

Inferenza Statistica a.a. 2010/2011

Docente

Dott.a Daniela Nappo
daniela.nappo@unina.it

Programma del corso

À Richiami delle variabili casuali

À Richiami di inferenza (stima e stimatore, stima puntuale ed a intervallo, verifica di ipotesi)

À Alcuni test statistici:

À Test sulla varianza

À Test sulla bontà di adattamento

À Test del chi quadro per la verifica dell'indipendenza

À Test sul coefficiente di correlazione

À Test non parametrici

À Il modello ANOVA ad un fattore

La probabilità

La probabilità è un concetto primitivo, perché innato e sempre presente nelle regole di comportamento dell'essere umano;

Essa è anche una misura, perché associa al concetto primitivo una valutazione numerica;

I suoi elementi caratterizzanti sono:

a. Incertezza del risultato b. Ripetibilità dell'esperimento c. Equiprobabilità dei risultati

Abbiamo tre definizioni di probabilità:

Definizione classica (esperimento in condizioni di perfetta uniformità)

La probabilità di un evento A è il rapporto tra il numero di esiti favorevoli e il numero di esiti possibili, posto che tutti i risultati siano ugualmente possibili (si verificano tutte le condizioni riportate sopra)

$$P(A) = \frac{n_A}{n}$$

Definizione frequentista

In n esperimenti, tutti effettuati nelle medesime condizioni, la probabilità di un evento A è il limite cui tende la frequenza relativa dell'evento al crescere del numero di prove (si verificano solo le condizioni 1 e 2)

$$P(A) = \lim_{n \rightarrow \infty} \frac{fr(A)}{n}$$

Definizione soggettivista (esperimento eventi futuri)

La probabilità di un evento A è una misura del grado di fiducia che una persona ripone sul verificarsi di un dato evento, avendo a disposizione informazioni sul fenomeno. Può essere quantificato nella somma che un individuo coerente è disposto a scommettere in un gioco equo nel quale, al verificarsi di A, egli riceve dal banco un importo unitario (si verifica solo la condizione 1)

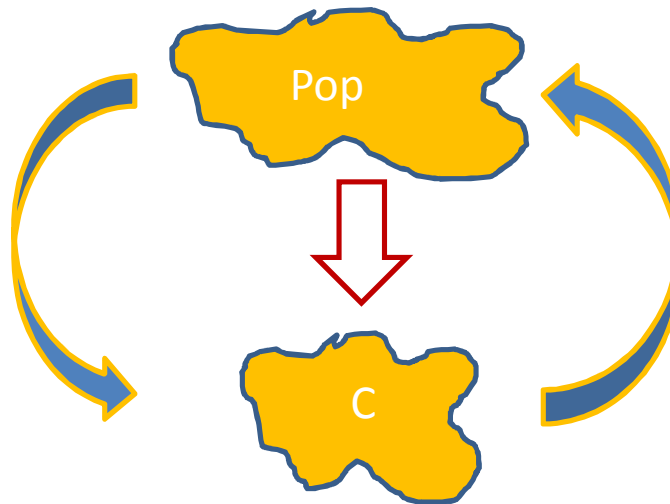
La probabilità in statistica

La teoria della probabilità consente di dedurre dal contenuto noto della popolazione il contenuto probabile del campione

Deduce, quindi, le proprietà di un processo fisico da un modello matematico.

L'inferenza statistica induce le caratteristiche della popolazione dall'analisi del contenuto del campione osservato

Inferisce, dunque, le proprietà del modello matematico a partire dall'analisi dei dati campionari osservati, andando oltre la teoria della probabilità, al fine di poter scegliere tra i modelli matematici alternativi che possono aver generato i risultati empirici osservati.



Alcune definizioni utili

Esperimento

→ Ogni operazione il cui risultato non è prevedibile
con certezza

Evento

→ Ogni risultato possibile di un evento

Spazio campionario (Ω)

→ L'insieme di tutti i possibili risultati
dell'esperimento

Eventi elementari ed eventi composti

Lancio di un dado

- $\Omega : \{1,2,\dots,6\}$ Eventi elementari
- $\Omega : \{\text{pari}; \text{dispari}\}$ Eventi composti

Algebra degli eventi

1. Somma logica o Unione

$$C = A \cup B$$

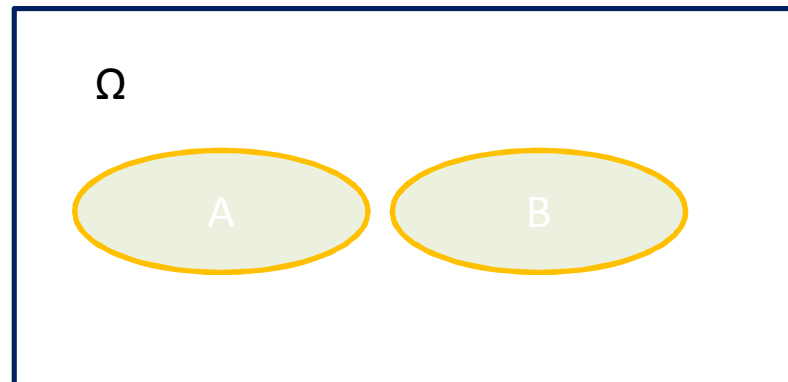
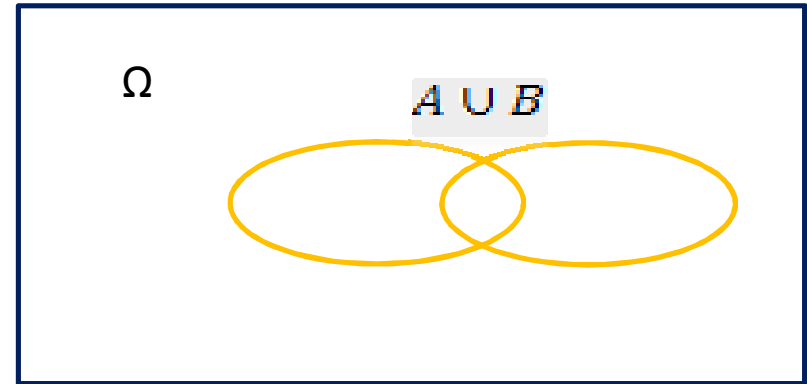
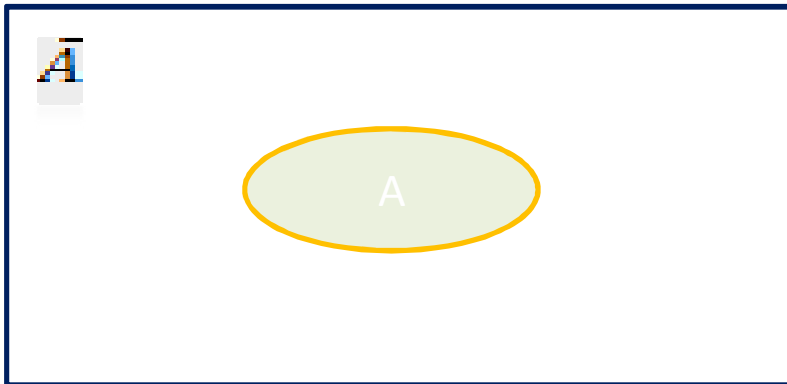
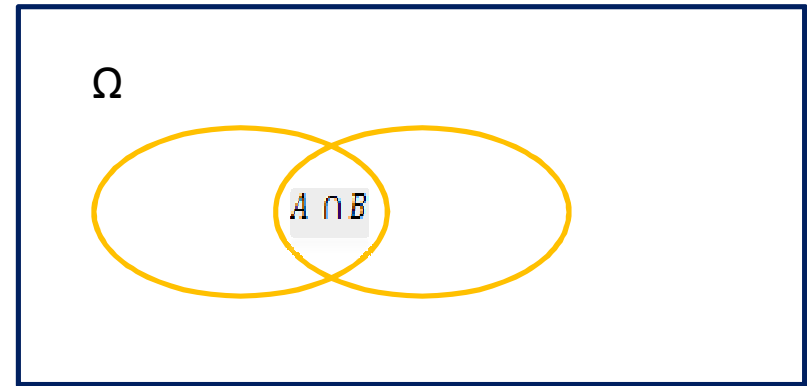
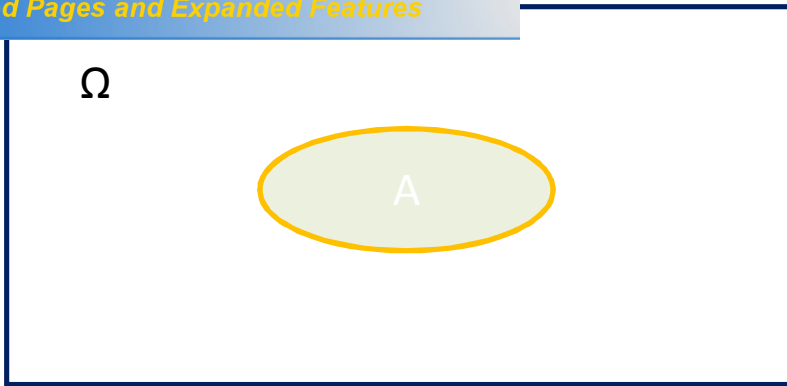
Definiamo Unione tra due eventi A e B l'evento C che si verifica quando si verifica almeno uno dei due eventi A e B;

2. Prodotto logico o intersezione

$$C = A \cap B$$

Definiamo Intersezione tra due eventi A e B l'evento C che si verifica se e solo se si verificano contemporaneamente sia A che B

Il diagramma di Venn



Gli assiomi della probabilità (Kolmogorof 1933)

1. $P(E_i) \geq 0 \quad \forall E_i \subset \Omega$
2. $P(\Omega) = 1$
3. $E_i \cap E_j = \emptyset \Rightarrow P(E_i \cup E_j) = P(E_i) + P(E_j)$

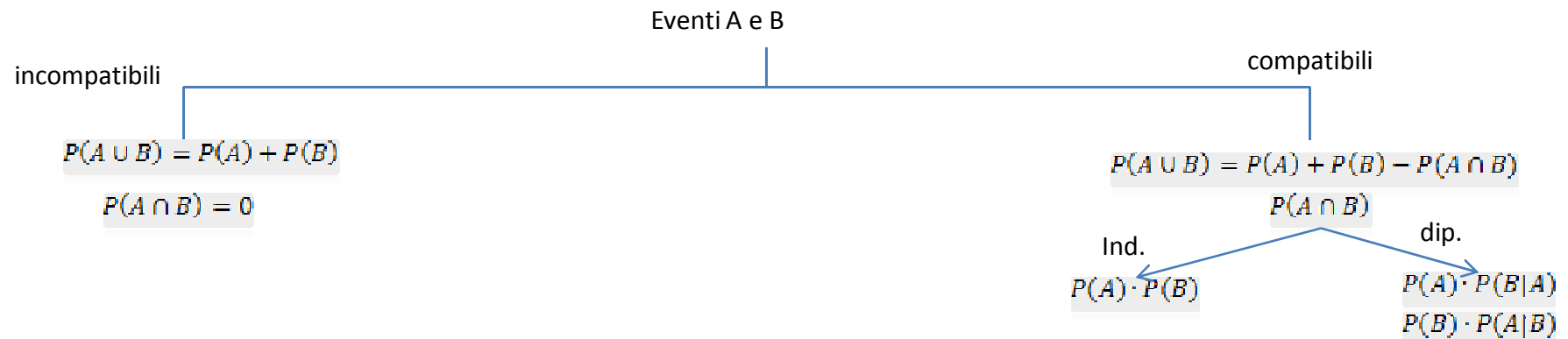
Probabilità
condizionata

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

Eventi
indipendenti

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

Incompatibilità e indipendenza



Alcuni esempi

Nell'ambito dell'esame di ammissione ad una Accademia teatrale si considerino i seguenti eventi:

- A il candidato ha meno di 35 anni
- B il candidato ha una buona dizione
- C il candidato ha già avuto esperienze nell'ambiente teatrale

Assumendo a caso uno tra i candidati, si scrivano i seguenti eventi:

1. Il candidato non ha una buona dizione
2. Ha meno di 35 anni ed ha una buona dizione
3. Ha meno di 35 anni ma non ha una buona dizione
4. Non ha una buona dizione ma ha già avuto esperienze teatrali
5. Ha più di 35 anni , una buona dizione ed ha avuto esperienze
6. Ha almeno una delle tre caratteristiche
7. Ha una buona dizione oppure ha avuto esperienze ma non entrambe
8. Ha una buona dizione dato che ha già avuto esperienze di lavoro

$$\bar{B}$$

$$A \cap B$$

$$A \cap \bar{B}$$

$$\bar{B} \cap C$$

$$\bar{A} \cap B \cap C$$

$$A \cup B \cup C$$

$$(B \cup C) - (B \cap C)$$

$$(B|C)$$

Alcuni esempi

il sabato pomeriggio ad una delle seguenti attività:

- | | | | |
|---|--------------------------------------|-----------------|------|
| A | <u>navigare in Internet</u> | con probabilità | 0.24 |
| B | <u>studiare</u> | “ “ | 0.15 |
| C | <u>uscire con le amiche</u> | “ “ | 0.33 |
| D | <u>ascoltare la musica preferita</u> | | 0.12 |
| E | <u>andare al cinema</u> | | 0.16 |

Qual è la probabilità che Sara trascorra il prossimo sabato pomeriggio:

1. Navigando in internet
2. Non studiando
3. A casa
4. Studiando, posto che debba rimanere a casa

$$P(A \cup D) = 0.24 + 0.12 = 0.36$$

$$P(\bar{B}) = 1 - 0.15 = 0.85$$

$$P(A \cup B \cup D) = 0.24 + 0.15 + 0.12 = 0.52$$

$$P(B | A \cup B \cup D) = \frac{0.15}{0.24 + 0.12 + 0.15} = 0.29$$

	Maschi	Femmine
Positivo	18	34
Negativo	22	26

Alcuni esempi

La tabella riporta i risultati di 100 candidati ad un Concorso pubblico, divisi per genere ed esito.

Si estrae a caso un candidato. Qual è la probabilità che:

1. Sia maschio
2. Abbia superato la prova
3. Abbia superato la prova posto che sia maschio
4. Sia maschio posto che abbia superato la prova
5. Abbia superato la prova oppure sia femmina
6. Abbia superato la prova ma non sia maschio

$$\frac{18 + 22}{100} = 0.4$$

$$\frac{18 + 34}{100} = 0.52$$

$$\frac{18}{40} = 0.45$$

$$\frac{18}{52} = 0.35$$

$$\frac{52}{100} + \frac{60}{100} - \frac{34}{100} = 0.78$$

$$\frac{34}{100} = 0.34$$

Alcuni esempi

In un cassetto ci sono 10 pile, di cui 7 sono funzionanti e 3 esaurite. Dal cassetto viene presa, a caso, una prima pila e poi, senza reintrodurre nel cassetto la prima, ne viene presa una seconda. Qual è la probabilità che le due pile siano:

1. Entrambe funzionanti (evento A)
2. Entrambe esaurite (evento B)
3. Una funzionante e una esaurita

1.

$$P[(A \cap B)] = \frac{7}{10} \times \frac{6}{9} = 0.467$$

2.

$$P(\bar{A} \cap \bar{B}) = \frac{3}{10} \times \frac{2}{9} = 0.067$$

3.

$$P[(A \cap \bar{B}) \cup P(\bar{A} \cap B)] = \frac{3}{10} \times \frac{7}{9} + \frac{7}{10} \times \frac{3}{9} = 0.466$$

Le variabili casuali

Una variabile casuale è strettamente legato a quello di esperimento, a quello, cioè, di una prova il cui esito è incerto.

È diverso dunque dal concetto di variabile definita su una popolazione, di cui posso conoscere o meno il valore che assume sulle singole unità, ma rispetto alla quale non c'è nulla di incerto.

È lo stesso concetto che possiamo associare al lancio di una moneta. Prima di lanciare la moneta, la probabilità che esca testa vale p e quella che esca croce $(1-p)$. Ma una volta lanciata la moneta, e cioè una volta che l'evento si realizza, questo può essere noto o meno (immaginiamo di avere la moneta nel pugno stretto), ma non c'è incertezza: la probabilità che il risultato sia testa è 1 (se è effettivamente uscito) e 0 (se è uscito croce).

Una variabile casuale X è associata ai possibili risultati $x_1, x_2, \dots, x_n$ di un esperimento o alla possibilità di associare ad ognuno di questi risultati la corrispondente probabilità

$$P(X=x)$$

Quando è possibile invece di specificare le singole probabilità $P(X)$ per i valori che la variabile X può assumere, si preferisce esprimere la relazione funzionale che lega queste probabilità e che viene sintetizzata attraverso la funzione $f(x)$

$$f(x) = P(X=x)$$

L'impiego della funzione $f(x)$ è indispensabile quando si ha a che fare con variabili di tipo continuo o con variabili di tipo discreto con un elevato numero di modalità.

Le variabili casuali discrete

Nel caso discreto la funzione $f(x)$ definisce la funzione di probabilità della v.c. X . Se X è discreta, anche $f(x)$ sarà discreta, nel senso che concentrerà la probabilità in un insieme finito di valori di X . La funzione di probabilità $f(x)$ di tipo discreto soddisfa le condizioni:

1.

$$f(x_i) \geq 0$$

2.

$$\sum_i f(x_i) = 1$$

In molti casi può essere necessario trovare la probabilità che la v.c. X assuma un valore inferiore o uguale ad un dato valore x_k . Tale probabilità è definita probabilità cumulata ed è descritta dalla funzione di ripartizione, che viene indicata con $F(x_k)$.

Quindi se $x_1, x_2, \dots, x_n$ sono i valori possibili di X ordinati in senso crescente, la probabilità cumulata sarà:

$$F(x_k) = f(x_1) + f(x_2) + \dots + f(x_k)$$

Esempio

Lancio di tre monete v.c. numero di teste uscite

$$f(x_0) = P(X=0) = P(C \cap C \cap C) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

$$f(x_1) = P(X=1) = P[(T \cap C \cap C) \cup (C \cap T \cap C) \cup (C \cap C \cap T)] = \frac{3}{8}$$

$$f(x_2) = P(X=2) = P[(T \cap T \cap C) \cup (T \cap C \cap T) \cup (C \cap T \cap T)] = \frac{3}{8}$$

$$f(x_3) = P(X=3) = P(T \cap T \cap T) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

$$F(x_0) = \frac{1}{8}$$

$$F(x_1) = \frac{1}{8} + \frac{3}{8}$$

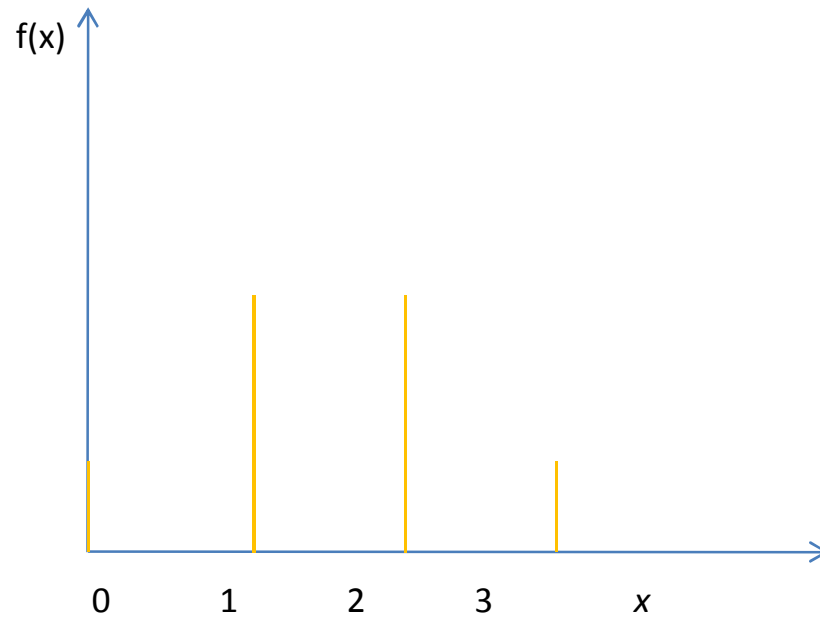
$$F(x_2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8}$$

$$F(x_3) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}$$

Le variabili casuali discrete

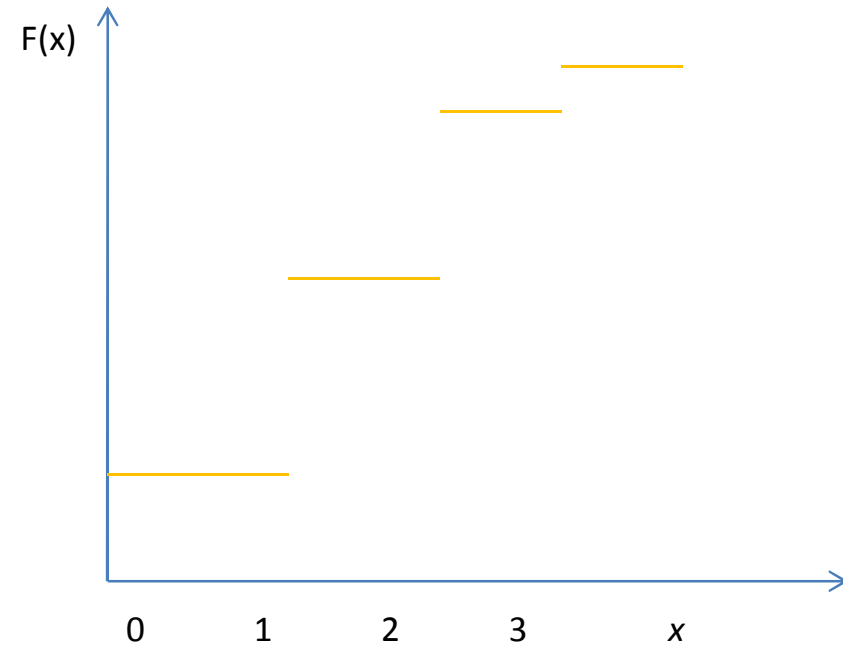
Esempio: lancio di tre monete

$$f(x_0) = 1/8 \quad f(x_1) = 3/8 \quad f(x_2) = 3/8 \quad f(x_3) = 1/8$$



v.c. Numero di teste uscite

$$F(x_0) = 1/8 \quad F(x_1) = 4/8 \quad F(x_2) = 7/8 \quad F(x_3) = 1$$



Le variabili casuali continue

Una variabile casuale continua è una v.c. che può assumere un numero infinito di valori compresi in un intervallo di ampiezza finita o infinita.

A differenza di quanto avviene per il caso discreto, non è possibile ottenere la probabilità che la variabile assuma qualsiasi valore interno all'intervallo sommando le probabilità dei singoli punti che lo compongono, in quanto i punti sono infiniti e una somma infinita di valori finiti non può dare l'unità.

Il c.d. paradosso della continuità viene risolto ricorrendo al concetto di area, assegnando probabilità a singoli intervalli piuttosto che a singoli punti e rappresentando le probabilità come delle aree su degli intervalli.

Una variabile casuale X è, allora, continua se esiste una funzione $f(x)$ tale che:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

dove a e b sono numeri reali qualsiasi con $a < b$.

Le variabili casuali continue

La funzione $f(x)$ viene definita funzione di densità di probabilità (f.d.p.) o densità di probabilità di X . In questo caso, tuttavia, la funzione non può essere interpretata come la $P(X=x)$, in quanto tale probabilità sarà sempre nulla, per v.c. di tipo continuo.

Si può però determinare la probabilità di osservare un valore compreso nell'intervallo $(x, x+\Delta x)$. Tale probabilità tende a zero quando $\Delta x > 0$ e, quindi la probabilità di ottenere esattamente il risultato x è generalmente nulla anche se l'evento x non è strettamente impossibile.

È invece possibile definire la funzione di ripartizione :

che conserva il suo significato

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

Speranza matematica (o valore atteso)

Una v.c. può essere interamente descritta dalla sua funzione di densità. È possibile definire dei parametri o grandezze caratteristiche di una distribuzione di probabilità, che hanno la capacità di riassumere in modo immediato e sintetico l'informazione relativa alla distribuzione.

Questi parametri vengono definiti in termini di valori attesi, o speranze matematiche, e rappresentano una sintesi dei diversi risultati dell'esperimento

Esempio 1 (da Orsi, 1995, pagg. 90-91)

Un amico ci propone un gioco i cui risultati possono essere A, B o C con probabilità rispettivamente a 0.1, 0.2 e 0.7. Se esce A si vincono 20 euro, se esce B si vincono 10 euro mentre se esce C si perdono 10 euro. Ci si chiede quale sarà il guadagno, o la perdita, che ci si deve attendere per un numero elevato di giocate.

È chiaro che il risultato del gioco sarà dato all'ammontare che si vince quando si presenta A o B, ognuno moltiplicato per le rispettive probabilità, sommato all'ammontare che si perde quando si presenta C, ponderato con la rispettiva probabilità.

Avremo dunque $(20 \cdot 0.1) + (10 \cdot 0.2) - (10 \cdot 0.7) = -3$

Il gioco ha un valore negativo, e più precisamente una perdita di 3 euro a partita. I 3 euro non rappresentano l'ammontare che si perde in una singola giocata ma ciò che si perderebbe in media, per partita, se si giocasse un numero elevato di volte (infatti in una singola giocata o si vincono 20 e 10 euro o se ne perdono 10, ma non se ne perderanno mai 3). Questa somma tuttavia rappresenta una sintesi dei diversi risultati del gioco, i quali portano a perdere, in media, 3 euro ogni giocata, e quindi non si avrà interesse a giocare perché il gioco non è equo. Il valore atteso, o speranza matematica, di una v.c. X discreta sarà quindi dato da:

$$E(X) = \sum_{i=1}^n x_i f(x_i)$$

Analogamente il valore atteso di una v.c. continua è dato da:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Esempio 2 (da Orsi, 1995, pagg. 92-93)

Una slot machine dispone di due quadranti: in ogni quadrante possono comparire 3 diversi tipi di figure: mele, campane e ciliegie. La macchina è strutturata in modo che i due quadranti girino in modo indipendente l'uno dall'altro. Dopo aver osservato attentamente il gioco, si stabilisce che le probabilità di uscita di ogni figura sono le seguenti:

$$P(\text{mele}) = 0.1 \quad P(\text{campane}) = 0.4 \quad P(\text{ciliegie}) = 0.5$$

Ogni giocata costa un euro. Il risultato sarà una delle 9 possibili coppie di figure, coppie che si verificano con probabilità diverse. La macchina paga i seguenti premi: i) 10 euro (mele, mele); ii) 2 euro (campane, campane); iii) 1 euro (ciliegie, ciliegie); iv) 0 per ogni altro risultato.

Qual è il guadagno atteso per ogni euro giocato?

Y = guadagno associato ad ogni risultato; Ω = spazio campionario $f(x)$ = funzione di probabilità

Risultato	(mele, mele)	(camp.,camp.)	(ciliegie, ciliegie)	altro
Guadagno(Y)	+9€	+1€	0€	-1€
$P(Y)$	0.01	0.16	0.25	0.58

Esempio 2 (da Orsi, 1995, pagg. 92-93)

Una slot machine dispone di due quadranti: in ogni quadrante possono comparire 3 diversi tipi di figure: mele, campane e ciliegie. La macchina è strutturata in modo che i due quadranti girino in modo indipendente l'uno dall'altro. Dopo aver osservato attentamente il gioco, si stabilisce che le probabilità di uscita di ogni figura sono le seguenti: $P(\text{mele}) = 0.1$ $P(\text{campane}) = 0.4$ $P(\text{ciliegie}) = 0.5$. Ogni giocata costa un euro. Il risultato sarà una delle 9 possibili coppie di figure, coppie che si verificano con probabilità diverse. La macchina paga i seguenti premi: i) 10 euro (mele, mele); ii) 2 euro (campane, campane); iii) 1 euro (ciliegie, ciliegie); iv) 0 per ogni altro risultato. Qual è il guadagno atteso per ogni euro giocato?

Y = guadagno associato ad ogni risultato; Ω = spazio campionario $f(x)$ = funzione di probabilità

Risultato	(mele, mele)	(camp.,camp.)	(ciliegie, ciliegie)	altro
Guadagno(Y)	+9€	+1€	0€	-1€
$P(Y)$	0.01	0.16	0.25	0.58

Il guadagno atteso per ogni giocata sarà dunque:

$$E(Y) = (9 \cdot 0.01) + (1 \cdot 0.16) + (0 \cdot 0.25) - (1 \cdot 0.58) = 0.09 + 0.16 - 0.58 = \underline{\underline{-0.33}}$$

per ogni euro giocato si perdono, in media, 33 centesimi, per cui per 100 giocate ci si aspetta di perdere 33 euro.

La perdita di 0.33 centesimi è quello che ci si attende in media per un numero elevato di prove. Questo valore dà una indicazione sul meccanismo di gioco, in questo caso un gioco non equo, poiché tende a produrre un guadagno sistematico per la macchina, pagando un premio non proporzionale alla posta pagata ma inferiore. Questo non vuol dire che non si possa vincere, ma se accade è un evento fortuito.

Coppie di variabili casuali

Siano X ed Y due v.c. che riguardano lo stesso esperimento. Si definisce funzione di ripartizione congiunta di X e Y $F(x,y)$ la funzione a due variabili:

$$F(x,y) := P(X \leq x, Y \leq y)$$

dove la virgola nella funzione $P()$ denota l'intersezione degli eventi. Ciò consente di calcolare le probabilità di tutti gli eventi che dipendono singolarmente o congiuntamente da X e Y .

Distribuzione congiunta per v.c. discrete

Se X e Y sono v.c. discrete che assumono i valori $x_1, x_2, \dots$ e $y_1, y_2, \dots$ rispettivamente, la funzione

$$p(x_i, y_j) := P(X = x_i, Y = y_j)$$

è la loro funzione di densità di probabilità congiunta.

Le funzioni di densità individuali possono essere ricavate da quella congiunta, considerando che siccome Y deve assumere uno dei valori y_j , l'evento $\{X=x_i\}$ può essere visto come l'unione al variare di j degli eventi $\{X=x_i, Y=y_j\}$, che sono mutuamente esclusivi.

$$P(X = x_i) = \sum_j p(x_i, y_j)$$

$$P(Y = y_j) = \sum_i p(x_i, y_j)$$

Distribuzione congiunta per v.c. continue

Due v.c. X e Y sono congiuntamente continue se esiste una funzione non negativa $f(x,y)$ definita per tutti gli x e y avente la proprietà che per ogni sottoinsieme C del piano cartesiano

$$P((X,Y) \in C) = \iint_{(x,y) \in C} f(x,y) dx dy$$

Tale è la funzione di densità di probabilità congiunta.

Analogamente a quanto detto per le v.c. discrete, le funzioni marginali sono:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx$$

Variabili casuali indipendenti

Due v.c. X e Y si dicono indipendenti se per ogni coppia di insiemi di numeri reali A e B è soddisfatta l'equazione

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

In caso contrario X e Y si dicono dipendenti.

Ciò equivale a dire che la funzione di ripartizione congiunta è uguale al prodotto delle funzioni marginali

$$F(a, b) = F_X(a)F_Y(b) \quad \forall a, b \in \mathbb{R}$$

La v.c. Normale $X \sim N(\mu, \sigma^2)$

Nello studio dei fenomeni reali, uno dei problemi che si pone è stabilire se, per la descrizione del fenomeno osservato, si debba costruire un'apposita variabile casuale oppure se non convenga fare riferimento a v.c. note per le loro caratteristiche e proprietà.

Tali variabili casuali note, sia discrete che continue, costituiscono altrettanti modelli probabilistici che consentono di descrivere la realtà con un sufficiente grado di approssimazione.

La distribuzione di probabilità più importante della Statistica è certamente la distribuzione Normale, proposta da Gauss nel 1809 come distribuzione in grado di descrivere gli errori accidentali commessi nel calcolo del cammino dei corpi celesti.

L'importanza della distribuzione Normale è legata alla considerazione che molti dei fenomeni osservabili si distribuiscono in un modo che può essere considerato Normale, ma anche al fatto che altre distribuzioni di probabilità, diverse dalla Normale, in molte circostanze possono essere approssimate in modo soddisfacente dalla distribuzione gaussiana.

La v.c. Normale $X \sim N(\mu, \sigma^2)$

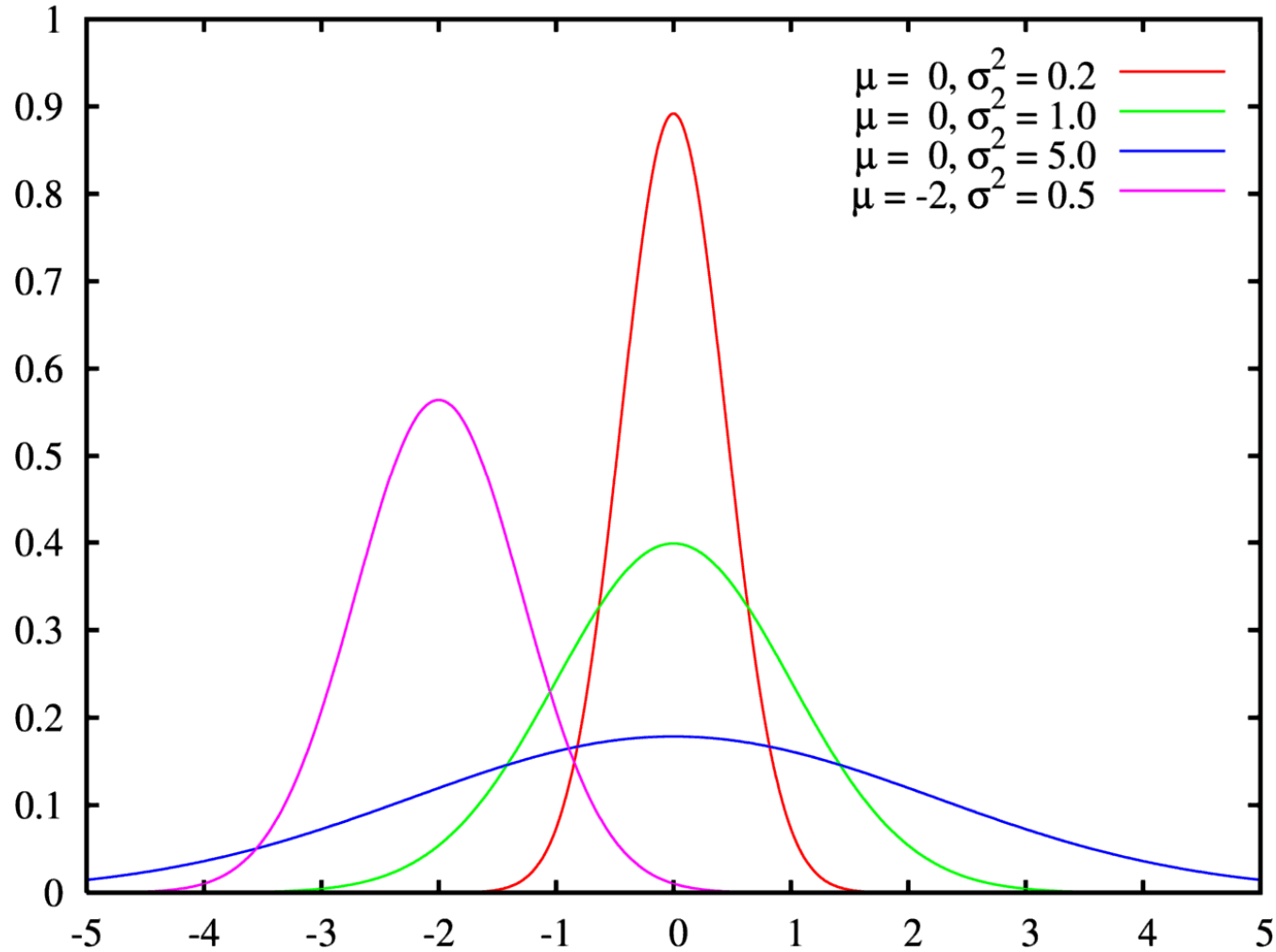
Una variabile casuale X segue una distribuzione Normale, con media μ e varianza σ^2 , se la sua funzione di densità di probabilità è data da:

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

Caratteristiche della distribuzione Normale:

- ⌘ Forma campanulare e simmetrica
- ⌘ Media, moda e mediana coincidenti
- ⌘ Punto di flesso a distanza σ dalla media
- ⌘ Circa il 68% dei casi è compreso nell'intervallo $\mu \pm \sigma$
- ⌘ Circa il 95% dei casi è compreso nell'intervallo $\mu \pm 2\sigma$
- ⌘ Circa il 99% dei casi è compreso nell'intervallo $\mu \pm 3\sigma$
- ⌘ Un aumento o una diminuzione della media determina uno slittamento della curva, a parità di forma, sull'asse delle X
- ⌘ Un aumento o una diminuzione della varianza determina, rispettivamente, una minore o una maggiore concentrazione di valori attorno al valore medio.

La v.c. Normale $X \sim N(\mu, \sigma^2)$



La v.c. Normale $X \sim N(\mu, \sigma^2)$

Una macchina produce biscotti il cui peso si distribuisce come una Normale, con media pari a 5 grammi e scarto quadratico medio pari a 0,2 grammi. Qual è la percentuale di biscotti il cui peso è compreso tra 5,12 grammi e 5,30 grammi?

$$X \sim N(5; 0.04)$$

$$\mu = 5 \quad \sigma = 0.2$$

$$P(5.12 \leq X \leq 5.30) = \int_{5.12}^{5.30} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Per calcolare questa probabilità si standardizza la v.c. X.

Data una v.c. X con media μ e scarto quadratico medio σ , è possibile costruire la variabile standardizzata Z, con media pari a zero e scarto quadratico medio pari a uno, mediante la trasformazione:

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

La v.c. Normale standardizzata

Qualsiasi distribuzione, e quindi anche una distribuzione Normale, può essere ricondotta ad una distribuzione con media nulla e varianza unitaria mediante la trasformazione:

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

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma} E(X - \mu) = \frac{1}{\sigma} [E(X) - E(\mu)] = \frac{1}{\sigma} (\mu - \mu) = 0$$

$$\text{Var}(Z) = \text{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2} \text{Var}(X - \mu) = \frac{1}{\sigma^2} [\text{Var}(X) - \text{Var}(\mu)] = \frac{1}{\sigma^2} [\sigma^2] = 1$$

$$Z \sim N \quad \mu = 0 \quad \sigma^2 = 1$$

$$X \longrightarrow Z$$

$$y = f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \Rightarrow \quad y = f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

La v.c. Normale standardizzata

$$X \sim N \quad \mu = 5 \quad \sigma = 0.2$$

Qual è la percentuale di casi compresi tra 5,12 e 5,30?

$$\text{Fr}(5.12 \leq X \leq 5.30)$$

$$Z = \frac{X - \mu}{\sigma} \sim N \quad \mu = 0 \quad \sigma = 1$$

Quali sono i valori standardizzati di $X = 5.12$ e $X = 5.30$?

$$Z = \frac{5.12 - 5}{0.2} = 0.6$$

$$Z = \frac{5.30 - 5}{0.2} = 1.5$$

$$\text{Fr}(0.6 \leq Z \leq 1.5) = 0.4332 - 0.2257 = 0.2075$$

e proprietà della distribuzione Normale

La proprietà riproduttiva:

“La combinazione lineare di v.c. Normali e indipendenti è ancora una v.c. Normale, con valore medio pari alla combinazione lineare dei valori medi e con varianza pari alla combinazione lineare delle varianze con i quadrati dei coefficienti.”

$$X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, 2, \dots, n \quad a_1, a_2, \dots, a_n$$

$$\sum_{i=1}^n a_i X_i \sim N(\mu, \sigma^2) \quad \text{con} \quad \mu = \sum_{i=1}^n a_i \mu_i \quad \sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2$$

“La somma di v.c. Normali e indipendenti è ancora una v.c. Normale, con valore medio pari alla somma dei valori medi e con varianza pari alla somma delle varianze.”

La v.c. Normale standardizzata

Esercizio

La durata delle lampadine di una certa marca è distribuita secondo una Normale con media pari a 2850 ore e scarto quadratico medio pari a 160 ore.

Si sceglie una lampadina a caso. Qual è la probabilità che questa:

- a. duri più di 3000 ore?
- b. duri meno di 2500 ore?
- c. duri più di 2600 ore?
- d. duri un numero di ore compreso tra 2500 e 2700?
- e. duri un numero di ore compreso tra 2500 e 3000?

Qual è il valore oltre il quale si trova il 15% delle lampadine più “longeve”?

a v.c. Normale standardizzata

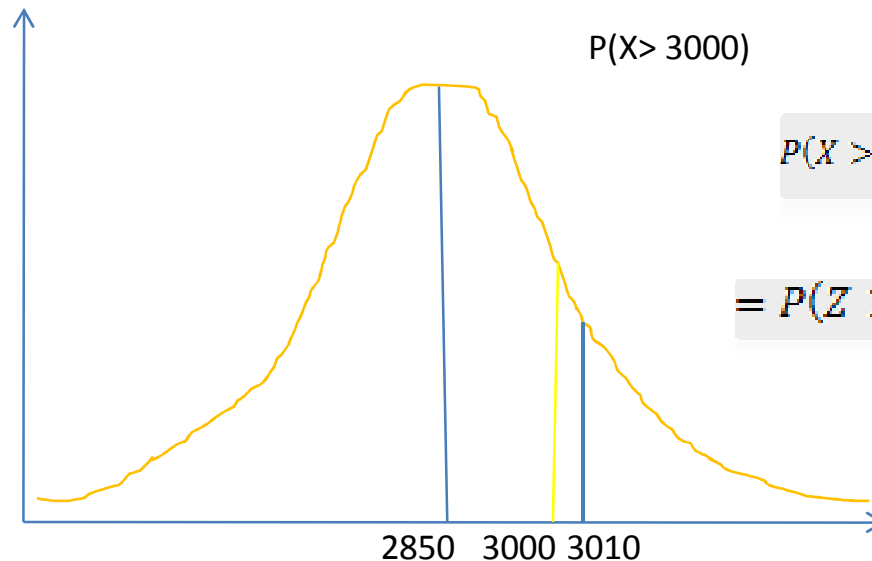
Esercizio

La durata delle lampadine di una certa marca è distribuita secondo una Normale con media pari a 2850 ore e scarto quadratico medio pari a 160 ore.

Si sceglie una lampadina a caso. Qual è la probabilità che questa:

- duri più di 3000 ore?
- duri meno di 2500 ore?
- duri più di 2600 ore?
- duri un numero di ore compreso tra 2500 e 2700?
- duri un numero di ore compreso tra 2500 e 3000?

Qual è il valore oltre il quale si trova il 15% delle lampadine più “longeve”?



$P(X > 3000)$

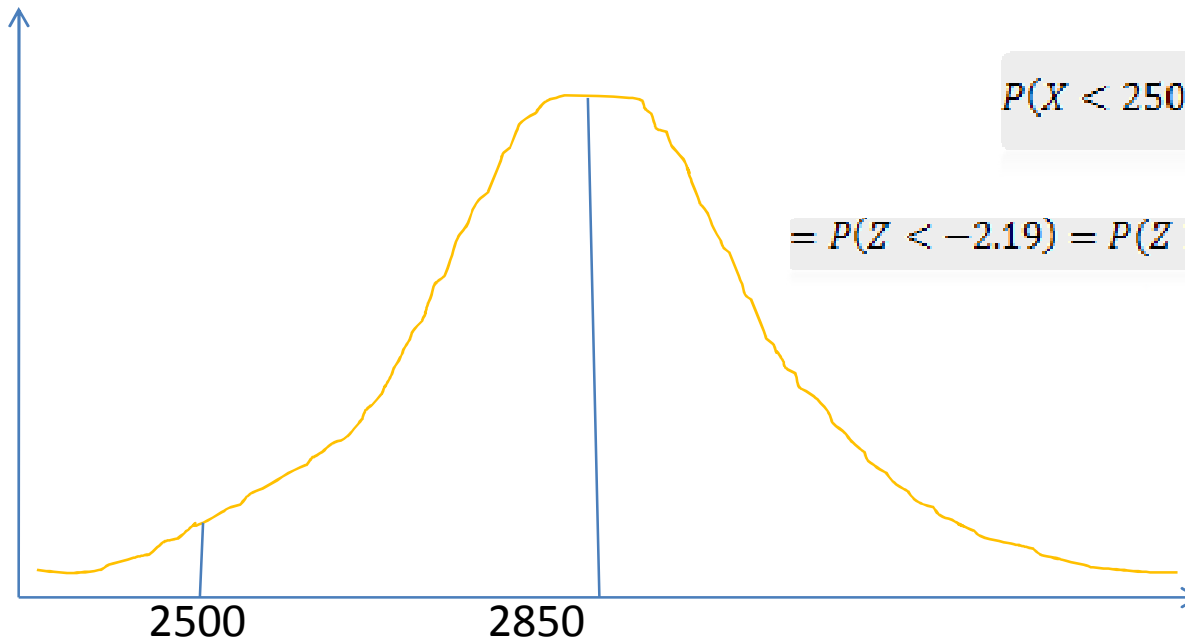
$$P(X > 3000) = P\left(\frac{X - \mu}{\sigma} > \frac{3000 - 2850}{160}\right)$$

$$= P(Z > 0.94) = 0.5 - 0.3264 = 0.1736$$

Esercizio

a v.c. Normale standardizzata

.duri meno di 2500 ore?



$$P(X < 2500) = P\left(\frac{X - \mu}{\sigma} < \frac{2500 - 2850}{160}\right)$$

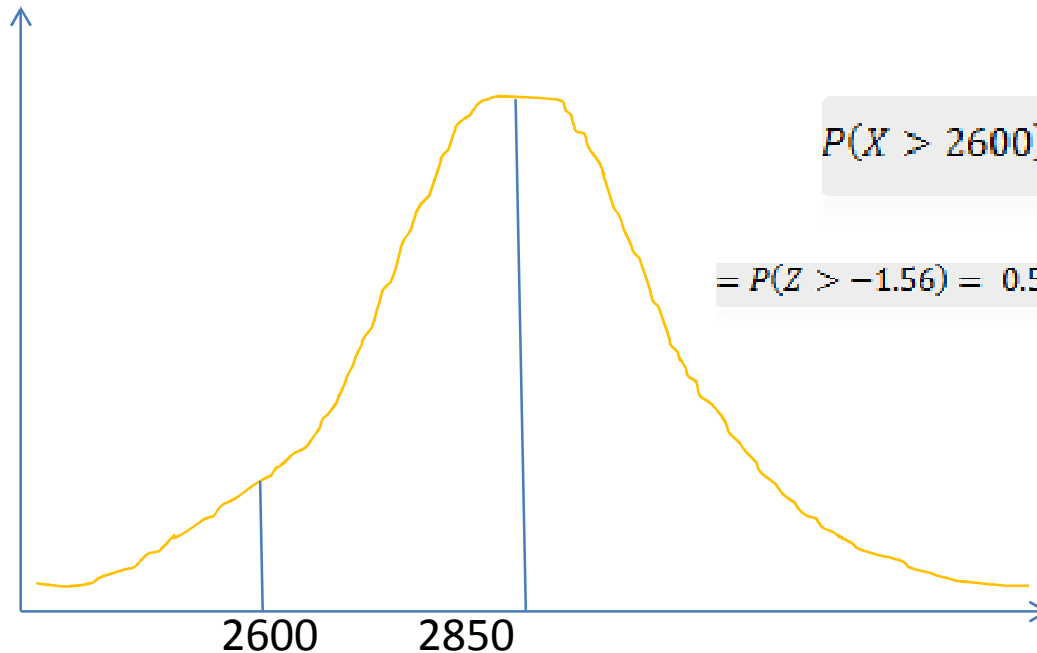
$$= P(Z < -2.19) = P(Z > 2.19) = 0.5 - 0.4857 = 0.0143$$

a v.c. Normale standardizzata

Esercizio

. duri meno di 2500 ore?

duri più di 2600 ore?



$$P(X > 2600) = P\left(\frac{X - \mu}{\sigma} > \frac{2600 - 2850}{160}\right)$$

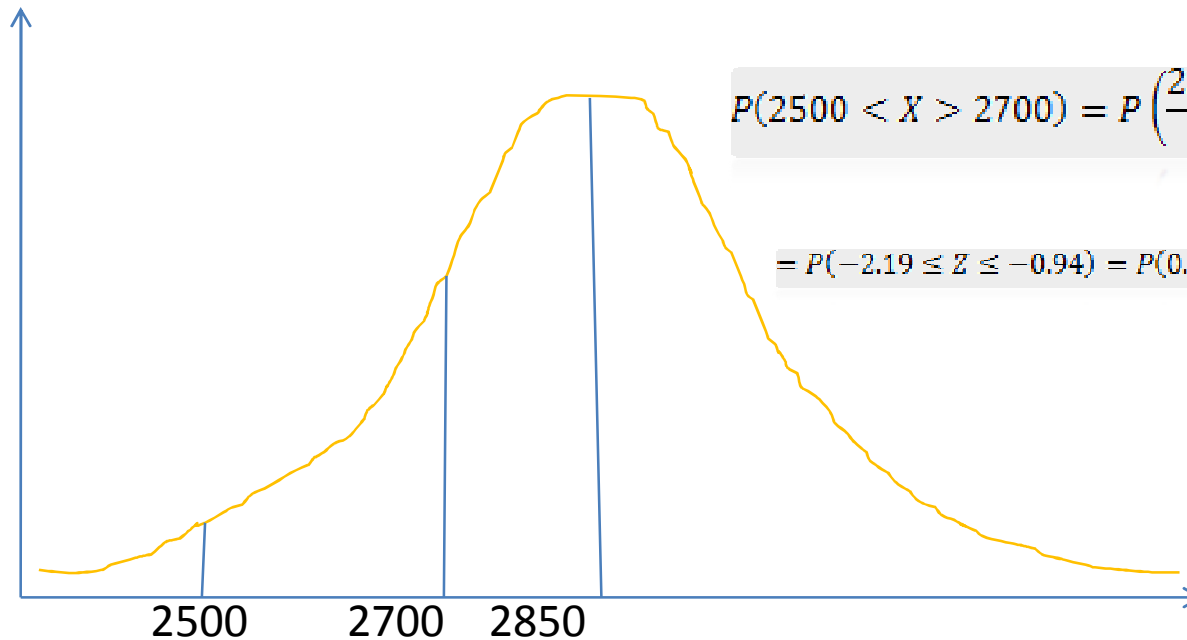
$$= P(Z > -1.56) = 0.5 + P(0 < Z < 1.56) = 0.5 + 0.4406 = 0.9406$$

Esercizio

a v.c. Normale standardizzata

. duri meno di 2500 ore?

duri un numero di ore compreso tra 2500 e 2700



$$P(2500 < X < 2700) = P\left(\frac{2500 - 2850}{160} < \frac{X - \mu}{\sigma} < \frac{2700 - 2850}{160}\right)$$

$$= P(-2.19 \leq Z \leq -0.94) = P(0.94 \leq Z \leq 2.19) = 0.4857 - 0.3264 = 0.1593$$

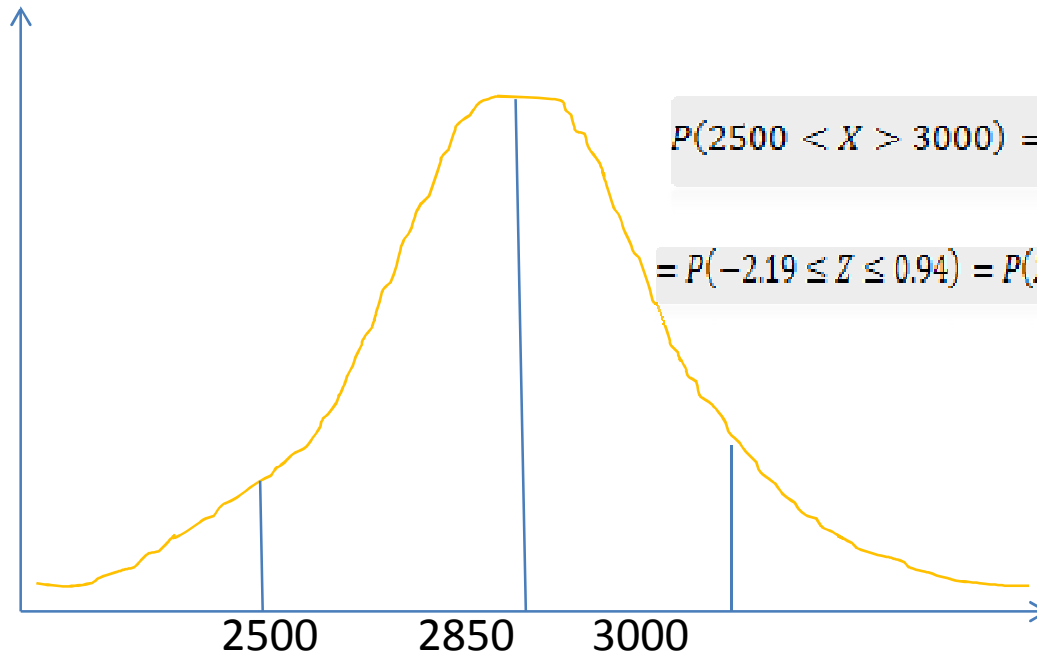
Esercizio

a v.c. Normale standardizzata

. duri meno di 2500 ore?

duri più di 2600 ore?

duri un numero di ore compreso tra 2500 e 3000?



$$P(2500 < X < 3000) = P\left(\frac{2500 - 2850}{160} < \frac{X - \mu}{\sigma} < \frac{3000 - 2850}{160}\right)$$

$$= P(-2.19 \leq Z \leq 0.94) = P(Z \leq 2.19) + P(Z \leq 0.94) = 0.4857 + 0.3264 = 0.8121$$

Esercizio

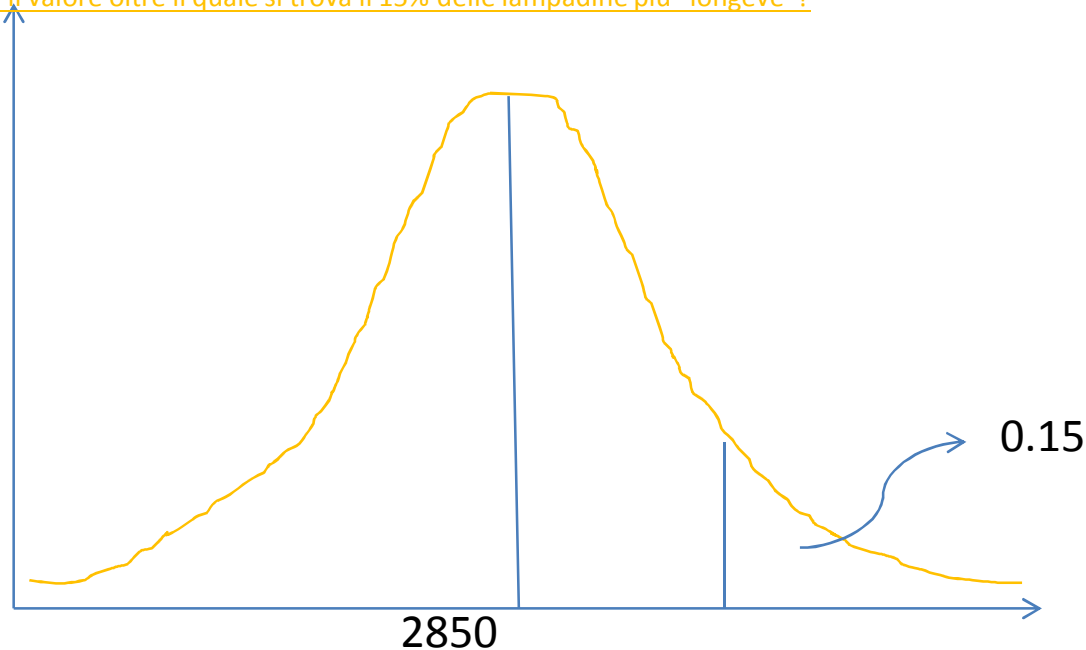
a v.c. Normale standardizzata

. duri meno di 2500 ore?

duri più di 2600 ore?

. duri un numero di ore compreso tra 2500 e 3000?

Qual è il valore oltre il quale si trova il 15% delle lampadine più “longeve”?

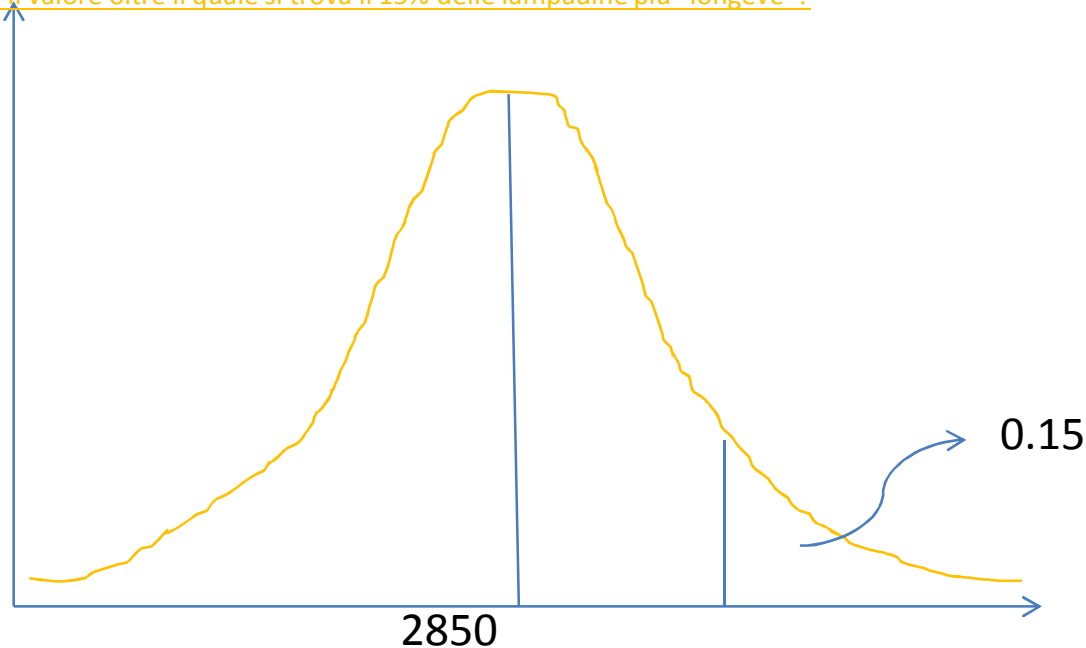


Qual è il valore di Z che
individua alla sua
destra un'area pari al
15%?

Esercizio

- . duri meno di 2500 ore?
- . duri più di 2600 ore?
- . duri un numero di ore compreso tra 2500 e 3000?

Qual è il valore oltre il quale si trova il 15% delle lampadine più “longeve”?



a v.c. Normale standardizzata

Qual è il valore di Z che
individua alla sua
destra un'area pari al
15%?

$$Z_{0.85} = 1.04$$

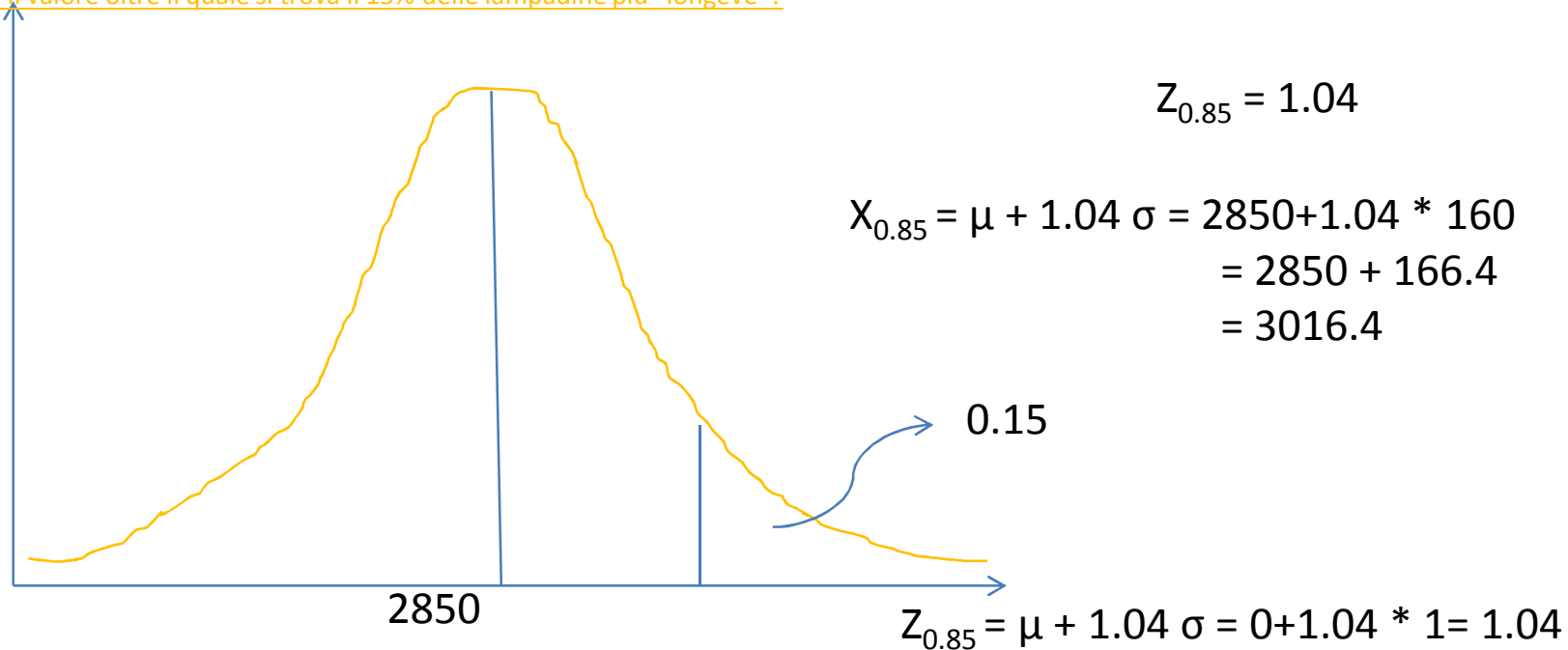
Qual è il valore
corrispondente al valore
standardizzato 1.04?

Esercizio

- . duri meno di 2500 ore?
- . duri più di 2600 ore?
- . duri un numero di ore compreso tra 2500 e 3000?

Qual è il valore
corrispondente al valore
standardizzato 1.04?

Qual è il valore oltre il quale si trova il 15% delle lampadine più "longeve"?



La v.c. Normale standardizzata

Esercizio

Supponiamo che un certo modello di computer portatile sia composto di due pezzi assemblati, la base e lo schermo. Il peso della base può essere ipotizzato seguire una distribuzione Normale con media $\mu=1,650$ kg e scarto quadratico medio $s=85$ grammi, mentre il peso dello schermo può essere ipotizzato anch'esso Normale con media $\mu=720$ grammi e scarto quadratico medio $s=11$ grammi.

La casa produttrice stabilisce che dovranno essere dichiarati “fuori qualità” i notebook con peso complessivo superiore a 2,5 kg. Relativamente, dunque, alla variabile peso complessivo:

1. Quali saranno le caratteristiche del prodotto, in termini di media e scarto quadratico medio?
2. Quale sarà la percentuale di notebook che presumibilmente sarà dichiarata “fuori qualità”?
3. Quale sarà il peso oltre il quale è compreso il 10% dei pezzi assemblati?
4. Quale sarà la percentuale di notebook con peso inferiore a 2 kg?

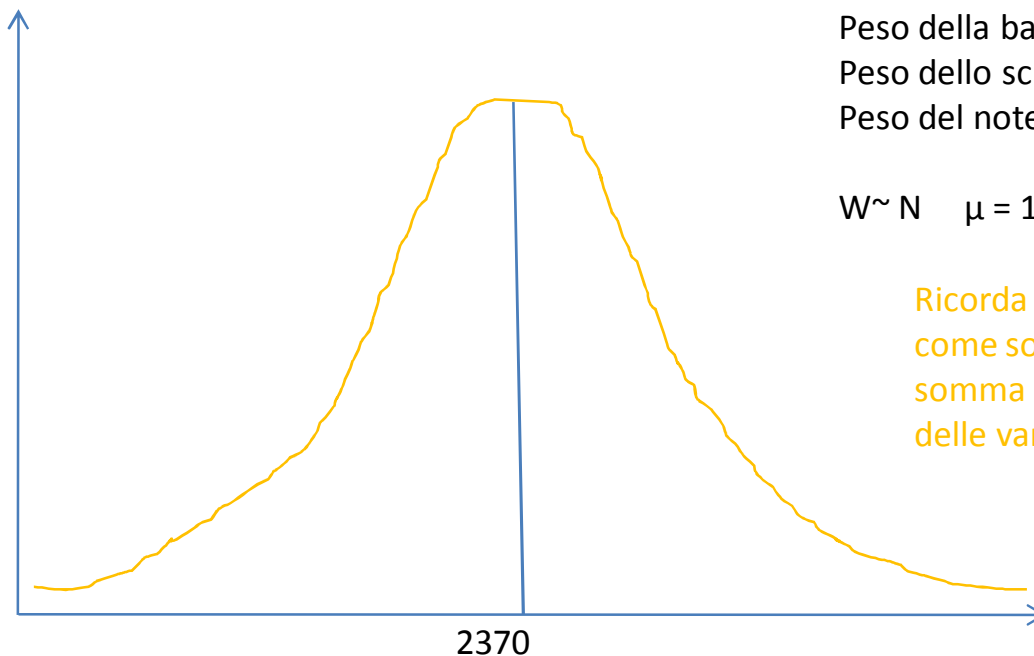
La v.c. Normale standardizzata

Esercizio

Supponiamo che un certo modello di computer portatile sia composto di due pezzi assemblati, la base e lo schermo. Il peso della base può essere ipotizzato seguire una distribuzione Normale con media $\mu=1,650$ kg e scarto quadratico medio $s=85$ grammi, mentre il peso dello schermo può essere ipotizzato anch'esso Normale con media $\mu=720$ grammi e scarto quadratico medio $s=11$ grammi.

La casa produttrice stabilisce che dovranno essere dichiarati “fuori qualità” i notebook con peso complessivo superiore a 2,5 kg. Relativamente, dunque, alla variabile peso complessivo:

1. Quali saranno le caratteristiche del prodotto, in termini di media e scarto quadratico medio?
2. Quale sarà la percentuale di notebook che presumibilmente sarà dichiarata “fuori qualità”?
3. Quale sarà il peso oltre il quale è compreso il 10% dei pezzi assemblati?
4. Quale sarà la percentuale di notebook con peso inferiore a 2 kg?



Peso della base $X \sim N(1650; 85^2)$

Peso dello schermo $Y \sim N(720; 11^2)$

Peso del notebook $W = (X+Y) \sim N(1650+720; 85^2 + 11^2)$

$W \sim N \quad \mu = 1650+720 = \underline{2370} \quad \sigma = \sqrt{85^2 + 11^2} = \underline{85.7}$

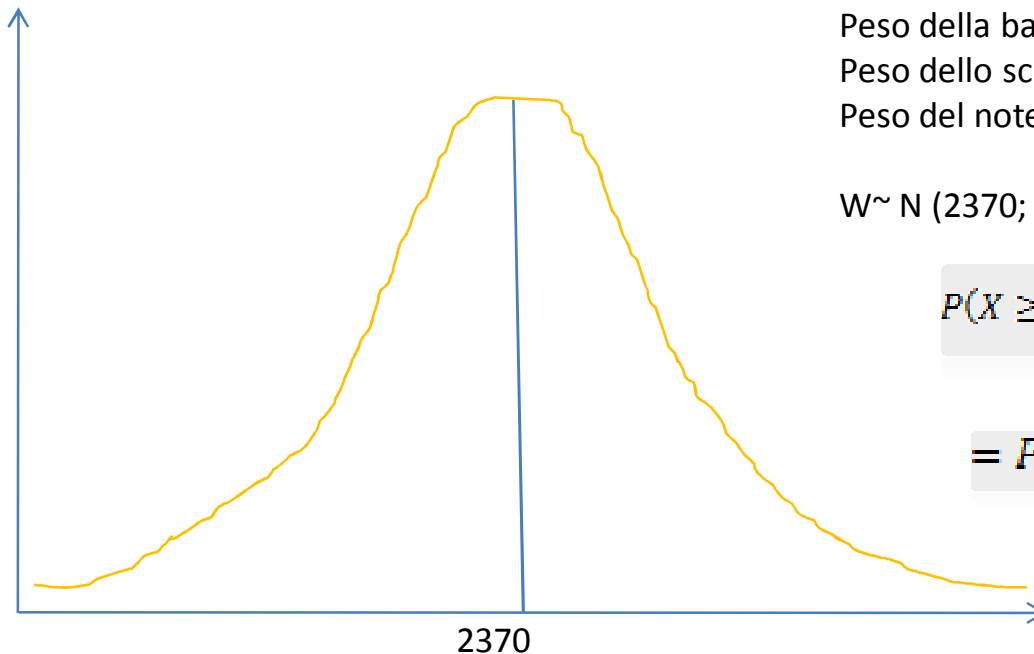
Ricorda che lo sqm di una distribuzione Normale ottenuta come somma di due distribuzioni Normali non è uguale alla somma degli scarti ma alla radice quadrata della somma delle varianze.

Esercizio

Supponiamo che un certo modello di computer portatile sia composto di due pezzi assemblati, la base e lo schermo. Il peso della base può essere ipotizzato seguire una distribuzione Normale con media $\mu=1,650$ kg e scarto quadratico medio $s=85$ grammi, mentre il peso dello schermo può essere ipotizzato anch'esso Normale con media $\mu=720$ grammi e scarto quadratico medio $s=11$ grammi.

La casa produttrice stabilisce che dovranno essere dichiarati "fuori qualità" i notebook con peso complessivo superiore a 2,5 kg. Relativamente, dunque, alla variabile peso complessivo:

1. Quali saranno le caratteristiche del prodotto, in termini di media e scarto quadratico medio?
2. Quale sarà la percentuale di notebook che presumibilmente sarà dichiarata "fuori qualità"?
3. Quale sarà il peso oltre il quale è compreso il 10% dei pezzi assemblati?
4. Quale sarà la percentuale di notebook con peso inferiore a 2 kg?



Peso della base $X \sim N(1650; 85^2)$

Peso dello schermo $Y \sim N(720; 11^2)$

Peso del notebook $W = (X+Y) \sim N(1650+720; 85^2 + 11^2)$

$W \sim N(2370; 85.7)$

$$P(X \geq 2500) = P\left(\frac{X - \mu}{\sigma} > \frac{2500 - 2370}{85.7}\right)$$

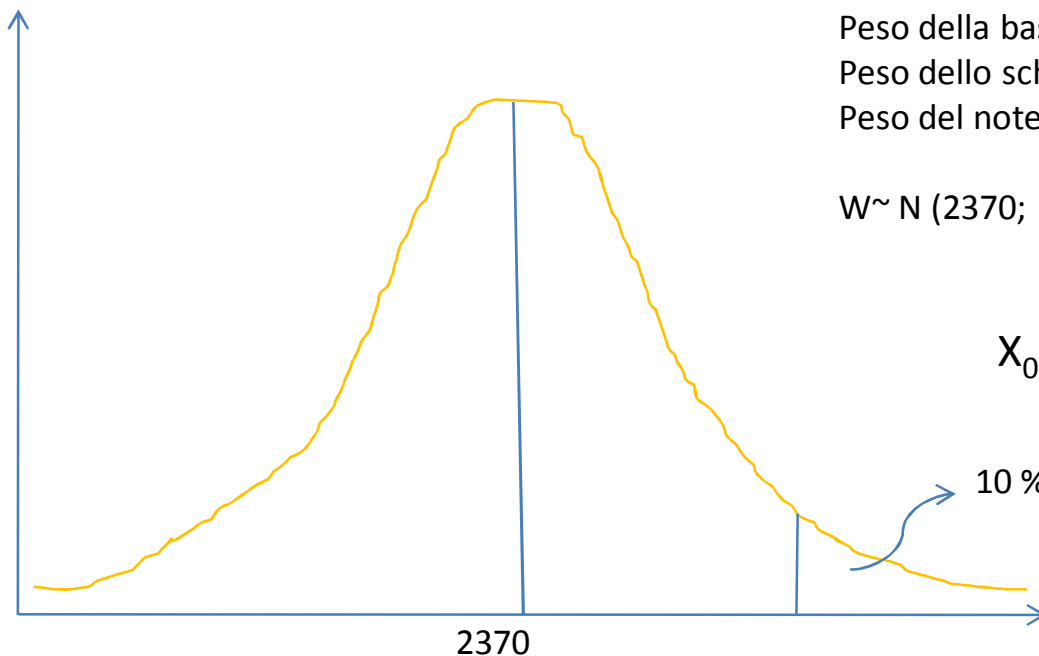
$$= P(Z \geq 1.52) = 0.5 - 0.4357 = 0.0643$$

Esercizio

Supponiamo che un certo modello di computer portatile sia composto di due pezzi assemblati, la base e lo schermo. Il peso della base può essere ipotizzato seguire una distribuzione Normale con media $\mu=1,650$ kg e scarto quadratico medio $s=85$ grammi, mentre il peso dello schermo può essere ipotizzato anch'esso Normale con media $\mu=720$ grammi e scarto quadratico medio $s=11$ grammi.

La casa produttrice stabilisce che dovranno essere dichiarati "fuori qualità" i notebook con peso complessivo superiore a 2,5 kg. Relativamente, dunque, alla variabile peso complessivo:

1. Quali saranno le caratteristiche del prodotto, in termini di media e scarto quadratico medio?
2. Quale sarà la percentuale di notebook che presumibilmente sarà dichiarata "fuori qualità"?
3. Quale sarà il peso oltre il quale è compreso il 10% dei pezzi assemblati?
4. Quale sarà la percentuale di notebook con peso inferiore a 2 kg?



Peso della base $X \sim N(1650; 85^2)$

Peso dello schermo $Y \sim N(720; 11^2)$

Peso del notebook $W = (X+Y) \sim N(1650+720; 85^2 + 11^2)$

$W \sim N(2370; 85.7)$

$$Z_{0.90} = 1.28$$

$$\begin{aligned} X_{0.90} &= \mu + 1.28 \sigma = 2370 + 1.28 * 85.7 \\ &= 2370 + 109.7 \\ &= \underline{\underline{2479.7}} \end{aligned}$$

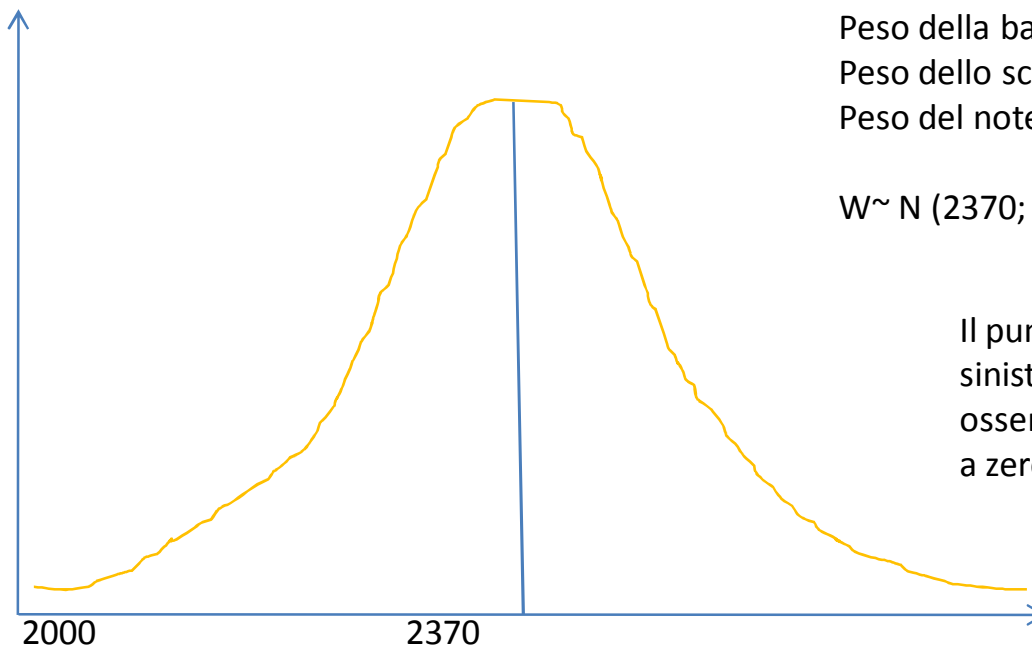
a v.c. Normale standardizzata

Esercizio

Supponiamo che un certo modello di computer portatile sia composto di due pezzi assemblati, la base e lo schermo. Il peso della base può essere ipotizzato seguire una distribuzione Normale con media $\mu=1,650$ kg e scarto quadratico medio $s=85$ grammi, mentre il peso dello schermo può essere ipotizzato anch'esso Normale con media $\mu=720$ grammi e scarto quadratico medio $s=11$ grammi.

La casa produttrice stabilisce che dovranno essere dichiarati “fuori qualità” i notebook con peso complessivo superiore a 2,5 kg. Relativamente, dunque, alla variabile peso complessivo:

1. Quali saranno le caratteristiche del prodotto, in termini di media e scarto quadratico medio?
2. Quale sarà la percentuale di notebook che presumibilmente sarà dichiarata “fuori qualità”?
3. Quale sarà il peso oltre il quale è compreso il 10% dei pezzi assemblati?
4. Quale sarà la percentuale di notebook con peso inferiore a 2 kg?



Peso della base $X \sim N(1650; 85^2)$

Peso dello schermo $Y \sim N(720; 11^2)$

Peso del notebook $W = (X+Y) \sim N(1650+720; 85^2 + 11^2)$

$W \sim N(2370; 85.7)$

Il punto $W=2.000$ si trova ad oltre 4 volte lo sqm a sinistra della media. La probabilità di trovare osservazioni oltre questo punto è, praticamente, pari a zero.

La v.c. chi quadrato

La somma di g variabili casuali Normali standardizzate, indipendenti e al quadrato, è una variabile casuale continua sul supporto $(0, +\infty)$ definita v.c. Chi quadrato con g gradi di libertà e indicata con:

$$X \sim \chi^2_{(g, \alpha)}$$

La funzione di densità di tale v.c. è asimmetrica positiva e tende alla simmetria al crescere di g .

I momenti caratteristici sono:

$$E(X) = g \quad Var(X) = 2g \quad Asym(X) = \sqrt{\frac{8}{g}} \quad Kurt(X) = 3 + \frac{12}{g}$$

La v.c. F di Fisher

Il rapporto tra due v.c. Chi quadrato, indipendenti tra loro e divise per i rispettivi gradi di libertà si definisce v.c. F di Fisher e si indica con:

$$X \sim F_{(g1, g2)}$$

I due parametri, $g1$ e $g2$, che caratterizzano questa distribuzione, vengono definiti, rispettivamente, gradi di libertà del numeratore e gradi di libertà del denominatore.

La v.c. T di Student

Il rapporto tra una v.c. Normale standardizzata e una v.c. indipendente Chi quadrato con g gradi di libertà, si distribuisce come una v.c. t di Student e si indica con:

$$X \sim T(g)$$

$$X \sim \frac{Z}{\sqrt{\frac{Y}{g}}} \sim T(g)$$

La funzione di densità della v.c. di Student è sempre simmetrica, con valore medio pari a 0, ed assume una forma molto simile a quello della Normale standardizzata alla quale tende assai velocemente al crescere dei gradi di libertà.

Per valori di g piccoli o moderati, la v.c. di Student si caratterizza per una curtosi leggermente più elevata e per code più “pesanti” della v.c. Normale.

La v.c. discreta $X \sim Ud(n)$

Una v.c. X si dice seguire una distribuzione Uniforme di parametro n se assume valori su un insieme finito $\{x_1, x_2, \dots, x_n\}$ e la sua funzione di probabilità è la seguente:

$$P(X = x_i) = \frac{1}{n}$$

I suoi momenti caratteristici risultano essere:

$$E(X) = \frac{n+1}{2} \quad Var(X) = \frac{n^2-1}{12} \quad Asym(X) = 0 \quad Kurt(X) = \frac{3(3n^2-7)}{5(n^2-1)}$$

La v.c. di Bernoulli $X \sim \text{Ber}(p)$

E' una v.c. che trae origine da una prova nella quale interessa verificare se l'evento E si è verificato o meno. E' legata a prove di tipo dicotomico (o dicotomizzabili) i cui due possibili risultati vengono indicati con i termini "successo" (1) e "insuccesso" (0), (senza per questo intendere che l'evento successo sia necessariamente un evento piacevole!...)

Formalmente, una v.c. X discreta si definisce v.c. di Bernoulli se assume il valore 1 con probabilità p e il valore 0 con probabilità $1-p$. La sua distribuzione di probabilità è:

$$P(X = x) = p^x (1 - p)^{1-x}$$

I suoi momenti caratteristici risultano essere:

$$E(X) = p \qquad \text{Var}(X) = p(1 - p)$$

N.B. – La varianza della v.c. di Bernoulli assume valore massimo (1/4) quando è $p=1/2$. E' questo, infatti, il caso di massima incertezza, in cui risulta più difficile prevedere il risultato.

La v.c. Binomiale $X \sim \text{Bin}(n, p)$

Consiste nel ripetere n volte, e nelle medesime condizioni, lo schema successo-insuccesso della v.c. di Bernoulli. Equivale all'estrazione con ripetizione di n palline da un'urna che ne contiene H , di cui b bianche e $(H-b)$ nere.

La sua distribuzione di probabilità è:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

con

$$\binom{n}{x} = \frac{n!}{x! (n-x)!}$$

numero di combinazioni in cui possono presentarsi x successi in n prove.

I suoi momenti caratteristici risultano essere:

$$E(X) = np \quad \text{Var}(X) = np(1-p) \quad \text{Asym}(X) = \frac{1-2p}{\sqrt{np(1-p)}} \quad \text{Kurt}(X) = 3 + \frac{1-6p+6p^2}{np(1-p)}$$

Proprietà della v.c. Binomiale

La v.c. Binomiale è simmetrica quando è $p=1/2$ e per n che tende ad infinito. È asimmetrica positiva per $p < 1/2$ e asimmetrica negativa per $p > 1/2$.

La v.c. Binomiale $X \sim \text{Bin}(n, p)$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x! (n-x)!}$$

10 lanci di una moneta

$n=10$ $p=1/2$

X = numero di teste

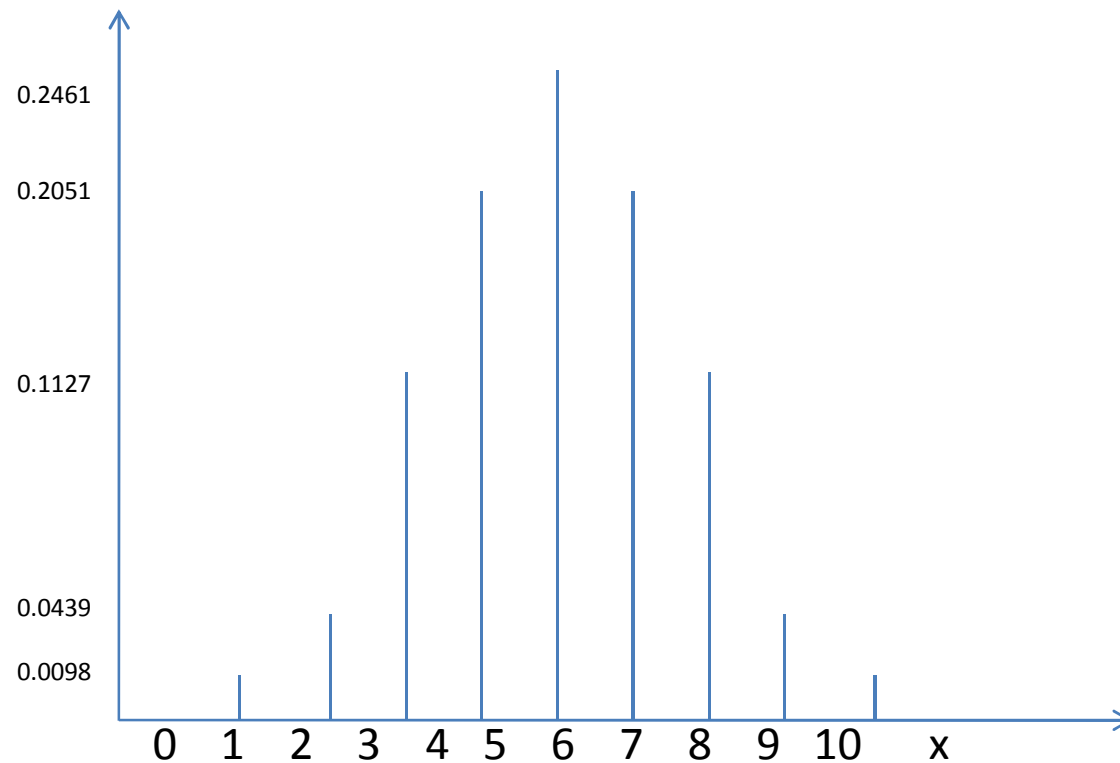
Numero di teste (X)	$\frac{n!}{x! (n-x)!}$	$p^x (1-p)^{1-x}$	P(X=x)
0	1	0.00098	0.00098
1	10	0.00098	0.00977
2	45	0.00098	0.04395
3	120	0.00098	0.11719
4	210	0.00098	0.20508
5	252	0.00098	0.24609
6	210	0.00098	0.20508
7	120	0.00098	0.11719
8	45	0.00098	0.04395
9	10	0.00098	0.00977
10	1	0.00098	0.00098

La v.c. Binomiale $X \sim \text{Bin}(n, p)$

10 lanci di una moneta

X = numero di teste uscite

$n=10$ $p=1/2$



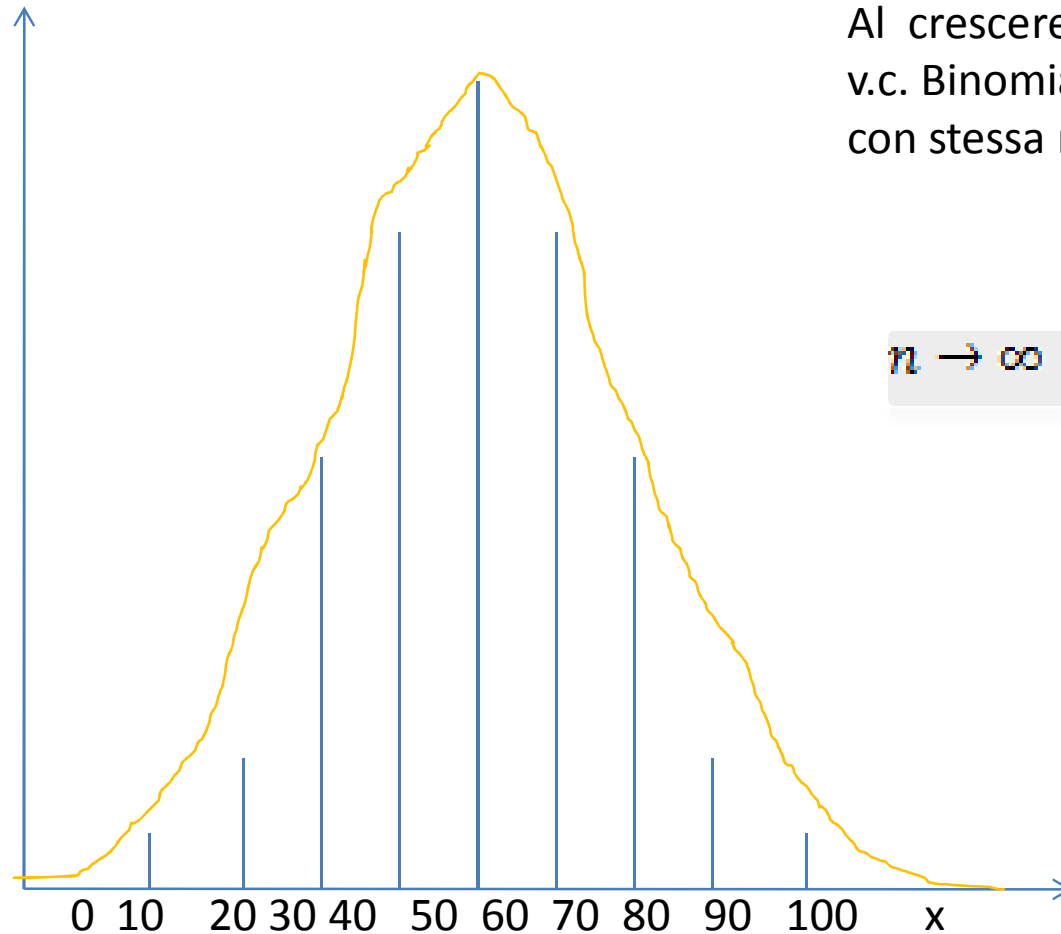
X	P(X=x)
0	0.00098
1	0.00977
2	0.04395
3	0.11719
4	0.20508
5	0.24609
6	0.20508
7	0.11719
8	0.04395
9	0.00977
10	0.00098

La v.c. Binomiale $X \sim \text{Bin}(n, p)$

100 lanci di una moneta

$n=10$ $p=1/2$

X = numero di teste uscite



Al crescere del numero delle prove la
v.c. Binomiale tende ad una v.c. Normale
con stessa media e varianza

$$n \rightarrow \infty \Rightarrow X \sim N[(np), np(1 - p)]$$

Il campionamento

Campione casuale

- ⌘ Non è un campione a casaccio
- ⌘ E' un campione scelto da una popolazione in cui ciascuna unità ha una probabilità non nulla di essere estratta.

Campione casuale semplice

- ⌘ E' un campione scelto da una popolazione in cui ciascuna unità ha la stessa probabilità di essere estratta.

Se il campione è formato da n elementi, ogni suo elemento può essere considerato come la realizzazione della variabile casuale X_i , indicando con X_i la i -esima estrazione della v.c. X .

Il campionamento casuale

Se il campione è formato da n elementi, ogni suo elemento può essere considerato come la realizzazione della variabile casuale X_i , indicando con X_i la i -esima estrazione della v.c. X .

	1 os	2 os	3 os	4 os	5 os	6 os	7 os	8 os
1 cam	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
2 cam	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}
.....								
Inf. cam	$X_{\infty 1}$	$X_{\infty 2}$	$X_{\infty 3}$	$X_{\infty 4}$	$X_{\infty 5}$	$X_{\infty 6}$	$X_{\infty 7}$	$X_{\infty 8}$

Ciasuna v.c osservazione campionaria X_i , avrà la stessa distribuzione e gli stessi parametri della v.c. X nella popolazione.

Il campionamento casuale

osservazione campionaria X_i è una variabile casuale, ogni funzione funzione che dipenda dalle osservazioni campionarie sarà essa stessa una variabile casuale e viene chiamata statistica statistica.

Å Quindi, i valori campionari ottenuti compiendo una qualsiasi trasformazione sui soli valori osservati vengono chiamate statistiche, mentre i rispettivi valori della popolazione, che sono delle costanti, vengono definiti parametri parametri.

$\bar{X} \rightarrow$ Media campionaria
statistica

$\mu \rightarrow$ Media della popolazione
parametro

Å Definiamo statistica $T_n = T(X_1, X_2, \dots, X_n)$ una qualsiasi funzione a valori reali del campione casuale $(X_1, X_2, \dots, X_n)$ che non dipende da quantità incognite.

Å Il valore della statistica T_n calcolata sul campione osservato $(x_1, x_2, \dots, x_n)$ costituisce la statistica calcolata statistica calcolata $t_n = T(x_1, x_2, \dots, x_n)$.

$T_n \rightarrow$ statistica
v.c.

$t_n \rightarrow$ statistica calcolata
realizzazione di una v.c.

La distribuzione campionaria

Dato un campione casuale:

$$(X_1, X_2, \dots, X_n)$$

e definita la statistica:

$$T_n = T(X_1, X_2, \dots, X_n)$$

ottenuta come elaborazione delle osservazioni campionarie, la distribuzione di probabilità della statistica $T_n(X_1, X_2, \dots, X_n)$ viene definita distribuzione Campionaria di T_n .

Ogni statistica è, dunque, una sintesi delle variabili casuali campionarie.

À Media campionaria

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

À Varianza campionaria

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

À Varianza campionaria corretta

$$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

La media campionaria

La variabile casuale **Media campionaria** è una combinazione lineare delle variabili casuali **Osservazioni campionarie** $X_i, i=1, \dots, n$.

Le n variabili casuali $(X_1, X_2, \dots, X_n)$ sono indipendenti e identicamente distribuite alla variabile X nella popolazione, con media μ e varianza σ^2 .

E' quindi importante definire la distribuzione della media campionaria ma, prima di questo, determinarne il **valore atteso** e la **varianza**.

La media campionaria

Esempio

Popolazione 7 5 10

$$N = 3 \quad \mu = 7.33 \quad \sigma^2 = 4.22$$

Campione

$$n = 2$$

1° caso: estrazione con reintroduzione

Media 7.33 7.33

Varianza 4.22 4.22

La media campionaria

Esempio

Popolazione 7 5 10

$$N = 3 \quad \mu = 7.33 \quad \sigma^2 = 4.22$$

Campione

$$n = 2$$

1° caso: estrazione con reintroduzione

campione	X_1	X_2	$\bar{X}$
1°	5	5	5
2°	5	7	6
3°	5	10	7.5
4°	7	5	6
5°	7	7	7
6°	7	10	8.5
7°	10	7	7.5
8°	10	5	8.5
9°	10	10	10

Media 7.33 7.33 7.33

Varianza 4.22 4.22 2.11

La media campionaria

Esempio

Popolazione 7 5 10

$N = 3$ $\mu = 7.33$ $\sigma^2 = 4.22$

Campione

$n = 2$

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

$$Var(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

La media campionaria

Esempio

Popolazione 7 5 10

$$N = 3 \quad \mu = 7.33 \quad \sigma^2 = 4.22$$

Campione

$$n = 2$$

1° caso: estrazione con reintroduzione

campione	X_1	X_2	$\bar{X}$
1°	5	5	5
2°	5	7	6
3°	5	10	7.5
4°	7	5	6
5°	7	7	7
6°	7	10	8.5
7°	10	7	7.5
8°	10	5	8.5
9°	10	10	10

$$E(\bar{X}) = \mu = 7.33$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{4.22}{2} = 2.11$$

Media 7.33 7.33 7.33

Varianza 4.22 4.22 2.11

La media campionaria

Esempio

Popolazione 7 5 10

$$N = 3 \quad \mu = 7.33 \quad \sigma^2 = 4.22$$

Campione

$$n = 2$$

2° caso: estrazione senza reintroduzione

campione	X_1	X_2	
1°	5	5	5
2°	5	7	6
3°	7	10	7.5
4°	7	5	6
5°	10	7	7
6°	10	10	8.5

Media 7.33 7.33 7.33
 Varianza 4.22 4.22 1.06

La media campionaria

2° caso: estrazione senza reintroduzione

$$Var(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) =$$

$$= \frac{1}{n^2} [Var(X_1) + Var(X_2) + \dots + Var(X_n)] =$$

$$= \frac{1}{n^2} \left[\sum_{i=1}^n Var(X_i) + 2 \sum_{i < j} Cov(X_i, X_j) \right] =$$

$$= \frac{1}{n^2} \left[n \cdot \sigma^2 + 2 \frac{n(n-1)}{2} \cdot \frac{-\sigma^2}{N-1} \right] =$$

$$= \frac{1}{n^2} \left[n \cdot \sigma^2 - \sigma^2 \frac{n(n-1)}{N-1} \right] =$$

$$= \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

La media campionaria

Esempio

Popolazione 7 5 10

$N = 3$ $\mu = 7.33$ $\sigma^2 = 4.22$

Campione

$n = 2$

2° caso: estrazione senza reintroduzione

campione	X_1	X_2	$\bar{X}$
1°	5	5	5
2°	5	7	6
3°	7	10	7.5
4°	7	5	6
5°	10	7	7
6°	10	10	8.5

Media 7.33 7.33 7.33
 Varianza 4.22 4.22 1.06

$$E(\bar{X}) = \mu = 7.33$$

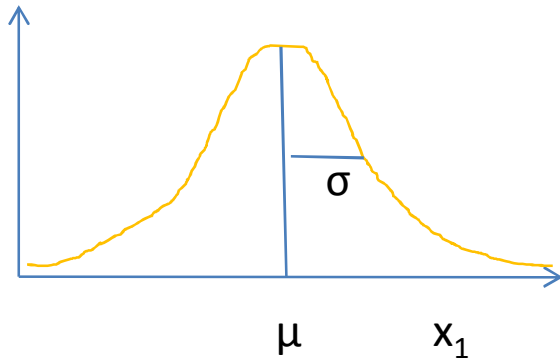
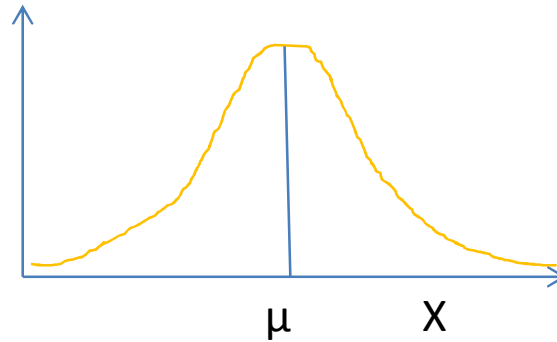
$$Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{4.22}{2} \cdot \frac{3-2}{3-1} = 1.06$$

La media campionaria

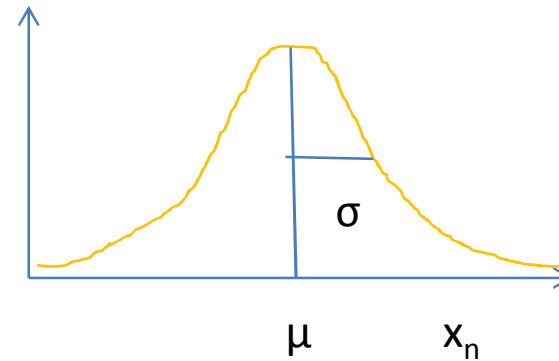
- Å Quando $n=1$ i risultati ottenuti con lo schema di campionamento con reintroduzione coincidono con quelli ottenuti nel campionamento senza reintroduzione;
- Å Quando è $n=N$ la varianza della media campionaria nello schema di campionamento senza reintroduzione è nulla. In questo caso, infatti, il campione coincide con la popolazione e non si ha più alcuna incertezza legata al campionamento;
- Å Nel caso più comune in cui è $n < N$, il fattore di correzione utilizzato nello schema senza reintroduzione è < 1 . Questo vuol dire che la varianza della media campionaria nello schema senza reintroduzione è minore di quella che si ottiene nello schema con reintroduzione;
- Å Quando è $n \ll N$, il fattore di correzione per lo schema di campionamento senza reintroduzione è prossimo a 1. La differenza tra i due schemi di campionamento può quindi essere considerata trascurabile.

La media campionaria distribuzione

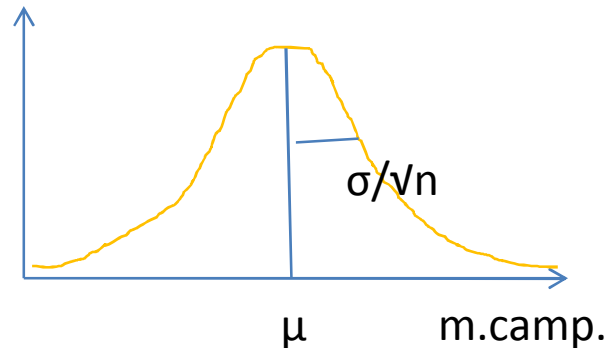
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



.....



$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$



La media campionaria

$$X_i \sim N(\mu, \sigma^2) \quad \forall i$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \sim N$$

Campionamento con reintroduzione:

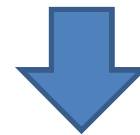
$$E(\bar{X}) = \mu \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$



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

Campionamento senza
reintroduzione:

$$E(\bar{X}) = \mu \quad Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$



$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}}} \sim N(0,1)$$

La media campionaria

Cosa succede se non si conosce la distribuzione della v.c. media campionaria ?

Il teorema del limite centrale

Tutte le volte che un fenomeno reale può essere interpretato come la somma, oppure la media, di un gran numero di cause indipendenti, nessuna delle quali ha prevalenza sulle altre, indipendentemente dai modelli probabilistici che generano le singole variabili casuali è ragionevole attendersi che la distribuzione di probabilità di quel fenomeno possa essere approssimabile mediante la distribuzione della v.c. Normale.

Sotto condizioni molto generali, la somma di n v.c. indipendenti è asintoticamente Normale, e questo è vero qualunque sia il tipo di distribuzione di ciascuna delle v.c. $X_i, i=1, \dots, n$.

Esempio (campionamento con reintroduzione)

Supponiamo che la durata delle lampadine prodotte da una certa macchina abbia distribuzione Normale con media $\mu=2.750$ ore e σ pari a 118 ore.

Estraendo un campione di 21 lampadine, qual è la probabilità che la durata media sia:

- a. Superiore a 2.800 ore;
- b. Superiore a 2.730 ore;
- c. Compresa tra 2.700 e 2.800 ore?

La media campionaria

Esempio (campionamento con reintroduzione)

Supponiamo che la durata delle lampadine prodotte da una certa macchina abbia distribuzione Normale con media $\mu=2.750$ ore e σ pari a 118 ore.

Estraendo un campione di 21 lampadine, qual è la probabilità che la durata media sia:

- Superiore a 2.800 ore;
- Superiore a 2.730 ore;
- Compresa tra 2.700 e 2.800 ore?

$$P(X > 2800) = P\left(X > \frac{X - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{2800 - 2750}{118/\sqrt{21}}\right) = P(Z > 1.94) = 0.5 - 0.4738 = 0.0262$$

$$P(X > 2800) = P\left(X > \frac{X - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{2730 - 2750}{118/\sqrt{21}}\right) = P(Z > -0.78) = 0.5 + 0.2823 = 0.7823$$

$$P(2700 \leq X \leq 2800) = P\left(\frac{2700 - 2750}{118/\sqrt{21}} \leq X \leq \frac{2800 - 2750}{118/\sqrt{21}}\right) = P(-1.94 \leq Z \leq 1.94) \\ = 0.4738 + 0.4738 = 0.9476$$

La media campionaria

Esempio

Le funi di sostegno di un ponte sono formate da cavi di acciaio.

La resistenza alla trazione di ogni cavo è una variabile casuale con media $\mu=0,1$ tonnellate e $\sigma^2=0,06$ tonnellate.

Assumendo che una fune abbia una resistenza alla trazione uguale alla somma delle resistenze dei cavi che la compongono, si determini la probabilità che una fune costituita da 100 cavi sopporti una trazione di 9 tonn.

La media campionaria

Esempio

Le funi di sostegno di un ponte sono formate da cavi di acciaio. La resistenza alla trazione di ogni cavo è una variabile casuale con media $\mu=0,1$ tonnellate e $\sigma=0,06$ tonnellate.

Assumendo che una fune abbia una resistenza alla trazione uguale alla somma delle resistenze dei cavi che la compongono, si determini la probabilità che una fune costituita da 100 cavi sopporti una trazione di 9 tonn.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \Rightarrow \sum_{i=1}^n X_i = n\bar{X}$$

$$\begin{aligned}
 P\left(\sum_{i=1}^n X_i \geq 9\right) &= P(100 \cdot \bar{X}) \geq 9 = P(\bar{X} \geq 0.09) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{0.09 - 0.1}{0.06/\sqrt{100}}\right) = P\left(Z \geq \frac{-0.01}{0.006}\right) \\
 &= P(Z \geq -1.67) = 0.9525
 \end{aligned}$$

$$= P(Z \leq 1.67) = 0.9525$$

Esempio (v.c. Binomiale)

Da studi interni è noto che il 35% dei clienti del supermercato GF.

Ad una cassa sono in fila 5 clienti. Qual è la probabilità che:

- a. Paghino tutti in contanti?
- b. Nessuno paghi in contanti?
- c. Due paghino in contanti, gli altri con Carta?
- d. I primi due paghino in contanti, gli altri con Carta?
- e. Almeno 1 paghi con la carta?

V.C. discrete

5 clienti del supermercato GF.

Ad una cassa sono in fila 5 clienti. Qual è la probabilità che:

- Paghino tutti in contanti?
- Nessuno paghi in contanti?
- Due paghino in contanti, gli altri con Carta?
- I primi due paghino in contanti, gli altri con Carta?
- Almeno 1 paghi con la carta?

- Calcolare la prob. che tutti paghino in contanti significa calcolare la prob. Che nessuno paghi con al carta, ovvero che in cinque prove ho sempre insuccessi, quindi $X=0$

$$P(X = 0) = \frac{5!}{0! 5!} 0.35^0 0.65^5 = 0.116$$

- Calcolare la prob. che nessuno paghi in contanti significa calcolare la prob. che $X=5$, ovvero in 5 prove ho sempre successi

$$P(X = 5) = \frac{5!}{5! 0!} 0.35^5 0.65^0 = 0.005$$

- In questo caso voglio sapere qual è la prob che in cinque prove ottengo 3 successi, quindi $X=3$

$$P(X = 3) = \frac{5!}{3! 2!} 0.35^3 0.65^2 = 0.182$$

- In questo caso significa considerare ogni prova singolarmente e valutare qual è la prob che nella la prima e seconda prova ho $X=0$, e $X=1$ per le 3 successive prove

$$\begin{aligned} &P(X = 0) \cap P(X = 0) \cap P(X = 1) \cap P(X = 1) \cap P(X = 1) \\ &= \frac{1!}{0! 1!} 0.35^0 0.65^1 \cdot \frac{1!}{0! 1!} 0.35^0 0.65^1 \cdot \frac{1!}{1! 1!} 0.35^1 0.65^0 \cdot \frac{1!}{1! 1!} 0.35^1 0.65^0 \cdot \frac{1!}{1! 1!} 0.35^1 0.65^0 \\ &= 0.35^3 \cdot 0.65^2 \end{aligned}$$

- Significa calcolare la prob che $X \geq 1$

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 1 - P(X = 0) = 1 - 0.116 \\ &= 0.884 \end{aligned}$$