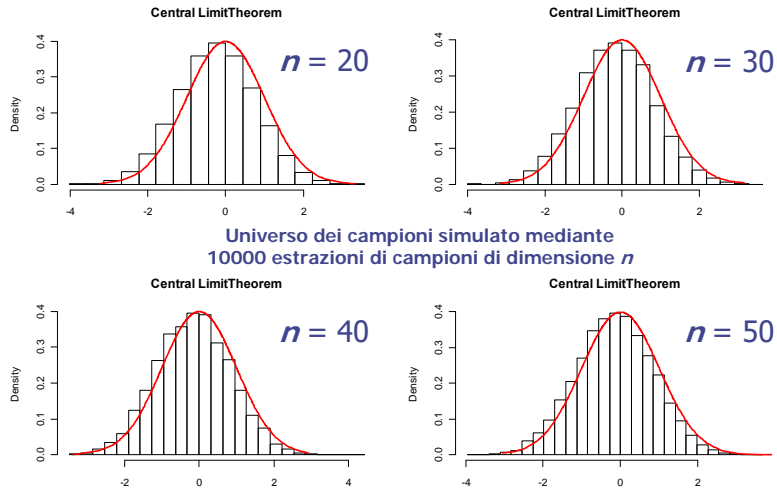
## Teorema Limite Centrale

Esempio: Popolazione di tipo Bernoulliano ( $p = 0.3$ )



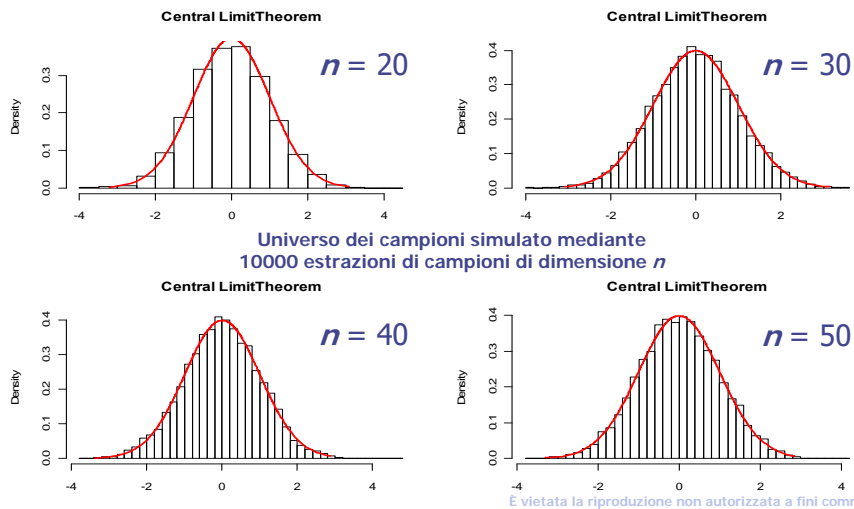
## Teorema Limite Centrale

Esempio: Popolazione di tipo Bernoulliano ( $p = 0.5$ )



## Teorema Limite Centrale

Esempio: Popolazione di tipo Uniforme (0,1)



## Intervalli di confidenza per $p (= \pi)$

Se la Popolazione è Bernoulliana e le v.c. sono I.I.D.:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \hat{p}$$

La media campionaria è la **proporzione campionaria di successi** osservati nel campione

$$E(\bar{Y}) = \mu \longleftrightarrow E(\hat{p}) = p$$

$$VAR(\bar{Y}) = \frac{\sigma^2}{n} \longleftrightarrow VAR(\hat{p}) = \frac{pq}{n}$$

} una Bernoulli ha media  $p$  e varianza  $pq$

Per il T.L.C.:

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \xrightarrow{n \rightarrow \infty} N(0,1)$$

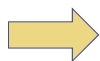
**NB:** questa **NON** è quantità **pivotal**, perché la distribuzione non è nota in quanto l'espressione contiene il parametro incognito sia a numeratore che a denominatore.

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $p$

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \xrightarrow{A} N(0,1)$$

Questa è **quantità pivotal**, perché la distribuzione è perfettamente nota (tabulata) e l'espressione contiene un unico parametro incognito



possiamo applicare il **Metodo del Pivot**

Partiamo da un'affermazione **probabilisticamente vera** relativa alla quantità **pivotal**:

$$\Pr \left\{ -z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \leq z_{\frac{\alpha}{2}} \right\} \approx 1 - \alpha$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza per $p$

Pivotiamo rispetto al parametro incognito  $p$  :

$$\Pr \left\{ -z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq \hat{p} - p \leq z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right\} \approx 1 - \alpha$$

...

$$\Pr \left\{ \hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right\} \approx 1 - \alpha$$

**Intervallo di confidenza per la proporzione di una Popolazione Bernoulliana**

## Intervalli di confidenza: riepilogo

$$\Pr \left\{ \bar{Y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha \quad \text{per } \mu, \sigma^2 \text{ noto}$$

$$\Pr \left\{ \bar{Y} - t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} \right\} = 1 - \alpha \quad \text{per } \mu, \sigma^2 \text{ non noto}$$

$$\Pr \left\{ \hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right\} \approx 1 - \alpha \quad \text{per } p$$

Livello di confidenza	$z_{\frac{\alpha}{2}}$	Livello di confidenza	$t_{\frac{\alpha}{2}, n-1}$
.90	1.645	.90	} dipende dai gradi di libertà della $t$
.95	1.96	.95	
.99	2.576	.99	



## Intervalli di confidenza: determinazione della dimensione campionaria


caso: intervallo per  $\mu$ ,  $\sigma^2$  noto

chiamiamo ME la semi-ampiezza dell'intervallo  $ME = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Intervallo confidenza = stima puntuale  $\pm$  ME  
 ME = MARGINE d'ERRORE

ME lo stabilisce il ricercatore nel momento in cui valuta l'informatività dell'intervallo in relazione al suo livello di confidenza.

Tali considerazioni consentono di determinare la numerosità campionaria adeguata in relazione a livello di confidenza e ME desiderati.


  $n = z_{\frac{\alpha}{2}}^2 \left( \frac{\sigma}{ME} \right)^2$

## Intervalli di confidenza: determinazione della dimensione campionaria

caso: intervallo per  $p$

Intervallo confidenza = stima puntuale  $\pm$  ME  
 dove:

$$ME = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

  $n = z_{\frac{\alpha}{2}}^2 \frac{\hat{p}(1 - \hat{p})}{ME^2}$

Però  $\hat{p}(1 - \hat{p})$  non è calcolabile se non dopo aver estratto il campione; e per estrarre il campione occorre conoscere  $n$ .

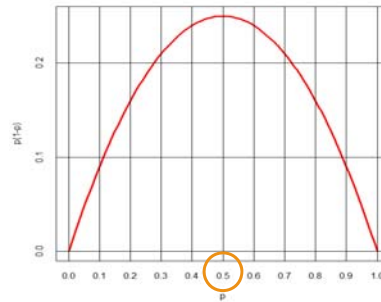
## Intervalli di confidenza: determinazione della dimensione campionaria

caso: intervallo per  $p$

Si sostituisce a  $\hat{p}(1 - \hat{p})$  il massimo valore assumibile dalla varianza nel caso di distribuzione di Bernoulli.

se  $p = 0.5 \Leftrightarrow pq = 0.25$

→ 
$$n = z_{\frac{\alpha}{2}}^2 \frac{0.25}{ME^2}$$



È vietata la riproduzione non autorizzata a fini commerciali.

## Intervalli di confidenza: determinazione della dimensione campionaria

caso: intervallo per  $p$

es: quante unità occorre selezionare da una popolazione bernoulliana (infinita o ad essa equiparabile) per stimare la proporzione di successi nella popolazione con un margine di errore del 4% ?

livello di confidenza	}	.90	$n = 1.645^2 \frac{0.25}{0.04^2} = 422.74 \rightarrow 423$
		.95	$n = 1.96^2 \frac{0.25}{0.04^2} = 600.25 \rightarrow 601$
		.99	$n = 2.576^2 \frac{0.25}{0.04^2} = 1036.70 \rightarrow 1037$

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 6

# Note di Inferenza parametrica (verifica di ipotesi)

**NB:** questa lezione non è nel programma di Statistica. È stata inserita solo per agevolare il ripasso di concetti acquisiti al corso di Psicometria.

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

La differenza rispetto alla teoria della stima consiste nel fatto che qualcuno ci informa che il parametro assume un certo valore:

### Ipotesi statistica sul parametro:

è un'affermazione che specifica **completamente** o **parzialmente** la legge di distribuzione di un fenomeno.

Per esempio, per ipotesi sulla media:

**ipotesi semplice:**  $X \sim N(5, \sigma^2)$

**ipotesi composta:**  $X \sim N(5 \leq \mu \leq 8, \sigma^2)$

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

### Def: Test d'ipotesi

è una regola attraverso la quale si accetta o meno l'ipotesi formulata sulla base dell'evidenza campionaria  
cioè in base al risultato campionario che ottengo si decide di accettare o respingere l'ipotesi formulata.

**NB:** se accettiamo una determinata ipotesi statistica non è detto che questa sia vera

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

In realtà le ipotesi sono due:
 
$$\begin{cases} H_0 : \theta \in \Theta_0 & \text{ipotesi nulla} \\ H_1 : \theta \in \Theta_1 & \text{ipotesi alternativa} \end{cases}$$

dove necessariamente:  $\Theta = \Theta_0 \cup \Theta_1$

Alcuni esempi: **reddito medio dei gioiellieri**  
(migliaia di euro annue)

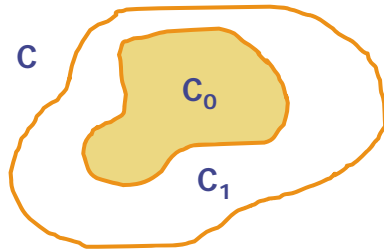
$$\begin{cases} H_0 : \mu = 20 \\ H_1 : \mu = 40 \end{cases} \quad \begin{cases} H_0 : \mu = 20 \\ H_1 : \mu > 20 \end{cases} \quad \begin{cases} H_0 : \mu \leq 20 \\ H_1 : \mu > 20 \end{cases}$$

L'**ipotesi nulla** è in generale l'ipotesi a cui non si crede; ovvero l'ipotesi che il ricercatore spera o crede sia falsa.

Si chiama **NULLA** perché se riusciamo a respingerla si fa qualcosa mentre se l'accettiamo, in generale, non si fa nulla.

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi



dato l'Universo dei Campioni  $C$ , un test delle ipotesi consiste nel bipartire tale Universo in due sottoinsiemi disgiunti  $C_0$  e  $C_1$  in modo tale che si decide di rifiutare l'ipotesi  $H_0$  se il punto campionario cade in  $C_1$  e viceversa di accettarla se cade in  $C_0$

$C_1$  prende il nome di **Regione Critica**.

È importante che  $C$  sia bipartito nel miglior modo possibile ovvero

**è importante individuare la miglior Regione Critica**

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

### Tavola decisionale

		STATI DI NATURA	
		$H_0$ vera	$H_1$ vera
AZIONI	respingo $H_0$	<b>Err I tipo</b>	OK
	non respingo $H_0$	OK	<b>Err II tipo</b>

Oggi non sappiamo quale di questi risultati si verificherà. Pertanto un test delle ipotesi è **sempre formato da decisioni giuste e da decisioni errate**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

### Come scegliere la miglior Regione Critica?

Si cerca di stabilire (**oggi**), prima di estrarre il campione (**domani**) un criterio di decisione in maniera tale da sapere a priori quando respingere l'ipotesi  $H_0$  e quando non respingerla.

Naturalmente ci farebbe piacere adottare a priori un criterio di comportamento tale che  
**la probabilità di commettere gli errori di primo e secondo tipo sia la più piccola possibile.**

**NB:** a priori si possono commettere entrambe gli errori, a posteriori si può commettere un solo tipo d'errore.

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

Def:

$$\alpha = \Pr(E_I) = \Pr\{\text{respingere } H_0/H_0 \text{ è vera}\}$$

$$\beta = \Pr(E_{II}) = \Pr\{\text{non respingere } H_0/H_1 \text{ è vera}\}$$

**Situazione ottimale:**  $\alpha, \beta \rightarrow 0$

ma questo vorrebbe dire esser certi di quello che affermiamo, e non è possibile esser certi sulla base dell'estrazione di un campione

Inoltre le due probabilità  $\alpha$  e  $\beta$  **variano in senso inverso per cui risulta impossibile minimizzarle entrambe.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

**Passi da seguire per l'individuazione della miglior Regione Critica:**

- ❑ si fissa la **probabilità di commettere l'errore più grave**, ovvero si fissa la  $\alpha = \Pr(E_I)$
- ❑ si sceglie la **variabile test** da utilizzare:  
 la **variabile test** è uno stimatore del parametro sottoposto a test oppure è una sua trasformazione (ad es. una standardizzazione)
- ❑ si **determina** la miglior regione critica **minimizzando** la probabilità di commettere l'errore di secondo tipo  $\beta = \Pr(E_{II})$

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

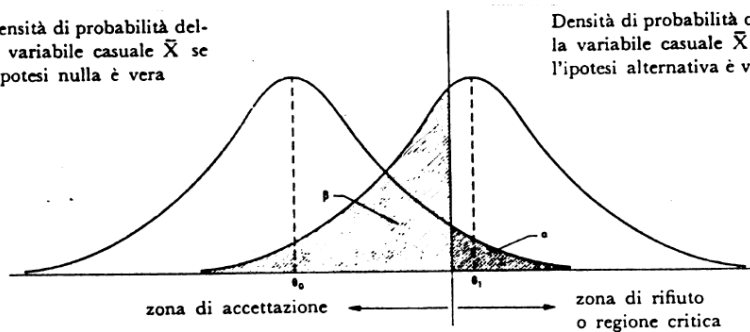
### Test sulla media con varianza nota

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{cases}$$

variabile test sotto  $H_0$ :  $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Densità di probabilità della variabile casuale  $\bar{X}$  se l'ipotesi nulla è vera

Densità di probabilità della variabile casuale  $\bar{X}$  se l'ipotesi alternativa è vera



È vietata la riproduzione non autorizzata a fini commerciali.

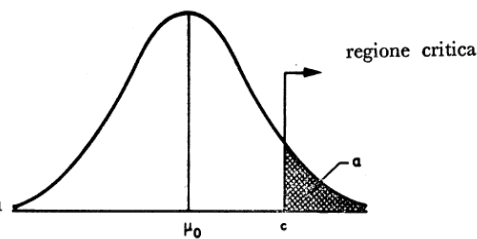
### Verifica (test) delle ipotesi

#### Test sulla media con varianza nota

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \quad \text{variabile test sotto } H_0: \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Caso b)

$$H_0 : \mu = \mu_0, H_1 : \mu > \mu_0; c = \mu_0 + 1,64 \sigma/\sqrt{n}$$



È vietata la riproduzione non autorizzata a fini commerciali.

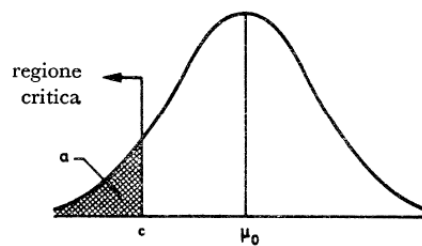
### Verifica (test) delle ipotesi

#### Test sulla media con varianza nota

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \quad \text{variabile test sotto } H_0: \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Caso c)

$$H_0 : \mu = \mu_0, H_1 : \mu < \mu_0; c = \mu_0 - 1,64 \sigma/\sqrt{n}$$



È vietata la riproduzione non autorizzata a fini commerciali.



## Verifica (test) delle ipotesi

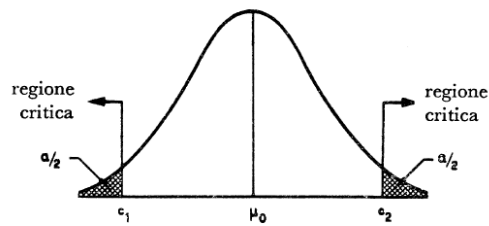
### Test sulla media con varianza nota

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

variabile test sotto  $H_0$ :  $\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Caso d)

$$\begin{aligned} H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0; \\ c_1 = \mu_0 - 1,96 \frac{\sigma}{\sqrt{n}}, c_2 = \\ = \mu_0 + 1,96 \frac{\sigma}{\sqrt{n}} \end{aligned}$$



È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

### Test sulla media con varianza NON nota

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

variabile test sotto  $H_0$ :  $\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi

### Test sulla proporzione

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p = p_1 \end{cases}$$

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$$

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases}$$

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

variabile test  
sotto  $H_0$ :

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \stackrel{n \rightarrow \infty}{\sim} N(0,1)$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi: POTENZA del TEST

$$\beta(H_1) = \Pr(E_{II}) = \Pr\{\text{non respingere } H_0 / H_1 \text{ è vera}\}$$



$$\gamma(H_1) = 1 - \beta(H_1) = \Pr\{\text{respingere } H_0 / H_1 \text{ è vera}\}$$

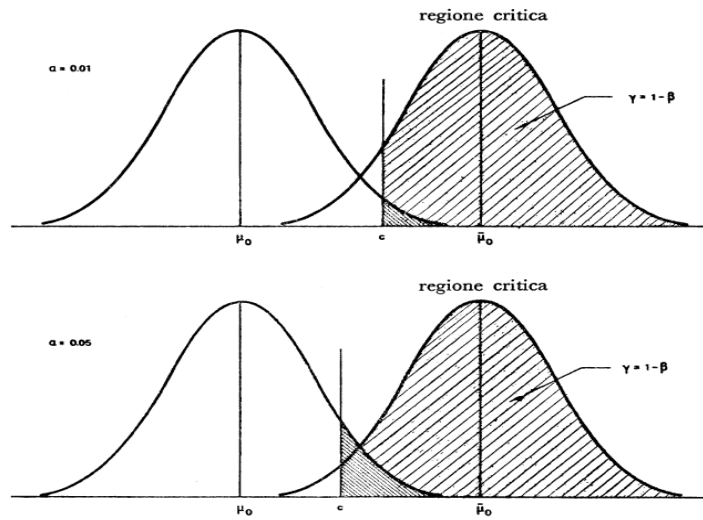
La **Potenza o Forza del TEST** è la probabilità di **NON** commettere un errore di seconda specie

Risulta influenzata da:

- livello di significatività  $\alpha$  prescelto;
- dalla specifica dell'ipotesi alternativa;
- dalla dimensione del campione.

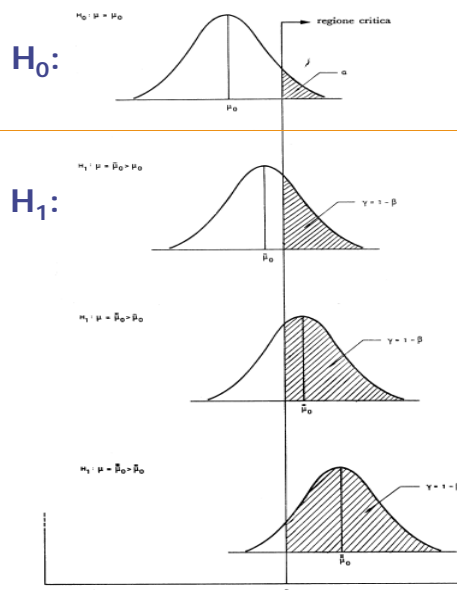
È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi: POTENZA del TEST



È vietata la riproduzione non autorizzata a fini commerciali.

## Verifica (test) delle ipotesi: POTENZA del TEST



man mano che la  
 specifica dell'ipotesi  
 alternativa  
 si sposta  
 verso destra,  
 la potenza cresce

ie non autorizzata a fini commerciali.

## Confronto fra campioni indipendenti

### Test sulla media con varianze $\sigma_X^2, \sigma_Y^2$ note

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$$

**variabile test**  
**sotto  $H_0$ :**

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

$$\frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

dove:

$n, m$  sono le dimensioni dei due campioni

È vietata la riproduzione non autorizzata a fini commerciali.

## Confronto fra campioni indipendenti

### Test sulla media con varianze NON note

ma  $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$$

**variabile test**  
**sotto  $H_0$ :**

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

$$\frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\frac{S^2}{n} + \frac{S^2}{m}}} = \frac{(\bar{X} - \bar{Y})}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

dove:

$n, m$  sono le dimensioni dei due campioni;

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n+m-2}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Confronto fra campioni indipendenti

### Test sulla media con varianze NON note

e  $\sigma_X^2 \neq \sigma_Y^2$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$$

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

variabile test  
sotto  $H_0$ : 
$$\frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \stackrel{n, m \rightarrow \infty}{\sim} N(0,1)$$

dove:

$n, m$  sono le dimensioni dei due campioni

**NB:**

se  $n, m$  sono piccoli, allora **non si può fare niente** perché non è nota la distribuzione della variabile test  
(**Behrens – Fisher problem**)

È vietata la riproduzione non autorizzata a fini commerciali.

## Confronto fra campioni indipendenti

### Test sulla proporzione:

adesso, l'ipotesi  $H_0$  specifica automaticamente l'uguaglianza tra le varianze

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 > p_2 \end{cases} \quad \text{variabile test sotto } H_0: \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\text{VAR}(\hat{p}_1 - \hat{p}_2)}} =$$

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 < p_2 \end{cases} \quad \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{pq}{n} + \frac{pq}{m}}} =$$

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 \neq p_2 \end{cases} \quad \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{pq \left( \frac{1}{n} + \frac{1}{m} \right)}} \stackrel{n, m \rightarrow \infty}{\sim} N(0,1)$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Confronto fra campioni indipendenti

### Test sulla proporzione

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 > p_2 \end{cases} \quad \text{variabile test} \\ \text{sotto } H_0: \quad \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n} + \frac{1}{m}\right)}} \stackrel{n,m \rightarrow \infty}{\sim} N(0,1)$$

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 < p_2 \end{cases}$$

$$\begin{cases} H_0 : p_1 = p_2 = p \\ H_1 : p_1 \neq p_2 \end{cases} \quad \text{stimando } p \text{ mediante lo stimatore non distorto}$$

$$\hat{p} = \frac{n \cdot \hat{p}_1 + m \cdot \hat{p}_2}{n + m}$$

perché l'ipotesi  $H_0$  specifica automaticamente l'uguaglianza tra le varianze

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 7

# Analisi dell'associazione tra variabili categoriali

È vietata la riproduzione non autorizzata a fini commerciali.

## Un breve ripasso...

**Variabile quantitativa:** assume valori che rappresentano i diversi ordini di grandezza (o livelli di intensità) del fenomeno misurato (es: *peso, altezza, reddito, temperatura, durata di una lampadina, ecc.*)

Il confronto a coppie dei possibili valori rilevati per una variabile quantitativa in generale produce una **scala di intervalli**. Se la scala presenta un'origine non convenzionale ma fissa si parla di **scala di rapporti** (es: *temperatura VS durata lampadina*).

**Variabile categoriale (o qualitativa):** assume valori che identificano un insieme di categorie (es: *genere, status occupazionale, credo religioso, preferenza politica, ecc.*)

Le categorie che non presentano nessun ordinamento formano una **scala nominale**.  
Le categorie che invece presentano un ordinamento naturale dei loro valori formano una **scala ordinale** (es: *titolo di studio*).  
Le variabili ordinali possiedono quindi una caratteristica delle scale quantitative: il concetto di «minore» o «maggiore» che ne determina l'ordinamento.

È vietata la riproduzione non autorizzata a fini commerciali.

## L'associazione tra variabili

In generale, si ha **associazione** tra due variabili se la distribuzione di una variabile varia al variare dell'altra variabile.

In questa lezione saranno presentati metodi per descrivere l'associazione tra variabili categoriali.

Tra due variabili categoriali, una assume generalmente il ruolo di **variabile risposta**, l'altra di **variabile esplicativa**.

Un modo per verificare se la distribuzione di una variabile varia al variare dell'altra variabile è attraverso l'analisi della cosiddetta **tavola di contingenza**.

es:

Genere	Area			Totale
	Umanistica	Ingegneristica	Medica	
Maschi	40	81	84	205
Femmine	96	72	110	278
Totale	136	153	194	<b>483</b>

È vietata la riproduzione non autorizzata a fini commerciali.

## Tavole di contingenza (riepilogo)

Se, in relazione allo studio di un certo fenomeno, si rilevano due variabili  $X$  (con  $s$  modalità) e  $Y$  (con  $r$  modalità), ciascuna delle  $n$  unità osservate sarà caratterizzata da un insieme di coppie di valori:

$$(x_i, y_j) \quad \text{con } i = 1 \dots s ; \quad j = 1 \dots r$$

a ciascuna delle quali è associata una certa frequenza assoluta  $n_{ij}$

o relativa  $f_{ij} = n_{ij} / n$

di osservazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## Tavole di contingenza (riepilogo)

	$y_1$	$y_2$	...	$y_j$	...	$y_r$	
$x_1$	$f_{11}$	$f_{12}$	...	$f_{1j}$	...	$f_{1r}$	$f_{1.}$
$x_2$	$f_{21}$	$f_{22}$	...	$f_{2j}$	...	$f_{2r}$	$f_{2.}$
...	...	...	...	...	...	...	...
$x_i$	$f_{i1}$	$f_{i2}$	...	$f_{ij}$	...	$f_{ir}$	$f_{i.}$
...	...	...	...	...	...	...	...
$x_s$	$f_{s1}$	$f_{s2}$	...	$f_{sj}$	...	$f_{sr}$	$f_{s.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.r}$	<b>1</b>

con:  $f_{rel}(X = x_i \cap Y = y_j) = f_{ij}$

$$f_{i.} = \sum_{j=1}^r f_{ij} \quad f_{.j} = \sum_{i=1}^s f_{ij} \quad \leftarrow \text{distribuzioni marginali}$$

È vietata la riproduzione non autorizzata a fini commerciali.



## Tavole di contingenza (riepilogo)

Calcolo delle **distribuzioni condizionate (relative)**

es:  $X/Y = y_1$

$x_1$	$f_{11}/f_{.1}$
$x_2$	$f_{21}/f_{.1}$
...	...
$x_i$	$f_{i1}/f_{.1}$
...	...
$x_s$	$f_{s1}/f_{.1}$
	1

$$f_{rel}(X = x_i / Y = y_1) = \frac{f_{rel}(X = x_i \cap Y = y_1)}{f_{rel}(Y = y_1)} = \frac{f_{i1}}{f_{.1}}$$

**NB:** la formula utilizzata è analoga al Principio delle Probabilità condizionate esposto nella Lezione 2

Analogamente, la distribuzione di  $Y$  condizionatamente alla  $i$ -esima modalità di  $X$ :  $f_{rel}(Y = y_j / X = x_i) = \frac{f_{ij}}{f_{i.}}$

## Indipendenza e dipendenza statistica

**Indipendenza statistica:**

nella Popolazione due variabili categoriali sono **statisticamente indipendenti** se tutte le distribuzioni condizionate di una variabile a ciascuna categoria dell'altra sono identiche.

Ovvero se e solo se:  $f_{rel}(X = x_i / Y = y_j) = f_{rel}(X = x_i) \quad \forall i, j$   
 $f_{rel}(Y = y_j / X = x_i) = f_{rel}(Y = y_j)$

in quanto il condizionamento non sortisce effetto.

Quindi, **in caso di indipendenza:**

$$\frac{f_{ij}}{f_{.j}} = f_{i.} \quad \text{oppure} \quad \frac{f_{ij}}{f_{i.}} = f_{.j} \Leftrightarrow f_{ij} = f_{i.} \cdot f_{.j}$$

## Indipendenza e dipendenza statistica

### Dipendenza statistica:

se, nella Popolazione, tutte le distribuzioni condizionate di una variabile a ciascuna categoria dell'altra **NON** sono identiche, allora esiste **associazione** tra due variabili che sono dette **statisticamente dipendenti**.

Casi estremi di dipendenza:

❑ **MASSIMA ASSOCIAZIONE (DIPENDENZA PERFETTA):**

La variabile  $Y$  dipende perfettamente da  $X$  se, in corrispondenza di ogni modalità di  $X$ , si verifica una sola modalità di  $Y$ .

❑ **INTERDIPENDENZA PERFETTA**

Ciascuna variabile dipende perfettamente dall'altra (dipendenza perfetta bilaterale – solo per tavole quadrate).

È vietata la riproduzione non autorizzata a fini commerciali.

## Indipendenza e dipendenza statistica

	y1	y2	y3
x1	0		0
x2		0	0
x3	0	0	
x4	0	0	

### Dipendenza perfetta

Comunque si osservi una  $x$ , siamo in grado di dire quale  $y$  si è verificata, per cui  $Y$  dipende perfettamente da  $X$ .  
Il viceversa non è vero.

	y1	y2	y3
x1	0		0
x2		0	0
x3	0	0	

### Interdipendenza perfetta

Adesso la dipendenza perfetta è bilaterale.

NB: la dipendenza perfetta è rara, e si osserva esclusivamente quando tra le due variabili esiste una dipendenza deterministica (ovvero una delle due variabili è funzione dell'altra).

È vietata la riproduzione non autorizzata a fini commerciali.

## Indipendenza e dipendenza statistica: Popolazione VS evidenza campionaria

**NB:** il concetto di **indipendenza** è analogo a quello definito nella lezione 2 «I Principi della Probabilità»;  
 la relazione si riferisce all'intera Popolazione.

Però si osservano dati di natura campionaria, che possono evidenziare una «forza» della relazione differente da quella che caratterizza l'intera Popolazione...

... in altre parole, a causa della variabilità campionaria, le distribuzioni condizionate nel campione saranno in generale diverse da quelle osservabili a livello di intera Popolazione.



**DOMANDA:** è plausibile ritenere che le differenze a livello di distribuzioni condizionate osservate nel campione siano dovute soltanto al caso?

## Indipendenza e dipendenza statistica: Popolazione VS evidenza campionaria

Ritornando al caso dell'esempio iniziale, (avendo osservato un campione di 483 individui), è possibile affermare che c'è associazione tra le variabili Area e Genere nella Popolazione?



Genere	Area			Base
	Umanistica	Ingegneristica	Medica	
Maschi	19.5%	39.5%	41.0%	205
Femmine	34.5%	25.9%	39.6%	278
				483

Le distribuzioni della variabile Area, condizionate ai due livelli della variabile Genere sono diverse, ma tale differenza, **riscontrata in questo campione**, è dovuta al caso o alla struttura della Popolazione?

## Test chi-quadrato di indipendenza

$\begin{cases} H_0 : \text{le variabili sono statisticamente indipendenti} \\ H_1 : \text{le variabili sono statisticamente dipendenti} \end{cases}$

**NB:** il test richiede che i dati siano ottenuti attraverso un campionamento casuale e che il campione sia sufficientemente grande.

statistica test

(chi-quadrato di Pearson):

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

con la sommatoria che agisce su tutte le celle della tavola di contingenza

dove:  $f_o$  frequenze osservate =  $n_{ij}$

$f_e$  frequenze attese (in caso di indipendenza) =  $n f_{i.} f_{.j} = \frac{n_{i.} \cdot n_{.j}}{n} \geq 5$  in tutte le celle

È vietata la riproduzione non autorizzata a fini commerciali.

## Test chi-quadrato di indipendenza

Quando  $H_0$  è vera, le frequenze osservate e attese tendono ad essere vicine in ogni cella e la statistica test assume valori relativamente piccoli.

Se  $H_0$  è falsa, alcune differenze saranno rilevanti, elevando il valore della statistica test. Più grande è il valore di  $\chi^2$ , maggiore è l'evidenza campionaria contro  $H_0$ .

es:

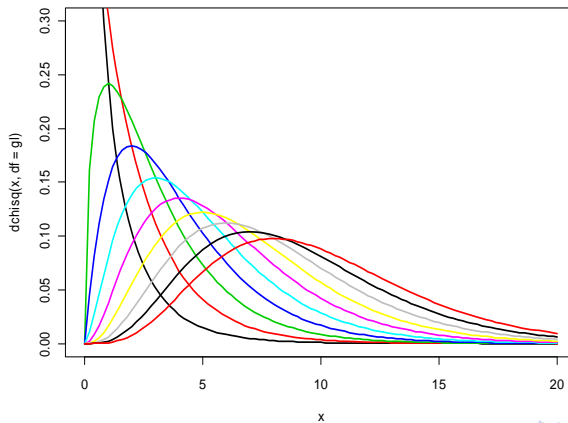
Genere	Area			Totale
	Umanistica	Ingegneristica	Medica	
Maschi	40 (57.7)	81 (64.9)	84 (82.3)	205
Femmine	96 (78.3)	72 (88.1)	110 (111.7)	278
Totale	136	153	194	483

È vietata la riproduzione non autorizzata a fini commerciali.

## Distribuzione di probabilità chi-quadrato

La distribuzione della statistica test  $\chi^2$  nell'universo dei campioni tende, per elevate numerosità campionarie, alla **distribuzione di probabilità chi-quadrato**.

Densità della variabile Chi-Quadrato al crescere dei g.l.



### Proprietà della distribuzione Chi-quadrato:

- è definita in  $\mathbb{R}^+$
- è asimmetrica positiva (coda allungata verso dx);
- la sua forma dipende dall'unico parametro «gradi di libertà»  $gdl$  ;
- la sua media è  $\mu = gdl$  ;
- la sua varianza è  $\sigma^2 = 2gdl$  ;
- all'aumentare dei  $gdl$  la distribuzione tende alla Normale

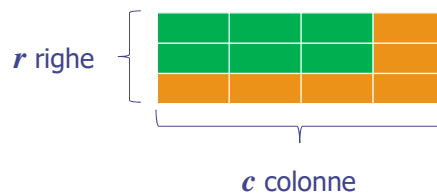
È vietata la riproduzione non autorizzata a fini commerciali.

## Test chi-quadrato di indipendenza

In una tavola di contingenza con  $r$  righe e  $c$  colonne, per sottoporre a verifica l'ipotesi "H<sub>0</sub>: indipendenza":

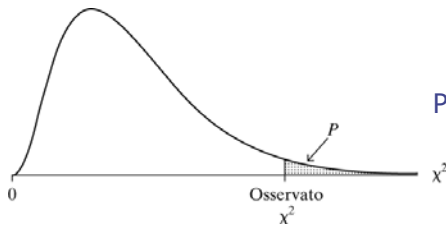
➡  $gdl = (r - 1)(c - 1)$

Questo perché, dati i vincoli imposti dalle distribuzioni marginali, sono solo  $(r - 1)(c - 1)$  le celle i cui valori possono essere liberamente attribuiti entro certi margini di «libertà» dettati dalle variabili oggetto di studio.



È vietata la riproduzione non autorizzata a fini commerciali.

## Test chi-quadrato di indipendenza



Poiché più grande è il valore di  $\chi^2$ , maggiore è l'evidenza campionaria contro  $H_0$ , ...

... è ragionevole collocare la **regione critica del test** nella coda destra della distribuzione Chi-quadro.

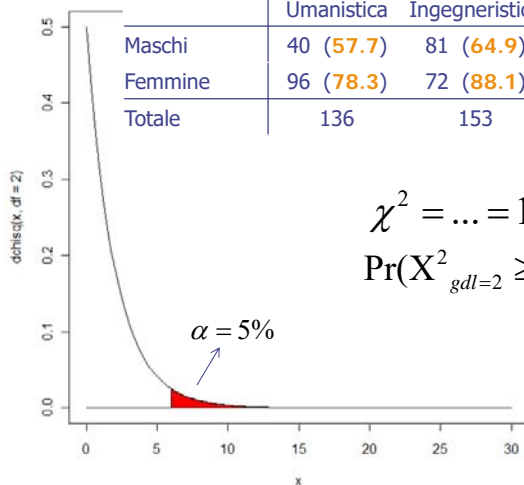
Il **p-value** misura quindi la probabilità, qualora sia vera  $H_0$ , che si verifichino valori almeno grandi quanto il valore di  $\chi^2$  effettivamente osservato.

se **p-value** <  $\alpha$  (livello di significatività prescelto)  $\Rightarrow$  si respinge  $H_0$

## Test chi-quadrato di indipendenza

es:

Genere	Area			Totale
	Umanistica	Ingegneristica	Medica	
Maschi	40 (57.7)	81 (64.9)	84 (82.3)	205
Femmine	96 (78.3)	72 (88.1)	110 (111.7)	278
Totale	136	153	194	483



$$\chi^2 = \dots = 16.4146$$

$$\Pr(X^2_{gdl=2} \geq 16.4146) = 0.0002726559$$

in caso di indipendenza, il valore osservato o uno ancor più estremo avrebbero una probabilità di verificarsi in 2 casi su 10000...

... **respingo l'ipotesi di indipendenza.**

## Ancora sul test chi-quadrato di indipendenza

Uno comodo strumento di calcolo su web:

<http://www.quantpsy.org/chisq/chisq.htm>

### NB:

- ❑ Il test  $\chi^2$  si applica generalmente a variabili nominali.  
Non usa la caratterizzazione aggiuntiva delle variabili ordinali.
- ❑ Non è necessario individuare una variabile risposta e una esplicativa.
- ❑ Il test  $\chi^2$  non dice nulla o quasi sulla forza dell'associazione.  
Se il  $p$  - *value* è molto piccolo, è evidentemente un segnale di una dipendenza importante.  
Che non siamo però in grado di quantificare.

È vietata la riproduzione non autorizzata a fini commerciali.

## Ancora sul test chi-quadrato di indipendenza

Il test  $\chi^2$  non dice nulla o quasi sulla forza dell'associazione.

Dimostrazione empirica:

Genere	Area			Totale
	Umanistica	Ingegneristica	Medica	
Maschi	400 (577.2)	810 (649.4)	840 (823.4)	2050
Femmine	960 (782.8)	720 (880.6)	1100 (1116.6)	2780
Totale	1360	1530	1940	4830

Prima:  $\chi^2 = \dots = 16.4146$

Ora:  $\chi^2 = \dots = 164.146$

Eppure, la moltiplicazione per 10 di tutte le celle non ha alterato la relazione tra le due variabili.  
In altre parole,  
le distribuzioni condizionate sono le stesse di prima.

È vietata la riproduzione non autorizzata a fini commerciali.

## La struttura dell'associazione: i residui

Una componente importante della statistica test  $\chi^2$  sono le differenze  $f_o - f_e$ .

Tali differenze, dette **residui**, consentono di comprendere se i casi osservati sono in misura maggiore o minore di quelli attesi.

I residui risentono però dell'ordine di grandezza delle frequenze osservate.

Per svincolarsi da tale effetto occorre calcolare i cosiddetti **residui standardizzati aggiustati (RSA)**:

$$RSA_{ij} = \frac{f_o - f_e}{\sqrt{f_e(1-f_{i.})(1-f_{.j})}} = \frac{n_{ij} - \frac{n_{i.}n_{.j}}{n}}{\sqrt{\frac{n_{i.}n_{.j}}{n} \left(1 - \frac{n_{i.}}{n}\right) \left(1 - \frac{n_{.j}}{n}\right)}}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La struttura dell'associazione: i residui

- ❑ Quando  $H_0$ : *le variabili sono indipendenti* è vera, i RSA seguono, per grandi campioni, una **distribuzione approssimativamente normale standardizzata** (quindi con media 0 e deviazione standard circa 1).
- ❑ Quindi, sempre se  $H_0$ : *le variabili sono indipendenti* è vera, le situazioni  $|RSA_{ij}| > 2$  dovrebbero verificarsi solo (circa) nel 5% dei casi (per le caratteristiche della Normale Standard).
- ❑ Le situazioni  $|RSA_{ij}| > 3$  sono poco verosimili sotto  $H_0$  e indice dell'esistenza di un **vero** (cioè non dovuto al caso, ovvero all'osservazione di un particolare campione) effetto associativo in quelle determinate celle.

È vietata la riproduzione non autorizzata a fini commerciali.



## Tavole 2 x 2: chi-quadro e differenza di proporzioni

Nel caso di **tavole** (o tabelle)  $2 \times 2$ , a una variabile dicotomica assume il ruolo di **variabile risposta** (genericamente **successo / insuccesso**) e si contrappone a una **variabile esplicativa** anch'essa dicotomica che generalmente rappresenta l'afferenza a due gruppi della stessa popolazione o a due distinte popolazioni.

	variabile risposta		
	successo	insuccesso	
grp1	$\pi_1$	$1-\pi_1$	1
grp2	$\pi_2$	$1-\pi_2$	1

} sono distribuzioni condizionate

$\pi_1$  : la probabilità di *successo* per la popolazione 1  
 $\pi_2$  : la probabilità di *successo* per la popolazione 2.

## Tavole 2 x 2: chi-quadro e differenza di proporzioni

Quindi, nel caso di **tavole**  $2 \times 2$ :

$H_0$ : *risposta e esplicativa sono indipendenti*  $\Leftrightarrow \begin{cases} H_0 : \pi_1 = \pi_2 = \pi \\ H_1 : \pi_1 \neq \pi_2 \end{cases}$

statistica test per il confronto tra proporzioni

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0}{\sqrt{\hat{\pi}_{pool} (1 - \hat{\pi}_{pool}) \left( \frac{1}{n_{1.}} + \frac{1}{n_{2.}} \right)}} \stackrel{n_{1.}, n_{2.} \rightarrow \infty}{\sim} N(0,1)$$

stimando  $\pi$  mediante lo stimatore non distorto:  $\hat{\pi}_{pool} = \frac{n_{1.} \cdot \hat{\pi}_1 + n_{2.} \cdot \hat{\pi}_2}{n_{1.} + n_{2.}}$

perché l'ipotesi  $H_0$  specifica automaticamente l'uguaglianza tra le varianze

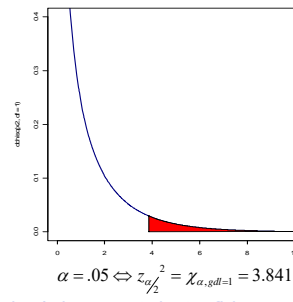
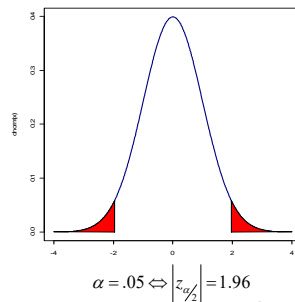
➔ (si vedano le ultime due diapositive della lezione 6).

## Tavole 2 x 2: chi-quadro e differenza di proporzioni

**NB.** Nel caso di tavole 2 x 2:

- esiste la relazione:  $z^2 = \chi^2$  ovvero il quadrato della statistica test  $z$  corrisponde al valore del test  $\chi^2$  di Pearson;
- A livello asintotico, il  $p$ -value ottenuto dalla distribuzione del chi-quadro è lo stesso di quello per il test bilaterale che usa la statistica  $z$ .

Elevando al quadrato un qualsiasi  $z$ -score associato ad una certa probabilità su due code si ottiene il valore del chi-quadro con  $gdl = 1$  corrispondente alla stessa probabilità sottesa alla coda destra della distribuzione.



È vietata la riproduzione non autorizzata a fini commerciali.

## Tavole 2 x 2: chi-quadro e differenza di proporzioni

esempio

	variabile risposta		
	test OK	test KO	
1° turno	138	43	181
2° turno	120	34	154

L'aver superato il test dipende dal turno in cui questo è stato svolto?

Pearson	→	expect freq.	Pearson	} $\chi^2_{Pearson} = 0.1325$
		1° turno	0.014001 0.046911	
test diff. prop.	→	$\pi_1$ -hat: 0,7624	} $\hat{\pi}_{pool} = 0.7701$	} $-0.3640^2 = 0.1325$
		$\pi_2$ -hat: 0,7792		

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\hat{\pi}_{pool} (1 - \hat{\pi}_{pool}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = -0.3640$$

si accetta l'ipotesi di indipendenza  $H_0$

È vietata la riproduzione non autorizzata a fini commerciali.

## Ancora sulle Tavole 2 x 2: l'Odds-Ratio

**Def: quota (odd)**  $\text{odd} = \frac{\text{probabilità di successo}}{\text{probabilità di insuccesso}}$

Se la probabilità di successo è maggiore di quella di insuccesso  $\text{odd} > 1$ , altrimenti  $0 \leq \text{odd} < 1$ .  
 $\text{odd} = 1$  implica che le due probabilità coincidono.

**Def: odd ratio (rapporto tra quote)**  $\theta = \frac{\text{odd riga1}}{\text{odd riga2}}$

è il rapporto tra gli odd delle due righe della tabella.

## Come interpretare l'odds-ratio

	variabile risposta		
	successo	insuccesso	
grp1	$\pi_1$	$1-\pi_1$	1
grp2	$\pi_2$	$1-\pi_2$	1

$\Rightarrow \theta = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}$

L'odds-ratio è una buona misura dell'associazione in tabelle  $2 \times 2$ .

- $\theta = 1 \Leftrightarrow$  l'odds nel grp1 è uguale all'odd nel grp2, cioè la variabile esplicativa non influenza la variabile risposta;
- $\theta > 1 \Leftrightarrow$  l'afferenza al grp1 è, o può essere, causa del verificarsi del «successo»;
- $\theta < 1 \Leftrightarrow$  l'afferenza al grp2 è, o può essere, causa del verificarsi del «successo».

## Ancora sulle Tavole 2 x 2: il Relative Risk

Def: Relative Risk

$$RR = \frac{\pi_1}{\pi_2} = \frac{a/(a+b)}{c/(c+d)}$$

variabile risposta

	successo	insuccesso
grp1	a	b
grp2	c	d

Si dimostra che:

$$se(\ln(RR)) = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$

per cui l'intervallo di confidenza al 95% per il **RR** è:

$$\left[ e^{\ln(RR) - 1.96 \cdot se(\ln(RR))}; e^{\ln(RR) + 1.96 \cdot se(\ln(RR))} \right]$$

Gli zero possono causare problemi nel calcolo dello standard error del  $\ln(RR)$ ; tale problema viene aggirato aggiungendo 0.5 a tutte le celle (a, b, c, d).

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali

Quando le variabili categoriali sono di tipo **ordinale** è possibile sfruttare l'informazione che proviene dall'ordinamento naturale delle loro modalità.

Supponiamo che  $X$  e  $Y$  siano due variabili ordinali.

In questo caso, si parla di:

- ❑ **Associazione positiva:** quando soggetti classificati con elevati valori di  $X$  tendono a manifestare anche elevati valori di  $Y$  e viceversa.
- ❑ **Associazione negativa:** quando soggetti classificati con elevati valori di  $X$  tendono a manifestare bassi valori di  $Y$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali: concordanza e discordanza

Poiché nelle tavole di contingenza bivariate ogni caso statistico è definito mediante una coppia di valori osservati  $(x, y)$  ...

### Definizione:

Una **coppia di casi statistici** è **concordante** quando uno dei due casi è superiore all'altro in entrambe le variabili osservate.

Una **coppia di casi statistici** è **discordante** quando uno dei due casi è superiore all'altro in una variabile, ma inferiore nella seconda variabile che compone l'osservazione.

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali: concordanza e discordanza

### Esempio:

Reddito Familiare	Felicità			Totale
	Non troppo Felice	Abbastanza Felice	Molto Felice	
Sotto la media	16 (24%)	36 (54%)	15 (22%)	67 (100.0%)
Nella media	11 (16%)	36 (53%)	21 (31%)	68 (100.0%)
Sopra la media	2 (9%)	12 (55%)	8 (36%)	22 (100.0%)
Totale	29	84	44	157

Calcoliamo le coppie di soggetti **concordanti (C)** e **discordanti (D)**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali: concordanza e discordanza

	NTF	AF	MF		NTF	AF	MF		NTF	AF	MF		NTF	AF	MF
Sotto la media	16					36									
Nella media		36	21				21		11					36	
Sopra la media		12	8				8			12	8				8

$$C = 16(36 + 21 + 12 + 8) + 36(21 + 8) + 11(12 + 8) + 36(8) = 2784$$

	NTF	AF	MF		NTF	AF	MF		NTF	AF	MF		NTF	AF	MF
Sotto la media			15							36					
Nella media	11	36					21		11					36	
Sopra la media	2	12			2	12			2				2		

$$D = 15(11 + 36 + 2 + 12) + 21(2 + 12) + 36(11 + 2) + 36(2) = 1749$$

Ad esempio, i 16 soggetti nella prima cella sono concordati quando appaiati con ciascuno dei  $(36 + 21 + 12 + 8)$  soggetti sotto e a destra che sono contraddistinti tutti dal mostrare categorie più alte per ciascuna delle due variabili oggetto di studio. Similmente, i 36 soggetti nella seconda cella della prima riga sono concordanti con i  $(21 + 8)$  soggetti che appartengono a categorie più elevate per ciascuna variabile.

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali: l'indice gamma

Se  $C - D > 0 \Leftrightarrow$  associazione positiva.

Se  $C - D < 0 \Leftrightarrow$  associazione negativa.

**NB:** C e D dipendono dalla dimensione campionaria.

Per eliminare tale effetto si standardizza la differenza  $C - D$  per il numero di coppie totali  $(C + D)$ :

$$\text{indice gamma} \quad \hat{\gamma} = \frac{C - D}{C + D}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione tra variabili ordinali: l'indice gamma

Proprietà di **gamma**:

- ❑ il valore di gamma varia tra  $-1$  e  $+1$ ;
- ❑ il segno di gamma indica se l'associazione è positiva o negativa;
- ❑ maggiore è il valore assoluto di gamma, più forte è l'associazione.

Per la tavola di contingenza *Reddito familiare VS Felicità*:

$$\hat{\gamma} = \frac{2784 - 1749}{2784 + 1749} = 0.228$$

il campione evidenzia una associazione positiva tra reddito familiare e felicità.

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 8

# Analisi dell'associazione tra variabili quantitative

È vietata la riproduzione non autorizzata a fini commerciali.

## Lo scatter plot

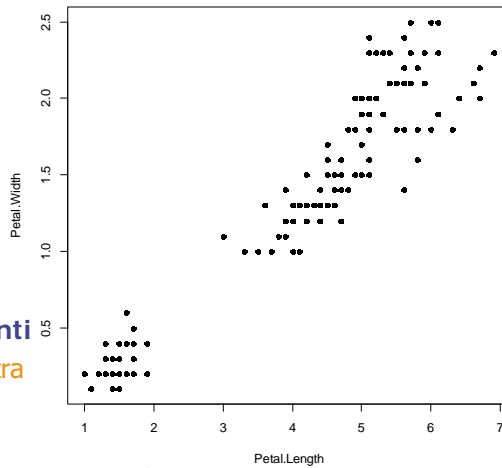
Un esempio: il data set IRIS

Il grafico evidenzia una associazione positiva tra la variabile  $X$  (lunghezza dei petali) e la variabile  $Y$  (larghezza del petalo).

All'aumentare di una variabile, aumentano in media anche i valori assunti dall'altra (ovvero anche l'altra tende ad aumentare).

Code R:

```
data(iris)
attach(iris)
plot(Petal.Length,Petal.Width,pch=16)
```



È vietata la riproduzione non autorizzata a fini commerciali.

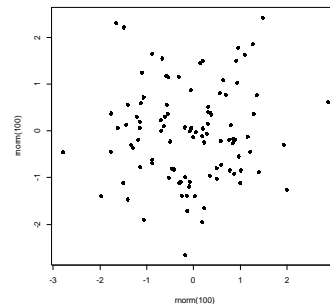
## Lo scatter plot

Nel caso di relazioni bivariate, lo **scatter plot** è uno strumento molto utile in quanto aiuta a comprendere se esiste una qualche associazione tra le variabili  $X$  e  $Y$ .

Ovvero, *al variare di una variabile l'altra tende ad aumentare?*

Oppure ... *a diminuire?*

Se al variare di una variabile l'altra non varia, ovvero tende a variare in maniera assolutamente casuale, allora siamo in assenza di associazione.



È vietata la riproduzione non autorizzata a fini commerciali.



## Una misura di co-variazione: la covarianza

Dalla lezione 7: si ha in generale **associazione** tra due variabili se la distribuzione di una variabile varia al variare dell'altra variabile.

In caso di una variabile quantitativa:

$$\text{Varianza: } \text{VAR}(X) = \sum_{i=1} (x_i - \mu)^2 f_i = \sum_{i=1} (x_i - \mu)(x_i - \mu) f_i$$

In caso di **DUE** variabili quantitative:

$$\text{COVARIANZA: } \text{COV}(X, Y) = \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) f_{ij}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La covarianza

Nel caso di  $N$  coppie di valori singoli, ovvero di dati non raggruppati secondo una tavola (discreta) doppia:

$$\text{COVARIANZA: } \text{COV}(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$$

**NB:**

nel continuo non avrei una tavola e non potrei ricorrere alla sommatoria.

Inoltre: se l'esperimento casuale deve essere ancora effettuato, è possibile definire la **covarianza tra due variabili casuali  $X$  e  $Y$**  come:

$$\text{COV}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La covarianza

Nel caso di **variabili statisticamente indipendenti**,  
si dimostra che:

$$\begin{aligned} COV(X,Y) &= \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) f_{ij} = \\ &= \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) f_{i.} \cdot f_{.j} = \\ &= \sum_i (x_i - \mu_x) f_{i.} \sum_j (y_j - \mu_y) f_{.j} = 0 \end{aligned}$$

in quanto:  $\sum_i (x_i - \mu_x) f_{i.} = \sum_i x_i f_{i.} - \mu_x \sum_i f_{i.} = \mu_x - \mu_x = 0$

$$\sum_j (y_j - \mu_y) f_{.j} = \sum_j y_j f_{.j} - \mu_y \sum_j f_{.j} = \mu_y - \mu_y = 0$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La covarianza

**NB:**  
se la **COV = 0**, non è detto che  $X$  e  $Y$  siano indipendenti

esempio:

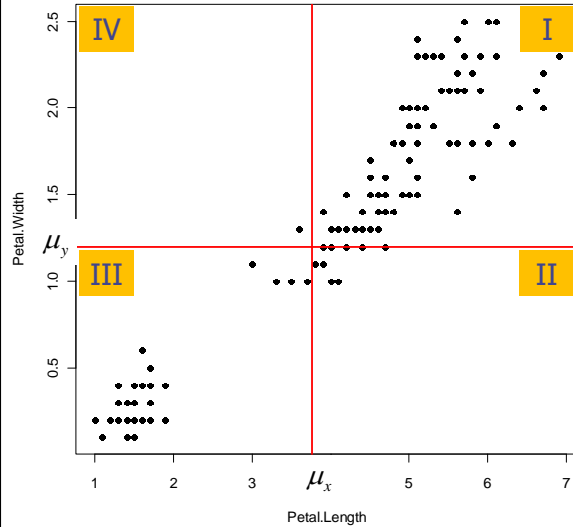
	14	15	16
2	0.25	0	0.25
4	0	0.50	0

caso di **dipendenza perfetta** di  $X$  da  $Y$

$$\begin{aligned} COV(X,Y) &= \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) f_{ij} = \\ &= (2-3)(14-15) \cdot .25 + (2-3)(16-15) \cdot .25 + (4-3)(15-15) \cdot .50 = \\ &= 0.25 - 0.25 + 0 = 0 \end{aligned}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## La covarianza: interpretazione



Quadranti:

- I:  $(x_i - \mu_x) > 0; (y_j - \mu_y) > 0$
- II:  $(x_i - \mu_x) > 0; (y_j - \mu_y) < 0$
- III:  $(x_i - \mu_x) < 0; (y_j - \mu_y) < 0$
- IV:  $(x_i - \mu_x) < 0; (y_j - \mu_y) > 0$

quindi, :

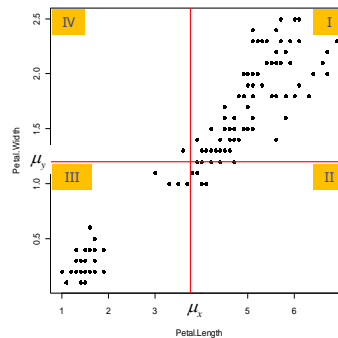
- punti in I e III: contribuiscono **positivamente** al calcolo della COV;
- punti in II e IV: contribuiscono **negativamente** al calcolo della COV.

È vietata la riproduzione non autorizzata a fini commerciali.

## La covarianza: interpretazione

Nel caso in cui, come nell'esempio, la nuvola di punti si trovi prevalentemente all'interno del I e III quadrante, allora la **covarianza è positiva**.

In tal caso, all'aumentare di una variabile, l'altra in media **aumenta**.



Se la **covarianza è negativa**, la nuvola dei punti si trova prevalentemente all'interno del II e IV quadrante.

In tal caso, all'aumentare di una variabile, l'altra in media **diminuisce**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Ancora sulla covarianza

Una proprietà:

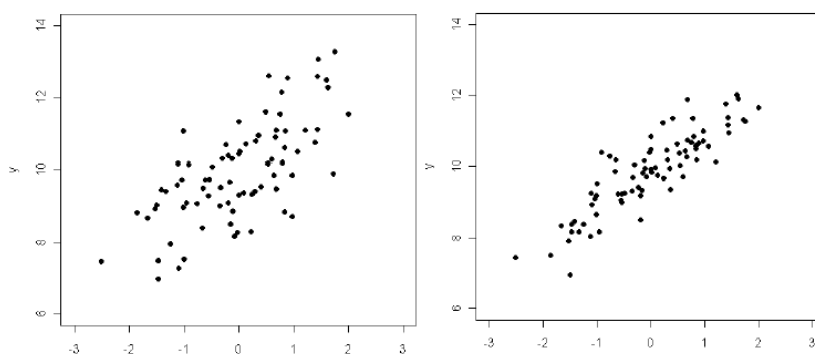
$$\begin{aligned} COV(aX, bY) &= \sum_i \sum_j (ax_i - a\mu_x)(by_j - b\mu_y) f_{ij} = \\ &= \sum_i \sum_j a(x_i - \mu_x)b(y_j - \mu_y) f_{ij} = \\ &= ab \sum_i \sum_j (x_i - \mu_x)(y_j - \mu_y) f_{ij} = abCOV(X, Y) \end{aligned}$$

ovvero, **cambiando unità di misura cambia il valore della covarianza**. Il suo valore, quindi, di per sé non è indicativo di niente.

**Solo il suo segno è informativo.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Confronto tra scatter plot



Entrambe i grafici illustrano situazioni a covarianza positiva. Dal confronto (condotto ovviamente **a parità di scala**) è però immediato comprendere quale sia il contesto in cui si osserva un'associazione più stretta.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Def:  $\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \cdot \sigma_Y}$  **coefficiente di correlazione**

Proprietà:

- 1)  $\rho_{X,Y}$  mantiene lo stesso segno di  $COV(X,Y)$  **con stesso significato;**
- 2) poiché  $VAR(aX) = a^2VAR(X)$ ,

$$\rho_{aX,bY} = \frac{COV(aX,bY)}{\sigma_{aX} \cdot \sigma_{bY}} = \frac{abCOV(X,Y)}{a\sigma_X \cdot b\sigma_Y} = \rho_{X,Y}$$

cioè  $\rho_{X,Y}$  non dipende dall'unità di misura;

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Proprietà:

- 3) i valori che assume sono:  $-1 \leq \rho_{X,Y} \leq 1$

In particolare, se  $Y = a + bX \Rightarrow VAR(Y) = b^2VAR(X) \Rightarrow \sigma_Y = |b|\sigma_X$   
 quindi:

$$\rho_{X,Y} = \frac{COV(X, a + bX)}{\sigma_X \cdot |b|\sigma_X} = \frac{bCOV(X, X)}{\sigma_X \cdot |b|\sigma_X} = \frac{bVAR(X)}{\sigma_X \cdot |b|\sigma_X} = \frac{b}{|b|} = \pm 1$$

a seconda del segno di  $b$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Teorema:  $|\rho_{X,Y}| \leq 1$

Dimostrazione:

consideriamo  $VAR(X + dY) = VAR(X) + d^2VAR(Y) + 2dCOV(X, Y) \geq 0$

Valutiamo:  $VAR(X + dY) = 0$

$$\Rightarrow X + dY = k \text{ costante} \Rightarrow Y = \frac{k}{d} - \frac{1}{d}X \Rightarrow \rho = \pm 1 \text{ a seconda del segno di } d$$

Valutiamo:  $VAR(X + dY) > 0$

in tal caso, il discriminante dell'equazione di secondo grado (parabola) è negativo

$$\Delta = 4COV(X, Y)^2 - 4VAR(Y)VAR(X) < 0$$

$$\Rightarrow \frac{COV(X, Y)^2}{VAR(Y)VAR(X)} < 1 \quad \text{ovvero} \quad \rho^2 < 1 \Leftrightarrow |\rho| < 1$$

CVD

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Quindi:

- valori di  $\rho$  prossimi a 1 indicano punti molto vicini alla **retta interpolante inclinata positivamente**;
- valori di  $\rho$  prossimi a -1 indicano punti molto vicini alla **retta interpolante inclinata negativamente**.

Per questo motivo

$\rho$  è un indice di **interdipendenza LINEARE**.

**Interdipendenza...** perché se la relazione di dipendenza fosse perfettamente LINEARE, la  $Y$  dipenderebbe perfettamente dalla  $X$  e viceversa la  $X$  perfettamente dalla  $Y$ , analogamente al caso esaminato delle tavole di contingenza quadrate del tipo:

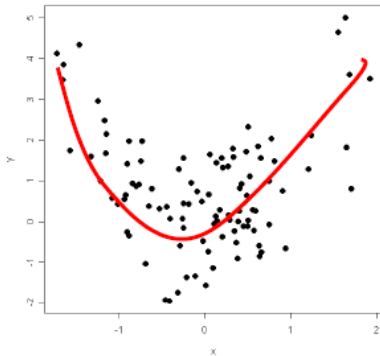
	y1	y2	y3
x1	0		0
x2		0	0
x3	0	0	

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Quindi:

se  $\rho = 0$ , non è detto che  $X$  e  $Y$  siano indipendenti

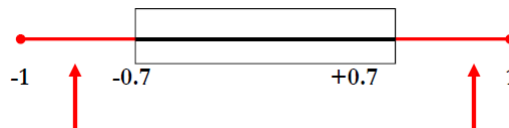


Nel grafico a sinistra si evidenzia ad esempio un **legame di tipo quadratico**, per cui si può **SOLO** concludere che **le variabili non sono LINEARMENTE interdipendenti**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione

Quali valori di  $\rho$  fanno ritenere che ci sia **forte** associazione LINEARE fra le variabili?



(forte) associazione negativa

(forte) associazione positiva

**NB:** le soglie dipendono però dal tipo di studio che si sta conducendo. In alcuni ambiti scientifici le variabili si considerano linearmente associate anche per valori assoluti di  $\rho$  inferiori a 0.7.

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 9

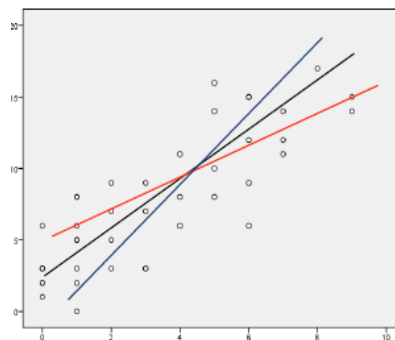
# Regressione lineare semplice

È vietata la riproduzione non autorizzata a fini commerciali.

## Introduzione

La Lezione 8 ha chiarito l'opportunità di approssimare una certa realtà di interesse mediante un modello matematico. Sintetizzare il trend di un certo insieme di osservazioni **mediante una retta**, significa **optare per un modello matematico molto semplice e di immediata interpretazione**.

**Ma quale retta scegliere?**



È vietata la riproduzione non autorizzata a fini commerciali.



## Modello classico di regressione lineare semplice

- ❑ **Modello:** concettualizzazione / costruzione finalizzata all'approssimazione di una certa realtà
- ❑ **Classico:** in riferimento alle ipotesi che stanno alla base del modello
- ❑ **Regressione:** vedi diapositiva successiva
- ❑ **Lineare:** il modello è caratterizzato da una combinazione lineare dei parametri che lo compongono
- ❑ **Semplice:** il modello è il più semplice possibile, ovvero si analizza la relazione esistente tra due sole variabili, la  $Y$  che assume il ruolo di variabile risposta o dipendente e la  $X$  che assume il ruolo di variabile esplicativa o indipendente.

È vietata la riproduzione non autorizzata a fini commerciali.

## Regressione...

Il termine regressione e la sua applicazione a problemi statistici furono introdotti verso la metà dell'ottocento, insieme con i concetti di base della correlazione, dall'inglese **Sir Francis Galton** (1822 - 1911).

Galton, di famiglia nobile inglese, era cugino di **Charles Darwin**. Il libro di Darwin del 1861 («*Origin of Species*») fu fonte di ispirazione per le sue ricerche.

Tra i tanti studi che condusse, Galton voleva verificare se la statura dei figli potesse essere prevista sulla base di quella dei genitori. Ed esprimere questa corrispondenza in una legge matematica. Se, conoscendo l'altezza dei genitori, è possibile predire quella dei figli, a maggior ragione è dimostrato che l'altezza è ereditaria. Il ragionamento del Galton genetista era: **nell'uomo esistono fattori ereditari fisici e psicologici?**

Il suo studio fu pubblicato nel 1886 su *Journal of the Anthropological Institute*, Vol. 15: ***Regression towards mediocrity in hereditary stature.***

In 309 casi, misurò l'altezza del figlio adulto e quella dei genitori. Rimase colpito dal fatto che a genitori alti corrispondevano mediamente figli di altezza leggermente inferiore. Simmetricamente, tra i genitori più bassi, osservò figli mediamente più alti. Chiamò questo fenomeno **regressione verso la mediocrità** corretta poi dagli statistici, con termini più appropriati, in **regressione verso la media**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Dalla teoria alla pratica: l'ipotesi *ceteris paribus*

Esempio:

È, in teoria, noto che la quantità di grano producibile per m<sup>2</sup> è certamente connessa alla fertilizzazione del terreno ma anche alla composizione dello stesso, e verosimilmente ai fattori metereologici, alla presenza di parassiti, ecc.

Si vogliono ora stabilire gli effetti di un nuovo fertilizzante a base di azoto nell'incremento della produzione di grano. Il ricercatore deve quindi adoperarsi per **mantenere «fisse»** tutte le altre variabili che possono influenzare la relazione:



L'osservazione empirica e la conseguente analisi devono quindi essere condotte sotto l'ipotesi *ceteris paribus* (= a parità di tutte le altre circostanze), **dal momento che risulta impossibile controllare tutte le variabili legate al problema.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Il modello di regressione lineare semplice

Per conciliare le diversità riscontrate tra teoria e pratica:

$$Y = f(X, K, Z, \dots)$$



**Primo livello di approssimazione:**  
 esistono altre variabili che possono avere un effetto sulla  $Y$  ma che sono impossibili da controllare.

$$Y = f(X) + \varepsilon'$$



**Secondo livello di approssimazione:**  
 la relazione tra  $Y$  e  $X$  è modellata in termini lineari.

$$Y = \alpha + \beta X + \varepsilon$$

**ERRORE  $\varepsilon$  :** include entrambe i livelli di approssimazione

È vietata la riproduzione non autorizzata a fini commerciali.

## Il modello di regressione lineare semplice

Approssimare la  $f$  mediante funzione lineare vuol dire ipotizzare che la relazione tra  $Y$  e  $X$  risulti lineare in media.

Ovvero immaginare che tutti i  $E(Y/X = x_i)$  siano disposti su una retta:

$$E(Y/X = x_i) = \alpha + \beta x_i$$

Tale funzione prende appunto il nome di funzione di regressione di  $Y$  su  $X$ .

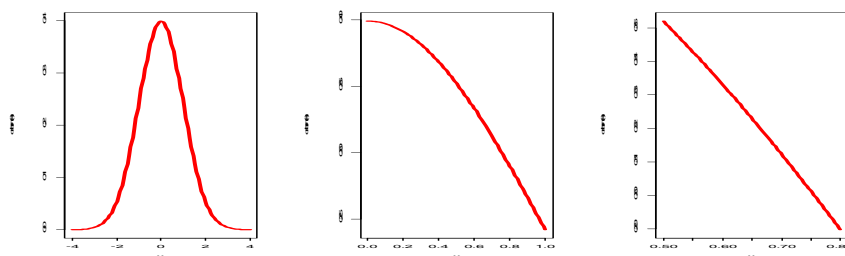
## Validità dell'approssimazione

**NB:**

a prima vista l'ipotesi di linearità può apparire poco realistica e, dunque, molto restrittiva.

In realtà occorre osservare che:

- anche se la  $f$  è molto distante dalla linearità, l'approssimazione lineare funziona abbastanza bene in intervalli limitati:



## Validità dell'approssimazione

Inoltre:

- un problema non lineare si può sempre analizzare mediante un modello lineare:

$$y = \alpha + \beta \frac{1}{x} \rightarrow y = \alpha + \beta w$$

$$y = \alpha + \beta x + \gamma x^2 \rightarrow y = \alpha + \beta x + \gamma w$$

$$y = \alpha x^\beta \rightarrow \log(y) = \log(\alpha) + \beta \log(x)$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello

**A posteriori:** le  $(x_i, y_i)$  sono coppie di valori osservati;

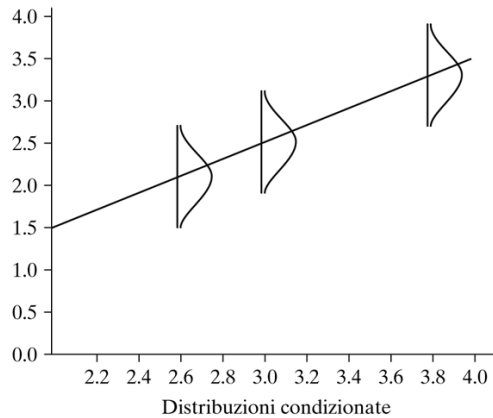
**A priori:** non sappiamo quale valore di  $Y$  si verificherà  
(es: non sappiamo quanto grano osserveremo in una  
particella di terreno trattata con un certo dosaggio di azoto)

**A priori** quindi:

- la  $X$  è una **variabile non stocastica** che assume fissati valori;
- l'analisi viene condotta **condizionatamente** ai vari valori di  $X$ ; in altre parole, si considerano i possibili valori di  $Y$  che possono verificarsi, fissato ciascuno dei valori  $X = x_i$

È vietata la riproduzione non autorizzata a fini commerciali.

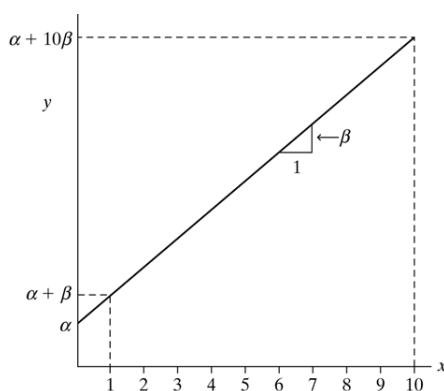
### Interpretazione del modello



Analogamente all'analisi dell'associazione per variabili categoriali, se le condizionate fossero tutte uguali tra loro (in media e in varianza) sarebbero tutte collocate su una retta parallela all'ascisse.  
 In tal caso, al variare di  $X$ , la  $Y$  non varierebbe in media, ovvero la  $X$  non si mostrerebbe correlata con la  $Y$ .

È vietata la riproduzione non autorizzata a fini commerciali.

### Interpretazione del modello



L'intercetta  $\alpha$  è il valore che assume  $E(Y/X = 0)$ .

La pendenza (*coefficiente angolare*)  $\beta$  esprime la variazione di  $Y$  per incrementi unitari di  $X$ .  
 Cioè, per due valori di  $x$  che differiscono di 1.0 (per esempio  $x = 0$  e  $x = 1$ ), i valori di  $y$  differiscono di una quantità  $\beta$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Le ipotesi classiche sul modello

- la  $X$  è una variabile non stocastica;
- $E(\varepsilon_i) = 0 \quad \forall i$
- $VAR(\varepsilon_i) = \sigma^2 \quad \forall i$
- $COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad i \neq j$

### NB:

$\varepsilon$  e  $Y$  sono due v.c. strettamente legate tra loro in quanto hanno:

- stessa forma;
- medie diverse  $E(\varepsilon_i) = 0 \Leftrightarrow E(y_i) = \alpha + \beta x_i$
- varianza uguale  $VAR(y_i) = VAR(\alpha + \beta x_i + \varepsilon_i) = VAR(\varepsilon_i) = \sigma^2$

È vietata la riproduzione non autorizzata a fini commerciali.

## Popolazione e campione: l'equazione di previsione

Se il modello  $Y = \alpha + \beta X + \varepsilon$  venisse in generale ritenuto realistico, occorre ricordare che **i suoi parametri incogniti hanno la funzione di descrivere una certa realtà d'interesse.**

Ma **in generale si osservano dati di natura campionaria**, tramite i quali è solo possibile pervenire ad una **stima** di  $(\alpha, \beta)$ :

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \quad \text{equazione di previsione}$$

Tale notazione rappresenta un'equazione che stima il modello ipotizzato ed è **in grado di fornire una previsione per la variabile risposta in relazione ad un qualsiasi valore di  $x$ .**

È vietata la riproduzione non autorizzata a fini commerciali.

## I residui

La distanza di un punto dalla retta di previsione:

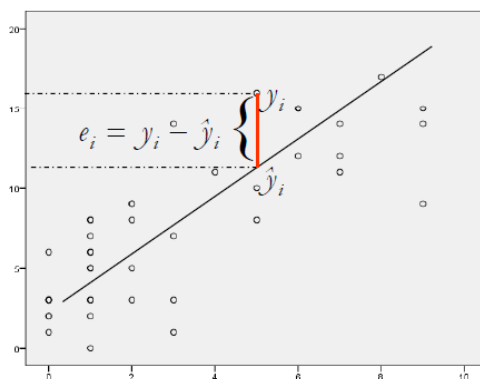
$$e_i = y_i - (\hat{\alpha} + \hat{\beta}x_i) \text{ prende il nome di } \mathbf{RESIDUO}.$$

Il residuo non è l'errore, ma solo **una sua stima**.

L'errore vero da modello infatti risulta:  $\varepsilon_i = y_i - (\alpha + \beta x_i)$

## Il metodo dei minimi quadrati

Tra tutte le possibili rette, la retta di previsione è quella che rende minima la somma dei quadrati dei residui:



$$Q = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

## Il metodo dei minimi quadrati

Le **stime**  $(\hat{\alpha}, \hat{\beta})$  che minimizzano  $Q$  sono:

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Regressione e correlazione

**NB:** 
$$\frac{S_x}{S_y} \hat{\beta} = \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{\sum (y_i - \bar{y})^2}} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} =$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = r_{X,Y} \text{ stimatore di } \rho_{X,Y}$$

⇒ 
$$r_{X,Y} = \frac{S_x}{S_y} \hat{\beta}$$

Il coefficiente di correlazione è il valore che assume la pendenza della retta di previsione **quando le due variabili hanno deviazioni standard uguali.**

È vietata la riproduzione non autorizzata a fini commerciali.



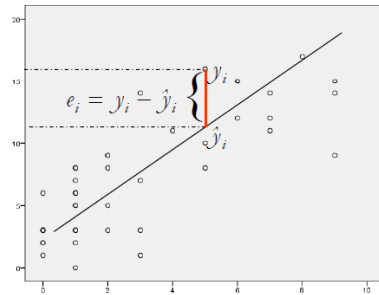
## Bontà d'adattamento

Domanda:

*Quanta parte della variabilità della Y è imputabile alla X?*

La retta stimata esprime il legame tra Y e X:  
 quindi, dato un certo  $x_i$ , il corrispondente valore  $y_i$  risulta in parte determinato da  $x_i$  ed in parte dall'errore  $e_i$ :

$$y_i = \hat{y}_i + e_i = (\hat{\alpha} + \hat{\beta}x_i) + e_i$$



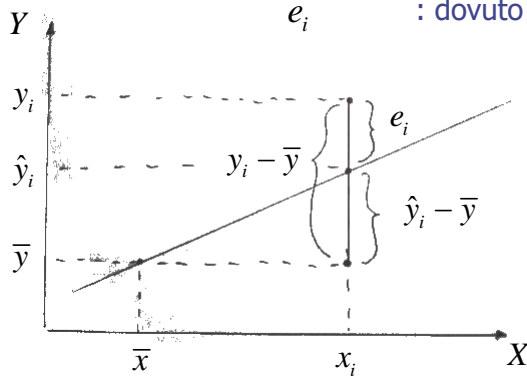
È vietata la riproduzione non autorizzata a fini commerciali.

## Bontà d'adattamento

Da ciò segue:  $y_i - \bar{y} = (\hat{y}_i + e_i) - \bar{y} = (\hat{y}_i - \bar{y}) + e_i$

dove  $(\hat{y}_i - \bar{y})$  : dovuto alla  $x_i$  a livello di stima

$e_i$  : dovuto all'errore a livello di stima



È vietata la riproduzione non autorizzata a fini commerciali.

## Bontà d'adattamento

Possiamo quindi scomporre l'indice di variabilità della  $Y$ :

$$\begin{aligned} \overbrace{\sum_i^n (y_i - \bar{y})^2}^{VT} &= \sum_i^n [(\hat{y}_i - \bar{y}) + e_i]^2 = \dots \\ &= \underbrace{\sum_i^n (\hat{y}_i - \bar{y})^2}_{VX} + \underbrace{\sum_i^n e_i^2}_{VE} \quad \Rightarrow \quad VT = VX + VE \end{aligned}$$

Ovvero, la **variabilità totale della  $Y$**  può essere scomposta in una **parte attribuibile alla  $X$**  e una **parte attribuibile all'errore**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione

➔ un importante indice di adattamento:

**coefficiente di determinazione**  $R^2 = \frac{VX}{VT} = \frac{VT - VE}{VT} = 1 - \frac{VE}{VT}$

➔  $0 \leq R^2 \leq 1$

$R^2$  indica quanta parte della variabilità di  $Y$  è spiegata dal modello; **in altre parole fornisce una idea dell'importanza di  $X$  nel determinare  $Y$ .**

casi limite:

$R^2 = 0$  :  $VX = 0$ , ovvero le variazioni della  $Y$  non sono dovute all'effetto della variabile indipendente;

$R^2 = 1$  :  $VE = 0$ , tutti gli errori sono zero, ovvero tra la  $Y$  e la  $X$  esiste un legame di interdipendenza lineare perfetta.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione

Nell'output di alcuni software statistici:

**TSS**: Total Sum of Squares = VT

**SSE**: Sum of Squared Errors = VE

È possibile dimostrare la relazione:  $R^2 = r_{X,Y}^2$

Ovvero **il coefficiente di determinazione è il quadrato del coefficiente di correlazione.**

È anche possibile dimostrare che:  $R^2 = r_{\hat{Y},Y}^2$

L'utilità di  $r_{\hat{Y},Y}$  verrà chiarita nella Lezione 11.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione

$$R^2 = r_{X,Y}^2$$

Dimostrazione:

$$\begin{aligned} R^2 &= \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} \\ \sum_i^n (\hat{y}_i - \bar{y})^2 &= \sum_i^n (\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2 = \sum_i^n (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{y})^2 = \\ &= \hat{\beta}^2 \sum_i^n (x_i - \bar{x})^2 = \left[ \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} \right]^2 \sum_i^n (x_i - \bar{x})^2 \\ &= \frac{(\sum_i^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_i^n (x_i - \bar{x})^2} \\ \Rightarrow R^2 &= \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2} = \frac{(\sum_i^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2} = r_{X,Y}^2 \quad \text{CVD} \end{aligned}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione

$$R^2 = r_{\hat{Y}, Y}^2$$

Dimostrazione:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \quad r_{\hat{Y}, Y}^2 = \frac{\left[ \sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y}) \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2 \sum_i (y_i - \bar{y})^2}$$

$$\bar{\hat{y}} = \frac{1}{n} \sum_i \hat{y}_i = \frac{1}{n} \sum_i (\alpha + \beta x_i) = \alpha + \beta \bar{x} = \bar{y}$$

$$\frac{\left[ \sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y}) \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2} = \frac{\left[ \sum_i (\hat{y}_i - \bar{y})(y_i - \bar{y}) \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2} =$$

$$= \frac{\left[ \sum_i (\hat{y}_i - \bar{y})(\hat{y}_i + e_i - \bar{y}) \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2} = \frac{\left[ \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i (\hat{y}_i - \bar{y}) \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2}$$

$$= \frac{\left[ \sum_i (\hat{y}_i - \bar{y})^2 \right]^2}{\sum_i (\hat{y}_i - \bar{y})^2} = \sum_i (\hat{y}_i - \bar{y})^2 \Rightarrow r_{\hat{Y}, Y}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2 \quad \text{CVD}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Stima della varianza condizionata

Come stimare la varianza condizionata

$$\sigma^2 = \text{VAR}(y_i) = \text{VAR}(\alpha + \beta x_i + \varepsilon_i) = \text{VAR}(\varepsilon_i) ?$$

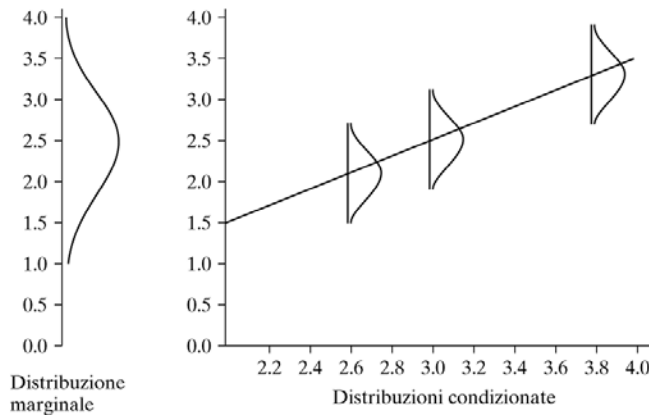
$$\Rightarrow s^2 = \frac{\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} =$$

$$= \frac{\sum_i e_i^2}{n-2} = \frac{SSE}{n-2}$$

↪ si perdono 2 *gdl* a causa del doppio vincolo legato alla stima dei due parametri che definiscono la retta.

È vietata la riproduzione non autorizzata a fini commerciali.

## Variabilità condizionata e variabilità marginale



$s^2$  è una stima della varianza condizionata.

La varianza condizionata **non deve essere confusa** con la varianza di  $Y$  (varianza della distribuzione marginale); questa è in generale più grande della varianza condizionata.

È vietata la riproduzione non autorizzata a fini commerciali.

## Regressione e inferenza

Un **intervallo di confidenza** per il coefficiente angolare di un modello di regressione lineare semplice informa sull'importanza dell'effetto di  $X$  su  $Y$ .

Un **test d'ipotesi** sul coefficiente angolare di un modello di regressione lineare consente di verificare se due variabili quantitative sono statisticamente indipendenti, e ha la stessa finalità di un test chi-quadro per variabili categoriali.

È vietata la riproduzione non autorizzata a fini commerciali.

## Regressione e inferenza

Notare che, in ottica inferenziale:

$$\hat{\alpha}, \hat{\beta}$$

sono **variabili casuali** (il campione verrà estratto domani), le cui distribuzioni sono caratterizzate dai differenti valori che potranno verificarsi nell'universo di tutti i possibili campioni.

## Regressione e inferenza

### Un'ipotesi aggiuntiva...

- la  $X$  è una variabile non stocastica;
- $E(\varepsilon_i) = 0 \quad \forall i$
- $VAR(\varepsilon_i) = \sigma^2 \quad \forall i$
- $COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad i \neq j$

□ le condizionate di  $Y$  a ciascun valore di  $X$  seguono una **distribuzione Normale**.

➔  $Y / x_i \sim N(\alpha + \beta x_i, \sigma^2) \Leftrightarrow \varepsilon_i \sim N(0, \sigma^2)$

## Regressione e inferenza

È possibile dimostrare che:

$$\left\{ \begin{array}{l} E(\hat{\beta}) = \beta \\ VAR(\hat{\beta}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \end{array} \right.$$

Poiché  $\hat{\beta}$  è **combinazione lineare di v.c. distribuite Normalmente**, per i teoremi visti nella Lezione 5 (una **combinazione lineare di distribuzioni Normali è ancora Normale**):

→  $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$

È vietata la riproduzione non autorizzata a fini commerciali.

## Regressione e inferenza

Standardizzando:

$$\frac{\hat{\beta} - \beta}{\sigma \sqrt{\sum_i (x_i - \bar{x})^2}} \sim N(0,1) \quad \text{se } \sigma \text{ fosse noto;}$$

$$\frac{\hat{\beta} - \beta}{s \sqrt{\sum_i (x_i - \bar{x})^2}} \sim t_{n-2}$$

in quanto  $\sigma$  deve essere stimato tramite:

$$s = \sqrt{\frac{\sum_i e_i^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per il coefficiente angolare

### Intervallo di confidenza per $\beta$

$$\Pr \left( \hat{\beta} - t_{\frac{\alpha}{2}; n-2} \cdot \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} \leq \beta \leq \hat{\beta} + t_{\frac{\alpha}{2}; n-2} \cdot \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} \right) = 1 - \alpha$$

Limiti dell'intervallo a livello  $1 - \alpha$  :

$$\hat{\beta} \pm t_{\frac{\alpha}{2}; n-2} \cdot \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per il coefficiente angolare

### Test d'ipotesi

$$\begin{cases} H_0 : \beta \leq \beta_0 \\ H_1 : \beta > \beta_0 \end{cases} \quad \begin{cases} H_0 : \beta = \beta_0 \\ H_1 : \beta \neq \beta_0 \end{cases}$$

Respingo  $H_0$  se:

$$\frac{\hat{\beta} - \beta_0}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}} > t_{\alpha; n-2} \quad \frac{|\hat{\beta} - \beta_0|}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}} > t_{\frac{\alpha}{2}; n-2}$$

È vietata la riproduzione non autorizzata a fini commerciali.

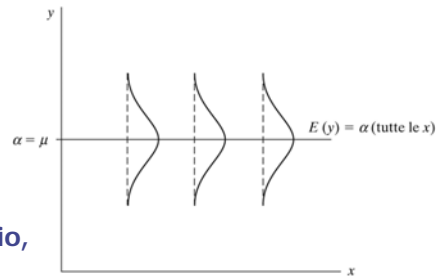


## Inferenza per il coefficiente angolare

### Il caso più frequente: test d'indipendenza

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

Infatti, sotto  $H_0$ ,  
**le distribuzioni condizionate hanno tutte stesso valor medio**,  
 ovvero tutte le distribuzioni condizionate sono identiche tra loro.



In tal caso, respingo  $H_0$  se:

$$\frac{|\hat{\beta} - 0|}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}} > t_{\frac{\alpha}{2}; n-2}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per la correlazione

L'assenza di correlazione si verifica quando la **pendenza della retta di previsione è nulla**.

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : \rho_{X,Y} = 0 \\ H_1 : \rho_{X,Y} \neq 0 \end{cases}$$

Quindi, si respinge  $H_0$  se:

$$\frac{|\hat{\beta} - 0|}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}} > t_{\frac{\alpha}{2}; n-2} \equiv \frac{|r_{X,Y} - 0|}{\sqrt{\frac{1 - r_{X,Y}^2}{n-2}}} > t_{\frac{\alpha}{2}; n-2}$$

Le due statistiche test sono **coincidenti**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per la correlazione

Dimostrazione:  $\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : \rho_{X,Y} = 0 \\ H_1 : \rho_{X,Y} \neq 0 \end{cases}$

$$\begin{aligned} \frac{|r_{X,Y} - 0|}{\sqrt{\frac{1 - r_{X,Y}^2}{n - 2}}} &= \frac{\left| \frac{S_X}{S_Y} \hat{\beta} - 0 \right|}{\sqrt{\frac{1 - R^2}{n - 2}}} = \frac{\left| \hat{\beta} - 0 \right|}{\frac{S_Y}{S_X} \sqrt{\frac{SSE/TSS}{n - 2}}} = \frac{\left| \hat{\beta} - 0 \right|}{\sqrt{\frac{TSS}{\sum_i (x_i - \bar{x})^2} \frac{SSE/TSS}{n - 2}}} = \\ &= \frac{\left| \hat{\beta} - 0 \right|}{\sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}} = \frac{\left| \hat{\beta} - 0 \right|}{s} \end{aligned}$$

CVD

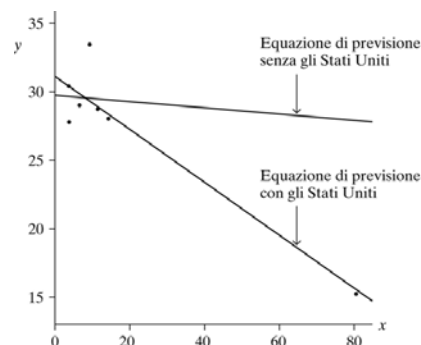
È vietata la riproduzione non autorizzata a fini commerciali.

## Il problema degli outliers

Uno svantaggio del metodo dei minimi quadrati è che singole osservazioni possono **condizionare** (talvolta pesantemente) **il processo di stima**.

esempio:  
 consideriamo le variabili  $Y$  = tasso di natalità (nati per 1000 ab.) e  $X$  = numero di televisioni per 100 abitanti, per diverse nazioni africane e asiatiche.

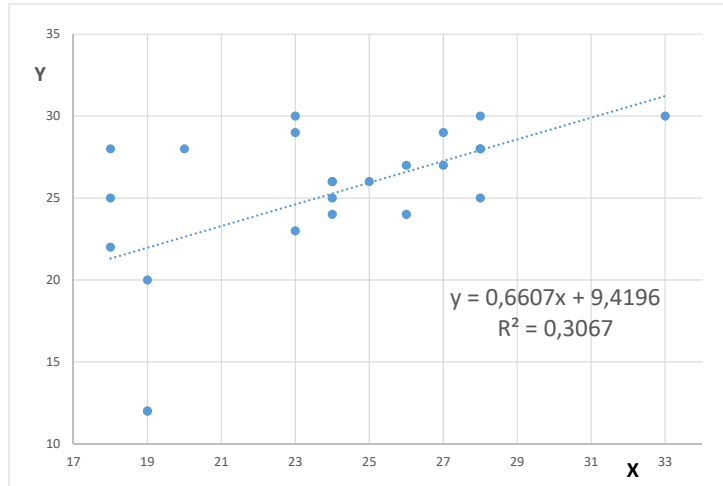
Il grafico illustra cosa succede inserendo nei dati gli USA.



È vietata la riproduzione non autorizzata a fini commerciali.

### Esempio: il voto alla prova intermedia di Statistica è un buon predittore del voto finale?

Y	X
22	18
28	18
25	18
12	19
12	19
20	19
28	20
29	23
30	23
23	23
26	24
26	24
25	24
24	24
26	25
24	26
27	26
27	27
29	27
28	28
30	28
28	28
25	28
30	33



È vietata la riproduzione non autorizzata a fini commerciali.

### Esempio: il voto alla prova intermedia di Statistica è un buon predittore del voto finale?

Y	X	Xmedio	Ymedio	DevX	DevY = VT	CoDevXY	Yhat	(Yhat-Ymedio)^2 = VX		
22	18	23,833	25,167	34,028	10,028	18,472	21,313	14,855		
28	18			34,028	8,028	-16,528	21,313	14,855	Beta-hat	0,66071
25	18			34,028	0,028	0,972	21,313	14,855	Alpha-hat	9,41964
12	19			23,361	173,361	63,639	21,973	10,198		
12	19			23,361	173,361	63,639	21,973	10,198	R^2	0,306731
20	19			23,361	26,694	24,972	21,973	10,198		
28	20			14,694	8,028	-10,861	22,634	6,415	VE	368,3571
29	23			0,694	14,694	-3,194	24,616	0,303	s^2	16,74351
30	23			0,694	23,361	-4,028	24,616	0,303	s	4,091883
23	23			0,694	4,694	1,806	24,616	0,303		
26	24			0,028	0,694	0,139	25,277	0,012	f(0.025;22)	2,074
26	24			0,028	0,694	0,139	25,277	0,012		
25	24			0,028	0,028	-0,028	25,277	0,012	Conf.Interv	
24	24			0,028	1,361	-0,194	25,277	0,012	Beta Linf	0,22149
26	25			1,361	0,694	0,972	25,938	0,594	Beta Lsup	1,09994
24	26			4,694	1,361	-2,528	26,598	2,049		
24	26			4,694	3,361	3,972	26,598	2,049	Test su Beta	
27	26			10,028	3,361	5,806	27,259	4,378	t.obs	3,11989
27	27			10,028	14,694	12,139	27,259	4,378	t.crit	2,074
29	27			17,361	8,028	11,806	27,920	7,579		
28	28			17,361	23,361	20,139	27,920	7,579		
30	28			17,361	8,028	11,806	27,920	7,579		
28	28			17,361	0,028	-0,694	27,920	7,579		
25	28			84,028	23,361	44,306	31,223	36,682		
30	33			Totale	373,333	531,333	246,667	162,976		

È vietata la riproduzione non autorizzata a fini commerciali.

## Esempio: il voto alla prova intermedia di Statistica è un buon predittore del voto finale?

```
> dati=read.table("c:\\users\\bruno\\desktop\\pino.txt",header=T)
> m=lm(Y~X,data=dati)
> summary(m)
```

```
lm(formula = Y ~ X, data = dati)
Residuals:
    Min       1Q   Median       3Q      Max
-9.9732 -1.3616  0.0804  1.8259  6.6875
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4196     5.1159   1.841  0.07911 .
X              0.6607     0.2118   3.120  0.00499 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.092 on 22 degrees of freedom
Multiple R-squared:  0.3067,    Adjusted R-squared:  0.2752
F-statistic: 9.734 on 1 and 22 DF,  p-value: 0.004988
```

```
> 2*(1-pt(3.12,df=22))
[1] 0.004986254
```

È vietata la riproduzione non autorizzata a fini commerciali.

## Lezione 10

# Relazioni multivariate

È vietata la riproduzione non autorizzata a fini commerciali.

## Associazione e causazione

Negli studi sperimentali, e ancor più in quelli osservazionali, raramente le manifestazioni di un fenomeno collettivo si limitano all'analisi di un solo carattere.

Quando si considerano due, o più caratteri, la ricerca NON può limitarsi all'esame delle singole variabili. **L'obiettivo è, soprattutto, quello di esaminare anche il tipo e l'intensità delle relazioni che sussistono tra i caratteri rilevati.**

Nella lezione 7, sono stati illustrate alcune tecniche statistiche per valutare tipo (e intensità laddove il contesto osservazionale lo consenta) dell'associazione tra variabili categoriali.

## Associazione e causazione

In molti ambiti scientifici, notevole importanza riveste la possibilità di individuare le cosiddette **relazioni di tipo causale** tra variabili.

Se esiste tra due variabili  $X$  e  $Y$  esiste una relazione che consente di verificare la sussistenza di:

- associazione tra le variabili;**
- appropriato ordine cronologico;**
- assenza di spiegazioni alternative.**

allora è possibile parlare di **relazione causale** tra  $X$  e  $Y$ , generalmente rappresentata secondo la simbologia:  $X \rightarrow Y$

Secondo questa rappresentazione:

**$X$  è una variabile esplicativa che ha un'influenza causale su  $Y$**  ( $X$  è la causa,  $Y$  la conseguenza).

## Associazione e causazione

### NB:

Verificare l'esistenza di un certo livello di associazione tra  $X$  e  $Y$  **NON** è quindi condizione sufficiente per potersi esprimere a favore della causazione.

Per poter interpretare adeguatamente i risultati, è importante, individuare correttamente quale variabile influenza l'altra (quale la causa, quale l'effetto e in tali accezioni si cela l'effetto tempo).

Ciò però potrebbe non bastare...

*es: danni causati dagli incendi e numero di pompieri impegnati nella loro estinzione*

**Occorre quindi escludere l'esistenza di spiegazioni alternative. e questo è forse il più rilevante dei problemi...**

È vietata la riproduzione non autorizzata a fini commerciali.

## Variabili controllate

In generale, comprendere se e come  $X$  influenzi  $Y$  non è semplice. Una tecnica molto utilizzata è quella del «controllo».

Una variabile è  $Z$  e detta **controllata** quando la sua possibile influenza viene rimossa suddividendo il campione in gruppi per i quali il valore (o gruppi di valori) della variabile da controllare è costante (es: stesso genere, stesso titolo di studio, stessa fascia d'età).

In altre parole, si suddivide il campione in base al numero di modalità di  $Z$ , e per ciascuno di essi si studia se e come  $X \rightarrow Y$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Variabili controllate

es: *le cicogne portano i bambini?*

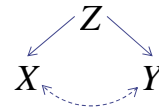
$Y$  = nascite

$X$  = numero nidi di cicogna

Se a «qualcuno» venisse in mente di instaurare una relazione causale tra  $X$  e  $Y$ , quel «qualcuno» si dovrebbe accorgere che la relazione sparisce controllando per i livelli di una terza variabile:

$Z$  = aree rurali / aree urbane

Nelle zone rurali, le famiglie sono più prolifiche e ci sono anche più nidi di cicogna.



In altre parole...

se, ad esempio, si ipotizza che una malattia  $Y$  sia dovuta al fattore  $X$ , non tenere conto dell'età  $Z$  (primo fattore di rischio per quasi tutte le malattie) è sbagliato.

È vietata la riproduzione non autorizzata a fini commerciali.

## Classificazione delle relazioni multivariate

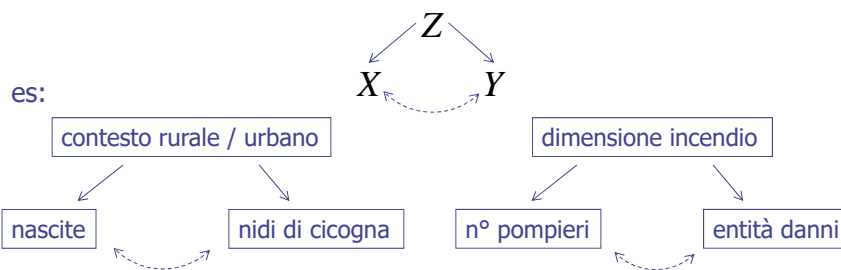
- Relazioni spurie
- Relazioni indirette o concatenate
- Cause multiple
- Variabili sopprimenti
- Interazione

È vietata la riproduzione non autorizzata a fini commerciali.

## Relazioni spurie

La relazione tra  $X$  e  $Y$  è **spuria** se entrambe le variabili dipendono da una terza variabile  $Z$  e se la loro associazione scompare quando  $Z$  è controllata.

È il caso classico di covariazione tra  $X$  e  $Y$  in assenza di causazione. I cambiamenti in  $Z$  producono modificazioni sia in  $Y$  sia in  $X$  che sono, quindi associate, ma solo in funzione della loro associazione con  $Z$ .



È vietata la riproduzione non autorizzata a fini commerciali.

## Relazioni concatenate

Si ha una relazione **indiretta** tra  $X$  e  $Y$  quando il loro legame è mediato da una terza variabile  $Z$ . In questo caso si parla anche di **concatenazione** delle relazioni.

$$X \rightarrow Z \rightarrow Y$$

$Z$  è detta variabile **interveniente** o **mediatrice**.

es:



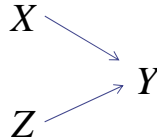
**NB:** l'associazione tra  $X$  e  $Y$  tende a scomparire controllando per  $Z$   
 (ad es: limitando l'analisi alla sola fascia ad alto reddito, la correlazione tra istruzione e lunghezza della vita dovrebbero risultare pressoché nulla)

È vietata la riproduzione non autorizzata a fini commerciali.



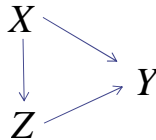
## Cause multiple

Se, come spesso accade, una variabile  $Y$  ha più di una causa, si parla di **cause multiple**.



**NB:**

nella ricerca sociale, le cause sono tra loro associate. Quindi una variabile  $X_1$  può esercitare un *effetto diretto* su  $Y$ , ma vi possono anche essere *effetti indiretti* dovuti alla presenza di variabili intervenienti.



È vietata la riproduzione non autorizzata a fini commerciali.

## Variabili sopprimenti

Vi sono casi in cui due variabili non mostrano alcuna associazione tra loro, fino a quando non viene considerata una terza variabile di controllo, definita **variabile sopprimente**.

esempio:

Istruzione	Reddito		Reddito		Istruzione			
	Alto	Basso	Età	Alto	Basso	Età	Alta	Bassa
Alta	250	250	Alta	350	150	Alta	150	350
Bassa	250	250	Bassa	150	350	Bassa	350	150

Ignorando l'età, la relazione tra Istruzione e Reddito è espressa dalla parte riquadrata della tabella.

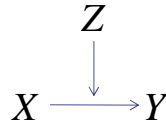
L'età è però positivamente associata con il Reddito e negativamente con l'Istruzione. Per cui controllando per Età, la relazione tra Istruzione e Reddito emerge chiaramente.

Istruzione	Reddito	Età = Bassa			Età = Alta		
		Alto	Basso	% Alto	Alto	Basso	% Alto
Alta	Alta	125	225	35.7%	125	25	83.3%
Bassa	Bassa	25	125	16.7%	225	125	64.3%

È vietata la riproduzione non autorizzata a fini commerciali.

## Interazione statistica

Se il vero effetto del predittore  $X$  su  $Y$  si modifica al variare dei valori assunti da un altro predittore  $Z$ , allora si parla di **interazione statistica** tra  $X$  e  $Z$  nei loro effetti su  $Y$ .



**NB:** la relazione tra  $X$  e  $Y$  potrebbe anche cambiare di direzione per effetto di  $Z$

## Lezione 11

# Regressione lineare multipla

## Modello classico di regressione lineare multipla

- ❑ **Modello:** concettualizzazione / costruzione finalizzata all'approssimazione di una certa realtà
- ❑ **Classico:** in riferimento alle ipotesi che stanno alla base del modello
- ❑ **Regressione:** → Galton
- ❑ **Lineare:** il modello è caratterizzato da una combinazione lineare dei parametri che lo compongono
- ❑ **Multipla:** si analizza la relazione esistente tra più variabili di cui una, la  $Y$ , assume il ruolo di variabile risposta o dipendente mentre le altre assumono il ruolo di variabili esplicative (predittive) o indipendenti.

È vietata la riproduzione non autorizzata a fini commerciali.

## Dalla teoria alla pratica: l'ipotesi *ceteris paribus*

### Esempio:

Riprendiamo l'esempio del grano visto per all'inizio della Lezione 9.

È molto più realistico che si vogliano stabilire gli effetti di un nuovo fertilizzante a base di azoto, fosforo e potassio nell'incremento della produzione di grano. È molto più realistico, perché azoto, fosforo e potassio sono gli elementi chimici che il ricercatore può **direttamente controllare, mantenendo «fisse»** tutte le altre variabili che possono influenzare la produzione di grano e che non sono controllabili (ipotesi « *ceteris paribus* »).

È vietata la riproduzione non autorizzata a fini commerciali.

## Il modello di regressione lineare multipla

Per conciliare le diversità riscontrate tra teoria e pratica:

$$Y = f(X_1, \dots, X_k, \dots, W, Z, \dots)$$



**Primo livello di approssimazione:**  
 esistono variabili che possono avere un effetto sulla  $Y$  ma che sono impossibili da controllare.

$$Y = f(X_1, \dots, X_k) + \varepsilon'$$



**Secondo livello di approssimazione:**  
 la relazione tra  $Y$  e le  $X_i$  è modellata in termini lineari.

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

**ERRORE  $\varepsilon$ :** include entrambe i livelli di approssimazione

È vietata la riproduzione non autorizzata a fini commerciali.

## Il modello di regressione lineare multipla

Approssimare la  $f$  mediante funzione lineare vuol dire ipotizzare che la relazione tra  $Y$  e le  $X_i$  risulti lineare in media.

Ovvero immaginare che tutti i valori attesi delle condizionate siano disposti su un piano:

$$E(Y/X_1 = x_{1i} \cap X_2 = x_{2i} \cap \dots \cap X_k = x_{ki}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello

**A posteriori:**  $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$  sono i valori che identificano la *i-esima* osservazione;

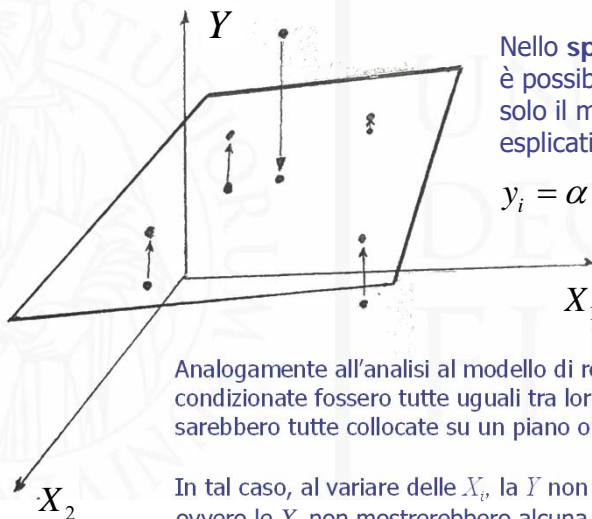
**A priori:** non sappiamo quale valore di  $Y$  si verificherà (es: non sappiamo quanto grano osserveremo in una particella di terreno trattata con un certo dosaggio di azoto, fosforo e potassio)

**A priori** quindi:

- la  $X$  è una **variabile non stocastica** che assume fissati valori;
- l'analisi viene condotta **condizionatamente** ai vari valori delle  $X_i$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello



Nello **spazio tridimensionale** è possibile rappresentare solo il modello con 2 variabili esplicative:

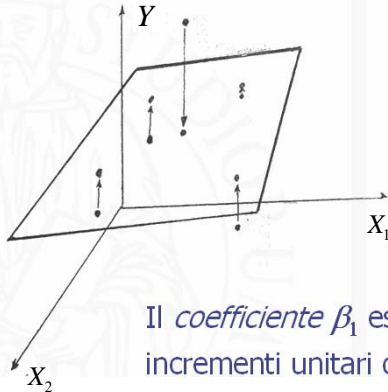
$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Analogamente all'analisi al modello di regressione semplice, se le condizionate fossero tutte uguali tra loro (in media e in varianza) sarebbero tutte collocate su un piano orizzontale.

In tal caso, al variare delle  $X_i$ , la  $Y$  non varierebbe in media, ovvero le  $X_i$  non mostrerebbero alcuna influenza nei confronti della  $Y$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello



L'**intercetta**  $\alpha$  è il valore che ci si attende per  $Y$  se tutte le  $X_i$  fossero uguali a zero.

Il **coefficiente**  $\beta_1$  esprime la variazione di  $Y$  per incrementi unitari di  $X_1$ , ferme restando le altre variabili che possono essere controllate (quindi a parte l'errore). Analogamente,  $\beta_2$  esprime la variazione di  $Y$  per incrementi unitari di  $X_2$ , ferme restando le altre variabili.

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello

Nel modello generale, i **parametri**  $(\beta_1, \beta_2, \dots, \beta_K)$  sono chiamati **coefficienti di regressione parziale**.

### NB1:

L'aggettivo **parziale** distingue questi parametri da quelli del modello di regressione lineare semplice in cui, **piuttosto che controllare, si ignora l'effetto delle altre variabili esplicative**.

Con il modello di regressione lineare multipla mettiamo in evidenza l'influenza di ciascuna variabile esplicativa sulla  $Y$ , separatamente rispetto alle altre.

È vietata la riproduzione non autorizzata a fini commerciali.

## Interpretazione del modello

### NB2:

In particolare, con due variabili esplicative:

**quando  $X_1$  e  $X_2$  sono «cause» indipendenti di  $Y$ , l'effetto di  $X_1$  su  $Y$  non cambia tenendo sotto controllo  $X_2$ .**

Quindi, se la correlazione tra  $X_1$  e  $X_2$  è pari a 0, le inclinazioni parziali e quelle che si otterrebbero stimando i modelli semplici sono identiche.

### Code R:

```
x1=rnorm(100)
x2=rnorm(100)
y=3*x1-2*x2+rnorm(100)
l=lm(y~x1+x2)
summary(l)
```

```
l1=lm(y~x1)
summary(l1)
l2=lm(y~x2)
summary(l2)
```

In generale, però un'inclinazione parziale in modello di regressione multipla è differente da quella che si otterrebbe per un modello di regressione semplice (considerando come esplicativa lo stesso predittore).

È vietata la riproduzione non autorizzata a fini commerciali.

## Le ipotesi classiche sul modello di regressione multipla

- ❑ le  $X_1, X_2, \dots, X_k$  sono variabili non stocastiche;
- ❑ le  $X_1, X_2, \dots, X_k$  sono tali che **nessuna è combinazione lineare delle altre**;
- ❑  $E(\varepsilon_i) = 0 \quad \forall i$
- ❑  $VAR(\varepsilon_i) = \sigma^2 \quad \forall i$
- ❑  $COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad i \neq j$

È vietata la riproduzione non autorizzata a fini commerciali.

## Le ipotesi classiche sul modello di regressione multipla

### NB1:

se  $X_4 = 2X_1 + 3X_2 - 4X_3$

allora  $X_4$  è combinazione lineare delle altre variabili.

In tal caso il modello non funziona perché non riusciamo a distinguere se quello che succede alla  $Y$  è dovuto alla  $X_4$  o alle altre variabili. In altre parole non riusciamo, ad esempio, a far muovere  $X_1$  e non  $X_4$ .

### NB2:

$\varepsilon$  e  $Y$  sono due v.c. strettamente legate tra loro in quanto hanno:

- stessa forma;
- medie diverse  $E(\varepsilon_i) = 0 \Leftrightarrow E(y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$
- varianza uguale  $VAR(y_i) = VAR(\varepsilon_i) = \sigma^2$

È vietata la riproduzione non autorizzata a fini commerciali.

## Popolazione e campione: l'equazione di previsione

Poiché **in generale si osservano dati di natura campionaria**, è solo possibile pervenire ad una **stima** dei parametri che definiscono il modello:  $(\alpha, \beta_1, \beta_2, \dots, \beta_k)$



### equazione di previsione

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki}$$

Tale notazione rappresenta un'equazione che stima il modello ipotizzato ed è **in grado di fornire una previsione per la variabile risposta in relazione a qualsiasi valore assunto dalle  $X_1, \dots, X_k$ .**

È vietata la riproduzione non autorizzata a fini commerciali.



## I residui

La distanza di un punto dal piano di previsione:

$$e_i = y_i - \hat{y}_i \text{ prende il nome di } \mathbf{RESIDUO}.$$

**Il residuo non è l'errore, ma solo una sua stima.**

L'errore vero da modello infatti risulta:

$$\varepsilon_i = y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki})$$

## Il metodo dei minimi quadrati

Analogamente al caso della regressione lineare semplice,  
**l'equazione di previsione è quella che rende minima la  
somma dei quadrati dei residui:**

$$Q = \sum_i^n (y_i - \hat{y}_i)^2 =$$

$$= \sum_i^n \left( y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \dots - \hat{\beta}_k x_{ki} \right)^2$$

NB: le formule per la stima dei parametri sono piuttosto complesse e non verranno illustrate in questo corso.

## Stima della varianza condizionata

Come stimare la varianza condizionata

$$\sigma^2 = \text{VAR}(y_i) = \text{VAR}(\varepsilon_i) \quad ?$$

$$\rightarrow s^2 = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{n - (k + 1)} = \frac{\sum_i^n e_i^2}{n - (k + 1)} = \frac{SSE}{n - (k + 1)}$$



si perdono  $k+1$  *gdl* a causa dei vincoli legati alla stima dei parametri che definiscono il modello.

**In caso di  $k$  variabili esplicative, i parametri da stimare sono  $k+1$**  (nella regressione semplice con una sola  $X$  si stimano  $\alpha$  e  $\beta$ ).

È vietata la riproduzione non autorizzata a fini commerciali.

## Scomposizione della variabilità totale

Analogamente al caso della regressione lineare semplice possiamo scomporre l'indice di variabilità della  $Y$ :

$$\begin{aligned} \overbrace{\sum_i^n (y_i - \bar{y})^2}^{\text{VT (TSS)}} &= \sum_i^n [(\hat{y}_i - \bar{y}) + e_i]^2 = \dots \\ &= \underbrace{\sum_i^n (\hat{y}_i - \bar{y})^2}_{\text{VX}} + \underbrace{\sum_i^n e_i^2}_{\text{VE (SSE)}} \quad \rightarrow \quad \text{VT} = \text{VX} + \text{VE} \end{aligned}$$

Overo, la **variabilità totale della  $Y$**  può essere scomposta in una parte attribuibile alle  $X_i$  e una parte attribuibile all'errore.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione multipla

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$
 coefficiente di determinazione multipla

$$0 \leq R^2 \leq 1$$

$R^2$  indica quanta parte della variabilità di  $Y$  è spiegata dal modello; in altre parole fornisce una idea dell'importanza delle variabili esplicative nel determinare  $Y$ .

casi limite:

$R^2 = 0$  :  $VX = 0$ , ovvero le variazioni della  $Y$  non sono dovute all'effetto delle variabili indipendenti;

$R^2 = 1$  :  $VE = 0$ , tutti gli errori sono zero, ovvero tra la  $Y$  e le  $X_i$  esiste un legame di dipendenza lineare perfetta.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di determinazione multipla

Come prevedere  $Y$ ...

□ senza ausilio di variabili esplicative?  $\longrightarrow \bar{y}$

□ usando le variabili esplicative?  $\longrightarrow \hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$

Il coefficiente di determinazione multipla  $R^2$  misura la riduzione proporzionale dell'errore che si commette impiegando l'equazione di previsione anziché  $\bar{y}$  per prevedere  $y$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione multipla

Ovviamente, non vale più la relazione  $R^2 = r_{X,Y}^2$  in quanto siamo in presenza di più variabili esplicative.

È però sempre possibile dimostrare che:  $R^2 = r_{\hat{Y},Y}^2$

Def:

$r_{\hat{Y},Y}$  è il **coefficiente di correlazione multipla**, e rappresenta la correlazione tra le  $y$  osservate e le  $y$  previste.

→  $0 \leq r_{\hat{Y},Y} \leq 1$

ovvero **i valori previsti non possono essere correlati negativamente con quelli osservati.**

## Il problema della multicollinearità

Quando in un modello ci sono molte variabili esplicative e le correlazioni tra queste sono (molto) forti, capita spesso che l'inserimento di altri predittori nel modello non produca incrementi significativi in  $R^2$ .

Questo fenomeno, particolarmente frequente nell'ambito delle Scienze Sociali, è noto con il termine **multicollinearità**.

Le difficoltà di ordine computazionale causate dalla multicollinearità sono meno stringenti quando si dispone di grandi campioni (idealmente, l'ampiezza campionaria dovrebbe essere almeno 10 volte il numero delle variabili esplicative).

## Regressione e inferenza: un'ipotesi aggiuntiva

- le  $X_1, X_2, \dots, X_k$  sono variabili non stocastiche;
- le  $X_1, X_2, \dots, X_k$  sono tali che **nessuna è combinazione lineare delle altre**;
- $E(\varepsilon_i) = 0 \quad \forall i$
- $VAR(\varepsilon_i) = \sigma^2 \quad \forall i$
- $COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0 \quad i \neq j$
- **le condizionate di  $Y$  a ciascun valore delle  $X_i$  seguono una distribuzione Normale.**

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per i coefficienti di regressione parziale

### Intervallo di confidenza per $\beta_i$

Limiti dell'intervallo a livello  $1 - \alpha$ :  $\hat{\beta}_i \pm t_{\frac{\alpha}{2}; n-(k+1)} \cdot \text{std.err}(\hat{\beta}_i)$

### Test d'ipotesi su $\beta_i$

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \Rightarrow \frac{|\hat{\beta}_i - 0|}{\text{std.err}(\hat{\beta}_i)} > t_{\frac{\alpha}{2}; n-(k+1)}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Esempio:

	num_esami	punteggio	voto_medio_esami	votomat
1	5	18.25	24.33333	80
2	14	12.25	24.00000	65
3	24	21.50	23.80952	60
4	16	15.25	24.14286	80
5	24	18.50	25.42857	74
6	22	20.50	26.90000	75
7	23	18.00	26.05000	72
8	11	16.00	26.00000	95
9	9	11.75	22.14286	71
10	10	15.50	24.88889	86
11	18	12.25	26.25000	83
12	24	19.25	28.90476	100
13	13	15.00	23.18182	90
14	10	11.50	22.75000	71
15	19	15.00	23.61111	72
16	21	17.50	23.78947	80
17	23	22.75	29.60000	93
18	24	20.25	28.40909	60
19	19	18.50	22.88235	72
20	22	19.50	26.14286	71
21	11	11.50	24.00000	73
22	20	19.25	26.83333	75
23	20	11.25	23.44444	80
24	18	15.75	24.25000	77
25	21	12.75	27.52632	82
26	24	17.25	26.85714	85
27	18	20.25	26.25000	74
28	22	15.75	24.05263	60

n = 28 studenti iscritti ad un CdS dell'Ateneo fiorentino, di cui sono noti:

- voto maturità;
- punteggio test ingresso;
- numero esami;
- voto medio esami.

È vietata la riproduzione non autorizzata a fini commerciali.

## Esempio:

```
l=lm(voto_medio_esami~votomat+punteggio, data=dati)
```

```
summary(l)
```

Call:

```
lm(formula = voto_medio_esami ~ votomat + punteggio, data = dati)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.6259 -0.8082 -0.0883  0.8456  3.1846
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.55721    2.65420    5.485 1.07e-05 ***
votomat      0.07015    0.02897    2.422 0.02302 *
punteggio    0.31892    0.08694    3.668 0.00116 **
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.521 on 25 degrees of freedom

Multiple R-squared: 0.4387, Adjusted R-squared: 0.3938

F-statistic: 9.769 on 2 and 25 DF, p-value: 0.0007331

È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza per l'insieme complessivo delle variabili esplicative

Le variabili esplicative hanno **nel loro complesso** un effetto statisticamente significativo sulla variabile dipendente?

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1 : \text{almeno un } \beta_i \neq 0 \end{cases} = \begin{cases} H_0 : \rho_{\hat{Y}, Y} = 0 \\ H_1 : \rho_{\hat{Y}, Y} > 0 \end{cases}$$

Sotto  $H_0$ :

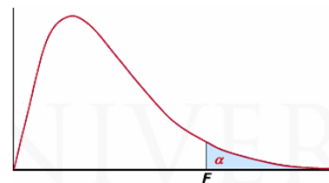
$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} \sim F_{k, n-(k+1)}$$

**F di Fisher**

È vietata la riproduzione non autorizzata a fini commerciali.

## La distribuzione F di Fisher

- ❑  $F$  assume solo valori non negativi;
- ❑ è asimmetrica a destra;
- ❑ la sua forma esatta dipende da 2 parametri:
  - $gdl_1 = k$  (numero di variabili esplicative nel modello)
  - $gdl_2 = n - (k + 1)$
- ❑ la sua media è approssimativamente uguale ad 1 ➡  $E(F) = \frac{gdl_2}{gdl_2 - 2}$
- ❑ ➡ **grandi valori della statistica test forniscono evidenza contro  $H_0$**
- ❑ le tavole della  $F$  elencano gli  $F$ -score che hanno, sulla coda destra della distribuzione, i  $p$ -value di 0.05, 0.01, 0.001 in relazione a diverse combinazioni di  $gdl_1$  e  $gdl_2$  (una tavola per ciascun livello di probabilità).



È vietata la riproduzione non autorizzata a fini commerciali.

## Inferenza complessiva VS singole variabili

### NB:

In presenza di multicollinearità in un modello con un elevato numero di predittori, è possibile che nessuno (o pochi) di essi evidenzino stime dei coefficienti di regressione parziale statisticamente diversi da zero.

Ciò nonostante è possibile che si possa osservare un  $R^2$  elevato, quindi un elevato valore per la statistica  $F$  nel test complessivo per tutti i  $\beta$ .

## Interazione tra predittori

Si parla di **interazione tra due variabili** se la relazione tra due variabili cambia al cambiare dei valori di una terza variabile (cfr. ultima diapositiva della Lezione 10).

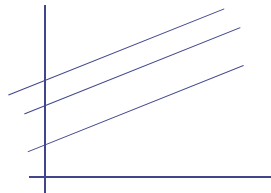
Quando il modello additivo è troppo semplicistico per risultare adeguato, è conveniente verificare la sussistenza di una qualche interazione tra i predittori considerati.

$$\begin{array}{l}
 y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\
 \Downarrow \\
 y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i
 \end{array}
 \left. \vphantom{\begin{array}{l} \\ \\ \end{array}} \right\} \begin{array}{l} \text{aggiunta di} \\ X_3 = X_1 X_2 \end{array}$$



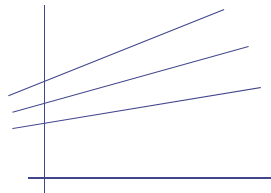
## Interazione tra predittori

In caso di **assenza di interazione** tra due variabili:



controllando per  $X_2$   
 si ottengono **rette parallele**  
 (ovvero con intercetta differente).

In caso di **presenza di interazione** tra due variabili:



controllando per  $X_2$   
 si ottengono **rette di pendenza ed intercetta differenti**.

È vietata la riproduzione non autorizzata a fini commerciali.

## Modelli a confronto

**Modello completo:** modello con tutti i predittori (comprese eventuali interazioni);

**Modello ridotto:** modello solo con alcuni di questi.

Quest'ultimo si dice nidificato all'interno del modello completo.

esempio:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon$$

VS

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

È vietata la riproduzione non autorizzata a fini commerciali.

### Modelli a confronto

Un test di confronto tra il modello completo e quello ridotto, nell'esempio precedente è:

$$\begin{cases} H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \\ H_1 : \text{almeno un } \beta_i \neq 0 \quad i = 4, 5, 6 \end{cases}$$

$$\Rightarrow F = \frac{(SSE_{rid} - SSE_{comp}) / gdl_1}{SSE_{comp} / gdl_2} = \frac{(R_{comp}^2 - R_{rid}^2) / gdl_1}{(1 - R_{comp}^2) / gdl_2}$$

dove:

$gdl_1$  = numero dei termini aggiuntivi per passare dal ridotto al completo;

$gdl_2$  =  $gdl$  per il modello completo  $gdl_2 = n - (k + 1)$

### Modelli a confronto

Una riduzione relativamente elevata del termine d'errore nel passaggio dal ridotto al completo porta ad un elevato valore della statistica  $F$  e ad un piccolo  $p$ -value.

Quindi ad un'evidenza contro  $H_0$  che induce ad optare per il modello superiore.

## Il test F: quadro sinottico (1/3)

One-way ANOVA ( $G$  gruppi):

$$F = \frac{\text{Between - groups var.}}{\text{Within - groups var.}}$$

$$= \frac{\sum_g^G n_g (\bar{y}_g - \bar{y})^2 / (G-1)}{\sum_g^G \sum_i^{n_g} (y_{ig} - \bar{y}_g)^2 / (n-G)} \sim F_{G-1, n-G}$$

NB: le  $G$  medie sono vincolate dalla media generale, solo  $G-1$  valori sono liberi.

$G$  gruppi  $\Rightarrow$   $G$  medie stimate, oltre alla media generale.  
 Si decompone la devianza TOT in devianza  $W + B$ .  
 Si respinge  $H_0$  se almeno una media di gruppo è statisticamente diversa dalle altre.  
 In tal caso la variabilità misurata tra le  $G$  medie dei vari gruppi è sensibilmente maggiore della variabilità interna ai gruppi.

È vietata la riproduzione non autorizzata a fini commerciali.

## Il test F: quadro sinottico (2/3)

Test sul modello di regressione:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{\frac{(TSS - SSE) 1}{TSS k}}{\frac{SSE 1}{TSS [n-(k+1)]}}$$

$$= \frac{(TSS - SSE)/k}{SSE/[n-(k+1)]} \sim F_{k, n-(k+1)}$$

Modello a  $k$  variabili  $\Rightarrow$   $(k+1)$  parametri (analogia con i  $G$  gruppi slide precedente).  
 Si decompone la devianza TOT delle  $Y$  in devianza spiegata dal modello + devianza attribuibile all'errore.  
 Si respinge  $H_0$  se almeno un coefficiente di regressione è statisticamente diverso da zero (OVVERO se l'iperpiano di regressione non è perfettamente orizzontale).

È vietata la riproduzione non autorizzata a fini commerciali.

## Il test F: quadro sinottico (3/3)

Test sul modello di regressione (modelli a confronto):

$$F = \frac{(SSE_{rid} - SSE_{comp}) / gdl_1}{SSE_{comp} / [n - (k + 1)]} \sim F_{gdl_1, n - (k + 1)}$$

Da un modello **ridotto** con  $(k - gdl_1)$  variabili si passa ad un modello **completo** con  $k$  variabili (ovvero  $k + 1$  parametri).

Si decompone la **devianza dell'errore del modello ridotto** ( $SSE_{rid}$ ) in **devianza dell'errore del modello completo** ( $SSE_{comp}$  che è più piccola, o al limite uguale, alla precedente per effetto della presenza di un numero superiore di variabili esplicative) + **termine di riduzione dell'SSE** nel passaggio dal ridotto al completo.

Si respinge  $H_0$  se **almeno** un coefficiente di regressione delle variabili esplicative aggiunte al modello ridotto è statisticamente diverso da zero.

**NB:** se si considerasse come caso limite di modello ridotto un modello a sola intercetta, si otterrebbe la statistica test illustrata nella slide precedente.

È vietata la riproduzione non autorizzata a fini commerciali.

## Correlazione parziale

I modelli di regressione multipla descrivono l'effetto di una variabile esplicativa sulla variabile risposta tenendo sotto controllo gli altri predittori.

**Come stabilire la forza di queste associazioni parziali?**



**Coefficiente di correlazione parziale**  
 tra  $Y$  e  $X_2$ , controllando per  $X_1$

$$r_{YX_2 \cdot X_1} = \frac{r_{Y, X_2} - r_{Y, X_1} r_{X_1, X_2}}{\sqrt{(1 - r_{Y, X_1}^2)(1 - r_{X_1, X_2}^2)}}$$

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione parziale

### Proprietà:

- $-1 \leq r_{YX_2 \cdot X_1} \leq 1$
- più grande è il suo valore assoluto, maggiore è l'associazione tra la  $Y$  e  $X_2$ , tenute sotto controllo le altre predittive;
- non dipende dall'unità di misura;
- ha lo stesso segno del coefficiente di regressione parziale  $\beta_2$ .

Tutto quanto detto vale ovviamente anche per  $r_{YX_1 \cdot X_2}$ .

È vietata la riproduzione non autorizzata a fini commerciali.

## Il coefficiente di correlazione parziale

Il **quadrato della correlazione parziale** è interpretabile in termini di riduzione proporzionale dell'errore (PRE).

### Correlazione parziale al quadrato

$$r_{YX_2 \cdot X_1}^2 = \frac{R^2 - r_{Y, X_1}^2}{1 - r_{Y, X_1}^2}$$

Il coefficiente evidenzia la **proporzione della variabilità residuale di  $Y$**  (ovvero non spiegata da  $X_1$ ) spiegata solo da  $X_2$ .

È vietata la riproduzione non autorizzata a fini commerciali.