Capitolo aggiuntivo 12

Dati statistici e scale di misura

La statistica è un insieme di metodi e tecniche per:

- raccogliere *informazioni* su un fenomeno (ad esempio i risultati di un esperimento di laboratorio)
- *sintetizzare* l'informazione (elaborare i dati)
- generalizzare i risultati ottenuti (ad esempio determinare un modello matematico che permetta di prevedere i risultati dell'esperimento)

Perché è importante?

- In generale, perché consente di valutare criticamente tutte le informazioni basate su rilevazioni e sondaggi (ad esempio: "il prodotto X è particolarmente efficace contro il raffreddore; infatti su 100 pazienti trattati, ben 95, pari quindi al 95% dei casi, ha mostrato completa remissione della malattia entro una settimana di cura" è un'affermazione priva di significato, perché di solito (nel 95% dei casi, appunto!) in una settimana il raffreddore passa anche senza nessun trattamento. Oppure: "il 75% delle persone intervistate si è dichiarata favorevole al tal partito politico" non costituisce un'informazione attendibile, in quanto non viene indicata la dimensione né la composizione del campione sottoposto all'intervista; se si tratta di 4 persone di cui 3 sono favorevoli e una no, non è possibile utilizzare questo dato ai fini di una previsione.

- In particolare, perché è elemento essenziale nell'applicazione del metodo scientifico (che consiste nel formulare un modello matematico per via induttiva di un dato problema; il modello viene quindi verificato deduttivamente mediante un esperimento; l'analisi dei dati ottenuti consente di validare il modello oppure formularne eventualmente un secondo).

In *ambito biologico* l'uso di appropriati metodi statistici consente di

- pianificare in modo adeguato la sperimentazione
- tenere sotto controllo l'errore sperimentale
- valutare l'affidabilità dei risultati ottenuti

Una prima distinzione: generalmente si parla di

- *statistica descrittiva*: insieme di metodi e tecniche per l'esplorazione e la sintesi dell'evidenza empirica (dati)
- *inferenza statistica*: insieme di metodi e principi per inferire le caratteristiche generali di un fenomeno mediante l'osservazione di un insieme limitato di manifestazioni dello stesso

Le due categorie differiscono principalmente per gli obiettivi che l'analisi dei dati si pone.

Dati statistici

Alcune definizioni

Definizione 12.1. *Si chiama* popolazione statistica *l'insieme di tutti gli elementi (individui, geni, cellule, ecc...) che si vogliono studiare.*

Una popolazione è dunque un insieme composto da un numero finito o infinito di elementi.

Esempio 12.2. L'insieme dei lupi del Parco Nazionale d'Abruzzo, degli abitanti di Milano, dei valori di temperatura rilevati a Roma alle ore 14 dal 1/6/1998 al 31/5/2001, l'altezza degli alunni di una classe di 30 bambini di 6 anni, sono alcuni esempi di popolazioni composte da un numero finito di elementi.

Gli esseri umani sulla terra, le larghezze delle corolle dei fiori, gli atomi o le molecole di un gas, sono elementi di insiemi finiti, ma il loro numero è così grande che, a volte, sarà utile considerare la popolazione come infinita.

Anche se l'oggetto di un indagine è un'intera popolazione statistica, in genere non è possibile raccogliere dati su ogni singolo elemento della popolazione, ma si può fare solo per un sottoinsieme, che si suppone sia *rappresentativo* dell'intera popolazione; questo sottoinsieme prende il nome di *campione*.

Definizione 12.3. *Si chiama* campione *qualunque sottoinsieme della popolazione, selezionato in modo opportuno.*

Definizione 12.4. *Fissata una popolazione, si chiamano* variabili statistiche *tutte quelle caratteristiche che variano al variare dei componenti della popolazione.*

Esempio 12.5. Il colore bianco, fulvo, nero, ecc..., della pelliccia degli esemplari di una certa specie, il sesso (maschio o femmina), sono esempi di variabili statistiche qualitative (dette anche *attributi*). L'età in mesi degli esemplari di lupo del Parco degli Abruzzi, oppure il numero di cuccioli nati da ogni femmina, sono variabili quantitative *discrete*. La temperatura di Roma rilevata alle ore 14 del primo Giugno di ogni anno è una variabile quantitativa *continua*. In generale le variabili discrete possono assumere solo un numero finito o una infinità numerabile di valori mentre quelle continue possono assumere tutti i valori compresi in un intervallo.

Matematicamente le variabili statistiche sono funzioni che si indicano generalmente con la lettera X (talvolta si usano anche Y e Z) che hanno come dominio il campione studiato e come codominio i valori osservati, detti (*determinazioni, valori o modalità*) e si indicano con lettere minuscole x, y, z.

Scale di misura

Risultati statistici affidabili si basano sempre su dati ottenuti dalle osservazioni o dalle misure, sia di un campione che di tutta la popolazione, ordinati e riassunti in poche informazioni semplici. Queste informazioni sono dette *empiriche*, per distinguerle da quelle *teoriche*, offerte da un eventuale modello che spieghi i dati osservati.

Le scale di misura più comunemente usate sono

$$\text{var. qualitative} \begin{cases} \text{nominale} \\ \text{ordinale} \end{cases} \qquad \text{var. quantitative} \begin{cases} \text{di intervallo} \\ \text{di rapporto} \end{cases}$$

Scala nominale. Se una variabile è misurata su scala nominale, si possono instaurare solo le seguenti relazioni tra le modalità

$$x_i = x_j \text{ oppure } x_i \neq x_j$$

Esempi: genere, gruppo sanguigno, sopravvivenza.

Scala ordinale. Se una variabile è misurata su scala ordinale, si possono instaurare le seguenti relazioni tra le modalità

$$x_i \leq x_j \text{ oppure } x_i \geq x_j$$

Le modalità della variabile possono quindi essere ordinate.

Esempi: titolo di studio, grado di soddisfazione, lunghezze. Ad esempio il giudizio sull'effetto di un fitofarmaco può essere espresso secondo la scala seguente:

1. peggioramento;
2. nessuna variazione;
3. lieve miglioramento;
4. deciso miglioramento;
5. guarigione.

Scala di intervallo. Si misurano così le variabili quantitative per le quali lo zero è *convenzionale* (arbitrario). In tal caso non ha senso rapportare le misure ottenute, ed è invece corretto confrontare per differenze. Esempio tipico: temperatura

Esempio 12.6. *In tre giorni diversi sono state rilevate le seguenti temperature:*

Giorno	*T °C*	*Diff. °C*	*T °F*	*Diff. °F*
1	6		42, 8	
		3		5, 4
2	9		48, 2	
		6		10, 8
3	15		59	

La variazione tra il secondo ed il terzo giorno è doppia di quella tra il primo ed il secondo, indipendentemente dalla scala utilizzata.

Scala di rapporto. Si misurano così le variabili quantitative per le quali lo zero è naturale. Esempio: peso, concentrazione, lunghezza. In questo caso le modalità possono essere confrontate per rapporto. Per esempio, si può affermare che la concentrazione di atrazina in un campione d'acqua è doppia rispetto a quella in un altro campione. Oppure, considerare il peso specifico di un oggetto significa considerare il rapporto tra il peso dell'oggetto e quello di un equivalente volume di acqua a $4°C$.

Capitolo aggiuntivo 13

Statistica descrittiva

Insieme di metodi e tecniche per sintetizzare l'informazione contenuta nei dati. Gli strumenti di sintesi sono essenzialmente di tre tipi:

- tabelle
- rappresentazioni grafiche
- indici sintetici

Attenzione! Quando sintetizziamo l'informazione contenuta nei dati, ne perdiamo una parte. Gli strumenti di sintesi devono essere scelti in modo tale da:

- preservare, per quanto possibile, l'informazione rilevante per il problema analizzato
- eliminare l'informazione non necessaria

Distribuzioni di frequenza

La **frequenza** misura quante volte una certa modalità è stata osservata nel campione studiato.

Tipica rappresentazione tabellare per variabili qualitative o per variabili quantitative discrete. Nella tabella sono riportate:

- le *modalità* della variabile
- le *frequenze* associate a ciascuna modalità

Frequenza assoluta

Definizione 13.1. *Sia C un campione di una popolazione Ω composto da N elementi, $M = \{x_1, \ldots, x_k\}$ un insieme finito di modalità e $X : C \to M$ una variabile statistica (ovviamente discreta). Si chiama* frequenza assoluta *della modalità x_i il numero*

$$n_i = \#\{c \in C \ : \ X(c) = x_i\} = \#X^{-1}(x_i), \quad i = 1, 2, \ldots, k.$$

Si chiama frequenza relativa *il rapporto*

$$p_i = \frac{n_i}{N}(\times 100), \quad i = 1, 2, \ldots, k.$$

In pratica, la frequenza assoluta misura quante volte una certa modalità è stata osservata nel campione studiato, mentre la frequenza relativa rappresenta la proporzione (talvolta in percentuale) di osservazioni che presentano una certa modalità della variabile analizzata.

Osservazione 13.2. Chiaramente si ha

$$\sum_{i=1}^{k} n_i = N \qquad \text{e} \qquad \sum_{i=1}^{k} p_i = 1.$$

Esempio 13.3. Su 50 soggetti è stato rilevato il gruppo sanguigno. I risultati sono stati riportati nella tabella seguente

Gruppo	n_i	p_i
A	20	$0,40$
B	5	$0,10$
AB	2	$0,04$
0	23	$0,46$
$Tot.$	50	$1,00$

Frequenza cumulata

Definizione 13.4. *Si chiama* frequenza cumulata assoluta *della modalità x_i il numero*

$$N_i = \#\{c \in C \ : \ X(c) \leq x_i\} = \#X^{-1}(]-\infty, x_i]), \quad i = 1, 2, \ldots, k.$$

Si chiama frequenza cumulata relativa *il rapporto*

$$P_i = \frac{N_i}{N}(\times 100), \quad i = 1, 2, \ldots, k.$$

La frequenza cumulata assoluta (relativa) associata ad una modalità della variabile indica il numero (la proporzione) di osservazioni che presentano un valore minore o uguale rispetto a quello della modalità. Si possono utilizzare solo se il carattere è misurato almeno su scala ordinale

Esempio 13.5. Nella tabella seguente è riportata la distribuzione dei giudizi all'esame di licenza media rilevati su un gruppo di studenti

Giudizio	n_i	p_i	N_i	P_i
$Suff.$	8	$0,1111$	8	$0,1111$
$Buono$	29	$0,4028$	37	$0,5139$
$Distinto$	30	$0,4167$	67	$0,9306$
$Ottimo$	5	$0,0694$	72	$1,0000$
$Tot.$	72	$1,00$		

Esempio 13.6. Numero di pizze difettose (troppo grandi) prodotte da una pressa in un'ora (6 giorni di osservazione)

Giorno	n_i	p_i	N_i	P_i
1	4	0.10	4	0.10
2	10	0.25	14	0.35
3	12	0.30	26	0.65
4	6	0.15	32	0.80
5	4	0.10	36	0.90
6	4	0.10	40	1.00
$Tot.$	40	1.00		

Vantaggi e svantaggi delle distribuzioni di frequenza:

+ Non si perde informazione rilevante (solo l'ordinamento va perduto)

- Scarso potere di sintesi se le modalità sono numerose

- Non utilizzabile per variabili continue.

In realtà l'ultimo punto non è del tutto vero ...

Distribuzione di frequenza per variabili continue

Se siamo disposti a rinunciare ad ulteriore informazione, la distribuzione di frequenza può essere costruita anche per variabili continue. Generalmente si opera nel modo seguente:

- si suddivide l'insieme dei valori che la variabile può assumere in intervalli, detti *classi*;

- si determina il numero di osservazioni che cadono all'interno di ciascuna classe.

Esempio 13.7. Aziende agricole secondo la superficie agricola totale. Provincia di Udine.

Superficie	n_i	p_i
$0 \dashv 1$	2406	0.085
$1 \dashv 2$	3404	0.120
$2 \dashv 3$	2857	0.101
$3 \dashv 5$	4415	0.155
$5 \dashv 10$	6856	0.241
$10 \dashv 20$	5708	0.201
$20 \dashv 30$	1365	0.048
$30 \dashv 50$	751	0.026
$50 \dashv 100$	410	0.014
> 100	238	0.008
Totale	28410	1.000

Esempio 13.8. 100 piante da fiore classificate in base alla larghezza della corolla

$x_i \dashv x_{i+1}$	n_i	p_i	N_i	P_i
$59,5 \dashv 62,5$	5	$0,05$	5	$0,05$
$62,5 \dashv 65,5$	18	$0,18$	23	$0,23$
$65,5 \dashv 68,5$	42	$0,42$	65	$0,65$
$68,5 \dashv 71,5$	27	$0,27$	92	$0,92$
$71,5 \dashv 74,5$	8	$0,08$	100	$1,00$

Come costruire le classi? Non esistono regole assolute per la costruzione delle classi. In generale è buona norma:

- evitare di costruire classi con frequenze molto basse;

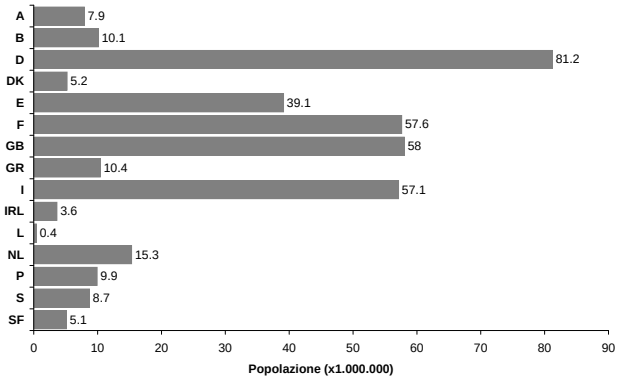
- modulare l'ampiezza delle classi in funzione della disponibilità di informazione "locale";

- *se possibile*, non variare l'ampiezza di classe (semplifica l'interpretazione).

Rappresentazioni grafiche

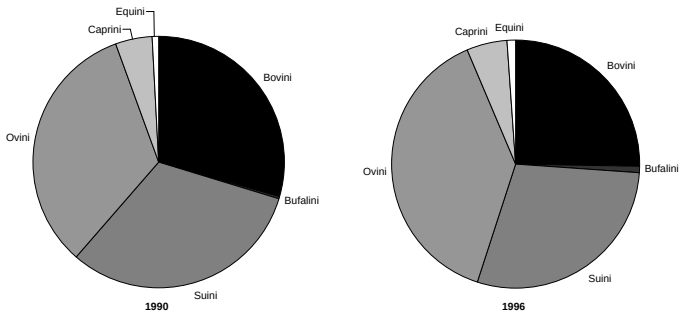
Strumenti molto utili per visualizzare le caratteristiche di una variabile. Ne esistono di svariati tipi, a seconda delle esigenze di analisi. Alcuni riproducono le stesse informazioni di una distribuzione di frequenza, altri riassumono caratteristiche difficilmente rappresentabili mediante tabelle.

Diagramma a barre - Popolazione Paesi UE 1993



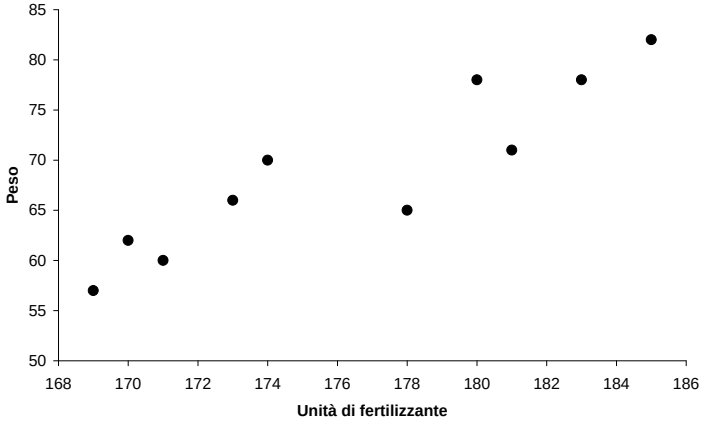
15

Diagramma a torta - Bestiame da allevamento per specie



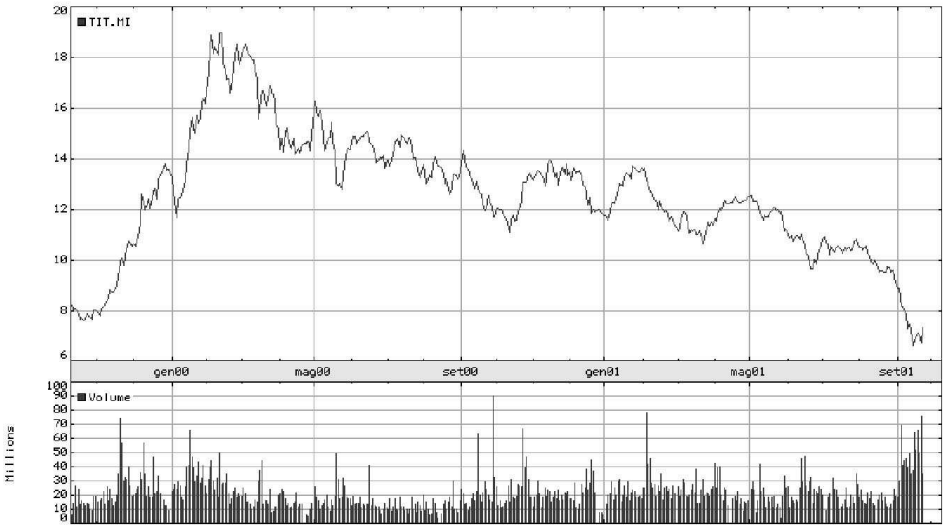
16

Diagramma di dispersione - Relazione dose-risposta



17

Serie storica



Grafici per variabili continue

Come rappresentare la distribuzione di frequenza di una variabile continua? Se le classi sono di ampiezza diversa, le frequenze *non sono* direttamente *confrontabili*. Per costruire un grafico che rappresenti in modo adeguato l'informazione è necessario eliminare l'effetto dell'ampiezza di classe.

Densità di frequenza

Il rapporto tra la frequenza e l'ampiezza (indicata con Δ_i) di una classe è detto *densità di frequenza.*

$$d_i = \frac{p_i}{\Delta_i}$$

Le densità di frequenza *sono* fra loro *confrontabili*. La densità di frequenza è assoluta o relativa a seconda del tipo di frequenza utilizzato nel calcolo.

Istogramma di frequenza

In un istogramma di frequenza ad ogni classe è associato un rettangolo:

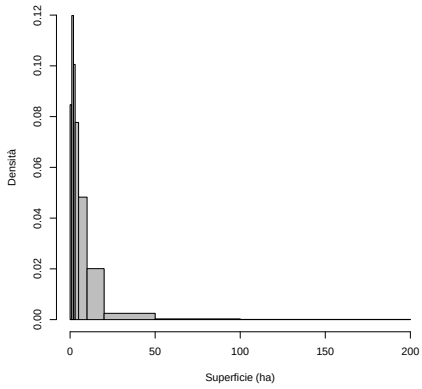
- la base del rettangolo è pari all'ampiezza di classe;
- l'altezza del rettangolo è pari alla densità di frequenza;
- l'area del rettangolo è per costruzione la frequenza (assoluta o relativa) associata alla classe;

Distribuzione delle aziende agricole per superficie agricola

Sup.	n_i	p_i	Δ_i	d_i
0 ⊣ 1	2406	0.085	1	0.08500
1 ⊣ 2	3404	0.120	1	0.12000
2 ⊣ 3	2857	0.101	1	0.10100
3 ⊣ 5	4415	0.155	2	0.07750
5 ⊣ 10	6856	0.241	5	0.04820
10 ⊣ 20	5708	0.201	10	0.02010
20 ⊣ 30	1365	0.048	10	0.00480
30 ⊣ 50	751	0.026	20	0.00130
50 ⊣ 100	410	0.014	50	0.00028
100+	238	0.008	100	0.00008
Totale	28410	1.000		

22

Istogramma - Aziende agricole per superficie agricola



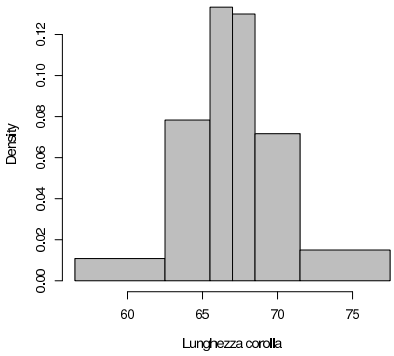
23

Distribuzione di piante in base alla lunghezza della corolla

$x_i \dashv x_{i+1}$	n_i	p_i	Δ_i	d_i
56,5 ⊣ 62,5	13	0,065	6,0	0,0108
62,5 ⊣ 65,5	47	0,235	3,0	0,0783
65,5 ⊣ 67,0	40	0,200	1,5	0,1333
67,0 ⊣ 68,5	39	0,195	1,5	0,1300
68,5 ⊣ 71,5	43	0,215	3,0	0,0717
71,5 ⊣ 77,5	18	0,090	6,0	0,0150

24

Istogramma - Piante in base alla lunghezza della corolla



25

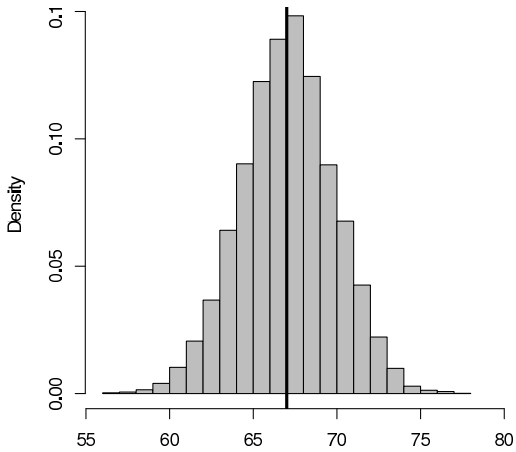
Caratteristiche dell'istogramma

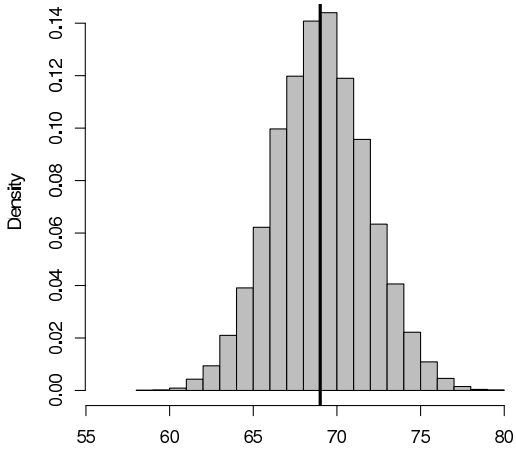
Da un istogramma è possibile desumere alcune rilevanti caratteristiche del fenomeno, per esempio:

- tendenza centrale
- dispersione
- grado di simmetria della distribuzione

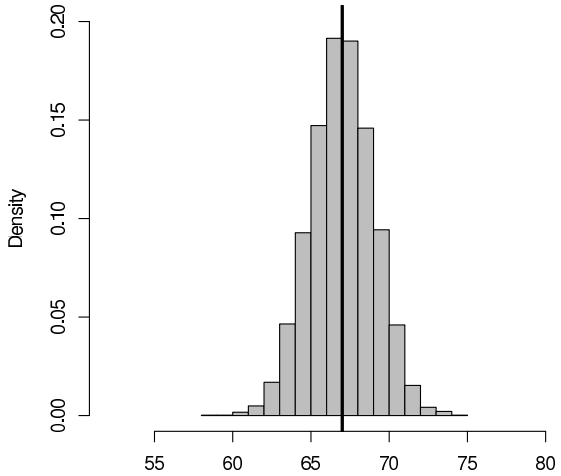
Illustriamo queste caratteristiche in alcuni esempi.

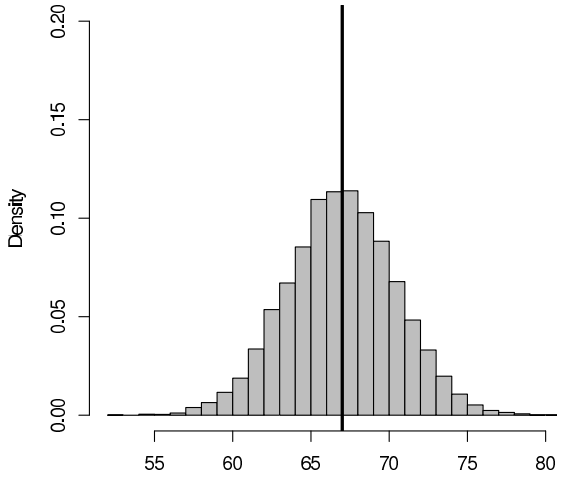
La tendenza centrale



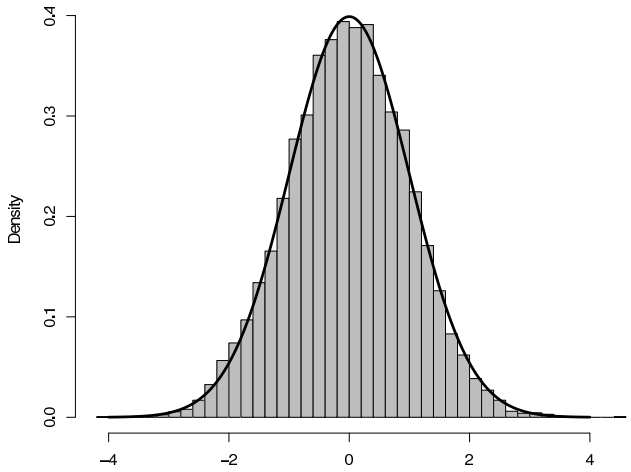


Il grado di dispersione



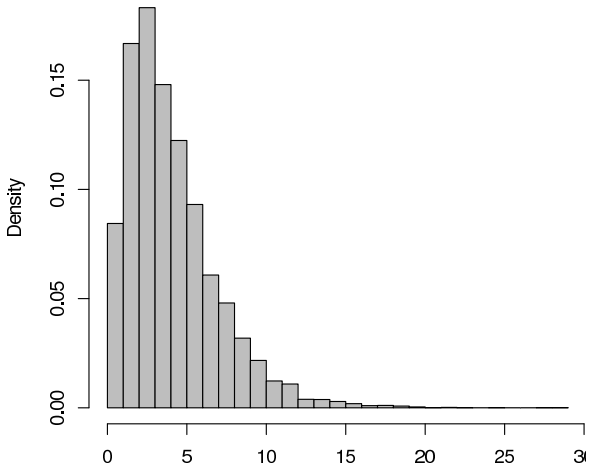


Simmetria ...



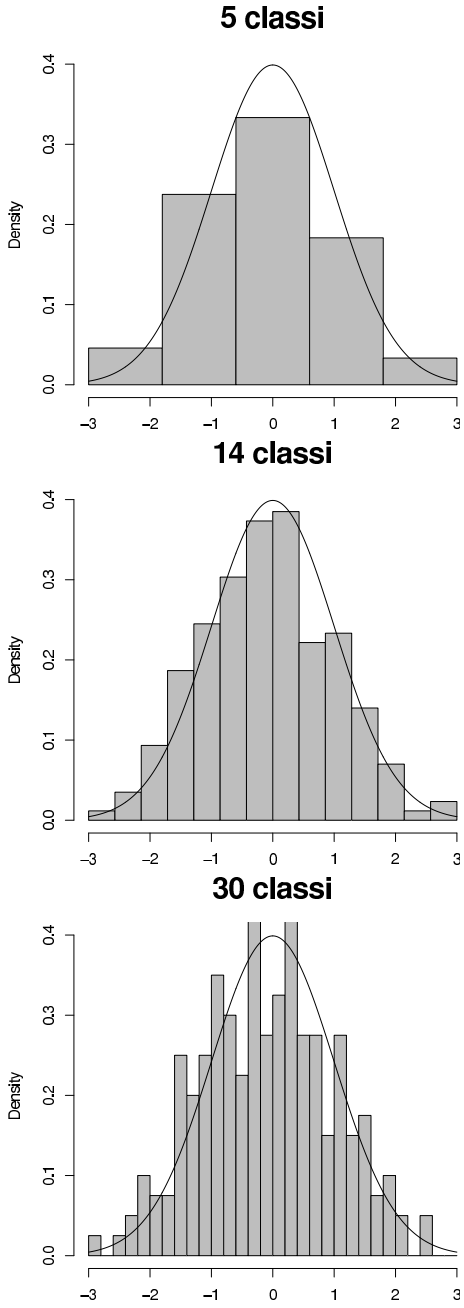
29

e asimmetria di una distribuzione



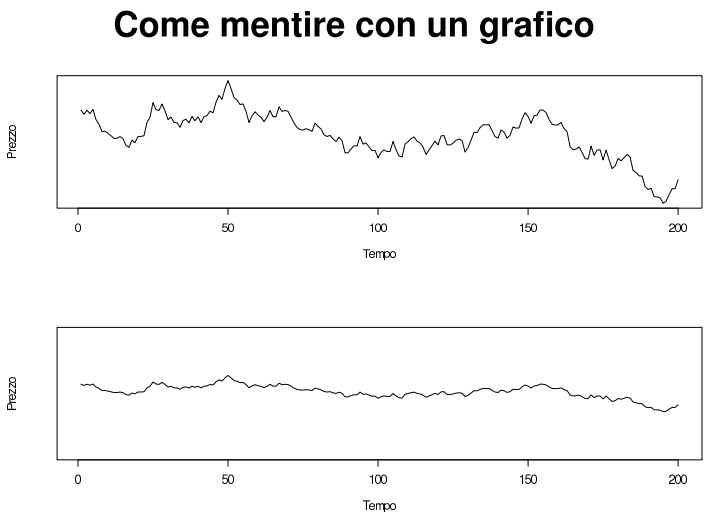
30

L'effetto dell'ampiezza di classe

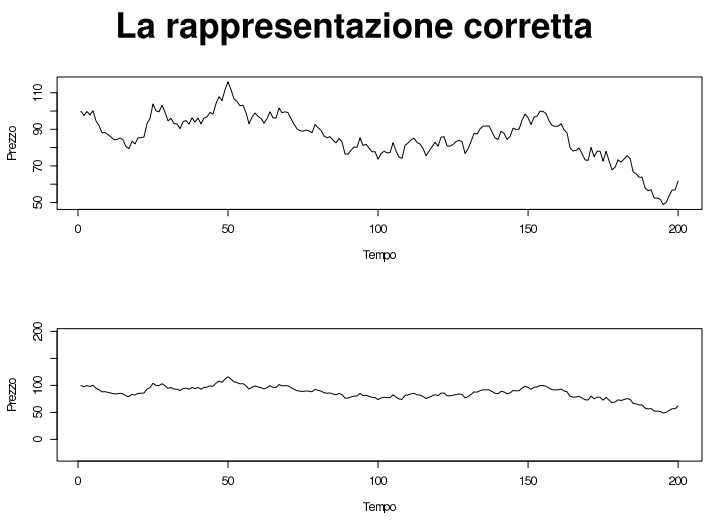


Vantaggi e svantaggi delle rappresentazioni grafiche

+ Conservano la maggior parte dell'informazione contenuta nei dati
+ Sono di immediata comprensione
- Nonostante la (presunta) semplicità, non sempre è chiaro quale sia la rappresentazione da utilizzare
- Possono essere usati in modo strumentale



33



34

Gli indici sintetici

Le caratteristiche più rilevanti di una distribuzione, per esempio

- la tendenza centrale del fenomeno
- il grado di dispersione
- la simmetria

possono essere rappresentate mediante numeri, detti *indici sintetici*.

Indici di posizione

Gli indici di posizione servono per individuare la tendenza centrale del fenomeno studiato. I più utilizzati sono:

- moda
- mediana
- media aritmetica

Ci riferiremo nel seguito sempre ad iuna variabile statistica $X : C \to M$ dove il campione C è composto da N elementi, cioè $C = \{c_1, c_2, \ldots, c_N\}$ mentre $M = \{x_1, \ldots, x_k\}$. Indicheremo come al solito con n_i la frequenze assoluta della modalità x_i e con p_i la frequenza relativa.

Moda

La *moda* di una distribuzione è la modalità più frequente (prevalente). Qualora si utilizzi una distribuzione in classi per variabili continue, la *classe modale* è quella con la *densità* di frequenza più elevata. Può essere utilizzata per qualunque tipo di variabile, ma è poco informativa.

Esempio 13.9 (continua)**.**

Gruppo	n_i	p_i
A	20	$0,40$
B	5	$0,10$
AB	2	$0,04$
0	23	$0,46$
Tot.	50	$1,00$

La moda (Mo) è il gruppo sanguigno 0.

Mediana

La *mediana* è il valore che occupa la posizione centrale nella distribuzione, tale che:

- metà delle osservazioni sono uguali o minori
- metà delle osservazioni sono uguali o superiori

La mediana divide in due parti di egual numero l'insieme dei valori osservati. Si può utilizzare solo per variabili misurate almeno su scala ordinale.

Calcolo della mediana. Per calcolare la mediana bisogna:

1. *ordinare* gli N valori osservati in ordine crescente
2. prendere il valore *centrale* nella graduatoria ordinata

Il modo di procedere per il secondo punto varia a seconda della numerosità del campione studiato.

Se N è dispari, esiste un unico valore che divide esattamente in due la distribuzione. Il valore centrale occupa la posizione

$$\frac{N+1}{2}$$

nella graduatoria ordinata. In tal caso

$$\text{Me} = X(\frac{N+1}{2}).$$

Se N è pari, si considerano valori centrali quelli che occupano le posizioni

$$\frac{N}{2} \text{ e } \frac{N}{2}+1$$

Esistono quindi due mediane

$$\text{Me}_1 = X(\frac{N}{2}) \text{ e } \text{Me}_2 = X(\frac{N}{2}+1)$$

Quando possibile (variabili *quantitative*) si usa come mediana la semisomma dei valori centrali

$$\text{Me} = \frac{X(\frac{N}{2}) + X(\frac{N}{2}+1)}{2}$$

Esempio 13.10 (di calcolo)**.** Nella tabella seguente sono riportati i giudizi (A, B, C o D) ottenuti ad un esame da 9 studenti.

Studente	1	2	3	4	5	6	7	8	9
Giudizio	*B*	*D*	*A*	*C*	*B*	*A*	*D*	*C*	*A*

Dovremo quindi ordinare i valori e scegliere come mediana quello che occupa la 5^a posizione

Posizione	1	2	3	4	5	6	7	8	9
Giudizio	*D*	*D*	*C*	*C*	$\boxed{B}$	*B*	*A*	*A*	*A*

Nel caso i valori osservati siano 10 (una D in più rispetto all'esempio precedente)

Posizione	1	2	3	4	5	6	7	8	9	10
Giudizio	*D*	*D*	*D*	*C*	$\boxed{C}$	$\boxed{B}$	*B*	*A*	*A*	*A*

bisogna considerare la 5^a e la 6^a posizione

$$\mathrm{Me}_1 = C, \quad \mathrm{Me}_2 = B$$

Calcolo su distribuzioni di frequenza

Qualora sia disponibile la distribuzione di frequenza cumulata, la mediana (classe mediana) corrisponde alla modalità (classe) associata alla prima frequenza cumulata relativa superiore al 50%.

Giudizio	n_i	p_i	N_i	P_i
Suff.	8	0,1111	8	0,1111
Buono	29	0,4028	37	0,5139
Distinto	30	0,4167	67	0,9306
Ottimo	5	0,0694	72	1,0000
Tot.	72	1,0000		

La mediana della distribuzione è "Buono".

Pregi e difetti della mediana

+ è un buon indicatore della tendenza centrale
+ risente poco di ciò che accade sulle code della distribuzione (è *robusta*)
- è difficile da trattare analiticamente

La media aritmetica

La *media aritmetica* è il più importante indice di posizione. La formula per il calcolo della media è:

$$\left.\begin{matrix}\bar{X}\\ \mu\\ M(X)\end{matrix}\right\} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N} = \frac{1}{N}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i p_i$$

ossia la media è la somma dei valori osservati divisa per la numerosità del campione. Si può utilizzare *solo* per variabili *quantitative*.

Nel caso particolare $k = N$ (cioè $n_i = 1$ per ogni i) si ha

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Proprietà della media aritmetica

- La media aritmetica è sempre compresa tra il minimo ed il massimo dei valori osservati (internalità)

$$x_{\min} \leq \bar{X} \leq x_{\max},$$

- la somma degli scarti dalla media è sempre pari a zero

$$\sum_{i=1}^{k} (x_i - \bar{X}) n_i = 0,$$

- è *lineare*, cioè se X e Y sono variabili statistiche e $a, b \in \mathbb{R}$ allora si ha $M(aX + bY) = aM(X) + bM(Y)$.

Pregi e difetti della media

\+ è un buon indicatore della tendenza centrale

\+ è semplice da trattare analiticamente

\- risente in misura rilevante di ciò che accade sulle code della distribuzione (è *poco robusta*)

Indici di dispersione o di variabilità

Variabilità: in quale misura i valori osservati differiscono tra loro
Dispersione: in quale misura i valori osservati differiscono da un valore di riferimento

In realtà i due concetti sono (almeno parzialmente) sovrapponibili e noi useremo i due termini come sinonimi.

Campo di variazione (range)

Il *campo di variazione o range o oscillazione* è la differenza tra il massimo ed il minimo valore osservati:

$$R = \max X - \min X = x_{\max} - x_{\min}$$

Il campo di variazione è poco usato perché:

- trascura la maggior parte dell'informazione disponibile
- risente eccessivamente dei valori estremi

Scarto interquartile

Per eliminare il problema dei valori estremi, talvolta si usa lo *scarto interquartile*, ossia la differenza tra il terzo ed il primo quartile.

Primo quartile: lascia alla sua sinistra il 25% delle osservazioni
Terzo quartile: Lascia alla sua sinistra il 75% delle osservazioni
Rimane inalterato il problema dello scarso sfruttamento dell'informazione

Come sfruttare tutta l'informazione?

Gli indici visti in precedenza sono poco informativi. È possibile costruire un indice che sfrutti al meglio il contenuto informativo dei dati? Il grado di dispersione delle singole osservazioni è misurato dagli scarti

$$x_i - \bar{X}$$

Un buon indice di dispersione deve essere una sintesi di queste quantità.

Devianza

La devianza è la somma degli scarti dalla media al quadrato

$$\mathrm{Dev}(X) = \sum_{i=1}^{k} |x_i - \bar{x}|^2 n_i$$

- Elevando al quadrato, trascuriamo il segno degli scarti
- La devianza dipende dalla numerosità del campione
- L'unità di misura è il quadrato di quella della variabile

Varianza

La varianza si usa per eliminare l'effetto della numerosità del campione. Si può calcolare in due modi, usando

- la numerosità del campione (*varianza campionaria*)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{k} |x_i - \bar{X}|^2 n_i = \sum_{i=1}^{k} |x_i - \bar{X}|^2 p_i$$

- i gradi di libertà (*varianza campionaria corretta*)

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{k} |x_i - \bar{X}|^2 n_i = \frac{N}{N-1} \sum_{i=1}^{k} |x_i - \bar{X}|^2 p_i$$

Gradi di libertà

Poiché la somma degli scarti dalla media è necessariamente uguale a zero, fissata la media solo $N-1$ scarti sono liberi di variare (ossia di assumere un qualunque valore). Lo scarto rimanente deve assumere l'unico valore che consente di soddisfare il vincolo.

Esempio 13.11 (di calcolo)**.** La tabella seguente si riferisce all'altezza rilevata su 10 soggetti.

X	$r(X)$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1.82	8	0.064	0.004096
1.84	10	0.084	0.007056
1.71	3	−0.046	0.002116
1.75	5	−0.006	0.000036
1.81	7	0.054	0.002916
1.72	4	−0.036	0.001296
1.82	9	0.064	0.004096
1.68	2	−0.076	0.005776
1.75	6	−0.006	0.000036
1.66	1	−0.096	0.009216
17.56			0.03664

$$\bar{x} = 1.756, \quad \mathrm{Me}_1 = 1.75, \ \mathrm{Me}_2 = 1.75, \quad S^2 = \frac{0.03664}{9} = 0.004071$$

Proprietà della varianza

- La varianza è sempre maggiore o uguale a zero
- La varianza è invariante per traslazioni

$$Y = a + X \Rightarrow \mathrm{Var}(Y) = \mathrm{Var}(X)$$

- La varianza non è invariante per cambiamenti di scala

$$Y = bX \Rightarrow \mathrm{Var}(Y) = b^2 \mathrm{Var}(X)$$

Scarto quadratico medio

Lo *scarto quadratico medio* o *deviazione standard* è la radice quadrata della varianza

$$\sigma = \sqrt{\sigma^2} \text{ oppure } S = \sqrt{S^2}.$$

È l'indice più frequentemente utilizzato perché è espresso nella stessa unità di misura della variabile d'interesse.

Coefficiente di variazione

Il coefficiente di variazione è dato da

$$CV = \frac{\sigma}{\bar{X}}$$

- È un numero puro (adimensionale)
- Elimina l'effetto dell'intensità media del fenomeno studiato.

Serve per fare confronti.

Il calcolo della varianza

La varianza può essere calcolata mediante una formula alternativa:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 n_i - \bar{X}^2 = \overline{X^2} - \bar{X}^2$$

oppure

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k} x_i^2 n_i - \frac{N}{N-1}\bar{X}^2$$

(dimostrazione: basta sviluppare il quadrato e usare la definizione di media aritmetica).

I vantaggi: l'uso della formula alternativa consente

- di ottenere il risultato con meno operazioni
- di ridurre gli errori dovuti ad arrotondamenti

Esempio 13.12 (di calcolo)**.** La varianza dell'altezza rilevata su 10 soggetti può essere calcolata più semplicemente.

X	X^2
1.82	3.3124
1.84	3.3856
1.71	2.9241
1.75	3.0625
1.81	3.2761
1.72	2.9584
1.82	3.3124
1.68	2.8224
1.75	3.0625
1.66	2.7556
17.56	30.872

$$S^2 = \frac{1}{9} \cdot 30.872 - \frac{10}{9} \cdot 1.756^2 = 0.004071$$

Esercizi consigliati

Esercizi da 12.1 a 12.9 del testo consigliato.

Capitolo aggiuntivo 14

Calcolo delle probabilità

Il calcolo delle probabilità è presupposto essenziale per il processo di inferenza statistica. In realtà il calcolo delle probabilità è una disciplina a sé stante:

- inizialmente sviluppata per lo studio dei giochi d'azzardo
- con applicazioni in numerosi campi della scienza (fisica, genetica, ...)

Definizioni

Esperimento: insieme di procedure volte a produrre un certo risultato
Esperimento aleatorio o casuale: esperimento il cui esito non può essere predetto con certezza
Spazio campionario o spazio degli eventi: insieme dei risultati possibili di un esperimento casuale. Si indica spesso con S o Ω.

Gli eventi

Un *evento* è un qualunque sottoinsieme dello spazio campionario.

Esempio 14.1. Nel lancio di un dado

$$S = \{1, 2, 3, 4, 5, 6\}$$

e alcuni eventi sono

$$A = \{1, 3\},\ B = \{5\},\ \varnothing,\ S$$

Nel caso del lancio di una moneta $S = \{T, C\}$ (dove T sta per testa e C per croce), nel caso di una partita di calcio $S = \{1, \times, 2\}$. Se invece l'esperimento è costituito da due lanci successivi di una moneta allora $S = \{(T,T), (T,C), (C,T), (C,C)\}$ cioè i risultati possibili sono coppie; se i lanci sono tre saranno terne e così via.

Cos'è la probabilità?

Le definizioni di probabilità sono molteplici. Le più rilevanti sono:

- definizione classica
- definizione frequentista
- definizione soggettiva
- definizione assiomatica

Definizione classica

La probabilità di un evento E è data dal rapporto tra:

- numero dei casi favorevoli ($\#E$) al verificarsi dell'evento
- numero di casi possibili ($\#S$), purché *ugualmente possibili*

$$P(E) = \frac{\#\text{ casi favorevoli}}{\#\text{ casi possibili}} = \frac{\#E}{\#S}.$$

Si osserva subito che $0 \leq P(E) \leq 1$ e che i casi $P(E) = 0$ e $P(E) = 1$ si verificano rispettivamente quando $E = \varnothing$ e $E = S$.

Esempio 14.2. *Supponiamo vi siano* 3 *diverse strade per andare dalla città* A *alla città* B *e* 5 *diverse strade per andare dalla città* B *alla città* C*; quante strade diverse si possono percorrere per andare da* A *a* C *passando per* B*?*

Indicato con S_3 l'insieme delle strade che vanno da A a B e con S_5 l'insieme delle strade che vanno da B a C, i risultati possibili sono gli elementi di $S = S_3 \times S_5$, cioè coppie ordinate in cui il primo elemento indica la strada seguita per andare da A a B e il secondo quella per andare da B a C. Il numero di strade possibili è quindi dato da

$$\#S = \#S_3 \times \#S_5 = 3 \cdot 5 = 15.$$

Esempio 14.3. *Se viene lanciata una moneta per 7 volte, quanti sono i possibili risultati?*

L'insieme dei risultati possibili è $S = \{T, C\}^7$ (settuple ordinate) e il loro numero è quindi $\#S = 2^7$.

Disposizioni e combinazioni

Disponendo di un insieme Ω di n elementi, che possiamo pensare come un'urna, da cui dobbiamo estrarne k abbiamo le seguenti modalità di scelta:

- *senza ripetizione*, cioè senza rimettere nell'urna l'elemento estratto,
- *con ripetizione*, cioè rimettendo ogni volta nell'urna l'elemento estratto.

Gli elementi scelti possono poi essere disposti nell'ordine in cui sono stati estratti oppure alla rinfusa, cioè la scelta può essere *ordinata* o *non ordinata*. Nel primo caso si parla di *disposizioni* e nel secondo di *combinazioni* (da non confondersi con quelle delle casseforti, che dovrebbero invece essere chiamate disposizioni, perché l'ordine in cui si introducono le cifre è essenziale).

Contiamo ora, nei vari casi, quanti sono i risultati possibili.

Disposizioni senza ripetizione. Fattoriale

Effettuiamo k estrazioni senza rimettere ogni volta nell'urna l'elemento estratto. Poiché la prima estrazione ha n possibili risultati, la seconda $n-1$, e così via fino alla k-esima che ha $n-k+1$ risultati, possiamo vedere i risultati possibili come il prodotto cartesiano di k insiemi

$$S_n \times S_{n-1} \times \cdots \times S_{n-k+1} \text{ con } \#S_i = i.$$

Allora i risultati possibili, cioè le disposizioni senza ripetizione di k elementi su un insieme di n sono in totale

$$D_{k,n} := n(n-1)(n-2)\cdots(n-k+1).$$

In questo caso dev'essere $k \leq n$.

Nel caso in cui $k = n$ si ha che tutti gli elementi sono stati estratti e disposti in maniera ordinata; qualunque risultato differisce allora dagli altri solo per l'ordine in cui gli elementi sono disposti, e quindi qualunque risultato può essere ottenuto partendo da un altro semplicemente permutando

l'ordine. Per questo motivo, se $k = n$ si parla di *permutazioni* ed il numero corrispondente

$$D_{n,n} = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$$

è *n fattoriale*; esso consiste nel prodotto di tutti i numeri naturali da n fino ad 1. È facile verificare che

$$D_{k,n} = \frac{n!}{(n-k)!}.$$

Esercizio 14.4. *Supponiamo di dover scegliere un presidente ed un segretario di una commissione di* 10 *membri. Quante sono le possibili scelte?*

$\boxed{R}$ È chiaro che non vi possono essere ripetizioni e poiché i ruoli non sono interscambiabili, l'ordine è importante. Si tratta quindi di disposizioni senza ripetizione di 2 oggetti su 10. Il loro numero è quindi

$$D_{2,10} = 10 \cdot 9 = 90.$$

Esercizio 14.5. *Supponiamo che* 5 *persone si dispongano allineate per fare un fotografia. Quante diverse fotografie possono essere fatte?*

$\boxed{R}$ Si tratta di contare le permutazioni di un insieme di 5 elementi, che sono $5! = 120$.

Disposizioni con ripetizione

In questo caso si ha che i k oggetti possono essere scelti in

$$D^r_{k,n} := n \cdot n \cdot n \cdots n = n^k$$

modi. Qui non occorre che $k \leq n$.

Esercizio 14.6. *Quante parole di* 5 *lettere si possono scrivere con le* 21 *lettere dell'alfabeto, indipendentemente dal loro significato?*

$\boxed{R}$ Si tratta di disposizioni con ripetizione di 5 elementi su un insieme di 21. Sono quindi 21^5.

Combinazioni senza ripetizione. Coefficienti binomiali

Quando l'ordine non è rilevante, ogni permutazione di k elementi viene considerata come una scelta equivalente (e dunque non distinguibile dalle altre): allorché scegliamo k oggetti, consideriamo dunque equivalenti le $k!$ permutazioni degli stessi. Se allora indichiamo con $C_{k,n}$ il numero delle combinazioni senza ripetizione, abbiamo

$$C_{k,n} = \frac{D_{k,n}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

(naturalmente in questo caso $k \leq n$). Ricordando che $0! = 1$ si ha

$$\binom{0}{0} = \binom{n}{0} = \binom{n}{n} = 1.$$

Esercizio 14.7. *In quanti modi è possibile pescare* 2 *carte da un mazzo di* 27*?*

R Siccome l'ordine in cui le carte vengono pescate non ha importanza, e non può esserci ripetizione, si tratta di combinazioni senza ripetizione di 2 elementi di un insieme di 27 che sono

$$\binom{27}{2} = \frac{27!}{2!25!} = \frac{27 \cdot 26}{2} = 351.$$

Combinazioni con ripetizione

Non si possono ottenere facilmente, come quelle senza ripetizione a partire dalle disposizioni, considerando che le permutazioni originano combinazioni equivalenti. Ciò è dovuto alla eventuale presenza di ripetizioni e per il fatto che se si permuta l'ordine di due elementi uguali non si ottiene una nuova disposizione. In questo caso il conto è più difficile e si potrebbe dimostrare che il numero complessivo delle combinazioni con ripetizione è

$$C^r_{k,n} := \binom{n+k-1}{k} = \binom{n+k-1}{n-1}.$$

Dimostrare per esercizio che vale la seconda uguaglianza.

Esercizi

Esercizio 14.8. *Siano A un insieme di k elementi e B un insieme di n elementi.*

1. *Quante sono le funzioni da A a B?*

2. *Quante sono le funzioni iniettive da A a B?*

3. *Sia $0 \leq k \leq n$. Quanti sono i sottoinsiemi di B con k elementi?*

4. *Quanti elementi ha l'insieme delle parti di B?*

R *1.* Sono tante quante le disposizioni con ripetizione di k elementi su n, cioè n^k.
2. Se $k > n$ non esistono funzioni iniettive da A a B. Se $k \leq n$ allora sono tante quante le disposizioni senza ripetizione di k elementi su n, cioè $\frac{n!}{(n-k)!}$.
3. Sono tanti quanti le combinazioni senza ripetizione di k elementi su un insieme di n, cioè $\binom{n}{k}$.
4. Sono $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ (cfr. formula del binomio).

Esercizio 14.9. *Ad un gran premio di Formula* 1 *partecipano* 22 *concorrenti. Sapendo che si classificano solo i primi* 6 *arrivati ed escludendo la possibilità di arrivi a pari merito, determinare*

1. *quante sono le possibili classifiche sapendo che al traguardo arriveranno almeno* 6 *concorrenti;*

2. *quante sono le possibili classifiche ammettendo che al traguardo possano arrivare, a causa di abbandoni, anche meno di* 6 *concorrenti.*

R *1.* $D_{22,6} = \dfrac{22!}{16!} = 53\,721\,360$; *2.* $\displaystyle\sum_{i=1}^{6} D_{22,i} = \sum_{i=1}^{6} \frac{22!}{(22-i)!} = 57\,066\,724$.

Esercizio 14.10. *Un computer genera numeri casuali di* 8 *cifre binarie (utilizzando cioè solo le cifre* 0 *e* 1*).*

1. *Quanti numeri è possibile generare in tal modo?*

2. *Quanti di questi numeri sono tali che la somma delle loro cifre è* 4*?*

R *1.* $2^8 = 256$; *2.* 70.

Esercizio 14.11. *Il codice genetico è costituito dalla corrispondenza degli aminoacidi con i codoni. Ogni codone è una disposizione con ripetizione di* 3 *fra le* 4 *differenti basi dell'RNA messaggero. Quanti sono i possibili codoni?*

R 64.

Critiche alla definizione classica

- di ordine teorico: la definizione è circolare (ugualmente possibili significa ugualmente probabili)
- di ordine pratico: non sempre è possibile enumerare tutti i casi possibili, oppure i casi possibili non sono ugualmente possibili

---

Definizione assiomatica

Dato un insieme S, indichiamo con $\wp(S)$ l'insieme delle parti di S. Gli assiomi del calcolo delle probabilità sono i seguenti.

La probabilità *è una funzione* $P : \wp(S) \to [0,1]$ *tale che*

1. $P(S) = 1$ *(cioè la probabilità dell'evento certo è pari a* 1*);*
2. $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$ *(cioè la probabilità è una funzione additiva).*

La definizione assiomatica

- stabilisce alcune regole (di carattere logico-formale) alle quali la probabilità deve sottostare
- è (quasi) universalmente accettata e condivisa
- non dà indicazioni su come assegnare probabilità agli eventi (vediamo con un esempio come questo dipenda dal contesto)

Esempio 14.12. Esistono molte funzioni P che soddisfano gli assiomi *1.* e *2.*. Il caso più semplice possibile è quello in cui S è costituito da due elementi, come nel caso del lancio di una moneta, cioè $S = \{T, C\}$. Assegnata una probabilità $p \in [0,1]$ all'evento $\{T\}$, cioè posto $P(\{T\}) := p$, da *1.* si ha che $P(\{T, C\}) = 1$ e dalla *2.* si ottiene

$$P(\{C\}) = P(\{T, C\}) - P(\{T\}) = 1 - p$$

e quindi la funzione P risulta completamente determinata. Esistono quindi in questo caso infinite probabilità, una per ciascun valore di $p \in [0,1]$. Tipicamente si sceglierà $p = 1/2$ se la moneta non è truccata, un valore diverso altrimenti.

Monotonia

Dagli assiomi segue subito la seguente proprietà di monotonia

$$\boxed{A \subset B \Rightarrow P(A) \leq P(B)}$$

infatti, in tal caso, osservando che $B = A \cup (B \setminus A)$ e $A \cap (B \setminus A) = \varnothing$ e usando la *2.* e fatto che P è non negativa si ha

$$P(B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A) \geq P(A).$$

Teorema delle probabilità totali

Teorema 14.13. *Dati due eventi A e B comunque scelti*

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

Attenzione: il teorema si differenzia dal terzo assioma in quanto gli eventi non sono necessariamente disgiunti.

DIMOSTRAZIONE Osservato che $A \cup B$ è unione disgiunta degli insiemi $A \setminus B$, $A \cap B$ e $B \setminus A$, per la additività si ha

$$P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A).$$

Sempre per l'additività si ha

$$P(A) = P(A \setminus B) + P(A \cap B), \quad P(B) = P(B \setminus A) + P(A \cap B)$$

e la tesi si ottiene sostituendo. □

Condizionamento

$A|B$ si legge "A condizionato (o dato) B"

Si suppone di aver osservato il verificarsi di B e ci si chiede se ed in quale misura questa informazione modifichi la valutazione di probabilità su A.

In generale

$$P(A|B) \neq P(A)$$

Esempio 14.14. Supponiamo di lanciare una moneta due volte, e che in ciascun lancio testa e croce abbiano la stessa probabilità di uscire. Come già osservato, lo spazio degli eventi è

$$S = \{(T,T),(T,C),(C,T),(C,C)\}$$

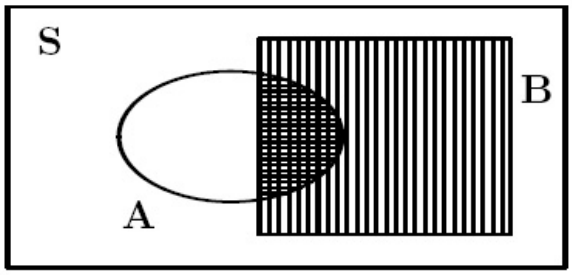
La probabilità che esca testa in entrambi i lanci è

$$P(\{(T,T)\}) = \frac{\#\{(T,T)\}}{\#\{(T,T),(T,C),(C,T),(C,C)\}} = \frac{1}{4}$$

Se sappiamo che nel primo lancio esce testa allora la probabilità che esca testa in entrambi i lanci diventa

$$P(\{(T,T)\}) = \frac{\#\{(T,T)\}}{\#\{(T,T),(T,C)\}} = \frac{1}{2}$$

Il condizionamento consiste dunque in una ridefinizione dello spazio campionario



che si riduce da S a B.

Definizione 14.15 (di probabilità condizionata)**.** *Sia $P(B) > 0$. La probabilità di un evento A condizionata al verificarsi di B si definisce nel modo seguente*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

Si riproporziona la probabilità di A in funzione della riduzione dello spazio campionario. Si osservi che $P(B|B) = 1$ e inoltre se $A \cap C = \varnothing$ allora $P(A \cup C|B) = P(A|B) + P(C|B)$. Dunque $P(\cdot|B)$ è una probabilità su B.

Indipendenza di eventi

Due eventi A e B si dicono *indipendenti* se e solo se

$$P(A|B) = P(A)$$

o, equivalentemente,

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

ossia se il verificarsi dell'evento B non modifica la valutazione di probabilità sull'evento A.

Esempio 14.16 (doppio lancio di una moneta)**.** *Mostriamo che nel lancio doppio di una moneta i risultati di ciascun lancio sono tra loro indipendenti.* Supponiamo che T e C siano equiprobabili. Lo spazio degli eventi è

$$S = \{(T,T),(T,C),(C,T),(C,C)\}.$$

Sia A l'evento "esce testa al primo lancio" e sia B l'evento "esce testa al secondo lancio", cioè

$$A = \{(T,T),(T,C)\}, \quad B = \{(T,T),(C,T)\}.$$

Si ha

$$P(A \cap B) = P(\{(T,T)\}) = \frac{1}{4},$$

$$P(A)P(B) = P(\{(T,T),(T,C)\}) \cdot P(\{(T,T),(C,T)\}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

quindi $P(A \cap B) = P(A)P(B)$ e i due eventi sono dunque indipendenti.

Variabili casuali o aleatorie

Esempio 14.17 (Famiglie con 4 figli)**.** Chiamiamo famiglie con quattro figli le disposizioni con ripetizione di due oggetti (M e F) su 4 posizioni ($n = 2$, $k = 4$), ovvero le quaterne ordinate di elementi dell'insieme $\{M, F\}$.

In una famiglia di 4 figli, ci si chiede qual'è la probabilità che 0, 1, 2, 3 o tutti i figli siano maschi (considerando equiprobabile la nascita di maschi e femmine). La popolazione S in tal caso è costituita da tutte le possibili famiglie con 4 figli.

Introduciamo una funzione X definita su S che conta i figli maschi. Data una famiglia x si avrà

$$X(x) = \#\text{ figli maschi di } x$$

Quando, come in questo caso, l'esito di un esperimento o di una prova si può rappresentare con un numero X e se ad ogni realizzazione dell'esperimento questo numero può assumere valori diversi, allora X prende il nome di *variabile aleatoria o casuale.*

In generale, dunque, una *variabile casuale o aleatoria* X è una funzione che assegna valori numerici agli eventi che compongono lo spazio campionario

$$X : S \to \mathbb{R}$$

Costituiscono gli analoghi probabilistici delle variabili statistiche e anch'esse si classificano in *discrete* (possono assumere solo un numero finito o una infinità numerabile di valori) e *continue* (possono assumere tutti i valori all'interno di un intervallo).

Variabili casuali discrete

I valori che la variabile può assumere si indicano con x_1, x_2, ...

Le variabili casuali discrete sono caratterizzate dalla *funzione di massa (o distribuzione) di probabilità*

$$f(x_i) = P(X = x_i) := P(\{s \in S \ : \ X = x_i\}) = P(X^{-1}(\{x_i\}))$$

che ad ogni valore x_i associa la probabilità che la v.c. X assuma il valore x_i.

Osservazione 14.18. La distribuzione delle frequenze relative p_i di una variabile statistica corrisponde alla funzione di distribuzione di probabilità nel caso in cui la popolazione abbia un numero finito N di elementi e le modalità siano equiprobabili. Infatti in tal caso

$$f(x_i) = P(X^{-1}(\{x_i\})) = \frac{\#X^{-1}(\{x_i\})}{N} = \frac{n_i}{N} = p_i$$

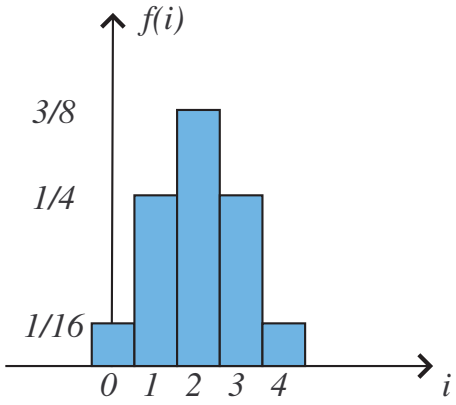
Esempio 14.19 (famiglie con 4 figli)**.** Nel caso della v.c. X introdotta nell'Esempio 14.17 si ha che la numerosità della popolazione S (possibili famiglie di 4 figli) è pari al numero di disposizioni con ripetizione con $n = 2$ e $k = 4$, che sono $2^4 = \mathbf{16}$. Le modalità sono 5 perché una famiglia può avere

da $x_0 = 0$ a $x_4 = 4$ figli. Quindi, indicate le varie modalità con $x_i = i$, si ha che il numero delle famiglie che hanno i-figli maschi sarà pari al numero di sottoinsiemi di $\{1, 2, 3, 4\}$ che hanno i elementi, ovvero di combinazioni con ripetione di i elementi su 4, cioè $\binom{4}{i}$. Si ha dunque

$$f(i) = P(X = i) = \frac{\binom{4}{i}}{16},$$

cioè, per esteso,

$$\begin{aligned}
&f(0) = P(X = 0) = \tfrac{1}{16};\\
&f(1) = P(X = 1) = 4 \cdot \tfrac{1}{16} = \tfrac{1}{4};\\
&f(2) = P(X = 2) = \tbinom{4}{2} \cdot \tfrac{1}{16} = \tfrac{3}{8};\\
&f(3) = P(X = 3) = \tbinom{4}{3} \cdot \tfrac{1}{16} = \tfrac{1}{4};\\
&f(4) = P(X = 4) = \tfrac{1}{16}
\end{aligned}$$



Media e varianza di una v.c. discreta

La *media o valore atteso* e la *varianza* di una variabile causale discreta sono date, rispettivamente, da

$$E(X) := \sum_i x_i f(x_i), \qquad \mathrm{Var}(X) := \sum_i [x_i - E(X)]^2 f(x_i). \tag{14.1}$$

Come si vede dalle definizioni questi indici dipendono solamente dalla distribuzione di probabilità; ne consegue che v.c. identicamente distribuite hanno la stessa media e la stessa varianza. È quindi naturale parlare di media e di varianza di una distribuzione di probabilità intendendo con ciò media e varianza di qualunque v.c. che abbia la distribuzione considerata.

Si verifica facilmente che vale la formula alternativa

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \sum_i x_i^2 f(x_i) - E(X)^2.$$

Inoltre valgono le proprietà di linearità della media e di invarianza per traslazioni della varianza viste nel capitolo precedente (dimostrarle per esercizio).

Esempio 14.20 (famiglie con 4 figli)**.** Nel caso della v.a. dell'esempio precedente si ha

$$E(X) = \sum_{i=0}^{4} i f(i) = f(1) + 2f(2) + 3f(3) + 4f(4) = \frac{1}{4} + 2\frac{3}{8} + 3\frac{1}{4} + 4\frac{1}{16} = 2,$$

$$\mathrm{Var}(X) = \sum_{i=0}^{4} [i - E(X)]^2 f(i) = \sum_{i=0}^{4} [i-2]^2 f(i) = 1$$

Variabile casuale binomiale o di Bernoulli

Si chiamano variabili casuali di Bernoulli quelle del tipo

$$X : S \to \{0, 1\}$$

cioè che assumono solo due valori, 0 (*insuccesso*) e 1 (*successo*), con probabilità pari rispettivamente a $1 - p$ e p, dove $p \in]0, 1[$ è detto *parametro*.

Il parametro p, pari alla probabilità di osservare un successo, rappresenta una caratteristica (generalmente incognita) del fenomeno rappresentato mediante la v.c. di Bernoulli (per es. la probabilità di sopravvivenza o che esca testa).

È indicata per descrivere fenomeni che si manifestano con due sole modalità possibili (per es. la sopravvivenza o il lancio di una moneta).

Distribuzione binomiale

La funzione di distribuzione di probabilità di qualunque v.c. di Bernoulli è

$$f(x;p) = p^x(1-p)^{1-x}$$

dove $x \in \{0,1\}$ ed è detta *distribuzione binomiale di parametro p.* Si ha

$$f(0;p) = P(X=0) = 1-p, \quad f(1;p) = P(X=1) = p.$$

Per indicare che un v.c. X ha distribuzione binomiale di parametro $p.$ si scrive

$$X \sim BI(1,p)$$

Si ha

$$E(X) = \sum_{i=0}^{1} if(i;p) = 0 \cdot (1-p) + 1 \cdot p = p,$$

e si può facilmente verificare che

$$Var(X) = p(1-p).$$

Se invece X è la variabile aleatoria che conta il numero di successi ottenuti in n prove indipendenti (esempio: n lanci di una moneta), allora X può assumere solo valori da 0 a n, cioè

$$X : S^n \to \{0,1,2,\ldots,n\}$$

e la sua distribuzione risulta

$$f_n(i;p) := P(X=i) = \binom{n}{i} p^i(1-p)^{n-i}, \quad i = 0,1,\ldots,n$$

e si scrive che

$$X \sim BI(n;p).$$

Ad esempio, la variabile che conta i figli maschi delle famiglie con 4 figli di un esempio precedente ha distribuzione $BI(4,1/2)$.

Si noti che dalla formula del binomio di Newton segue che

$$\sum_{k=0}^{n} P(X=i) = \sum_{i=0}^{n} \binom{n}{i} p^i(1-p)^{n-i} = (p+1-p)^n = 1$$

in accordo col fatto che la probabilità totale deve valere 1.

Si può dimostrare che media e varianza della distribuzione $BI(n,p)$ sono date da

$$E(X) = np; \quad \mathrm{Var}(X) = np(1-p).$$

Marcatura e distribuzione ipergeometrica

Esempio 14.21 (Esempio 11.21 del Testo)**.** *Di una popolazione di 15 lupi, 5 vengono catturati, marcati con un collare e rilasciati nel loro ambiente.*

Successivamente, 3 lupi vengono catturati sperando che tra essi ve ne siano alcuni di quelli marcati, in modo da osservare le differenze con l'analisi precedente.

Qual'è la probabilità che esattamente 2 tra i 3 animali catturati siano già marcati?

Spazio campionario S: tutti i sottoinsiemi di 3 lupi che sono

$$\#S = \binom{15}{3} = 455$$

Casi favorevoli: terne in cui almeno due lupi sono marcati, che sono in totale

$$\begin{gathered}\#(\text{sottoinsiemi di 2 lupi tra i 5 marcati}) \\ \times \\ \#(\text{modi di scegliere un lupo tra i 10 rimanenti}) \\ = \\ \binom{5}{2} \cdot \binom{10}{1}\end{gathered}$$

Indicata con X la v.c. che conta i lupi marcati si ha

$$P(X=2) = \frac{\binom{5}{2} \cdot \binom{10}{1}}{\binom{15}{3}} = \frac{20}{91} \simeq 0.22$$

Procedendo in maniera analoga si trova che

$$P(X=i) = \frac{\binom{5}{i} \cdot \binom{10}{3-i}}{\binom{15}{3}}$$

e si calcola allora facilmente tutta la distribuzione di probabilità di X

$$P(X=0) = \frac{\binom{5}{0} \cdot \binom{10}{3}}{\binom{15}{3}} = \frac{10!}{3!7!}\frac{3!12!}{15!} = \frac{10!12!}{7!15!} = \frac{8 \cdot 9 \cdot 10}{13 \cdot 14 \cdot 15} = \frac{24}{91} \simeq 0.26$$

$$P(X=1) = \frac{\binom{5}{1} \cdot \binom{10}{2}}{\binom{15}{3}} = 5\frac{10!}{2!8!}\frac{3!12!}{15!} = \frac{45}{91} \simeq 0.49$$

$$P(X=3) = \frac{\binom{5}{3} \cdot \binom{10}{0}}{\binom{15}{3}} = \frac{5!}{3!2!}\frac{3!12!}{15!} = \frac{2}{91} \simeq 0.02$$

Generalizzando al caso di una popolazione di N elementi di cui $0 \leq K \leq N$ marcati e supponendo di pescarne a caso n, la v.c. che conta gli esemplari marcati ha la seguente distribuzione:

$$\boxed{P(X=i) = \frac{\binom{K}{i} \cdot \binom{N-K}{n-i}}{\binom{N}{n}}}$$

detta *distribuzione ipergeometrica*. Media e varianza di X valgono

$$E(X) = \frac{K}{N}n, \quad \mathrm{Var}(X) = \frac{K}{N}\Big(1 - \frac{K}{N}\Big)$$

Esercizio 14.22 (Esempio 11.22 del Testo)**.** *In una popolazione di 20 lupi ne vengono marcati 4. Determinare il numero minimo di animali da ricatturare per essere sicuri al 90% di prenderne almeno uno marcato.*

Indicata con X la v.c. che conta i lupi marcati tra quelli catturati, la probabilità che catturando n lupi ve ne sia almeno uno marcato è

$$P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$$

ma è meno calcoloso osservare che

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X=0).$$

Quindi basta calcolare $P(X=0)$. Sapendo che $K=4$ e $N=20$ si ha

$$P(X=0) = \frac{\binom{K}{i} \cdot \binom{N-K}{n-i}}{\binom{N}{n}} = \frac{\binom{4}{0} \cdot \binom{16}{n}}{\binom{20}{n}} = \frac{(20-n)(19-n)(18-n)(17-n)}{20 \cdot 19 \cdot 18 \cdot 17}.$$

Dobbiamo ora imporre che

$$P(X \geq 1) = 1 - P(X=0) \geq 0.9$$

cioè

$$1 - \frac{(20-n)(19-n)(18-n)(17-n)}{20 \cdot 19 \cdot 18 \cdot 17} \geq 0.9$$

e trovare il minimo n tale che la disuguaglianza sia sodisfatta. Si tratta di una disequazione di grado 4 in n, ma calcolando esplicitamente i valori del primo membro per i diversi valori di n a partire da $n=1$ si ottiene che la condizione è sodddisfatta se $n \geq 9$.

Numeri ritardatari e distribuzione geometrica

Esempio 14.23 (Gioco del Lotto - Esempio 11.24 Testo)**.** Il gioco del Lotto consiste nell'eseguire 5 estrazioni contemporanee da un'urna contenente sfere identiche numerate da 1 a 90. Le estrazioni vengono eseguite 11 volte, una per ogni "ruota". Chiaramente, la probabilità di uscita di un singolo numero giocando su di una sola ruota è di $5/90 = 1/18 \simeq 0.05$. Qual'è la probabilità di realizzare un ambo giocando 3 numeri su una ruota sola?

Pensando di marcare i numeri su cui giochiamo, possiamo ricorrere alla distribuzione ipergeometrica con $K = 3$, $n = 5$, $N = 90$. Si ha

$$P(X = 2) = \frac{\binom{K}{i} \cdot \binom{N-K}{n-i}}{\binom{N}{n}} = \frac{\binom{3}{2} \cdot \binom{87}{3}}{\binom{90}{5}} \simeq 0.007$$

Puntando 1 euro su questo risultato se ne vincono 78.33 (fonte sito web Lottomatica). Quanto possiamo sperare di vincere? Se giochiamo 1000 volte spendiamo 1000 euro e vinciamo mediamente 7 volte, cioè $7 \cdot 78.33 = 548.41$ euro. Su mille giocate da un euro perderemo allora mediamente 451.49 euro.

Esempio 14.24 (numeri ritardatari)**.** I giocatori del lotto hanno molto interesse per i numeri "ritardatari", quelli cioè che non escono da molte estrazioni. Un ritardo di un centinaio di estrazioni è, in genere, piuttosto raro. Calcoliamo la probabilità di ritardo di un dato numero.

Sia X la v.c. che conta a quale estrazione esce il numero. Sia $p = 1/18$. Come visto nell'esempio precedente la probabilità che il numero esca alla prima estrazione (ritardo 0) è p. Di conseguenza quella che non esca è $1 - p$. Allora la probabilità che esca alla seconda (evento A) e non alla prima (evento B) è

$$P(X = 2) = P(A \cap B) = P(A)P(B) = p(1 - p)$$

In generale, se si verificano $k-1$ insuccessi e il numero esce alla k-estrazione, si ha

$$\boxed{P(X = k) = p(1 - p)^{k-1}}$$

detta *distribuzione geometrica* di parametro p.

In particolare la probabilità che esca esattamente alla 101-esima estrazione è

$$P(X = 101) = \frac{1}{18}(1 - \frac{1}{18})^{100} \simeq 2 \cdot 10^{-5}$$

La probabilità che il numero ritardi almeno di 100 estrazioni (ma non esca necessariamente alla 101-esima) è

$$\begin{aligned} P(X \geq 101) &= 1 - P(X < 101) = 1 - \sum_{k=1}^{100} P(X = k) \\ &= 1 - \sum_{k=1}^{100} p(1-p)^{k-1} = 1 - p\sum_{k=0}^{99}(1-p)^k \\ &= (1-p)^{100} = (\frac{17}{18})^{100} \simeq 0.003 \end{aligned}$$

Siccome le ruote sono 11 e i numeri 90, vi saranno, in media, $90 \cdot 11 \cdot 0.003 \simeq 3$ numeri con ritardi superiori alle 100 estrazioni.

Eventi rari e distribuzione di Poisson

Esempio 14.25 (Esempio 11.26 Testo)**.** *In una zona pianeggiante di* 10 km^2 *sono distribuite* 40.000 *querce. Con quale probabilità analizzando una zona limitata, per esempio di* 1000 m^2 *possiamo trovare i querce?*

Immaginiamo che sia possibile suddividere la zona di 10 km^2 in quadrati di area pari a 1000 m^2. Siccome 10 $km^2 = 10 \cdot (10^3 m)^2 = 10^7$ m^2 il numero di quadrati della suddivisione è $10^7/10^3 = 10^4$.

Sia X la v.c. che conta le querce che cadono in uno dei quadrati.

Condiderata una certa quercia, questa avrà probabilità $p = 1/10^4$ di appartenere al quadrato (successo) e $1 - p$ di non appartenere (insuccesso). Quindi la probabilità che una quercia appartenga al quadrato ha distribuzione binomiale $BI(1, p)$. Ripetendo l'esperimento aleatorio per ogni quercia, cioè 40000 volte, si ha che la v.c. X avrà distribuzione binomiale $BI(n; p)$ con $n = 4 \cdot 10^4$ e $p = 10^{-4}$ ovvero

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} = \binom{4 \cdot 10^4}{i} 10^{-4i} (1 - 10^{-4})^{4 \cdot 10^4 - i}.$$

Si ha dunque

$$
\begin{aligned}
&P(X=0) = \tbinom{4\cdot 10^4}{0}(1-10^{-4})^{4\cdot 10^4} = 0.9999^{40000} \simeq 0.018,\\
&P(X=1) = \tbinom{4\cdot 10^4}{1}10^{-4}(1-10^{-4})^{4\cdot 10^4-1} = 40.9999^{39999} \simeq 0.073,\\
&P(X=2) = \tbinom{4\cdot 10^4}{2}10^{-8}(1-10^{-4})^{4\cdot 10^4-2} \simeq 2\cdot 3.9999\cdot 0.9999^{39998} \simeq 0.146,\\
&P(X=3) \simeq 0.195,\\
&P(X=4) \simeq 0.195,\\
&P(X=5) \simeq 0.156,\\
&\ldots
\end{aligned}
$$

Ne risulta un conto disagevole a causa degli alti valori di n. Anche senza fare questi conti possiamo però calcolare la media e la varianza della distribuzione, che sono date da

$$E(X) = np = 4\cdot 10^4\cdot 10^{-4} = 4, \quad \mathrm{Var}(X) = np(1-p) = 4(1-10^{-4}) \simeq 4$$

Come osservato il conto di $P(X=i)$ non è agevole, ma è possibile ottenerne facilmente un'approssimazione. Anzitutto, si osserva che il prodotto np è costante e uguale alla media m della distribuzione (nel caso in esame $m=4$). Si pone dunque $np=m$, da cui $p=m/n$ e si sostituisce p nell'espressione della distribuzione di probabilità, ottenendo

$$P(X=i) = \binom{n}{i}p^i(1-p)^{n-i} = \binom{n}{i}(\frac{m}{n})^i(1-\frac{m}{n})^{n-i}$$

A questo punto si passa al limite per $n\to\infty$ ottenendo

$$
\begin{aligned}
\lim_{n\to\infty} P(X=i) &= \lim_{n\to\infty}\Big[\binom{n}{i}(\frac{m}{n})^i(1-\frac{m}{n})^{n-i}\Big]\\
&= \lim_{n\to\infty}\Big[\frac{n(n-1)\ldots(n-i+1)}{i!}\frac{m^i}{n^i}\,\mathrm{e}^{(n-i)\log(1-\frac{m}{n})}\Big]\\
&= \lim_{n\to\infty}\Big[\frac{n(n-1)\ldots(n-i+1)}{n^i}\frac{m^i}{i!}\,\mathrm{e}^{n\log(1-\frac{m}{n})-i\log(1-\frac{m}{n})}\Big]\\
&= \frac{m^i}{i!}\,\mathrm{e}^{-m}\,.
\end{aligned}
$$

poiché $\lim_{n\to\infty} n\log(1-\frac{m}{n}) = \lim_{n\to\infty} -m\Big[-\frac{n}{m}\log(1-\frac{m}{n})\Big] = -m$ mentre $\lim_{n\to\infty}\frac{n(n-1)\ldots(n-i+1)}{n^i} = \lim_{n\to\infty}\frac{n}{n}\frac{(n-1)}{n}\ldots\frac{(n-i+1)}{n} = 1.$

Utilizzando la formula con $m=4$

$$P(X=i) \simeq \frac{4^i}{i!}\,\mathrm{e}^{-4}$$

nel caso dell'esempio e approssimando alla terza cifra decimale si riottengono, con meno fatica, gli stessi valori calcolati in precedenza per $i = 1, 2, 3, 4, 5$.

Definizione 14.26. *Si chiama* distribuzione di Poisson *di media m, la distribuzione di probabilità*

$$\boxed{P(X = i) = \frac{m^i}{i!}\,\mathrm{e}^{-m}}$$

Osservazione 14.27. *La distribuzione di Poisson descrive la distribuzione di probabilità di una variabile aleatoria X che può assumere un numero infinito di valori interi (i), con $E(X) = \mathrm{Var}(X) = m$.*

Essa si utilizza in particolare per calcolare la probabilità di ripetizione di eventi, nel caso in cui il singolo evento abbia una piccola probabilità di realizzarsi e le prove ripetute siano molte; per questi motivi viene anche detta *legge degli eventi rari*.

La proprietà di uguaglianza tra media e varianza permette, nella pratica, di decidere se sia verosimile l'ipotesi che una v.c. sia distribuita con legge di Poisson. A partire dai dati sperimentali, infatti, si possono determinare il valore atteso e la varianza di X. Solo se questi valori sono simili, si può ipotizzare che i dati siano distribuiti con legge di Poisson.

Esercizi consigliati e consigli per risolverli

Esercizi su variabili discrete: Testo da 11.1 a 11.8.

Quale distribuzione usare?

Dipende dal contesto. Ecco tre esempi che riassumono quanto osservato in precedenza:

- Nella *ripetizione di eventi indipendenti* la variabile X *che conta il numero di successi* è distribuita con legge Binomiale. Al crescere del numero di ripetizioni il calcolo si complica; se però la probabilità di successo è piccola (evento raro) e il numero di ripetizioni è grande allora X è distribuita con legge di Poisson (media e varianza sono uguali).

- *Se in un insieme di N elementi di cui k posseggono una caratteristica che li distingue dagli altri*, allora la distribuzione di probabilità della v.c. X *che conta gli elementi con quella caratteristica in un campione casuale di cardinalità n* è distribuita con legge ipergeometrica.

- Nel caso di *prove ripetute e indipendenti, la v.c. che conta a quale ripetizione un dato evento si verifica per la prima volta* è distribuita con legge geometrica.

Esercizio 14.28 (Esempio 11.27 Testo)**.** *In un dipartimento si usano vari microscopi elettronici prodotti da una stessa ditta. In 10 anni, ogni microscopio ha avuto in media 6 guasti, con una deviazione standard $\sigma = 2.5$.*

Determinare la probabilità che in 10 anni si abbiano più di 10 guasti.

Indichiamo con X il numero di rotture. Possiamo escludere di usare le leggi geometrica e ipergeometrica. Anche la binomiale semplice è da escludere perchè X non assume solo due valori. Chiediamoci se possiamo ipotizzare che X sia distribuita con legge di Poisson.

L'ipotesi non è irragionevole poiché $\mathrm{Var}(X) = \sigma^2 = 2.5^2 = 6.25$ quasi uguale alla media (6). Dunque

$$P(X \geq 10) = 1 - \sum_{i=0}^{9} P(X = i) = 1 - \sum_{i=0}^{9} \frac{6^i}{i!}\,\mathrm{e}^{-6} \simeq 1 - 0.96 = 0.04$$

Variabili casuali continue

Per le variabili casuali continue è poco utile assegnare probabilità ai singoli valori. Consideriamo infatti il seguente esempio.

Esempio 14.29 (La distribuzione uniforme nell'intervallo $[0, 1[$)**.** Sia X la v.c. che sceglie "a caso" un numero $x \in [0, 1[$. Si ha

$$X : S \to [0, 1[$$

Si tratta dunque di una v.c. continua. Osserviamo che

- dividendo l'intervallo $[0, 1[$ in n intervalli la probabilità che x cada in uno di essi è $1/n$;

- in generale, dati x_1, $x_2 \in [0, 1[$ si ha $P(X \in [x_1, x_2[) = x_2 - x_1$;

- i singoli valori hanno probabilità nulla, infatti

$$0 \leq P(X = x_0) \leq P(X \in [x_0, x_0 + \frac{1}{n}[) = \frac{1}{n} \to 0$$

quindi $P(X = x_0) = 0$.

Funzione di ripartizione e densità

Ha più senso invece assegnare probabilità agli intervalli. Perciò si definisce la funzione di ripartizione.

Definizione 14.30. *La funzione* $F(x) := P(X \leq x)$ *si chiama* funzione di ripartizione (o di distribuzione cumulativa) della v.c. X *rispetto alla probabilità* P.

Esempio 14.31. Nel caso della variabile X dell'Esempio 14.29 si ha

$$F(x) = P(X \leq x) = x \quad \text{per ogni } x \in [0, 1[$$

detta *distribuzione uniforme* nell'intervallo $[0, 1[$.

Definizione 14.32. *Se la funzione di ripartizione* F *è derivabile, la sua derivata*

$$f(x) := F'(x)$$

è detta densità di probabilità.

Se, come può accadere, la funzione di ripartizione F non è derivabile in uno o più punti, allora la v.c. X non è descrivibile in termini di una funzione di densità di probabilità. Occorrerà in tal caso ricorrere a strumenti più sofisticati di analisi matematica (Teoria delle Distribuzioni).

Osserviamo che, per definizione di derivata

$$f(x) = \lim_{h \to 0} \frac{P(X \leq x + h) - P(X \leq x)}{h} = \lim_{h \to 0^+} \frac{P(X \in]x, x + h])}{h}$$

e ciò si può interpretare dicendo che *la probabilità di trovare* X *in un intervallo di ampiezza* h *piccola intorno ad* x *è approssimativamente uguale a* $f(x)h$. In questo senso la densità di probabilità è l'analogo probabilistico della densità di frequenza di una variabile statistica (vedi capitolo precedente).

Valore atteso e varianza di una v.c. continua

Anche per le v.c. continue $X : S \to \mathbb{R}$ possiamo definire il valore atteso

$$E(X) := \int_{-\infty}^{+\infty} x f(x)\, dx := \lim_{b \to +\infty} \int_{-\infty}^{b} x f(x)\, dx$$

e la varianza

$$\mathrm{Var}(X) := \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)\, dx$$

Si osservi l'analogia con le definizioni (14.1) date nel caso di una v.c. discreta. In effetti le nuove definizioni si possono ottenere dalle vecchie sostituendo alla distribuzione di probabilità discreta la densità di probabilità e alla sommatoria l'integrale. Valgono anche in questo caso le proprietà di linearità del valore atteso e di invarianza per traslazioni della varianza (dimostrarle per esercizio).

La v.c. Normale

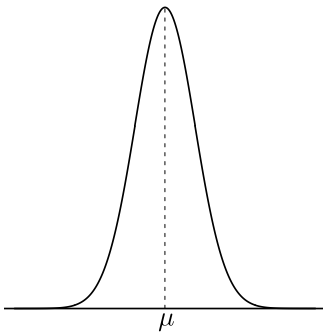
La v.c. continua più importante è la cosiddetta v.c. *Normale o Gaussiana*

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

La funzione di densità ha un andamento campanulare simmetrico.

Grafico della funzione di densità



31

Importanza della v.c. Normale

La v.c. Normale riveste un ruolo fondamentale perché

- descrive bene il manifestarsi di molti fenomeni, per esempio:
  - errori di misura (genesi della Normale)
  - caratteristiche morfologiche (altezza, lunghezza)
- gode di importanti proprietà (aspetto tecnico rilevante)

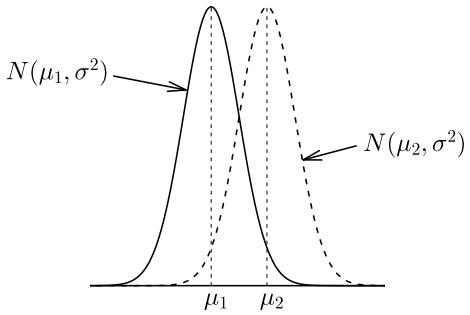
I parametri

La funzione di densità di probabilità della v.c. Normale dipende da due parametri:

- μ rappresenta il centro (valore atteso o media) della distribuzione; si ha infatti $E[X] = \mu$;
- σ^2 modula il grado di dispersione dei valori; si ha infatti $\text{Var}[X] = \sigma^2$;

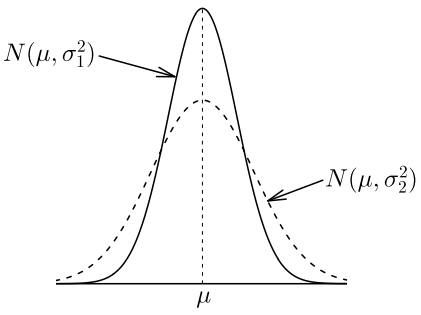
Anche in questo caso i parametri rappresentano caratteristiche incognite del fenomeno studiato.

V.c. Normali a media diversa



34

V.c. Normali a varianza diversa



35

$\sigma_1 < \sigma_2$

Errori di misura e genesi della Normale

Le misure di una grandezza non sono mai veramente precise. Gli errori di misurazione si dividono in

- *sistematici*, dovuti ad esempio ad imperfezioni o starature dello strumento di misura; sono *eliminabili*;

- *aleatori*, dovuti alle condizioni in cui si replica l'esperimento (es. temperatura, pressione, umidità, umore dello sperimentatore, ecc.); sono variabili da misura a misura e quindi *non eliminabili*.

Proprio per il fatto che gli errori aleatori non sono eliminabili, è importante conoscerne le proprietà statisticamente rilevanti in modo da tenerne conto negli esperimenti.

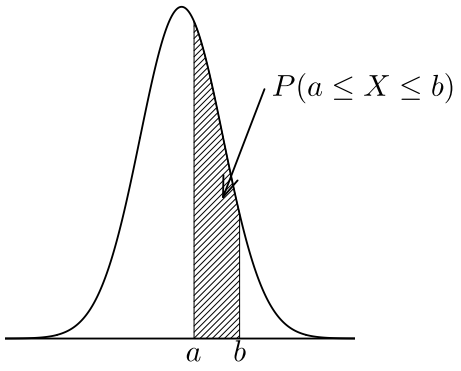
Supponiamo di misurare una grandezza X e di ottenere N misure X_1, X_2,..., X_N. Questi valori non sono generalmente identici, ma oscillano attorno alla loro media m_N. Queste medie, a loro volta, al crescere di N oscillano attorno ad un valore limite m. Molti studiosi, tra cui Gauss, indagando a fondo il comportamento di queste oscillazioni, hanno concluso che la v.c. m_N *è distribuita con una legge gaussiana con valore atteso uguale ad* m (da considerarsi il vero valore della grandezza).

Questo risultato è spiegato da uno dei principali teoremi del calcolo delle probabilità, il *Teorema del limite centrale* di cui parleremo più avanti.

Calcolo della probabilità per un intervallo

Per una qualunque v.c. continua con densità f, la probabilità è rappresentata dall'area sotto la curva di densità all'interno dell'intervallo.

Area=Probabilità



36

L'area si determina mediante l'operazione di integrazione. Infatti, per il Teorema Fondamentale del Calcolo Integrale si ha

$$\int_a^b f(x)\,dx = \int_a^b \frac{d}{dx}P(X \leq x)\,dx = P(X \leq b) - P(X \leq a) = P(a < x \leq b).$$

Osserviamo che affinché f sia una densità di probabilità occorre dunque che

- $f \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1.$

Caso della distribuzione Normale

Sfortunatamente, il calcolo dell'integrale spesso non è agevole. Nel caso della distribuzione Normale, l'integrale non è calcolabile analiticamente. Consideriamo ad esempio una variabile Z con distribuzione Normale con media 0 e varianza 1, detta *distribuzione Normale standard*

$$Z \sim \mathcal{N}(0, 1).$$

Si ha, ad esempio,

$$P(-1 < Z \leq 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \mathrm{e}^{-\frac{x^2}{2}}\, dx,$$

ma non è possibile procedere nel calcolo dell'integrale cercando una primitiva della funzione integranda. Si potrebbe infatti dimostrare che non esiste una primitiva esprimibile in termini finiti come somma, prodotto o composizione di funzioni elementari.

È tuttavia possibile effettuare un calcolo approssimato dell'integrale sostituendo l'esponenziale con funzioni più semplici da integrare, ad esempio con opportuni polinomi.

Questa idea ha consentito ai matematici di redarre delle *tavole numeriche* e di predisporre *programmi* per il calcolo numerico degli integrali.

Per le necessità pratiche di calcolo, quindi, tutti i programmi di statistica per computer più comuni dispongono di funzioni per il calcolo numerico di questi integrali. In mancanza di un computer si può ricorrere alle *tavole* che riportano i risultati dei calcoli più comunemente utilizzati.

V.c. Normale standardizzata e tavole

Le tavole della distribuzione Normale (o i computer; ad esempio la funzione DISTRIB.NORM.ST di Excel) consentono di risolvere il seguente problema

$$P(Z \leq a) = ?$$

dove Z indica la Normale standardizzata.

Altri problemi

Qualora si debba calcolare la probabilità per un intervallo di forma diversa, si applicano le seguenti regole

$$P(a < Z \leq b) = P(Z \leq b) - P(Z < a), \qquad P(Z > a) = 1 - P(Z \leq a)$$

Gli estremi dell'intervallo sono irrilevanti (la probabilità di un punto è pari a zero).

Tornando all'esempio precedente si ha dunque

$$P(-1 < Z \leq 1) = P(Z \leq 1) - P(Z < -1) \simeq 0.841 - 0.159 = 0.682$$

Per le altre distribuzioni normali ...

Per una variabile casuale $X \sim \mathcal{N}(\mu, \sigma^2)$ si applica l'operazione di standardizzazione:

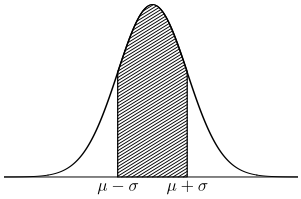
$$Z = \frac{X - \mu}{\sigma}$$

Si ha $Z \sim \mathcal{N}(0, 1)$ e

$$P(X \leq a) = P(Z \leq \frac{a - \mu}{\sigma})$$

Le situazioni seguenti si riferiscono ad una distribuzione $N(\mu, \sigma^2)$.

Alcune situazioni particolari - 1

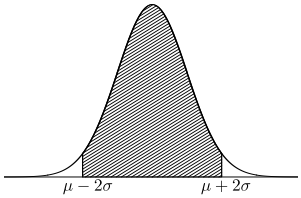


$P(\mu - \sigma \leq X \leq \mu + \sigma) \simeq 0.67$

=

$P(|Z| \leq 1) \simeq 0.68$

Alcune situazioni particolari - 2

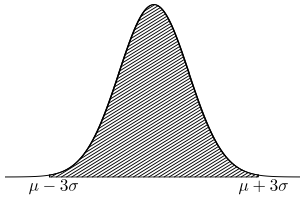


$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \simeq 0.95$

=

$P(|Z| \leq 2)$

Alcune situazioni particolari - 3



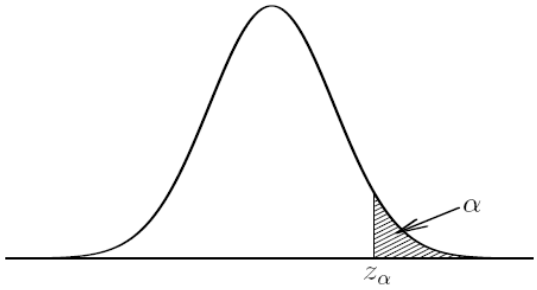
$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \simeq 0.997$

=

$P(|Z| \leq 3)$

Il problema inverso

Molto spesso si deve calcolare il valore che lascia alla sua destra (o sinistra) un'area prefissata α



Anche per questo si può ricorrere alle tavole dove il problema si trova risolto per valori tipici di α

$$\begin{array}{lcl} \alpha = 0.05 & \Rightarrow & z_\alpha = 1.6449 \\ \alpha = 0.025 & \Rightarrow & z_\alpha = 1.9600 \\ \alpha = 0.01 & \Rightarrow & z_\alpha = 2.3263 \\ \alpha = 0.005 & \Rightarrow & z_\alpha = 2.5758 \end{array}$$

oppure si può usare la funzione INV.NORM.ST di Excel.

Esempio 14.33 (Distribuzione delle altezze)**.** *L'altezza media dei maschi adulti di una certa popolazione è* $m = 175\,\text{cm}$*. Supponendo che le altezze siano distribuite con legge Normale e che lo scarto dalla media sia* $\sigma = 10\,\text{cm}$*, calcoliamo l'intervallo, centrato intorno alla media, in cui, con probabilità del 99.7% sono distribuite le altezze. Inoltre, preso a caso un individuo della popolazione, determiniamo con quale probabilità esso avrà un'altezza compresa tra 175 e 195 cm.*

Dobbiamo determinare α tale che

$$P(m - \alpha < X \leq m + \alpha) = 0.997$$

In base a quanto osservato nella "situazione particolare - 3" si ha $\alpha = 3\sigma = 30$. Dunque l'intervallo cercato è $[145, 205]$. Solo un 3 per mille della popolazione avrà quindi un'altezza inferiore a 145 cm o superiore a 205.

Osserviamo poi che

$$\begin{aligned} P(175 < X \leq 195) &= P(m \leq X \leq m + 2\sigma) = \tfrac{1}{2} P(m - 2\sigma < X \leq m + 2\sigma) \\ &= \tfrac{1}{2} 0.95 = 0.475 \end{aligned}$$

Riassumendo ...

- le v.c. sono utilizzate come modelli teorici per rappresentare fenomeni reali
- la distribuzione di probabilità di una v.c. dipende da parametri, generalmente incogniti
- i parametri rappresentano caratteristiche intrinseche del fenomeno studiato

Esercizi

Esercizi consigliati: da 11.10 a 11.14 del testo consigliato.

Capitolo aggiuntivo 15

Inferenza statistica

Spesso l'informazione a disposizione deriva da un'osservazione parziale del fenomeno studiato. In questo caso lo studio di un fenomeno mira solitamente a trarre, sulla base di ciò che si è osservato, considerazioni di carattere generale. Per sua natura il processo di inferenza è soggetto ad errore, che può essere tenuto sotto controllo, o almeno quantificato, mediante criteri e tecniche di tipo statistico.

Fasi del processo di inferenza:

- definizione del problema
- individuazione di un opportuno modello teorico
- estrazione del campione
- raccolta e analisi dei dati
- generalizzazione

Il modello

Generalmente descriviamo la distribuzione di un fenomeno mediante una opportuna distribuzione di probabilità. La forma (il tipo) della distribuzione è assunta nota, mentre sono considerati incogniti i parametri della distribuzione. In questo schema logico, i parametri (costanti caratteristiche del fenomeno studiato) sono l'oggetto di interesse del processo di inferenza (inferenza parametrica).

Il campionamento

Come selezionare il campione da osservare?

Possiamo distinguere:

- campionamento ragionato: il campione è scelto ad hoc in quanto rappresentativo della popolazione
- campionamento casuale: il campione è estratto mediante procedimenti di selezione casuale

Il campione ragionato

Il ricercatore cerca di costruire una buona "immagine" della popolazione sulla base di caratteristiche note e spera che il campione sia rappresentativo anche per le variabili oggetto di studio. È usato molto di frequente per i sondaggi, rarissimamente (mai) in ambito sperimentale. È uno strumento potente ma molto delicato (il rischio di introdurre distorsioni è elevato), inoltre è difficile quantificare l'errore. Ad esempio, se considerassimo l'altezza media di un campione di 500 individui adulti di un piccolo comune utaliano, difficilmente questo valore sarebbe espressivo dell'altezza media di tutti gli italiani.

Il campione casuale

Il campionamento dovrebbe essere sempre *casuale*, cioè ogni campione dovrebbe avere la stessa probabilità di essere scelto che hanno tutti gli altri possibili campioni della popolazione. Soddisfare questo criterio di scelta equivale a fare una "estrazione probabilistica" (ovvero "casuale") del campione, che praticamente si può realizzare nei modi seguenti

- popolazione finita: il campione viene estratto mediante etichettatura e sorteggio;
- popolazione infinita: le osservazioni campionarie (dati) derivano dalla ripetizione dell'esperimento casuale nelle medesime condizioni. È il caso tipico degli esperimenti scientifici. Esempio: lancio di una moneta infinite volte (risultati possibili infiniti; un campione di dimensione N è costituito da N lanci ripetuti nelle medesime condizioni).

Osserviamo che

- la casualità non garantisce la rappresentatività del campione: la procedura è rappresentativa, non necessariamente il campione estratto;

- il grado di rappresentatività del campione non è determinabile, ma è possibile stimare l'errore dovuto al campionamento.

Il campione casuale semplice

Definizione 15.1. *Un* campione casuale semplice *di* dimensione (o numerosità) N *è una* N*-upla di v.c.* X_1*, ...,* X_N *(i cui valori sono detti* osservazioni o determinazioni campionarie o dati*)*

- indipendenti*, cioè tali che per ogni scelta di intervalli* I_1*, ...,* I_N *si ha*

$$P(X_1 \in I_1,\, X_2 \in I_2,\, ..., X_N \in I_N) = \\ = P(X_1 \in I_1)P(X_2 \in I_2)\cdots P(X_N \in I_N),$$

- identicamente distribuite*, cioè*

$$X_i \sim X, \quad i = 1, \ldots, N$$

dove X *è una distribuzione adottata come modello per la popolazione.*

Sintesi dell'informazione campionaria

L'informazione campionaria può essere sintetizzata mediante gli indici sintetici già visti in statistica descrittiva. In particolare, possiamo definire:

- $m_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} X_i$ media campionaria

- $S_N^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N} |X_i - m_N|^2$ varianza campionaria corretta*

Entrambe le quantità sono v.c., in quanto somme e prodotti di v.c., e variano (cioè assumono valori) nell'universo dei campioni da cui selezioniamo in modo casuale.

*Sul testo chiamata semplicemente "varianza campionaria".

Distribuzioni campionarie

Indicando con μ e σ^2 rispettivamente media e varianza della popolazione, cioè di X (e coincidenti quindi con media e varianza delle X_i), è possibile dimostrare che, qualunque sia la numerosità del campione,

1. $E[m_N] = \mu$, cioè la media campionaria è uno *stimatore non distorto* della media della popolazione
2. $\mathrm{Var}[m_N] = \sigma^2/N$ (importantissimo)
3. $E[S_N^2] = \sigma^2$, cioè la varianza campionaria corretta è uno *stimatore non distorto* della varianza della popolazione

In particolare, la 1. segue dal fatto che le X_i sono identicamente distribuite e quindi hanno la stessa media. Infatti per la proprietà di linerità

$$E(\frac{1}{N}\sum_{i=1}^{N} X_i) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu.$$

La dimostrazione di 2. è molto più complicata e gioca un ruolo essenziale il fatto che le variabili X_1, ..., X_n sono indipendenti, da cui segue che la varianza della somma è uguale alla somma delle varianze (vedi ad esempio Proposizione 11.12 e Corollario 11.5 di [NPA]).

Osservazione 15.2. Ricordiamo che μ e σ^2 sono incognite da determinarsi.

Dato un campione X_1, ..., X_N potremmo pensare di ottenerle da 1 e 3 calcolando i valori attesi. Purtroppo per fare questo calcolo occorrerebbe conoscere la distribuzione delle v.c. X_1, ..., X_N che è incognita al pari di quella della popolazione.

Ci viene in aiuto, a questo punto la 2 che dice che se il campione è sufficientemente grande allora $\mathrm{Var}[m_N]$ è molto piccola, ma questo significa che in tal caso m_N è con probabilità prossima ad 1 coincidente col proprio valore atteso, cioè μ! In sostanza la 2 dice che se il campione è molto numeroso allora la media della popolazione μ si può stimare con la media campionaria m_N.

Questa osservazione è resa più precisa dal seguente teorema

Teorema 15.3 (Legge gei grandi numeri)**.** *Per ogni $\eta > 0$ si ha*

$$\lim_{N\to\infty} P(|m_N - \mu| > \eta) = 0$$

Esercizi

Da 12.6 a 12.9 del testo.

La legge dei grandi numeri

Esempio 15.4 (lancio ripetuto di una moneta)**.** Schematizziamo l'esperimento con una successione di v.c. X_i che valgono 1 se esce testa e 0 se esce croce. Supponiamo che la moneta non sia truccata. Nel caso del singolo lancio si ha

$$P(X_i = k) = \frac{1}{2}, \qquad k = 0, 1$$

quindi le X_i sono equidistribuite con distribuzione binomiale *uniforme*, e si ha

$$E(X_i) = \sum_{k=0}^{1} kP(X_i = k) = \frac{1}{2}$$
$$\mathrm{Var}(X_i) = \sum_{k=0}^{1}(k - \frac{1}{2})^2 P(X_i = k) = \frac{1}{2}\sum_{k=0}^{1}\frac{(2k-1)^2}{4} = \frac{1}{4}$$

Un campione di dimensione $N = 1$ è costituito da una sola v.c. X_1 con distribuzione binomiale uniforme. Si ha in tal caso

$$m_1 = X_1, \quad E(m_1) = E(X_1) = \frac{1}{2} = \mu$$

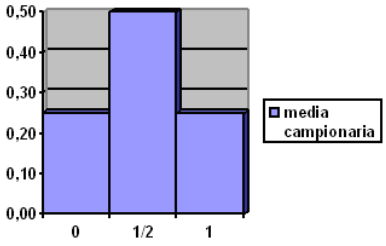
ma, evidentemente m_1, assumendo solo i valori 0 o 1, non assume mai il valore $\mu = \frac{1}{2}$.

Un campione di dimensione $N = 2$, è costituito da due v.c. X_1 e X_2 con distribuzione discreta uniforme. Si ha in tal caso

$$m_2 = \frac{X_1 + X_2}{2}, \quad E(m_2) = \frac{E[X_1] + E[X_2]}{2} = \frac{1}{2} = \mu$$

La distribuzione di m_2, cioè $f(x) = P(\frac{X_1+X_2}{2} = x)$, è la seguente

x	coppie a media x	$f(x)$
0	$(0,0)$	$f(0)=\frac{1}{4}$
$\frac{1}{2}$	$(1,0)$ $(0,1)$	$f(\frac{1}{2})=\frac{1}{2}$
1	$(1,1)$	$f(1)=\frac{1}{4}$
	Tot. 4	



Si osserva che

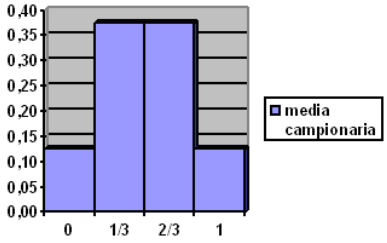
- la distribuzione della media non è più uniforme, cioè $P(\frac{X_1+X_2}{2}=x)$ non è più costante;
- la media campionaria più probabile coincide con la media di popolazione $\mu=1/2$;
- risulta (esercizio) $\mathrm{Var}[m_2]=1/8$.

Un campione di dimensione $N=3$, è costituito da 3 v.c. X_1, X_2 e X_3. Si ha in tal caso

$$m_3=\frac{X_1+X_2+X_3}{3},\quad E(m_3)=\frac{1}{2}=\mu$$

La distribuzione di m_3, cioè $f(x)=P(\frac{X_1+X_2+X_3}{3}=x)$, è la seguente

x	terne a media x	$f(x)$
0	$(0,0,0)$	$f(0)=\frac{1}{8}$
$\frac{1}{3}$	$(1,0,0)$ $(0,1,0)$ $(0,0,1)$	$f(\frac{1}{3})=\frac{3}{8}$
$\frac{2}{3}$	$(1,1,0)$ $(1,0,1)$ $(0,1,1)$	$f(\frac{2}{3})=\frac{3}{8}$
1	$(1,1,1)$	$f(1)=\frac{1}{8}$
	Tot. 8	



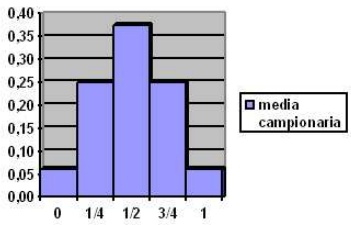
Osserviamo che $\mathrm{Var}[m_3]=1/12$, quindi è diminuita la dispersione.

Un campione di dimensione $N=4$, è costituito da 4 v.c. X_1, X_2, X_3 e X_4. Si ha in tal caso

$$m_4=\frac{X_1+X_2+X_3+X_4}{4},\quad E(m_4)=\frac{1}{2}=\mu$$

La distribuzione di m_4, cioè $f(x)=P(m_4=x)$, è la seguente

x	quaterne a media x	$f(x)$
0	$\binom{4}{0} = 1$	$f(0) = \frac{1}{16}$
$\frac{1}{4}$	$\binom{4}{1} = 4$	$f(\frac{1}{4}) = \frac{1}{4}$
$\frac{1}{2}$	$\binom{4}{2} = 6$	$f(\frac{1}{2}) = \frac{3}{8}$
$\frac{3}{4}$	$\binom{4}{3} = 4$	$f(\frac{3}{4}) = \frac{1}{4}$
1	$\binom{4}{4} = 1$	$f(1) = \frac{1}{16}$
	Tot. 16	

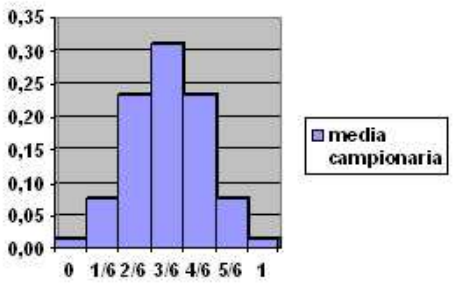


Osserviamo che

- la media campionaria maggiormente probabile corrisponde alla media di popolazione $\mu = 1/2$
- $\mathrm{Var}[m_4] = 1/16$, quindi, come si vede anche bene dal grafico, la dispersione è ulteriormente diminuita
- la distribuzione della media comincia ad avere un andamento a campana.

Il caso $N = 6$:

x	6-uple a media x	$f(x)$
0	$\binom{6}{0} = 1$	$f(0) = \frac{1}{2^6}$
$\frac{1}{6}$	$\binom{6}{1} = 6$	$f(\frac{1}{6}) = \frac{6}{2^6}$
$\frac{2}{6}$	$\binom{6}{2} = 15$	$f(\frac{2}{6}) = \frac{15}{2^6}$
$\frac{3}{6}$	$\binom{6}{3} = 20$	$f(\frac{3}{6}) = \frac{20}{2^6}$
$\frac{4}{6}$	$\binom{6}{4} = 15$	$f(\frac{4}{6}) = \frac{15}{2^6}$
$\frac{5}{6}$	$\binom{6}{5} = 6$	$f(\frac{5}{6}) = \frac{6}{2^6}$
1	$\binom{6}{6} = 1$	$f(1) = \frac{1}{2^6}$
	Tot. 2^6	



La tendenza ad assumere una forma a campana si accentua sempre di più al crescere di N. Si ha $\mathrm{Var}[m_N] = \dfrac{\sigma^2}{N} = \dfrac{1}{4N}$ quindi la campana diventa sempre più stretta.

Ciò significa che al crescere della dimensione del campione aumenta sempre più la probabilità che la media campionaria sia vicina ad 1/2, cioè che testa e croce escano lo stesso numero di volte.

Esempio 15.5 (lancio ripetuto di un dado)**.** Indichiamo con X_i la variabile il cui valore coincide col numero uscito nel lancio i-esimo. Nel caso del singolo lancio si ha

$$P(X_i = k) = \frac{1}{6}, \qquad k = 1, ..., 6$$

quindi le X_i sono equidistribuite con distribuzione *discreta uniforme*, e si ha

$$E(X_i) = \sum_{k=1}^{6} kP(X_i = k) = \frac{1}{6}\sum_{k=1}^{6} k = \frac{1}{6}\frac{6(6+1)}{2} = \frac{7}{2},$$
$$\mathrm{Var}(X_i) = \sum_{k=1}^{6}(k - \frac{7}{2})^2 P(X_i = k) = \frac{1}{6}\sum_{k=1}^{6}\frac{(2k-7)^2}{4} = \frac{35}{12} \simeq 2,9$$

Un campione di dimensione $N = 1$, in questo caso è costituito da una sola v.c. X_1 con distribuzione discreta uniforme. Si ha in tal caso

$$m_1 = X_1, \quad E(m_1) = E(X_1) = \frac{7}{2} = \mu$$

ma, evidentemente m_1, assumendo solo valori interi (k) non assume mai il valore $\mu = \frac{7}{2}$.

Nel caso di due lanci il risultato è espresso dalla variabile $X_1 + X_2$ che assume valori interi tra $x = 2$ e $x = 12$, ma questi non sono più equiprobabili. Infatti la situazione si può schematizzare nel modo seguente

x	coppie di somma x	$f(x) = P(X_1 + X_2 = x)$
2	$(1,1)$	$f(2) = 1/36$
3	$(1,2)$ $(2,1)$	$f(3) = 2/36$
4	$(1,3)$ $(2,2)$ $(3,1)$	$f(4) = 3/36$
5	$(1,4)$ $(2,3)$ $(3,2)$ $(4,1)$	$f(5) = 4/36$
6	$(1,5)$ $(2,4)$ $(3,3)$ $(4,2)$ $(5,1)$	$f(6) = 5/36$
7	$(1,6)$ $(2,5)$ $(3,4)$ $(4,3)$ $(5,2)$ $(6,1)$	$f(7) = 6/36 = 1/6$
8	$(2,6)$ $(3,5)$ $(4,4)$ $(5,3)$ $(6,2)$	$f(8) = 5/36$
9	$(3,6)$ $(4,5)$ $(5,4)$ $(6,3)$	$f(9) = 4/36$
10	$(4,6)$ $(5,5)$ $(6,4)$	$f(10) = 3/36$
11	$(5,6)$ $(6,5)$	$f(11) = 2/36$
12	$(6,6)$	$f(12) = 1/36$
	Tot. 36	

Per quanto riguarda la distribuzione della media $\frac{X_1+X_2}{2}$, la situazione è la seguente

x	coppie a media x	$f(x) = P(\frac{X_1+X_2}{2} = x)$
1	$(1,1)$	$f(1) = 1/36$
3/2	$(1,2)$ $(2,1)$	$f(3/2) = 2/36$
2	$(1,3)$ $(2,2)$ $(3,1)$	$f(2) = 3/36$
5/2	$(1,4)$ $(2,3)$ $(3,2)$ $(4,1)$	$f(5/2) = 4/36$
3	$(1,5)$ $(2,4)$ $(3,3)$ $(4,2)$ $(5,1)$	$f(3) = 5/36$
7/2	$(1,6)$ $(2,5)$ $(3,4)$ $(4,3)$ $(5,2)$ $(6,1)$	$f(7/2) = 6/36 = 1/6$
4	$(2,6)$ $(3,5)$ $(4,4)$ $(5,3)$ $(6,2)$	$f(4) = 5/36$
9/2	$(3,6)$ $(4,5)$ $(5,4)$ $(6,3)$	$f(9/2) = 4/36$
5	$(4,6)$ $(5,5)$ $(6,4)$	$f(5) = 3/36$
11/2	$(5,6)$ $(6,5)$	$f(11/2) = 2/36$
6	$(6,6)$	$f(6) = 1/36$
	Tot. 36	

Si nota che la distribuzione della media non è più uniforme, cioè $P(\frac{X_1+X_2}{2} = x)$ non è più costante. Si nota anche che il risultato maggiormente probabile corrisponde alla media $\mu = 7/2$).

Continuando con un campione di dimensione $N > 2$, come abbiamo fatto nel caso dei lanci della moneta si noterebbe che la distribuzione della media comincia ad assumere una forma a campana. Un esperimento simulato al computer di lancio di dadi si trova sul sito
http://www.stat.sc.edu/∼west/javahtml/CLT.html

Il Teorema del Limite Centrale

Il fenomeno di convergenza della distribuzione delle medie ad una distribuzione Normale osservato negli esempi precedenti è del tutto generale e riassunto nel *Teorema del Limite Centrale* (TLC). Esso afferma che, sotto opportune condizioni abbastanza generali (la più forte è l'indipendenza), la standardizzata della distribuzione della media di variabili casuali aventi tutte la medesima distribuzione, converge, in un senso che andrebbe meglio precisato, alla distribuzione $\mathcal{N}(0,1)$ quando la numerosità tende ad infinito.

Vale a dire

$$\frac{m_N - \mu}{\sigma/\sqrt{N}} \to \mathcal{N}(0,1), \quad \text{per } N \to \infty.$$

Importanza del Teorema del Limite Centrale

Il TLC è importantissimo, perché ci consente di utilizzare la distribuzione Normale anche quando la popolazione non è distribuita normalmente, purché il campione sia sufficientemente grande. Non esiste una regola per stabilire quando l'approssimazione basata sul TLC è buona: in alcuni casi anche poche osservazioni sono sufficienti, mentre in altri la numerosità campionaria deve essere dell'ordine delle centinaia.

Le applicazioni alla statistica si basano sul seguente **principio**: [†]

se X_1, X_2, ..., X_N sono v.c. che rappresentano i dati di un campione di dimensione N estratto da una popolazione con media (di popolazione) μ e varianza σ^2, la media campionaria m_N è distribuita, approssimativamente, come una variabile aleatoria gaussiana di media μ e varianza σ^2/N, cioè

$$m_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

(formula appunto solo "approssimativamente" vera, perché in effetti m_N potrebbe anche essere discreta, come visto negli esempi precedenti).

Servendoci delle relazioni di pagina 55, si ha

$$|m_N - \mu| \leq \frac{\sigma}{\sqrt{N}} \text{ con probabilità } 0.682$$
$$|m_N - \mu| \leq \frac{2\sigma}{\sqrt{N}} \text{ con probabilità } 0.954$$
$$|m_N - \mu| \leq \frac{3\sigma}{\sqrt{N}} \text{ con probabilità } 0.997$$

e la stima di μ con m_N diventa via via più accurata al crescere di N.

[†]Si ha, in generale, che se X ha media μ e varianza σ^2 allora

$$m_N - \mathcal{N}(\mu, \sigma^2/N) \to 0.$$

Infatti, se $Y \sim \mathcal{N}(\mu, \sigma^2/N)$ allora la standardizzata $\frac{Y-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$, e quindi

$$\frac{m_N - \mu}{\sigma/\sqrt{N}} - \frac{Y-\mu}{\sigma/\sqrt{N}} \to 0$$

perché ambo i termini a primo membro tendono allo stesso limite, il primo per il TLC e il secondo in quanto costante; d' altra parte, semplificando si ha

$$\sqrt{N}(m_N - Y) \to 0$$

e quindi $m_M - Y = \frac{1}{\sqrt{N}}\sqrt{N}(m_N - Y) \to 0$, c.v.d.

Stima e test delle ipotesi

Il problema di inferenza può essere impostato in modi diversi.

- *Stima* sulla base dell'evidenza empirica: si assegna
  - un valore (stima *puntuale*)
  - un insieme di valori (stima *per intervallo*) al parametro di interesse
- *Test delle ipotesi*: si formulano ipotesi alternative sul valore del parametro di interesse e si valuta quale è maggiormente supportata dall'evidenza empirica

Stima puntuale

Il parametro incognito viene stimato mediante un'opportuna funzione dei dati campionari, detta *stimatore*.

Solitamente si usa:

- la media campionaria per stimare la media della popolazione
- la varianza campionaria per stimare la varianza della popolazione
- la frequenza relativa di successo per stimare la probabilità di successo

Stimatore e stima

La *stima* è il valore che lo stimatore assume nel campione osservato.

Lo stimatore è una v.c., la stima è un numero.

Mentre siamo in grado di valutare la qualità dello stimatore in base alle sue caratteristiche nell'universo dei campioni, non possiamo dire nulla della stima ottenuta in corrispondenza del singolo campione osservato.

In particolare, non siamo in grado, sulla base della sola stima (un numero), di valutare l'errore dovuto al campionamento.

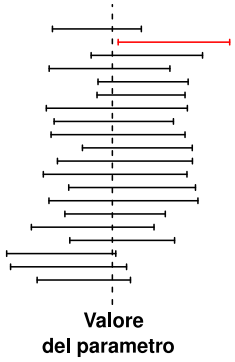
Stima per intervallo

Il parametro viene stimato mediante un intervallo (detto *intervallo di confidenza*) i cui estremi dipendono dal campione estratto (sono casuali). Un intervallo di confidenza è quindi un insieme di valori plausibili per il parametro incognito sulla base dell'evidenza empirica. Se il campione è rappresentativo (ovviamente è impossibile saperlo), allora l'intervallo contiene il valore del parametro da stimare.

Gli estremi dell'intervallo vengono individuati in modo tale che la probabilità di estrarre un campione che fornisce un risultato corretto (leggi l'intervallo contiene il valore del parametro) sia fissata pari a $1 - \alpha$ (*livello di confidenza*). Attenzione: il livello di confidenza rappresenta il grado di affidabilità della procedura, non il grado di affidabilità del risultato corrispondente al singolo campione estratto. Generalmente si usa come livello di confidenza il 95% ($\alpha = 5\%$).

Ripetendo l'operazione di stima . . .

su più campioni, potrebbe capitare la cosa seguente



18

Stima per intervallo della media

Indicando con μ e σ^2 la media e la varianza di X (incognite), una stima per intervallo del parametro μ può essere ottenuta sfruttando il fatto che:

$$\frac{m_N - \mu}{\sigma/\sqrt{N}} \to \mathcal{N}(0,1)$$

oppure

$$\frac{m_N - \mu}{S_N/\sqrt{N}} \sim t_{N-1}$$

dove t_{N-1} indica la *distribuzione t di Student con* $N-1$ *gradi di libertà.* Solitamente la varianza della popolazione è incognita (mentre la varianza campionaria S è nota) e si deve quindi necessariamente ricorrere alla seconda espressione.

La distribuzione t di Student

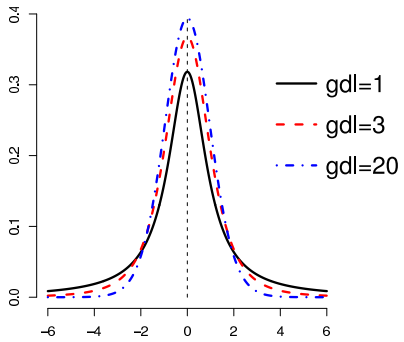
W.S. Gossett (1876-1937) ha mostrato che la variabile aleatoria

$$\frac{m_N - \mu}{S_N/\sqrt{N}}$$

ha una precisa distribuzione di probabilità detta "t di Student" in quanto lo statistico inglese firmò il proprio lavoro scientifico con lo pseudonimo "Student" (per aggirare il divieto di pubblicare, imposto dalla birreria Guinness, di cui Gossett era dipendente, per limitare lo spionaggio industriale).

La distribuzione t di Student ha un andamento simile a quello della distribuzione Normale (campanulare simmetrico). Rispetto alla Normale, la t ha le code più alte ("pesanti"), perché rappresenta una situazione di maggiore variabilità (incertezza), derivante dalla stima (soggetta quindi ad errore) della varianza della popolazione. Le tavole della distribuzione t di Student consentono di trovare $t_{N-1;\alpha}$, ossia il valore che lascia sulla coda di destra un'area prefissata α.

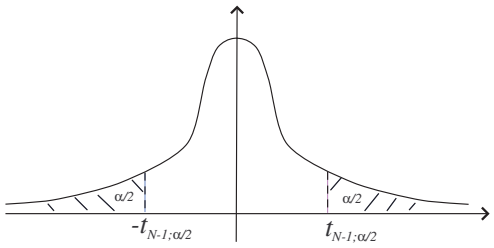
Densità della distribuzione t



gdl = gradi di libertà

L'intervallo che stima la media

Sapendo che $\dfrac{m_N - \mu}{S_N/\sqrt{N}} \sim t_{N-1}$ e che



$$P\big(\frac{m_N - \mu}{S_N/\sqrt{N}} \in [-t_{N-1;\alpha/2}, t_{N-1;\alpha/2}]\big) = 1 - \alpha,$$

si trova che l'intervallo di confidenza per la stima della media μ di una distribuzione a varianza incognita e livello di confidenza $1 - \alpha$ ha la forma seguente:

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}]$$

ossia gli estremi dell'intervallo sono dati da

$$m_N \pm \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}$$

Esempio 15.6. Si vuole stimare per intervallo ($1 - \alpha = 95\%$) la lunghezza media della spiga di una nuova varietà di mais. I valori osservati e i calcoli necessari sono riportati nella tabella che segue:

X	X^2
17.2	295.84
20.1	404.01
18.4	338.56
16.3	265.69
15.0	225.00
14.8	219.04
19.2	368.64
16.7	278.89
15.8	249.64
17.8	316.84
171.3	2962.15

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$S_{10}^2 = \frac{1}{10-1} \sum_{i=1}^{10} x_i^2 - \frac{10}{10-1} m_{10}^2 = 3.0868$$

$$t_{9;0.025} = 2.2622$$

$$[m_N - \tfrac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \tfrac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] =$$
$$= [m_{10} - \tfrac{S_{10}}{\sqrt{10}} t_{9;0.025}, m_{10} + \tfrac{S_{10}}{\sqrt{10}} t_{9;0.025}] = [15.87; 18.39]$$

L'ampiezza dell'intervallo

L'ampiezza dell'intervallo è molto rilevante. Quanto più l'intervallo è stretto, tanto maggiore è il grado di precisione che caratterizza lo strumento statistico utilizzato.

Nella stima della media, l'ampiezza dell'intervallo è pari a

$$\Delta = 2t_{N-1;\alpha/2} S_N / \sqrt{N}$$

NB: usando S_N, l'ampiezza dell'intervallo è una v.c., in quanto dipende dal campione estratto.

L'ampiezza dell'intervallo dipende quindi da

- α: al diminuire di α (al crescere del livello di confidenza $1 - \alpha$) l'ampiezza dell'intervallo aumenta
- S_N: misura la variabilità del fenomeno studiato. Al crescere della variabilità, cresce anche l'incertezza e quindi l'ampiezza dell'intervallo aumenta
- N: al crescere di N aumenta la quantità di informazione disponibile e quindi l'ampiezza dell'intervallo diminuisce

Il dimensionamento del campione

In fase di pianificazione dello studio, è importante determinare la numerosità campionaria in modo tale che gli strumenti statistici utilizzati abbiano certe

caratteristiche (per es. elevata precisione o bassa probabilità di errore). Nel caso di stima per intervallo, l'obiettivo da raggiungere si individua fissando a priori un certo grado di precisione, ossia una certa ampiezza dell'intervallo.

Dimensionamento per la stima della media

Indicando con Δ^* l'ampiezza dell'intervallo prefissata, si ottiene

$$N = \left(\frac{2t_{N-1;\alpha/2}}{\Delta^*}\right)^2 S_N^2$$

Per calcolare il valore di N bisogna risolvere due problemi:

1. S_N^2 non è nota prima di estrarre il campione
2. $t_{N-1;\alpha/2}$ dipende da N (l'espressione non è in forma chiusa)

Soluzioni:

1. usare un valore presunto per S_N^2 (indicato con S^{*2}, derivandolo da studi precedenti, indagini pilota o valutazioni di esperti
2. usare un algoritmo iterativo, calcolando ripetutamente N usando di volta in volta i gradi di libertà ottenuti al passo precedente

L'algoritmo iterativo

L'algoritmo procede nel modo seguente:

1. $N_0 = \infty$ (inizializzazione)
2. $N_1 = \left(\frac{2t_{\infty;\alpha/2}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2z_{\alpha/2}}{\Delta^*}\right)^2 S^{*2}, \quad z_{\alpha/2} = \text{coda della } \mathcal{N}(0,1)$
3. $N_2 = \left(\frac{2t_{N_1-1;\alpha/2}}{\Delta^*}\right)^2 S^{*2}$
4. .....

terminando quando si ottiene lo stesso valore in due passi successivi.

Esempio 15.7. Si vuole calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 la lunghezza media della

spiga di una nuova varietà di mais (livello di confidenza 95%). Su varietà simili si è osservata una varianza pari a 3.

$$
\begin{aligned}
N_0 &= \infty \quad \Rightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96 \\
N_1 &= \left(\tfrac{2t_{\infty;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\tfrac{2\cdot 1.96}{1.5}\right)^2 \cdot 3 = 20.49 \simeq 20 \\
N_2 &= \left(\tfrac{2t_{19;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\tfrac{2\cdot 2.093}{1.5}\right)^2 \cdot 3 = 23.36 \simeq 23 \\
N_3 &= \left(\tfrac{2t_{22;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\tfrac{2\cdot 2.0739}{1.5}\right)^2 \cdot 3 = 22.94 \simeq 23
\end{aligned}
$$

La regola di arresto è soddisfatta e possiamo quindi fermarci. Ripetendo il passo ancora una volta otterremmo lo stesso risultato.

Stima per intervallo di una probabilità

Se la popolazione è descritta mediante una distribuzione di Bernoulli (fenomeno dicotomico), il parametro da stimare è la probabilità di successo p. Se il campione è sufficientemente grande, possiamo sfruttare il TLC che ha la seguente conseguenza (vedi il principio di pagina 66)

$$m_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) = \mathcal{N}(p, \frac{p(1-p)}{N})$$

cosa solo "approssimativamente" vera, perché in effetti m_N è discreta.

In modo analogo a quanto visto per la media della Normale, otteniamo il seguente intervallo di confidenza per p (livello di confidenza $1-\alpha$)

$$\boxed{p \in \Big[m_N - z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}\Big]} \tag{15.1}$$

dove la varianza è stata stimata sostituendo a p lo stimatore m_N.

Infatti, standardizzando si ha

$$\frac{m_N - E[m_N]}{\sqrt{\mathrm{Var}[m_M]}} = \frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \sim \mathcal{N}(0,1),$$

quindi

$$P\Big(\frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \in I\Big) = P(\mathcal{N}(0,1) \in I).$$

D'altra parte $P(\mathcal{N}(0,1) \in I) = 1-\alpha$ se $I = [-z_{\alpha/2}, z_{\alpha/2}]$. Affinché $P(\frac{m_N-p}{\sqrt{p(1-p)}}\sqrt{N} \in I) = 1-\alpha$ è quindi sufficiente che

$$\frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \in [-z_{\alpha/2}, z_{\alpha/2}],$$

cioè che

$$-z_{\alpha/2} \leq \frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \leq z_{\alpha/2}.$$

Per determinare un intervallo di confidenza per p è dunque sufficiente risolvere quest'ultimo sistema di disuguaglianze nell'incognita p. Il problema si semplifica sostituendo il denominatore $\sqrt{p(1-p)}$ con $\sqrt{m_N(1-m_N)}$.

Esempio 15.8. *Volendo valutare l'effetto della conservazione in atmosfera modificata dell'insalata, su 200 confezioni è stata rilevata la presenza di foglie avvizzite dopo 5 giorni trascorsi in un banco frigo. Si sono osservate 158 confezioni integre, mentre 42 presentano segni di degrado. Se $X = 1$ se la confezione è integra e $X = 0$ altrimenti, allora $X \sim BI(1,p)$ dove p rappresenta la probabilità che una confezione si mantenga integra. Problema: determinare un intervallo di confidenza per p con livello di confidenza del 95%.*

Stima puntuale

$$m_N = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{200}158 = 0.79$$

Stima intervallare

$$\alpha = 0.05, \quad z_{\alpha/2} = 1.96,$$

$$\left[m_N - z_{\alpha/2}\sqrt{\tfrac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\tfrac{m_N(1-m_N)}{N}}\right] =$$
$$= \left[0.79 - 1.96\sqrt{\tfrac{0.79(1-0.79)}{200}}, 0.79 + 1.96\sqrt{\tfrac{0.79(1-0.79)}{200}}\right] = [0.7335, 0.8465]$$

Vediamo ora un paio di esempi in un campo nel quale gli intervalli di confidenza rivestono particolare interesse, quello dei sondaggi di opinione.

Esempio 15.9. *100 persone vengono intervistate su come voteranno ad un referendum: 42 dichiarano di votare NO e 58 dichiarano di votare SI. Determiniamo un'intervallo di confidenza al 95% per la percentuale di SI al referendum.*

Si ha $m_{100} = 58/100 = 0.580$ e $\alpha = 0.05$). Dunque

$$\begin{aligned} p \in & \left[m_N - z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}\right] = \\ & = \left[0.58 - 1.96\sqrt{\frac{0.58 \cdot 0.42}{100}}, 0.58 + 1.96\sqrt{\frac{0.58 \cdot 0.42}{100}}\right] = \\ & = \Big[0.580 - 1.960 \cdot 0.049, 0.580 + 1.960 \cdot 0.049\Big] = [0.484, 0.670] \end{aligned}$$

Il risultato non da risposte conclusive sull'esito del referendum. Con un livello di confidenza del 99% si avrebbe $\bar{p} = [0.45, 0.71]$.

Esempio 15.10. *1000 persone vengono intervistate su come voteranno ad un referendum: 420 dichiarano di votare NO e 580 dichiarano di votare SI. Determiniamo un'intervallo di confidenza al* 95% *per la percentuale di SI al referendum.*

Si ha $m_{1000} = 580/1000 = 0.58$ e $\alpha = 0.05$. Pertanto

$$\begin{aligned} p \in & \left[m_N - z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}\right] = \\ & = \left[0.58 - 1.96\sqrt{\frac{0.58 \cdot 0.42}{1000}}, 0.58 + 1.96\sqrt{\frac{0.58 \cdot 0.42}{1000}}\right] = \\ & = \Big[0.58 - 1.96 \cdot 0.016, 0.58 + 1.96 \cdot 0.016\Big] = [0.549, 0.611] \end{aligned}$$

Esercizio 15.11. *In relazione all'esempio precedente, calcolare quanto deve essere grande N per essere sicuri al* 99% *che vinceranno i SI, se si osserva una frequenza del* 58% *di SI sul campione.*

Stima per intervallo della varianza

Supponendo che $X \sim \mathcal{N}(\mu, \sigma^2)$, una stima per intervallo del parametro σ^2 può essere ottenuta sfruttando il fatto che:

$$\frac{(N-1)S_N^2}{\sigma^2} \sim \chi^2_{N-1}$$

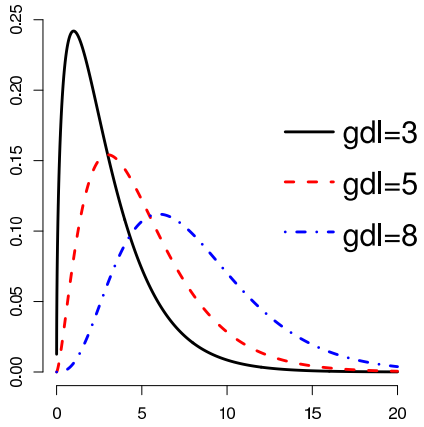
dove χ^2_{N-1} indica la distribuzione χ^2 (chi quadro) con $N-1$ gradi di libertà.

La distribuzione χ^2

La v.c. χ^2 assume valori nell'intervallo $[0, +\infty[$ ed ha distribuzione *asimmetrica*.

Le tavole della distribuzione χ^2 consentono di determinare $\chi^2_{N-1;\alpha}$, ossia il valore che lascia sulla coda di destra un'area prefissata α.

Densità della distribuzione χ^2



36

L'intervallo che stima la varianza

L'intervallo di confidenza per la stima della varianza a livello di confidenza $1-\alpha$ ha la forma seguente:

$$\left[\frac{(N-1)S_N^2}{\chi^2_{N-1;\frac{\alpha}{2}}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\frac{\alpha}{2}}}\right]$$

Mentre l'intervallo per la media è simmetrico rispetto alla media campionaria, quello per la varianza è asimmetrico.

Esempio 15.12. Nella tabella seguente sono riportati i risultati ottenuti da un tecnico in 10 misurazioni della concentrazione di un certo principio attivo in una soluzione. Stimare per intervallo $(1-\alpha = 95\%)$ la varianza

delle misure prodotte dal tecnico.

X	X^2
14.8	219.04
14.7	216.09
14.8	219.04
15.0	225.00
14.6	213.16
14.7	216.09
14.5	210.25
14.8	219.04
14.8	219.04
14.7	216.09
147.4	2172.84

Stima puntuale della varianza

$$N = 10, \quad m_N = \frac{1}{N}\sum_{i=1}^{N} x_i = 14.74, \quad S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{N}{N-1} m_N^2 = 0.0182$$

Stima per intervallo della varianza

$$\chi^2_{9;0.025} = 19.0228, \quad \chi^2_{9;0.975} = 2.7004$$

$$\left[\frac{(N-1)S_N^2}{\chi^2_{N-1;\alpha/2}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\alpha/2}}\right] = \left[\frac{9 \cdot 0.0182}{\chi^2_{9;0.025}}, \frac{9 \cdot 0.0182}{\chi^2_{9;0.975}}\right] = [0.0086, 0.0607]$$

Esercizi

Esercizio 15.13 (12.10 del testo)**.** *Si sospetta che un campo di mais sia stato contaminato da semi transgenici oltre la soglia dello* 0.1%*. Superata questa soglia è obbligatorio dichiarare la percentuale di OGM presente nelle farine ricavate dal mais.*

Viene analizzato un campione di 8000 *semi, di cui* 6 *risultano della varietà transgenica. A un livello di fiducia del* 95%*, qual'è l'intervallo di confidenza della frazione di semi transgenici sul totale della piantagione.*

Si chiede di stimare per intervallo con un livello di confidenza del 95% ($\alpha = 0.05$) la probabilità che un seme sia transgenico.

Stima puntuale

$$m_N = \frac{6}{8000} = \frac{3}{4000} \simeq 0.00075$$

Stima intervallare della probabilità

$$\begin{aligned} m_N \pm z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}} &= 0.00075 \pm 1.96\sqrt{\frac{0.00075(1-0.00075)}{8000}} \\ &= 0.00075 \pm 0.0006 \end{aligned}$$

quindi

$$p \in [0.00015, 0.00135].$$

Poiché lo 0.1% corrisponde a $p = 0.001$ l'estremo superiore di questo intervallo è superiore al limite di legge.

Capitolo aggiuntivo 16

Test delle ipotesi

Il test delle ipotesi consente di verificare se, e quanto, una determinata ipotesi (di carattere biologico, medico, economico,...) è supportata dall'evidenza empirica. Il fenomeno studiato deve essere rappresentato mediante una distribuzione di probabilità e l'ipotesi sulle caratteristiche del fenomeno studiato è tradotta in ipotesi su uno o più parametri della distribuzione (test parametrico).

Test statistico e verifica di ipotesi

Cominciamo col mostrare un semplice esempio.

Esempio 16.1. Nel lancio di una moneta si vince se esce testa e si perde se esce croce. Il lanciatore garantisce che la moneta non è truccata. Prima di giocare stiamo un po' a vedere e osserviamo che su 20 lanci esce testa solo 6 volte, un numero un po' basso rispetto alla media teorica di 10.

Ci chiediamo se il lanciatore ci sta ingannando o se il valore osservato sia un ragionevole frutto del caso. Al di là di risposte soggettive ed opinabili, è possibile attuare un *test statistico* per decidere se denunciare o meno il lanciatore.

Scopo del test è verificare se il dato osservato sia probabilisticamente credibile, assumendo che la moneta non sia truccata.

La probabilità che in 20 lanci esca testa solo 6 volte è data da (usando ad esempio il fatto che la distribuzione è binomiale $B(20, 1/2)$)

$$P(X = 6) = \frac{\binom{20}{6}}{2^{20}} \simeq 0.037$$

Dunque il risultato è decisamente poco probabile. Questo però non ci autorizza a concludere nulla; infatti tutti i valori di $P(X = k)$ con $k = 0, ..., 20$ sono piuttosto piccoli. In particolare, ad esempio

$$P(X = 10) = \frac{\binom{20}{10}}{2^{20}} \simeq 0.18$$

ed in effetti, su 20 lanci ci aspettiamo che testa esca "circa" 10 volte, e non "esattamente" 10.

Pensandoci bene, per rispondere al nostro dubbio dobbiamo calcolare qual'è la probabilità di osservare un risultato "sospetto" o "estremo" quanto e anche più di quello osservato. In questo senso i possibili risultati *estremi* sono $X = 6, 5, 4, 3, 2, 1, 0$ e anche $X = 14, 15, 16, 17, 18, 19, 20$ (infatti 6 e 14 sono alla stessa distanza dal valore atteso $E = 10$). La probabilità complessiva di questi eventi è

$$p = P(X = 0) + \cdots + P(X = 6) + P(X = 14) + \cdots + P(X = 20) \simeq 0.12$$

Quindi, assumendo la moneta non truccata, la probabilità di ottenere un risultato estremo è del 12%, una percentuale abbastanza alta per non avere forti dubbi che la moneta sia truccata.

Lo stesso risultato su 30 lanci avrebbe dato

$$p = P(X = 0)+\cdots+\cdots+P(X = 6)+P(X = 24)+\cdots+P(X = 30) \simeq 0.0014$$

In questo caso il sospetto che la moneta sia truccata sarebbe seriamente fondato.

Definizione 16.2. *Si chiama* test statistico *ogni procedura atta a verificare se un dato è in accordo con una teoria e si articola nelle seguenti fasi:*

- *formulazione dell'ipotesi da verificare, detta* ipotesi nulla *e indicata con H_0;*
- *calcolo della probabilità p di ottenere un risultato estremo come e più di quello osservato, nell'ipotesi che H_0 sia vera; p è detta* valore p del test *o* p-value*,*
- *valutazione di p; se p è troppo piccolo si rifiuta l'ipotesi H_0, se è grande la si accetta.*

Osserviamo che, detto x il risultato osservato, il valore p è dato dalla formula

$$\boxed{p = P(|X - E| \geq |x - E|)}$$

Livello di significatività del test

Nella pratica statistica i valori critici di p, detti *livelli di significatività del test* sono fissati dalla seguente convenzione.

Definizione 16.3.
- *Se* $p \geq 0.05$*, la discrepanza tra dato osservato e valore atteso* non è statisticamente significativa *(cioè può trattarsi di un effetto casuale del campionamento) e* H_0 *viene accettata.*
- *Se* $p < 0.05$*,* H_0 *viene, in genere, rifiutata e in particolare*
  - *se* $0.01 \leq p < 0.05$ *la discrepanza è* statisticamente significativa*;*
  - *se* $0.001 \leq p < 0.01$ *la discrepanza è* molto significativa*;*
  - *se* $p < 0.001$ *la discrepanza è* estremamente significativa*;*

Attenzione: Il p-value non è la probabilità che H_0 sia vera (cosa che non ha senso), ma la probabilità del verificarsi di eventi estremi assumendo H_0 vera, cioè rappresenta un *livello di confidenza* del test.

Una moneta un po' truccata

Il test effettuato nell'Esempio 16.1 non ci permette di rifiutare l'ipotesi che la moneta sia non truccata. Vediamo però nel seguente esempio che non possiamo nemmeno escludere che la moneta sia *un po' truccata*, addirittura a nostro favore.

Esempio 16.4. Con riferimento all'Esempio 16.1, assumiamo come H_0 che $p(T) = 0.3$. Siccome il fenomeno è descritto dalla binomiale $B(N, p(T))$ con $N = 20$ allora si ha

$$E = N \cdot p(T) = 20 \cdot 0.3 = 6$$

Siccome tutti i valori di k da 0 a 20 sono estremi come o più di 6, si avrà $p = 1$. Siccome $p > 0.05$ non possiamo escludere quindi nemmeno che la moneta sia molto truccata a favore del lanciatore.

Supponiamo ora che $p(T) = 0.52$. Ipotizziamo cioè che la moneta sia un po' truccata in modo che la probabilità dell'uscita di testa sia un po' più grande (0.52) di quella dell'uscita di croce. Il trucco in questo caso sarebbe quindi a nostro favore.

Eseguiamo i conti fatti nell'esempio precedente, ma usando questa volta la distribuzione $B(20, 0.52)$ (anziché la $B(20, 1/2)$). Il valore atteso è $E = Np(T) = 20 \cdot 0.52 = 10.4$ e $10.4 - 6 = 4.4$, quindi sono da considerare

estremi i casi $k = 6,5,4,3,2,1,0$ e $k = 20,19,18,17,16,15$ ($k = 14$ no perché $14 - 10.4 = 3.6 < 4.4$). Si ha allora

$$\begin{aligned} p &= \sum_{k=0}^{6} \binom{20}{k} 0.52^k 0.48^{20-k} + \sum_{k=15}^{20} \binom{20}{k} 0.52^k 0.48^{20-k} \\ &= 0.48^{20} + 20 \cdot 0.52 \cdot 0.48^{19} + 190 \cdot 0.52^2 \cdot 0.48^{18} + 1140 \cdot 0.52^3 \cdot 0.48^{17} \\ &\quad +4845 \cdot 0.52^4 \cdot 0.48^{16} + 15504 \cdot 0.52^5 \cdot 0.48^{15} + 38760 \cdot 0.52^6 \cdot 0.48^{14} \\ &\quad +0.52^{20} + 20 \cdot 0.48 \cdot 0.52^{19} + 190 \cdot 0.48^2 \cdot 0.52^{18} + 1140 \cdot 0.48^3 \cdot 0.52^{17} \\ &\quad +4845 \cdot 0.48^4 \cdot 0.52^{16} + 15504 \cdot 0.48^5 \cdot 0.52^{15} \\ &\simeq 0.07 \end{aligned}$$

calcolo che può anche essere eseguito automaticamente nel modo seguente

$$\begin{aligned} p &= P(X \leq 6) + P(X \geq 15) = P(X \leq 6) + (1 - P(X \leq 14)) \\ &= DISTRIB.BINOM(6; 20; 0,52; 1) + \\ &\quad +(1 - DISTRIB.BINOM(14; 20; 0,52; 1)) \\ &\simeq 0.04 + (1 - 0.97) = 0.04 + 0.03 = 0.07 \end{aligned}$$

Poichè $p > 0.05$ dobbiamo accettare anche che la moneta possa essere un po' truccata addirittura a nostro favore.

Con $p(T) = 0.6$ si ha $E = 20 \cdot 0.6 = 12$ e dunque sono estremi i valori $k = 0 - 6$ e $k = 18, 19, 20$. Si ha in tal caso

$$\begin{aligned} p &= P(X \leq 6) + P(X \geq 18) = P(X \leq 6) + (1 - P(X \leq 17)) \\ &= DISTRIB.BINOM(6; 20; 0,6; 1) + \\ &\quad +(1 - DISTRIB.BINOM(17; 20; 0,6; 1)) \\ &\simeq 0.01 \end{aligned}$$

che indica un notevole scostamento dal valore atteso. In tal caso la discrepanza è statisticamente significativa e l'ipotesi $p(T) = 0.6$ va rifiutata.

Z-test

Consente di effettuare il calcolo del p-value in maniera *approssimata* ma molto *più veloce*. L'idea è di approssimare la distribuzione discreta con una Normale. Siccome si usa il TLC, il risultato sarà tanto più accurato quanto più è alto il numero di prove N.

Indicato con N il numero di prove, k il numero di successi e q il valore ipotizzato del parametro (H_0), siccome

$$X \sim \mathcal{N}\big(Nq, Nq(1-q)\big)$$

allora, standardizzando, si ha

$$Z = \frac{X - Nq}{\sqrt{Nq(1-q)}} \sim \mathcal{N}(0,1)$$

quindi

$$p = P(|X - E| \geq |k - E|) \tag{16.1}$$
$$= P(|X - Nq| \geq |k - Nq|) \tag{16.2}$$
$$= P(|Z| \geq \frac{|k - Nq|}{\sqrt{Nq(1-q)}}) \tag{16.3}$$

In definitiva

$$\boxed{p = P\big(|Z| \geq \frac{|k - Nq|}{\sqrt{Nq(1-q)}}\big)}$$

dove $Z \sim \mathcal{N}(0,1)$. Il valore $s = \dfrac{|k - Nq|}{\sqrt{Nq(1-q)}}$ con cui va confrontata la normale è detto *statistica del test*.

Esempio 16.5 (Z-test su H_0 = moneta non truccata ($q = 0.5$))**.** Con $N = 20$, $k = 6$ e $q = 0.5$ (H_0) si ha

$$s = \frac{|k - Nq|}{\sqrt{Nq(1-q)}} = \frac{|6 - 10|}{\sqrt{10(1 - 0.5)}} = \frac{4}{\sqrt{5}} \simeq 1.79$$

quindi

$$p = P\big(|Z| \geq s)\big) = P\big(|Z| \geq 1.79\big) = 2P\big(Z \geq 1.79)\big)$$
$$= 2 \cdot \big(1 - P(Z \leq 1.79)\big) = 2 \cdot (1 - 0.9633) = 0.0734 > 0.05$$

quindi l'ipotesi H_0: $q = 1/2$ non si può rigettare.

Esempio 16.6 (Z-test su H_0 = moneta un po' truccata ($q = 0.52$))**.** *Con $N = 20$, $k = 6$ e $q = 0.52$ (H_0), si ha $s \simeq 1.96$ e $p = 0.05$, quindi l'ipotesi H_0: $q = 0.52$ non si può rigettare.*

Esercizi

Esercizio 16.7 (12.11 del testo)**.** *Viene analizzato un campione di 1235 semi importati. Di essi 22 risultano transgenici. La ditta produttrice garantisce che la percentuale di semi transgenici tra i suoi prodotti è dell'*1%*. Si testi l'ipotesi nulla che la ditta affermi il vero.*

Indicata con $p(T)$ la percentuale di semi transgenici si ha

$$H_0) \qquad p(T) = 0.01$$

Possiamo procedere in 3 modi

1. calcolando il p-value con la formula binomiale
2. calcolando il p-value con lo Z-test
3. determinando un intervallo di confidenza al 95%

La quantità di calcoli da effettuare nei tre casi è molto diversa; nel primo caso il risultato è più accurato, ma è indispensabile l'ausilio di un calcolatore, mentre il secondo e il terzo sono meno precisi ma i calcoli si possono fare anche a mano.

1. Calcolo del p-value con formula binomiale.
   Si ha

$$E = N \cdot p(T) = 1235 \cdot 0.01 = 12.35, \qquad x - E = 22 - 12.35 = 9.65$$

Sono quindi da considerare estremi tutti i valori di k che distano almeno 9.65 dal valore atteso 12.35, cioè

$$k \in \{0, 2, 22, 23, \ldots, 1235\}.$$

Dunque

$$\begin{aligned} p &= P(X \leq 2) + P(X \geq 23) = P(X \leq 2) + 1 - P(X \leq 22) \\ &= DISTRIB.BINOM(22; 1235; 0,01; 1) \\ &\quad +1 - DISTRIB.BINOM(22; 1235; 0,01; 1) \\ &= 0.0004 + 1 - 0.9959 = 0.0045 \end{aligned}$$

Siccome $p < 0.05$ l'ipotesi nulla è da rifiutare. In effetti $0.001 < p < 0.01$, quindi lo scostamento dal valore medio è statisticamente *molto significativo*.

2. Calcolo del p-value con Z-test.
   Con $N = 1235$, $k = 22$ e $q = 0.01$, si ha $s \simeq 2.957$ e $p = 0.0102$. Siccome $p < 0.05$ l'ipotesi nulla è da rifiutare.

3. Intervallo di confidenza al 95%.
   Si ha

$$\bar{q} = x/N = 22/1235 \simeq 0.018, \quad z_{\alpha/2} = z_{0.025} = 1.96,$$

$$\bar{q} \pm z_{\alpha/2}\sqrt{\frac{\bar{q}(1-\bar{q})}{N}} = 0.0178 \pm 1.96\sqrt{\frac{0.0178(1-0.0178)}{1235}} = 0.018 \pm 0.006$$

quindi

$$p(T) \in [0.012, 0.024]$$

Poiché l'estremo inferiore dell'intervallo è maggiore di 0.01 l'ipotesi nulla è da rifiutare.

Esercizio 16.8. *Calcolare*

1. *media e varianza della distribuzione uniforme nell'intervallo* $[0, 1]$*;*
2. *media e varianza della distribuzione uniforme nell'intervallo* $[a, b]$*, con* $a < b$*.*

1. Per definizione X ha distribuzione uniforme in $[0, 1]$ se

$$F(x) = P(X \leq x) = x, \quad x \in [0, 1].$$

Si ha dunque $f(x) = F'(x) = 1$ e

$$\mu = E[X] = \int_0^1 x f(x)\, dx = \int_0^1 x\, dx = \Big[\frac{x^2}{2}\Big]_0^1 = \frac{1}{2}$$

mentre

$$\begin{aligned} \sigma^2 = Var[X] &= \int_0^1 |x-\mu|^2 f(x)\, dx = \int_0^1 |x - \frac{1}{2}|^2\, dx \\ &= \Big[\frac{(x-\frac{1}{2})^3}{3}\Big]_0^1 = \frac{2}{3}(\frac{1}{2})^3 = \frac{1}{12} \end{aligned}$$

2. X ha distribuzione uniforme in $[a, b]$ se

$$F(x) = P(X \leq x) = \frac{x-a}{b-a}, \quad x \in [a, b].$$

Si ha dunque $f(x) = F'(x) = \frac{1}{b-a}$ e

$$\begin{aligned} \mu = E[X] &= \int_a^b x f(x)\, dx = \frac{1}{b-a}\int_a^b x\, dx == \frac{1}{b-a}\Big[\frac{x^2}{2}\Big]_a^b \\ &= \frac{1}{b-a}\Big[\frac{b^2}{2} - \frac{a^2}{2}\Big] = \frac{1}{b-a}\frac{(b-a)(b+a)}{2} = \frac{a+b}{2} \end{aligned}$$

mentre

$$\begin{aligned}\sigma^2 = Var[X] &= \int_a^b |x-\mu|^2 f(x)\,dx = \frac{1}{b-a}\int_a^b |x-\frac{a+b}{2}|^2\,dx \\ &= \frac{1}{b-a}\Big[\frac{(x-\frac{a+b}{2})^3}{3}\Big]_a^b = \frac{1}{b-a}\frac{1}{3}\big[(\frac{b-a}{2})^3 - (\frac{a-b}{2})^3\big] \\ &= \frac{1}{b-a}\frac{1}{3}2(\frac{b-a}{2})^3 = \frac{(b-a)^2}{12}\end{aligned}$$

Esercizio 16.9 (12.8 e 12.12 del testo)**.** *Viene effettuato un test di durata su un campione casuale di 100 lampadine a incandescenza. I dati vengono raggruppati in classi secondo la seguente tabella.*

T	$Frequenza$
$(0, 2.5)$	8
$(2.5, 5)$	27
$(5, 7.5)$	15
$(7.5, 10)$	17
$(10, 12.5)$	27
$(12.5, 15)$	6

con la durata T espressa in centinaia di ore. Si sottoponga a test l'ipotesi nulla che la durata delle lampadine sia uniformemente distribuita negli intervalli di tempo considerati.

Anzitutto, osserviamo che possiamo, a tutti gli effetti sostituire ad ogni classe il punto medio, ottenendo la tabella seguente

T	$Frequenza$
1.25	8
3.75	27
6.25	15
8.75	17
11.25	27
13.75	6
45	

Una v.c. X ha distribuzione uniforme se ad ogni classe associa la medesima probabilità, cioè $1/6$, visto che le classi sono 6, cioè se la frequenza relativa è costantemente $p_k = 1/6$. Si ha allora

$$\mu = \sum_{k=1}^{6} x_k p_k = \frac{1}{6}\sum_{k=1}^{6} x_k = \frac{1}{6}45 = \frac{15}{2} = 7.5$$

La varianza è invece data da

$$\begin{aligned}\sigma^2 &= \sum_{k=1}^{6} x_k^2 p_k - \mu^2 \\ &= \frac{1}{6}[1.25^2 + 3.75^2 + 6.25^2 + 8.75^2 + 11.25^2 + 13.75^2] - 7.5^2 = 18.23\end{aligned}$$

Dunque

$$H_0) \quad \mu = 7.5,\ \sigma^2 = 18.23$$

Eseguiamo una stima intervallare su base campionaria della media e della varianza della popolazione a livello di confidenza del 95%.

Stima della media. Si ha

$$\begin{aligned}m_N &= \sum_{k=1}^{6} x_k p_k \\ &= \frac{1}{100}[1.25 \cdot 8 + 3.75 \cdot 27 + 6.25 \cdot 15 + 8.75 \cdot 17 + 11.25 \cdot 27 + 13.75 \cdot 6] \\ &= 7.4\end{aligned}$$

$$\begin{aligned}S_N^2 &= \frac{1}{N-1}\sum_{i=1}^{6} x_i^2 n_i - \frac{N}{N-1} m_N^2 \\ &= \frac{1}{99}[1.25^2 \cdot 8 + 3.75^2 \cdot 27 + 6.25^2 \cdot 15 + 8.75^2 \cdot 17 + 11.25^2 \cdot 27 + \\ &\qquad + 13.75^2 \cdot 6] - \frac{100}{99} 7.4^2 = \\ &= \frac{1}{99} 6831.25 - 55.31 = 13.69\end{aligned}$$

$$m_N \pm \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2} = 7.4 \pm \frac{\sqrt{13.69}}{\sqrt{100}} t_{99;0.025} =\simeq 7.4 \pm \frac{3.7}{10} 1.984 = 7.4 \pm 0.73$$

quindi

$$\mu \in [6.67, 8.13]$$

Stima della varianza. Si ha

$$\chi^2_{N-1;\alpha/2} = \chi^2_{99;0.025} = 129.56, \quad \chi^2_{N-1;1-\alpha/2} = \chi^2_{99;9.975} = 74.22$$

$$\Big[\frac{(N-1)S_N^2}{\chi^2_{N-1;\alpha/2}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\alpha/2}}\Big] = \Big[\frac{99 \cdot 13.69}{129.56}, \frac{99 \cdot 13.69}{74.22}\Big] = [10.46, 18.26]$$

Poiché $7.5 \in [6.67, 8.13]$ e $18.23 \in [10.46, 18.26]$ non si può rifiutare l'ipotesi che la distribuzione sia uniforme.

Vantaggi e svantaggi dello Z-test

+ i conti sono più semplici
  − è meno preciso
  − si applica a modelli che coinvolgono una sola variabile

Test di adattamento del χ^2 (Pearson)

È utile quando il modello coinvolge più di una variabile aleatoria. In questo test la **statistica** viene confrontata con la v.a. χ^2. Cominciamo dal seguente semplice esempio.

Esempio 16.10 (esperimento di Mendel)**.** Incrociando tra loro piante di piselli di due linee pure, a *fiore rosso* e a *fiore bianco*, Mendel osservò

705 piante a fiore rosso
224 piante a fiore bianco

Testiamo l'ipotesi

H_0) i dati sono in accordo con la seguente legge di Mendel

Prima legge di Mendel. La proporzione tra piante di fenotipo dominante (rosso) e fenotipo recessivo (bianco) è $3:1$ (i.e. $p(R) = 3/4$ e $p(B) = 1/4$).

Potremmo procedere in 3 modi:

1. test binomiale (esercizio)
2. Z - test (esercizio)
3. test del χ^2

Illustriamo il terzo. Consideriamo la seguente tabella

	R	B	*totale*
frequenze reali	$F_R = \mathbf{705}$	$F_B = \mathbf{224}$	929
valori attesi teorici	$E_R = 929\frac{3}{4} = \mathbf{696.75}$	$E_B = 929\frac{1}{4} = \mathbf{232.5}$	929
scarti	$F_R - E_R = \mathbf{8.25}$	$F_B - E_B = \mathbf{-8.25}$	0

La statistica del test del χ^2 è

$$s = \frac{|F_R - E_R|^2}{E_R} + \frac{|F_B - E_B|^2}{E_B} = \frac{8.25^2}{696.75} + \frac{(-8.25)^2}{232.5} \simeq 0.39$$

È da ritenere che se s è piccolo allora vi sia un buon accordo con l'ipotesi nulla. Se è grande l'ipotesi va rifiutata.

Questione: come determinare se s è grande o piccolo?

Ciò è determinato dal p-value

$$p = P(\chi_1^2 \geq s) = P(\chi_1^2 \geq 0.39) > P(\chi_1^2 \geq 1.64) = 0.2$$

dove il numero di gradi di libertà è scelto pari a 1 perché in effetti c'è realmente una sola variabile indipendente. Il valore p del test è dunque molto alto e l'ipotesi è confermata

In generale

Il test del χ^2 confronta l'accordo (o adattamento) tra frequenza osservata e frequenza attesa di dati organizzati in n categorie qualitative.

Supponiamo di estrarre da una popolazione un campione di dimensione N e di osservare nel campione le frequenze F_1, F_2, ..., F_n.

Se le frequenze relative delle diverse categorie sono q_1, q_2, ..., q_n (ipotesi nulla), i valori attesi di tali frequenze sono $E_i = Nq_i$.

La statistica del test è il numero

$$s = \sum_{i=1}^{n} \frac{|F_i - E_i|^2}{E_i}$$

e l'ipotesi nulla va valutata in relazione al p-value

$$p = P(\chi_{n-1}^2 \geq s)$$

Esempio 16.11. Per testare l'efficacia di un principio attivo si preparano 3 farmaci V_1, V_2 e V_3 dove

V_1 non contiene il principio attivo
V_2 contiene il principio in quantità q
V_3 contiene il principio in quantità $2q$

Si osservano i seguenti risultati nella sperimentazione

	V_1	V_2	V_3	*totale*
pazienti migliorati	12	5	29	46
pazienti non migliorati	114	80	90	284
totale	126	85	119	330

Frequenze reali

Sottoponiamo a test l'ipotesi nulla

$$\begin{aligned} H_0 &= \text{il farmaco è inefficace} \\ &= \text{gli eventi "miglioramento" e "assunzione del farmaco" sono indipendenti} \end{aligned}$$

Osserviamo che

$$\begin{aligned} H_0 \Rightarrow\ & \text{valore atteso teorico di pazienti migliorati che hanno assunto } V_3 \\ &= \text{probabilità che un paziente migliorato abbia assunto } V_3 \cdot 330 \\ &= \frac{46}{330} \cdot \frac{119}{330} \cdot 330 = \frac{46}{330} 119 \simeq 16.6 \end{aligned}$$

Analogamente si calcolano gli altri valori attesi "teorici" che riportiamo nella seguente tabella

	V_1	V_2	V_3
pazienti migliorati	17.6	11.8	16.6
pazienti non migliorati	108.4	73.2	102.4

Valori attesi teorici

Sottraendo alle frequenze reali i valori attesi teorici otteniamo la seguente tabella degli scarti

	V_1	V_2	V_3
pazienti migliorati	-5.6	-6.8	12.4
pazienti non migliorati	5.6	6.8	-12.4

Scarti

La statistica del test è data da

$$s = \frac{(-5.6)^2}{17.6} + \frac{(-6.8)^2}{11.8} + \frac{(12.4)^2}{16.6} + \frac{(5.6)^2}{108.4} + \frac{(6.8)^2}{73.2} + \frac{(-12.4)^2}{102.4} \simeq 17.4$$

che, confrontata con la distribuzioine χ^2 con 5 gradi di libertà (perché le variabili in gioco sono 6, come i valori attesi) fornisce il seguente valore p

$$0.001 = P(\chi^2_5 \geq 20.5) < p = P(\chi^2_5 \geq 17.4) < P(\chi^2_5 \geq 16.7) = 0.005$$

L'ipotesi H_0 va rigettata. I risultati osservati sono statisticamente molto significativi e inducono a ritenere che il farmaco abbia effetto.

Esercizi di ricapitolazione

Esercizio 16.12. *Data*

$$f(x) = \begin{cases} 0 & \text{se } x < 1/2 \\ \frac{3}{(4x-1)^{\frac{7}{4}}} & \text{se } x \geq 1/2 \end{cases}$$

i) verificare che è una funzione di densità di probabilità di una variabile aleatoria X;

ii) calcolare la probabilità $P(X > 0)$;

iii) calcolare la probabilità $P(X < 17/4)$.

R *i)* 1; *ii)* 7/8.

Esercizio 16.13. *Data*

$$f(x) = \begin{cases} 0 & \text{se } x \leq 1 \\ \frac{32}{(3x-2)^{\frac{5}{3}}} & \text{se } x > 1 \end{cases}$$

i) verificare che è una funzione di densità di probabilità di una variabile aleatoria X;

ii) calcolare la probabilità $P(X > 0)$;

iii) calcolare la probabilità $P(X < 10/3)$.

R *i)* 1; *ii)* 3/4.

Esercizio 16.14. *In una scatola ci sono 12 caramelle, 5 delle quali sono alla ciliegia. Scegliendone 8 a caso,*

i) qual'è la probabilità P_1 di averne preso solo una alla ciliegia?

ii) qual'è la probabilità P_2 di averne preso almeno una alla ciliegia?

R *i)* $P_1 = 1/99$; *ii)* $P_2 = 1$.

Esercizio 16.15. *In un cassetto ci sono 8 pile, 5 delle quali sono scariche. Scegliendone 3 a caso,*

i) qual'è la probabilità P_1 di averle prese tutte scariche?

ii) qual'è la probabilità $P2$ di averne preso solo una carica?

$\boxed{R}$ *i)* $P_1 = 5/28$; *ii)* $P_2 = 15/36$.

Esercizio 16.16. *La ditta Paneburro vende burro in panetti il cui peso è rappresentabile con una variabile aleatoria X con distribuzione normale di media $\mu = 250g$ e deviazione standard $\sigma = 1.5g$.*

i) Qual'è la probabilità $P_1 = P(X > 249)$ che un panetto di burro preso a caso pesi più di $249g$?

ii) Qual'è la probabilità $P_2 = P(250 < X < 253)$ che un panetto di burro preso a caso pesi più di $250g$ e meno di $253g$?

iii) Qual'è la probabilità $P_3 = P(X = 250)$ che un panetto di burro preso a caso pesi esattamente $250g$?

$\boxed{R}$ *i)* $P_1 \simeq 0.7486$; *ii)* $P_2 \simeq 0.47$; *ii)* $P_3 = 0$.

Esercizio 16.17. *Una macchina produce viti la cui lunghezza può essere rappresentata con una variabile aleatoria X che si distribuisce normalmente, con media $\mu = 1.62cm$ e deviazione standard $\sigma = 0.4cm$.*

i) Qual'è la probabilità $P_1 = P(X < 1.61)$ che una qualunque vite prodotta misuri meno di $1.61cm$?

ii) Su 5 viti prodotte (indipendentemente), qual'è la probabilità P_2 che esattamente 2 viti misurino più di $1.62cm$?

$\boxed{R}$ *i)* $P_1 \simeq 0.49$; *ii)* $P_2 = 5/16$.

Esercizio 16.18. *Lanciando (in modo indipendente) 10 volte due dadi,*

i) qual'è la probabilità p che la somma dei dadi dia 9 al primo lancio?

ii) qual'è la probabilità P_1 che la somma dei dadi dia 9 esattamente 4 volte sui 10 lanci?

ii) qual'è il valor medio μ della somma dei dadi sui 10 lanci?

Esercizio 16.19. *Sia X_1, X_2, X_3, X_4 un campione casuale estratto da una popolazione X distribuita secondo una legge normale con media μ e varianza σ^2, entrambe incognite.*

Determinare un intervallo di confidenza per μ e per σ^2 al livello del 95% avendo osservato il campione $x_1 = 8$, $x_2 = x_3 = x_4 = 4$.

a) Si ha

$$m_N = \frac{x_1 + x_2 + x_3 + x_4}{4} = 5,$$

$$S_N^2 = \frac{(x_1-5)^2 + (x_2-5)^2 + (x_3-5)^2 + (x_4-5)^2}{3} = 4 \Rightarrow S_N = 2.$$

$$m_N \pm \frac{S^N}{\sqrt{N}} t_{N-1;\alpha/2} = 5 \pm \frac{2}{\sqrt{2}} t_{3;0.025} = 5 \pm 3.1824$$

quindi

$$\mu \in [1.8176, 8.1824]$$

Per la varianza si ha invece

$$\begin{aligned} \sigma^2 \in \Big[\frac{(N-1)S_N^2}{\chi^2_{N-1;\frac{\alpha}{2}}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\frac{\alpha}{2}}}\Big] &= \Big[\frac{12}{\chi^2_{3;0.025}}, \frac{12}{\chi^2_{3;0.975}}\Big] \\ &= \Big[\frac{12}{9.3484}, \frac{12}{0.2158}\Big] = \Big[1.2836, 55.6070\Big] \end{aligned}$$

Esercizio 16.20. *In un acquario ci sono 4 pesciolini rossi ed uno nero. Prendendone due a caso,*

i) quale 'e la probabilit'a p che abbiano colori diversi?

ii) sapendo che uno dei due 'e rosso, qual 'e la probabilit'a q che abbiano colori diversi?