NoiStudiamo

DISPENSA DI TEORIA

Metodi Matematici e Statistici

Probabilità, statistica descrittiva e inferenziale, regressione, numeri pseudocasuali, metodo Monte Carlo e catene di Markov

A CURA DI

Damiano La Pusata

Ottobre 2026

INDICE

Indice dei capitoli

I capitoli della dispensa, nell'ordine in cui conviene studiarli.

1. Calcolo delle probabilità 3

2. Variabili aleatorie 9

3. Legge dei grandi numeri e approssimazione normale 16

4. Statistica descrittiva 21

5. Statistica inferenziale: stimatori e intervalli di confidenza 25

6. Verifica di ipotesi 30

7. Regressione lineare 35

8. Numeri pseudocasuali e metodo Monte Carlo 40

9. Catene di Markov 44

10. Formulario 49

CAPITOLO 1

Calcolo delle probabilità

Gli strumenti per ragionare in modo rigoroso sull'incertezza: eventi, definizioni di probabilità, calcolo combinatorio, probabilità condizionata e teorema di Bayes.

Esperimenti aleatori ed eventi

Un **esperimento aleatorio** è un fenomeno il cui esito non è prevedibile con certezza prima di osservarlo: il lancio di un dado, il numero di richieste che arrivano a un server in un minuto, la durata di una batteria. L'insieme di tutti gli esiti possibili si chiama **spazio campionario** e si indica con Ω; i suoi elementi ω sono gli **esiti elementari**.

DEFINIZIONE

Un **evento** è un'affermazione sull'esito di un esperimento aleatorio che, a esperimento concluso, risulta vera oppure falsa. In termini insiemistici un evento è un sottoinsieme $E \subseteq \Omega$: l'evento si verifica se l'esito osservato appartiene a E.

L'evento Ω si verifica sempre ed è detto **evento certo**; l'insieme vuoto $\varnothing$ non si verifica mai ed è detto **evento impossibile**. Un evento formato da un solo esito si dice **evento elementare**.

Nel lancio di un dado, per esempio, $\Omega = \{1, 2, 3, 4, 5, 6\}$ e l'affermazione "esce un numero pari" corrisponde all'evento $E = \{2, 4, 6\}$. Gli eventi si combinano tra loro esattamente come le proposizioni della logica, e per questo le operazioni hanno anche un nome logico.

OPERAZIONE	NOTAZIONE	SI VERIFICA QUANDO
Unione (somma logica)	$A \cup B$	si verifica almeno uno tra A e B ("A oppure B")
Intersezione (prodotto logico)	$A \cap B$	si verificano sia A sia B ("A e B")
Contrario (complementare)	A^c	non si verifica A ("non A")
Implicazione	$A \subseteq B$	ogni volta che si verifica A si verifica anche B
Differenza	$A \setminus B = A \cap B^c$	si verifica A ma non B

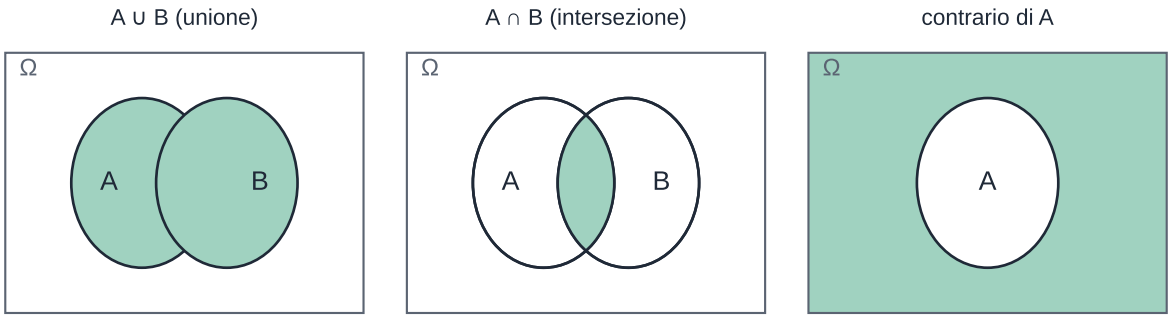

Figura 1.1. Le operazioni tra eventi: la zona più scura è l'evento risultante.

DEFINIZIONE

Due eventi A e B si dicono **incompatibili** (o disgiunti) se non possono verificarsi contemporaneamente, cioè se $A \cap B = \varnothing$. Più eventi sono **a due a due incompatibili** se lo è ogni coppia di essi.

Due regole permettono di passare dal contrario di un'unione all'intersezione dei contrari, e viceversa. Si chiamano **leggi di De Morgan** e sono lo strumento principale per tradurre frasi negative in simboli:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

A parole: "non è vero che si verifica almeno uno dei due" equivale a "non si verifica nessuno dei due"; "non si verificano entrambi" equivale a "almeno uno dei due non si verifica".

ESEMPIO: TRADURRE UNA FRASE IN SIMBOLI

Un'applicazione funziona se il database è raggiungibile e almeno uno dei due server di calcolo è attivo. Poniamo A = "il primo server è guasto", B = "il secondo server è guasto", D = "il database non è raggiungibile". Procediamo in tre passi.

Passo 1. Scriviamo l'evento "l'applicazione funziona". Il database deve essere raggiungibile, cioè D^c; almeno un server deve essere attivo, cioè $A^c \cup B^c$. Le due condizioni devono valere insieme, quindi:

$$\text{Funziona} = D^c \cap (A^c \cup B^c)$$

Passo 2. L'evento "l'applicazione non funziona" è il contrario. Applichiamo De Morgan al prodotto esterno:

$$\text{Funziona}^c = (D^c)^c \cup (A^c \cup B^c)^c = D \cup (A^c \cup B^c)^c$$

Passo 3. Applichiamo De Morgan anche alla parentesi: $(A^c \cup B^c)^c = A \cap B$. Si ottiene

$$\text{Funziona}^c = D \cup (A \cap B)$$

cioè l'applicazione si ferma se il database non è raggiungibile oppure se entrambi i server sono guasti, che è esattamente ciò che ci si aspetta dal buon senso.

Le definizioni di probabilità

La probabilità di un evento è un numero tra 0 e 1 che misura quanto ci aspettiamo che l'evento si verifichi. Storicamente sono state proposte quattro definizioni: le prime tre nascono da situazioni concrete, la quarta le riassume tutte in un sistema di assiomi.

1. DEFINIZIONE CLASSICA

Se l'esperimento ha n esiti elementari **tutti ugualmente possibili** e l'evento E si verifica in h di essi, la probabilità di E è il rapporto tra casi favorevoli e casi possibili:

$$P(E) = \frac{h}{n} = \frac{\text{numero di casi favorevoli}}{\text{numero di casi possibili}}$$

Il limite di questa definizione è che presuppone l'equiprobabilità, cioè usa il concetto di probabilità per definire la probabilità stessa, e non si applica quando gli esiti non sono simmetrici (un dado truccato) o sono infiniti.

2. DEFINIZIONE FREQUENTISTA

Se l'esperimento si può ripetere molte volte nelle stesse condizioni, la probabilità di E è il valore a cui tende la **frequenza relativa** k/n, dove k è il numero di volte in cui E si è verificato su n prove. È l'idea su cui si basano tutte le simulazioni al calcolatore: per stimare una probabilità si ripete l'esperimento tante volte e si conta. Il limite è che richiede esperimenti ripetibili, cosa che non sempre accade (la probabilità che un certo progetto venga consegnato

in tempo non si può misurare ripetendo il progetto).

3. DEFINIZIONE SOGGETTIVA

Secondo l'impostazione di De Finetti, la probabilità è il **grado di fiducia** che una persona coerente attribuisce al verificarsi di un evento. Si rende operativa con una scommessa: P(E) = p è la quota che si è disposti a pagare, in proporzione a una somma S, per ricevere S se E si verifica e nulla altrimenti. Se E si verifica il guadagno è $G_1 = S(1 - p)$; se non si verifica è $G_2 = -pS$. La scommessa è **coerente** se non garantisce né una vincita certa né una perdita certa, cioè se G_1 e G_2 non hanno lo stesso segno: questo accade esattamente quando $0 \le p \le 1$.

4. DEFINIZIONE ASSIOMATICA

Kolmogorov ha mostrato che tutte le definizioni precedenti soddisfano le stesse poche regole, e ha preso quelle regole come assiomi. La probabilità diventa così un oggetto matematico definito dalle sue proprietà, indipendentemente da come la si interpreta.

DEFINIZIONE

Uno **spazio di probabilità** è una terna (Ω, F, P) in cui Ω è lo spazio campionario, F è la famiglia degli eventi e P è una funzione che associa a ogni evento la sua probabilità.

La famiglia F deve essere una **σ algebra**, cioè deve contenere Ω, deve contenere il contrario di ogni suo elemento e deve contenere l'unione di ogni successione (finita o numerabile) di suoi elementi.

La funzione $P: F \to [0, 1]$ deve soddisfare i tre **assiomi di Kolmogorov**: $P(A) \ge 0$ per ogni evento A (non negatività); $P(\Omega) = 1$ (normalizzazione); se $A_1, A_2, \ldots$ sono a due a due incompatibili, allora la probabilità della loro unione è la somma delle probabilità (additività numerabile).

Dagli assiomi discendono tutte le regole di calcolo che si usano nella pratica. Le più importanti sono riassunte qui sotto; ciascuna si dimostra in una riga scrivendo l'evento come unione di pezzi incompatibili.

DA RICORDARE

Probabilità del contrario: $P(A^c) = 1 - P(A)$. Infatti A e A^c sono incompatibili e la loro unione è Ω.

Monotonia: se $A \subseteq B$ allora $P(A) \le P(B)$.

Probabilità dell'unione di due eventi qualsiasi:

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

Si sottrae l'intersezione perché, sommando P(A) e P(B), i casi comuni vengono contati due volte. Se A e B sono incompatibili l'intersezione è vuota e resta la semplice somma.

La formula dell'unione si generalizza a più eventi con il **principio di inclusione ed esclusione**: si sommano le probabilità dei singoli eventi, si sottraggono quelle delle intersezioni a due a due, si aggiungono quelle delle intersezioni a tre a tre, e così via alternando i segni. Per tre eventi:

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

Richiami di calcolo combinatorio

Con la definizione classica calcolare una probabilità significa contare: quanti sono i casi possibili, quanti i favorevoli. Il calcolo combinatorio fornisce le formule per contare senza elencare. Tutto si riduce a due domande da porsi sempre: **l'ordine conta?** e **si possono ripetere gli elementi?** Le quattro risposte possibili danno le quattro formule della tabella, dove si scelgono r elementi da un insieme di n.

TIPO	ORDINE	RIPETIZIONE	NUMERO DI MODI
Disposizioni con ripetizione	conta	sì	n^r
Disposizioni semplici	conta	no	$n! / (n - r)!$
Combinazioni semplici	non conta	no	$n! / (r!\,(n - r)!)$, il coefficiente binomiale
Combinazioni con ripetizione	non conta	sì	$(n + r - 1)! / (r!\,(n - 1)!)$
Permutazioni (r = n)	conta	no	$n!$

Il **coefficiente binomiale** si legge "n su r" e conta i sottoinsiemi di r elementi di un insieme di n elementi:

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

Più in generale, il numero di modi di suddividere n oggetti in k gruppi di numerosità $n_1, \dots, n_k$ (con $n_1 + \dots + n_k = n$) è dato dal **coefficiente multinomiale** $n! / (n_1!\, n_2! \cdots n_k!)$.

ESEMPIO: ESTRAZIONI DA UN'URNA

Un'urna contiene 4 palline rosse e 6 blu. Se ne estraggono 3 insieme, cioè senza reinserimento e senza tener conto dell'ordine. Calcoliamo la probabilità che esattamente 2 siano rosse.

Passo 1, casi possibili. L'ordine non conta e non c'è ripetizione: sono le combinazioni di 10 palline prese 3 alla volta.

$$\binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$$

Passo 2, casi favorevoli. Bisogna scegliere 2 rosse tra 4 e, indipendentemente, 1 blu tra 6; i modi si moltiplicano.

$$\binom{4}{2}\binom{6}{1} = 6 \cdot 6 = 36$$

Passo 3, probabilità. Si fa il rapporto: P = 36/120 = 0,3. Questo schema di calcolo è proprio quello della distribuzione ipergeometrica, che incontreremo nel prossimo capitolo.

Probabilità condizionata e indipendenza

Spesso si conosce un'informazione parziale sull'esito. Sapere che un evento B si è verificato restringe lo spazio dei casi possibili da Ω a B, e la probabilità di A va ricalcolata dentro questo nuovo universo.

DEFINIZIONE

Dati due eventi A e B con P(B) > 0, la **probabilità condizionata** di A dato B è

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

e si legge "probabilità di A sapendo che si è verificato B". L'evento condizionato A | B è vero se B è vero e anche A lo è, falso se B è vero e A è falso, e indeterminato se B è falso.

Rigirando la definizione si ottiene la **regola del prodotto**, $P(A \cap B) = P(A \mid B)\, P(B)$, che è il modo più comune di calcolare la probabilità che due eventi si verifichino insieme quando uno dei due avviene "dopo" l'altro: prima si estrae una carta, poi una seconda dal mazzo rimasto.

DEFINIZIONE

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

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

Se P(B) > 0 questo equivale a P(A | B) = P(A): sapere che B si è verificato non cambia la probabilità di A.

ATTENZIONE

Indipendenti e incompatibili sono concetti molto diversi. Se A e B sono incompatibili e hanno probabilità positive, allora P(A ∩ B) = 0 mentre P(A) P(B) > 0: due eventi incompatibili sono quindi **fortemente dipendenti**, perché il verificarsi dell'uno esclude l'altro.

Con tre o più eventi bisogna distinguere due livelli. Gli eventi A, B, C sono **indipendenti a due a due** se la regola del prodotto vale per ogni coppia; sono una **famiglia di eventi indipendenti** se vale anche P(A ∩ B ∩ C) = P(A) P(B) P(C). Il primo livello non implica il secondo.

ESEMPIO: INDIPENDENZA A COPPIE MA NON CONGIUNTA

Si lanciano due monete equilibrate e si considerano A = "la prima dà testa", B = "la seconda dà testa", C = "le due monete danno lo stesso risultato". Ognuno dei tre eventi ha probabilità 1/2. Ogni intersezione a due corrisponde al solo esito "testa, testa" oppure a un solo esito, e ha probabilità 1/4 = 1/2 · 1/2: le tre coppie sono indipendenti. Però A ∩ B ∩ C è ancora l'esito "testa, testa", con probabilità 1/4, mentre il prodotto P(A) P(B) P(C) vale 1/8. I tre eventi non formano una famiglia indipendente: sapere che A e B si sono verificati rende C certo.

Teorema delle probabilità totali e teorema di Bayes

Molti problemi hanno questa struttura: un evento B può essere prodotto da diverse "cause" alternative, di cui si conosce la frequenza, e si sa con che probabilità ciascuna causa produce B. Le cause formano una partizione dello spazio campionario.

TEOREMA

Probabilità totali. Siano $A_1, \dots, A_n$ eventi a due a due incompatibili, con probabilità positive, la cui unione è Ω (una **partizione** di Ω). Allora per ogni evento B

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)\,P(A_i)$$

DIMOSTRAZIONE

Poiché gli A_i ricoprono Ω, l'evento B si scrive come unione dei pezzi $B \cap A_i$, che sono a due a due incompatibili perché lo sono gli A_i. Per l'additività e poi per la regola del prodotto:

$$P(B) = P\left(\bigcup_i (B \cap A_i)\right) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i)\,P(A_i)$$

TEOREMA

Bayes. Dati due eventi A e B con probabilità positive,

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Se $A_1, \ldots, A_n$ è una partizione di Ω, usando le probabilità totali al denominatore si ottiene la forma che si usa negli esercizi:

$$P(A_k \mid B) = \frac{P(B \mid A_k)\,P(A_k)}{\sum_i P(B \mid A_i)\,P(A_i)}$$

DIMOSTRAZIONE

Per la definizione di probabilità condizionata $P(A \cap B) = P(A \mid B)\,P(B)$ e anche $P(A \cap B) = P(B \mid A)\,P(A)$. Uguagliando i due secondi membri e dividendo per $P(B)$ si ottiene la tesi.

Il teorema di Bayes permette di "invertire" il condizionamento: dalle probabilità che la causa produca l'effetto, $P(B \mid A_k)$, si ricava la probabilità che, osservato l'effetto, la causa sia proprio A_k. Le $P(A_k)$ si chiamano probabilità **a priori**, le $P(A_k \mid B)$ probabilità **a posteriori**.

ESEMPIO: UN FILTRO ANTISPAM

Il 20% delle email ricevute da un'azienda è spam. Un filtro segnala come sospetto il 95% delle email di spam e, per errore, il 4% delle email legittime. Calcoliamo la probabilità che un'email segnalata sia davvero spam.

Passo 1, le cause. S = "l'email è spam" e S^c = "l'email è legittima" formano una partizione, con $P(S) = 0{,}2$ e $P(S^c) = 0{,}8$.

Passo 2, l'effetto. Sia R = "l'email viene segnalata". I dati sono $P(R \mid S) = 0{,}95$ e $P(R \mid S^c) = 0{,}04$.

Passo 3, probabilità totali.

$$P(R) = 0{,}95 \cdot 0{,}2 + 0{,}04 \cdot 0{,}8 = 0{,}19 + 0{,}032 = 0{,}222$$

Passo 4, Bayes.

$$P(S \mid R) = \frac{0{,}95 \cdot 0{,}2}{0{,}222} = \frac{0{,}19}{0{,}222} \approx 0{,}856$$

Circa l'86% delle email segnalate è spam. La probabilità è salita dal 20% a priori all'86% a posteriori: l'informazione "segnalata" ha reso lo spam più di quattro volte più probabile. Se lo spam fosse raro, per esempio l'1% delle email, la stessa procedura darebbe circa il 19%: con eventi rari anche un buon filtro produce molti falsi allarmi.

DA RICORDARE

Schema per i problemi di Bayes: individuare le cause che formano la partizione; scrivere le probabilità a priori $P(A_i)$ e le probabilità dell'effetto $P(B \mid A_i)$; calcolare $P(B)$ con le probabilità totali; applicare Bayes. Un risultato fuori dall'intervallo [0, 1] segnala sempre un errore di impostazione.

CAPITOLO 2

Variabili aleatorie

Dal singolo evento ai numeri: variabili discrete e continue, funzione di ripartizione, valore atteso, varianza e momenti, e le distribuzioni notevoli che descrivono i fenomeni più comuni.

La definizione

Nella maggior parte dei problemi non interessa l'esito in sé, ma un numero associato all'esito: la somma dei punti di due dadi, il numero di pezzi difettosi in un lotto, il tempo di attesa a uno sportello. Una variabile aleatoria è proprio la regola che trasforma ogni esito in un numero.

DEFINIZIONE

Dato uno spazio di probabilità (Ω, F, P), una **variabile aleatoria** è una funzione $X: \Omega \to \mathbb{R}$ che associa a ogni esito elementare ω un numero reale $X(\omega)$, in modo che per ogni $x \in \mathbb{R}$ l'insieme $\{\omega \in \Omega : X(\omega) \le x\}$ sia un evento, cioè appartenga a F.

Nonostante il nome, una variabile aleatoria non è una variabile ma una **funzione**: il suo argomento è l'esito dell'esperimento e il suo valore è un numero. La condizione sugli insiemi $\{X \le x\}$ garantisce che abbia senso calcolarne la probabilità.

Esempio. Si lanciano due monete e X conta le teste. Lo spazio campionario è $\Omega = \{TT, TC, CT, CC\}$ e X è la funzione che vale $X(TT) = 2$, $X(TC) = X(CT) = 1$, $X(CC) = 0$. L'evento $\{X = 1\}$ è il sottoinsieme $\{TC, CT\}$ e ha probabilità 1/2.

Variabili discrete e continue

Una variabile aleatoria si dice **discreta** se assume un numero finito o un'infinità numerabile di valori (si possono elencare: 0, 1, 2, …); si dice **continua** se i suoi valori riempiono un intervallo e la probabilità si distribuisce come una massa spalmata, descritta da una densità.

DEFINIZIONE

La **densità discreta** (o funzione di probabilità) di una variabile discreta X con valori $x_1, x_2, \ldots$ è la funzione $p(x_i) = P(X = x_i)$. Deve soddisfare $p(x_i) \ge 0$ per ogni i e $\Sigma\, p(x_i) = 1$.

Una variabile X è **continua** se esiste una funzione $f(x) \ge 0$, detta **densità**, con integrale su tutto $\mathbb{R}$ uguale a 1, tale che per ogni $a < b$

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

ATTENZIONE

Per una variabile continua la probabilità di un singolo valore è sempre zero, $P(X = x) = 0$, perché l'integrale su un punto è nullo. Quindi per le continue scrivere $\le$ oppure $<$ è indifferente.

La densità $f(x)$ **non è una probabilità** e può superare 1: l'uniforme sull'intervallo $[0; 0{,}5]$ ha densità costante uguale a 2. Solo l'area sotto la densità è una probabilità.

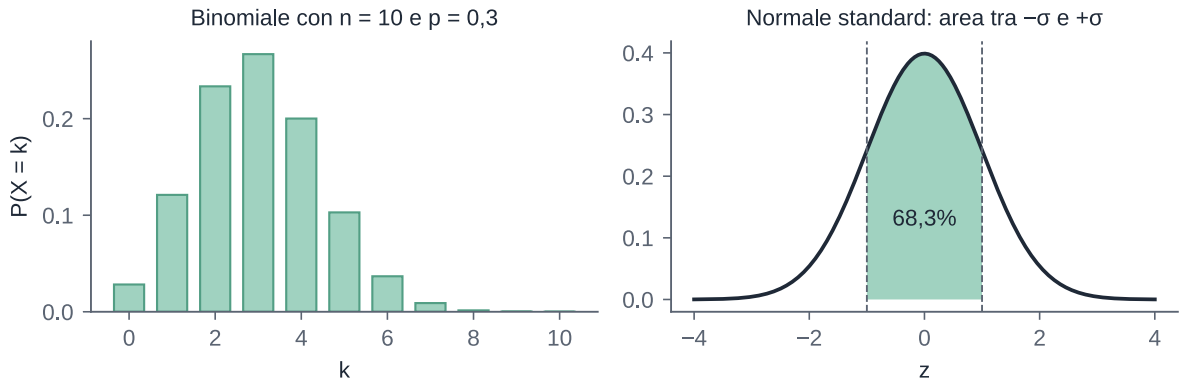

Figura 2.1. A sinistra una densità discreta (barre), a destra una densità continua (curva).

DEFINIZIONE

La **funzione di ripartizione** di una variabile aleatoria X (discreta o continua) è F(x) = P(X ≤ x). È non decrescente, vale 0 verso meno infinito e 1 verso più infinito, ed è continua a destra.

Permette di calcolare la probabilità di qualunque intervallo: P(a < X ≤ b) = F(b) − F(a). Nel caso continuo la densità è la derivata della ripartizione, f(x) = F′(x).

Per le variabili discrete la ripartizione è una funzione a gradini, che salta di $p(x_i)$ in corrispondenza di ogni valore x_i; per le continue è una curva senza salti. L'inversa della funzione di ripartizione si chiama **funzione quantile**: il quantile di ordine q è il valore x tale che F(x) = q. Servirà per gli intervalli di confidenza e per la generazione di numeri casuali.

Valore atteso, varianza e momenti

Una distribuzione contiene molta informazione; spesso basta riassumerla con due numeri, uno che dice **dove** si trova il centro e uno che dice **quanto** i valori si disperdono attorno al centro. Il valore atteso è il baricentro della distribuzione: se si immaginano le probabilità come pesi appoggiati su un'asta, è il punto in cui l'asta sta in equilibrio.

DEFINIZIONE

Il **valore atteso** (o media, o speranza matematica) di X è

$$E[X] = \sum_i x_i\, p(x_i) \quad \text{(discreta)}, \qquad E[X] = \int_{-\infty}^{+\infty} x\, f(x)\, dx \quad \text{(continua)}$$

La **varianza** è il valore atteso del quadrato dello scarto dalla media, con μ = E[X]:

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right] = E[X^2] - \mu^2$$

La radice quadrata della varianza si chiama **deviazione standard** (o scarto quadratico medio) e si indica con σ: ha la stessa unità di misura di X.

La seconda forma della varianza, $E[X^2] - \mu^2$, si ottiene sviluppando il quadrato e usando la linearità; è quasi sempre la più comoda per i calcoli. Le proprietà da conoscere sono poche e si usano di continuo.

DA RICORDARE

Linearità: $E[aX + b] = a\,E[X] + b$, e $E[X + Y] = E[X] + E[Y]$ sempre, anche se X e Y non sono indipendenti.

Varianza di una trasformazione lineare: $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$. Aggiungere una costante sposta la distribuzione ma non la disperde; moltiplicare per a ne cambia la scala.

Varianza di una somma: $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, dove $\mathrm{Cov}(X, Y) = E[XY] - E[X]\,E[Y]$ è la **covarianza**. Se X e Y sono indipendenti la covarianza è nulla e le varianze si sommano.

DEFINIZIONE

Il **momento di ordine k** di X è $E[X^k]$; il **momento centrato di ordine k** è $E[(X - \mu)^k]$.

Il momento centrato di ordine 1 vale sempre 0, perché $E[X - \mu] = E[X] - \mu = 0$; quello di ordine 2 è la varianza. I momenti centrati di ordine 3 e 4, divisi per una potenza di σ, danno gli indici di forma della distribuzione.

L'**indice di asimmetria** (skewness) è $\gamma_1 = E[(X - \mu)^3] / \sigma^3$: vale zero per le distribuzioni simmetriche, è positivo se la coda destra è più lunga, negativo se lo è la sinistra. L'**indice di curtosi** è $\gamma_2 = E[(X - \mu)^4] / \sigma^4 - 3$: confronta il peso delle code con quello della normale, per cui vale zero; un valore positivo indica code più pesanti e un picco più appuntito. Il meno 3 serve proprio a rendere la normale il termine di paragone; alcuni testi chiamano curtosi direttamente il rapporto $\beta_2 = E[(X - \mu)^4] / \sigma^4$, che per la normale vale 3.

ESEMPIO: MEDIA E VARIANZA DA UNA DENSITÀ DISCRETA

Sia X con densità $p(k) = c \cdot k$ per $k = 1, 2, 3, 4$. Determiniamo c, la media e la varianza.

Passo 1. Le probabilità devono sommare a 1: $c(1 + 2 + 3 + 4) = 10c = 1$, quindi $c = 1/10$.

Passo 2. La media è la somma dei valori pesati con le probabilità:

$$E[X] = 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{4}{10} = \frac{30}{10} = 3$$

Passo 3. Il momento secondo e poi la varianza:

$$E[X^2] = \frac{1 + 8 + 27 + 64}{10} = 10, \qquad \mathrm{Var}(X) = 10 - 3^2 = 1$$

Distribuzioni notevoli discrete

Alcuni schemi si ripetono in situazioni molto diverse, e per questo le loro distribuzioni hanno un nome. Per ciascuna conviene ricordare **che cosa conta**, prima ancora della formula: è da lì che si riconosce quale usare.

BERNOULLI

$X \sim \mathrm{Ber}(p)$ descrive un singolo esperimento con due esiti, successo ($X = 1$) con probabilità p e insuccesso ($X = 0$) con probabilità $q = 1 - p$. Si ha $E[X] = p$ e, poiché $X^2 = X$, $\mathrm{Var}(X) = p - p^2 = pq$.

DEFINIZIONE

La **binomiale** $X \sim \mathrm{Bin}(n, p)$ conta il numero di successi in n prove **indipendenti**, ciascuna con la stessa probabilità di successo p. La densità è

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}, \qquad k = 0, 1, \ldots, n$$

e si ha $E[X] = np$ e $\mathrm{Var}(X) = np(1 - p)$.

La formula si legge così: $p^k(1-p)^{n-k}$ è la probabilità di una particolare sequenza con k successi e n − k insuccessi, e il coefficiente binomiale conta in quanti modi si possono collocare i k successi tra le n prove.

DERIVAZIONE DELLA MEDIA E DELLA VARIANZA

La strada più semplice è scrivere X come somma: $X = X_1 + X_2 + \dots + X_n$, dove X_i vale 1 se la prova i è un successo. Le X_i sono Bernoulli di parametro p e indipendenti. Per la linearità del valore atteso

$$E[X] = \sum_{i=1}^{n} E[X_i] = n\,p$$

e, poiché le X_i sono indipendenti, le varianze si sommano:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\,p\,(1-p)$$

Lo stesso risultato si ottiene anche direttamente dalla densità. Per la media si usa l'identità $k\cdot C(n,k) = n\cdot C(n-1,k-1)$, si raccoglie np e si pone $j = k-1$: la somma che resta è lo sviluppo del binomio $(p+q)^{n-1} = 1$.

$$E[X] = \sum_{k=1}^{n} k\binom{n}{k}p^k q^{n-k} = np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j q^{n-1-j} = np\,(p+q)^{n-1} = np$$

Per la varianza si calcola allo stesso modo $E[X(X-1)] = n(n-1)p^2$, da cui $E[X^2] = n(n-1)p^2 + np$ e infine $\mathrm{Var}(X) = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)$.

DEFINIZIONE

L'**ipergeometrica** conta il numero di successi in n estrazioni **senza reinserimento** da una popolazione di N elementi, di cui K hanno la caratteristica cercata:

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad E[X] = n\frac{K}{N}, \qquad \mathrm{Var}(X) = n\frac{K}{N}\left(1-\frac{K}{N}\right)\frac{N-n}{N-1}$$

La differenza con la binomiale è tutta nel reinserimento: estraendo con reinserimento le prove sono indipendenti e si usa la binomiale, estraendo senza reinserimento la composizione dell'urna cambia a ogni estrazione e si usa l'ipergeometrica. Il fattore $(N-n)/(N-1)$ si chiama **correzione per popolazione finita**: tende a 1 quando N è molto più grande di n, e in quel caso le due distribuzioni quasi coincidono.

DEFINIZIONE

La **geometrica** $X \sim \mathrm{Geom}(p)$ conta il numero di prove necessarie per ottenere il **primo successo**, in prove indipendenti con probabilità di successo p:

$$P(X=k) = (1-p)^{k-1}p, \qquad k = 1, 2, \dots, \qquad E[X] = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

Alcuni testi contano invece gli insuccessi prima del primo successo, $Y = X - 1$, con valori 0, 1, 2, …; in tal caso $E[Y] = (1-p)/p$ e la varianza è la stessa. Conviene dichiarare sempre quale convenzione si usa.

Il legame con la binomiale è diretto: servono più di k prove se e solo se nelle prime k prove non c'è nessun successo, quindi $P(X > k) = (1-p)^k$, che è la probabilità binomiale di zero successi su k prove. La geometrica ha una proprietà caratteristica, l'assenza di memoria.

PROPRIETÀ

Assenza di memoria. Per ogni n, k ≥ 0 vale P(X > n + k | X > n) = P(X > k): sapere di aver già fallito n volte non cambia la distribuzione del numero di prove ancora necessarie.

DIMOSTRAZIONE

Si usa $P(X > m) = (1 - p)^m$. Poiché l'evento {X > n + k} implica {X > n}, l'intersezione dei due è {X > n + k} e

$$P(X > n + k \mid X > n) = \frac{P(X > n + k)}{P(X > n)} = \frac{(1-p)^{n+k}}{(1-p)^n} = (1-p)^k = P(X > k)$$

A parole: chi ha lanciato un dado dieci volte senza ottenere un sei ha, per il lancio successivo, esattamente le stesse probabilità di chi comincia ora. Il dado non ricorda i lanci precedenti.

DEFINIZIONE

La **Poisson** X ∼ Pois(λ) conta il numero di eventi che si verificano in un intervallo fissato di tempo o di spazio, quando gli eventi sono rari, indipendenti e avvengono in media λ volte per intervallo:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots, \qquad E[X] = \mathrm{Var}(X) = \lambda$$

La Poisson è il limite della binomiale quando il numero di prove n è grande, la probabilità p è piccola e il prodotto np = λ resta fisso: in pratica Bin(n, p) si approssima bene con Pois(np) per n ≥ 50 e p ≤ 0,05. Esempi tipici sono il numero di accessi a un sito in un minuto o il numero di errori di stampa in una pagina.

Esempio. Gli errori in una pagina seguono una Poisson con λ = 3. La probabilità di esattamente 2 errori è $3^2 e^{-3}/2! = 4{,}5 \cdot e^{-3} \approx 0{,}224$; quella di al massimo 2 errori è $e^{-3}(1 + 3 + 4{,}5) = 8{,}5 \cdot e^{-3} \approx 0{,}423$.

UNIFORME DISCRETA

X uniforme su {1, 2, …, n} assegna a ogni valore la stessa probabilità 1/n, come il lancio di un dado equilibrato. Si ha E[X] = (n + 1)/2 e $\mathrm{Var}(X) = (n^2 - 1)/12$; per il dado, n = 6, la media è 3,5 e la varianza 35/12 ≈ 2,92.

Distribuzioni notevoli continue

UNIFORME CONTINUA

X ∼ U(a, b) ha densità costante 1/(b − a) sull'intervallo [a, b] e nulla altrove: ogni sottointervallo di uguale lunghezza ha la stessa probabilità. Si ha E[X] = (a + b)/2 e $\mathrm{Var}(X) = (b - a)^2/12$. L'uniforme su [0, 1] è il punto di partenza di tutte le simulazioni al calcolatore.

DEFINIZIONE

L'**esponenziale** X ∼ Exp(λ), con λ > 0, descrive il tempo di attesa fino al prossimo evento in un processo in cui gli eventi arrivano in media λ volte per unità di tempo:

$$f(x) = \lambda e^{-\lambda x} \ (x \geq 0), \qquad F(x) = 1 - e^{-\lambda x}, \qquad E[X] = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Il parametro λ è un **tasso**, non una probabilità: se in media arrivano 4 richieste al minuto, λ = 4 al minuto e il tempo medio tra due richieste è 1/4 di minuto. L'esponenziale è la controparte continua della geometrica e, come quella, non ha memoria: $P(X > s + t \mid X > s) = P(X > t) = e^{-\lambda t}$. Se il numero di arrivi in un intervallo di ampiezza t segue una Poisson di parametro λt, il tempo tra due arrivi consecutivi è esponenziale di parametro λ.

Esempio. Un componente ha durata esponenziale con media 2000 ore, quindi $\lambda = 1/2000$. La probabilità che duri più di 3000 ore è $e^{-3000/2000} = e^{-1,5} \approx 0,223$. Se ha già funzionato per 1000 ore, la probabilità che funzioni altre 3000 ore è ancora $e^{-1,5}$: per l'assenza di memoria il tempo trascorso non conta.

DEFINIZIONE

La **normale** (o gaussiana) $X \sim N(\mu, \sigma^2)$ ha densità

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}$$

con $E[X] = \mu$ e $\mathrm{Var}(X) = \sigma^2$. La densità ha la forma di una campana simmetrica attorno a μ, con due punti di flesso in $\mu - \sigma$ e $\mu + \sigma$.

La **normale standard** $Z \sim N(0, 1)$ ha media 0 e varianza 1; la sua funzione di ripartizione si indica con Φ e i suoi valori si leggono dalle tavole o si calcolano al computer. Ogni normale si riconduce alla standard con la **standardizzazione**:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \qquad P(X \le x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Per simmetria $\Phi(-z) = 1 - \Phi(z)$. Circa il 68,3% della probabilità cade entro una deviazione standard dalla media, il 95,4% entro due e il 99,7% entro tre.

DA RICORDARE

I quantili della normale standard che si usano più spesso: $\Phi(1,645) = 0,95$, $\Phi(1,96) = 0,975$, $\Phi(2,576) = 0,995$. Corrispondono ai livelli di confidenza del 90%, 95% e 99%.

Esempio. Il peso di una confezione è normale con $\mu = 500$ g e $\sigma = 4$ g. La probabilità che una confezione pesi meno di 494 g è $\Phi((494 - 500)/4) = \Phi(-1,5) = 1 - \Phi(1,5) \approx 1 - 0,9332 = 0,0668$, cioè circa il 6,7%.

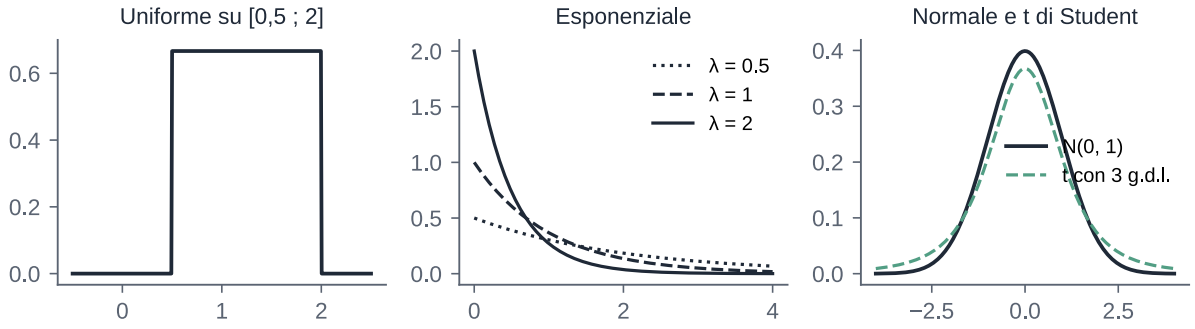

Figura 2.2. Densità continue notevoli. La t ha code più pesanti della normale (linea tratteggiata).

DEFINIZIONE

Se $Z_1, \ldots, Z_n$ sono normali standard indipendenti, la variabile $X = Z_1^2 + \ldots + Z_n^2$ ha distribuzione **chi quadro con n gradi di libertà**, $\chi^2(n)$. Assume solo valori positivi, ha $E[X] = n$ e $\mathrm{Var}(X) = 2n$.

La chi quadro è asimmetrica a destra, e la sua asimmetria diminuisce al crescere dei gradi di libertà. Interviene nella statistica della varianza campionaria e nei test chi quadro, che vedremo nel capitolo sulla verifica di ipotesi.

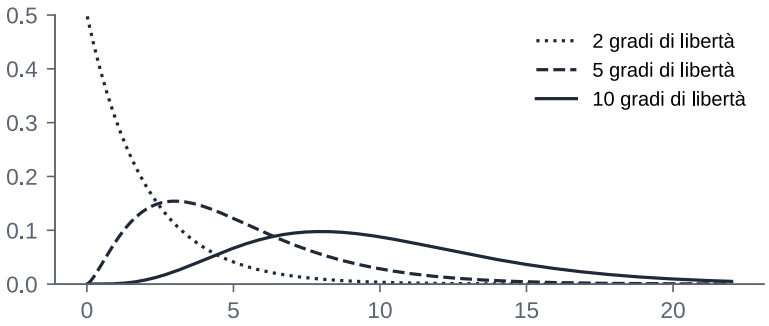

Figura 2.3. Densità chi quadro: asimmetrica a destra, si sposta verso destra al crescere dei gradi di libertà.

DEFINIZIONE

Se $Z \sim N(0, 1)$ e $V \sim \chi^2(n)$ sono indipendenti, la variabile

$$T = \frac{Z}{\sqrt{V/n}}$$

ha distribuzione **t di Student con n gradi di libertà**. È simmetrica attorno a 0, ha varianza $n/(n-2)$ per $n > 2$ e code più pesanti della normale; al crescere di n tende alla normale standard, e già per n maggiore di 30 le due sono quasi indistinguibili.

La t di Student compare quando si standardizza la media campionaria usando la deviazione standard **stimata** dai dati invece di quella vera: l'incertezza in più dovuta alla stima si traduce nelle code più pesanti.

Quadro riassuntivo

DISTRIBUZIONE	CHE COSA DESCRIVE	MEDIA	VARIANZA
Bernoulli Ber(p)	un singolo successo o insuccesso	p	$p(1-p)$
Binomiale Bin(n, p)	successi in n prove indipendenti	np	$np(1-p)$
Ipergeometrica	successi in n estrazioni senza reinserimento	nK/N	$nK/N\ (1-K/N)(N-n)/(N-1)$
Geometrica Geom(p)	prove fino al primo successo	$1/p$	$(1-p)/p^2$
Poisson Pois(λ)	eventi rari in un intervallo	λ	λ
Uniforme discreta	n valori equiprobabili da 1 a n	$(n+1)/2$	$(n^2-1)/12$
Uniforme U(a, b)	valore a caso in un intervallo	$(a+b)/2$	$(b-a)^2/12$
Esponenziale Exp(λ)	tempo di attesa senza memoria	$1/\lambda$	$1/\lambda^2$
Normale N(μ, σ²)	somme di molti contributi	μ	σ^2
Chi quadro $\chi^2(n)$	somma di n normali standard al quadrato	n	$2n$
t di Student t(n)	media standardizzata con σ stimata	0	$n/(n-2)$

CAPITOLO 3

Legge dei grandi numeri e approssimazione normale

Perché la media di molte osservazioni è affidabile e perché la normale compare ovunque: media campionaria, disuguaglianze di Markov e Chebyshev, convergenze, legge dei grandi numeri e teorema del limite centrale.

La media campionaria

Ripetere un esperimento n volte in modo indipendente produce n variabili aleatorie $X_1, \dots, X_n$ con la stessa distribuzione: si dice che sono **indipendenti e identicamente distribuite** (in breve i.i.d.). Il modo naturale di riassumerle è farne la media.

DEFINIZIONE

Siano $X_1, \dots, X_n$ variabili i.i.d. con media μ e varianza σ^2 finite. La **media campionaria** è

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

Essendo una funzione di variabili aleatorie, la media campionaria è a sua volta una **variabile aleatoria**: campioni diversi danno medie diverse. Valgono

$$E[\bar{X}_n] = \mu, \qquad \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

DIMOSTRAZIONE

Per la media basta la linearità del valore atteso, che vale anche senza indipendenza:

$$E[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

Per la varianza la costante 1/n esce al quadrato e, grazie all'indipendenza, la varianza della somma è la somma delle varianze:

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Il risultato dice due cose. La media campionaria è centrata sul valore vero μ, e quindi in media non sbaglia. La sua dispersione si riduce al crescere di n: la deviazione standard della media, $\sigma/\sqrt{n}$, si chiama **errore standard**. Per dimezzare l'errore standard occorre quadruplicare il numero di osservazioni.

Le disuguaglianze di Markov e di Chebyshev

Queste due disuguaglianze permettono di limitare la probabilità che una variabile assuma valori lontani dalla media conoscendo solo la media e la varianza, senza sapere nulla della distribuzione.

TEOREMA

Disuguaglianza di Markov. Se $X \geq 0$ e $\varepsilon > 0$, allora

$$P(X \geq \varepsilon) \leq \frac{E[X]}{\varepsilon}$$

DIMOSTRAZIONE (CASO CONTINUO)

Poiché $X \geq 0$, l'integrale che definisce la media parte da 0. Togliendo il tratto da 0 a ε l'integrale non aumenta, e nel tratto rimanente $x \geq \varepsilon$:

$$E[X] = \int_0^{+\infty} x f(x)\,dx \;\geq\; \int_\varepsilon^{+\infty} x f(x)\,dx \;\geq\; \varepsilon \int_\varepsilon^{+\infty} f(x)\,dx = \varepsilon\, P(X \geq \varepsilon)$$

Dividendo per ε si ottiene la tesi. Nel caso discreto il ragionamento è identico con le somme al posto degli integrali.

TEOREMA

Disuguaglianza di Chebyshev. Se X ha media μ e varianza σ^2 finite, per ogni $\varepsilon > 0$

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Ponendo $\varepsilon = k\sigma$ si ottiene la forma equivalente $P(|X - \mu| \geq k\sigma) \leq 1/k^2$.

DIMOSTRAZIONE

Si applica Markov alla variabile non negativa $Y = (X - \mu)^2$ con soglia ε^2. L'evento $\{|X - \mu| \geq \varepsilon\}$ coincide con $\{(X - \mu)^2 \geq \varepsilon^2\}$, quindi

$$P(|X - \mu| \geq \varepsilon) = P\big((X - \mu)^2 \geq \varepsilon^2\big) \leq \frac{E[(X - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

Chebyshev afferma che **nessuna** distribuzione può mettere troppa probabilità lontano dalla media: oltre due deviazioni standard c'è al massimo il 25%, oltre tre al massimo l'11%. Sono limiti larghi (per la normale oltre due deviazioni standard c'è solo il 4,6%), ma hanno il pregio di valere sempre.

ESEMPIO: QUANTO È LARGO IL LIMITE

Per il lancio di un dado $\mu = 3{,}5$ e $\sigma^2 = 35/12 \approx 2{,}917$. Chebyshev dà $P(|X - 3{,}5| \geq 2) \leq 2{,}917/4 \approx 0{,}729$, mentre il valore esatto, corrispondente agli esiti 1 e 6, è $2/6 \approx 0{,}333$. Per la media di 100 lanci la varianza diventa $2{,}917/100$ e lo stesso scostamento ha probabilità al massimo $0{,}0292/4 \approx 0{,}0073$: la stessa distanza dalla media diventa molto improbabile. È il meccanismo della legge dei grandi numeri.

Convergenza in probabilità e convergenza in legge

Per parlare di cosa succede alla media campionaria quando n cresce servono due nozioni di convergenza per successioni di variabili aleatorie, più deboli della convergenza dei numeri.

DEFINIZIONE

Una successione di variabili aleatorie Y_n **converge in probabilità** alla costante c se, per ogni $\varepsilon > 0$, la probabilità di uno scostamento maggiore di ε tende a zero:

$$\lim_{n \to \infty} P(|Y_n - c| \geq \varepsilon) = 0$$

Una successione Y_n **converge in legge** (o in distribuzione) a una variabile Y se le funzioni di ripartizione $F_n(x)$ tendono alla ripartizione F(x) di Y in ogni punto x in cui F è continua.

La prima dice che Y_n si concentra sempre di più attorno a un numero; la seconda dice che la **forma** della distribuzione di Y_n si avvicina a quella di Y. La convergenza in probabilità implica quella in legge, ma non viceversa.

La legge dei grandi numeri

TEOREMA

Legge debole dei grandi numeri. Siano X_1, X_2, … variabili i.i.d. con media μ e varianza σ^2 finita. Allora la media campionaria converge in probabilità a μ: per ogni $\varepsilon > 0$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| \geq \varepsilon) = 0$$

DIMOSTRAZIONE

Si applica Chebyshev alla media campionaria, che ha media μ e varianza σ^2/n:

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2}$$

Con σ e ε fissati, il secondo membro tende a zero quando n tende all'infinito, e quindi anche il primo, che è una probabilità e non può essere negativo.

Il significato è intuitivo: al crescere del numero di prove la media dei risultati osservati si avvicina alla media teorica. È ciò che giustifica la definizione frequentista (la frequenza relativa di un evento è la media di variabili di Bernoulli e tende alla sua probabilità) e tutti i metodi di simulazione: quando si ripete un esperimento al calcolatore N volte e si fa la media, il risultato è affidabile proprio grazie a questo teorema. In statistica la legge dice che la media del campione è una stima sempre più precisa della media della popolazione.

ATTENZIONE

La legge dei grandi numeri non dice che la media campionaria diventa uguale a μ, né che "i risultati si compensano": dice che la probabilità di uno scostamento fissato diventa trascurabile. Per ogni n finito uno scostamento resta sempre possibile.

Il teorema del limite centrale

La legge dei grandi numeri dice dove va la media campionaria; il teorema del limite centrale dice **come** fluttua attorno a quel valore per n grande, ed è il motivo per cui la normale compare in moltissimi fenomeni che di per sé non sono normali.

TEOREMA

Teorema del limite centrale. Siano X_1, X_2, … variabili i.i.d. con media μ e varianza σ^2 finita e positiva. Allora la media campionaria standardizzata converge in legge alla normale standard:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \longrightarrow N(0,1)$$

In pratica, per n grande la media campionaria è approssimativamente $N(\mu, \sigma^2/n)$ e la somma è approssimativamente $N(n\mu, n\sigma^2)$, **qualunque** sia la distribuzione delle X_i.

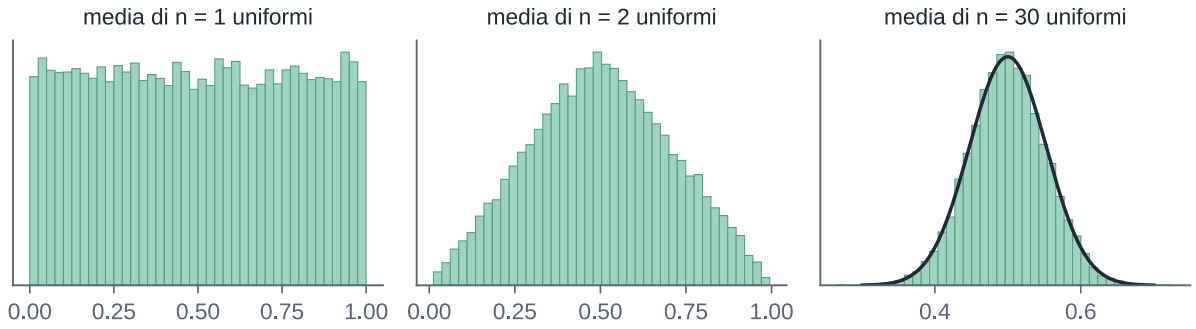

Figura 3.1. Il teorema del limite centrale al lavoro: la media di 30 uniformi ha già la forma della normale (curva continua).

La figura mostra la media di n numeri uniformi su [0, 1], simulata molte volte. Con n = 1 l'istogramma è piatto, con n = 2 è triangolare, con n = 30 è già una campana che si sovrappone alla normale con media 1/2 e varianza 1/(12n). Di solito l'approssimazione è considerata buona per n maggiore di 30; serve un n più grande se la distribuzione di partenza è molto asimmetrica.

APPROSSIMAZIONE NORMALE DELLA BINOMIALE

Una binomiale Bin(n, p) è la somma di n Bernoulli indipendenti, quindi per il teorema del limite centrale è approssimativamente normale con media np e varianza np(1 − p). Una regola pratica per usare l'approssimazione è che np e n(1 − p) siano entrambi almeno 5 (alcuni testi chiedono 10). Poiché si approssima una variabile discreta con una continua, conviene la **correzione di continuità**: il valore intero k viene sostituito dall'intervallo che va da k − 0,5 a k + 0,5.

ESEMPIO: RISPOSTE CORRETTE A CASO

Un quiz ha 100 domande a quattro opzioni e uno studente risponde a caso. Il numero di risposte corrette è $X \sim \text{Bin}(100; 0{,}25)$. Stimiamo $P(X \geq 30)$.

Passo 1, parametri. $\mu = np = 25$ e $\sigma^2 = np(1-p) = 18{,}75$, quindi $\sigma \approx 4{,}33$. Le condizioni $np \geq 5$ e $n(1-p) \geq 5$ sono soddisfatte.

Passo 2, correzione di continuità. L'evento $\{X \geq 30\}$ corrisponde, per la variabile continua, all'evento $\{X \geq 29{,}5\}$.

Passo 3, standardizzazione.

$$P(X \geq 30) \approx P\left(Z \geq \frac{29{,}5 - 25}{4{,}33}\right) = P(Z \geq 1{,}04) = 1 - \Phi(1{,}04) \approx 0{,}149$$

Il valore esatto della binomiale è circa 0,150, quindi l'approssimazione è ottima. Senza correzione di continuità si otterrebbe $1 - \Phi(1{,}15) \approx 0{,}124$, un errore molto più grande.

DA RICORDARE

Legge dei grandi numeri: la media campionaria tende al valore vero μ (convergenza in probabilità verso un numero).

Teorema del limite centrale: le fluttuazioni della media attorno a μ sono approssimativamente normali, con ampiezza dell'ordine di σ/√n (convergenza in legge verso una distribuzione).

Conseguenza pratica: per guadagnare una cifra decimale di precisione servono cento volte più osservazioni.

Statistica descrittiva e statistica inferenziale

I risultati di questo capitolo fanno da ponte tra il calcolo delle probabilità e la statistica, che si divide in due grandi rami. La **statistica descrittiva** organizza e riassume i dati osservati con tabelle, grafici e indici, senza pretendere di dire nulla oltre i dati stessi. La **statistica inferenziale** usa i dati di un campione per trarre conclusioni sull'intera popolazione da cui il campione è estratto, dichiarando il margine di errore: stime, intervalli di confidenza, test. La legge dei grandi numeri e il teorema del limite centrale sono esattamente gli strumenti che permettono di quantificare quel margine.

	DESCRITTIVA	INFERENZIALE
Oggetto	il campione osservato	la popolazione
Strumenti	frequenze, indici, grafici	stimatori, intervalli di confidenza, test
Domanda	com'è fatto questo insieme di dati?	che cosa posso dire della popolazione?
Risultato	esatto, riguarda solo i dati	probabilistico, con un margine di errore

Esempio. Calcolare il voto medio dei 40 studenti di un corso è statistica descrittiva; usare quei 40 voti per stimare, con un intervallo di confidenza al 95%, il voto medio di tutti gli studenti che sostengono quell'esame negli anni è statistica inferenziale.

CAPITOLO 4

Statistica descrittiva

Come organizzare e riassumere un insieme di dati: frequenze, raggruppamento per valori singoli e per classi, rappresentazioni grafiche, indici di posizione, di dispersione e di forma.

Popolazione, campione e caratteri

La statistica studia fenomeni collettivi osservandone le singole unità. L'insieme di tutte le unità di interesse si chiama **popolazione**; la parte effettivamente osservata è il **campione**, che ha **numerosità** N. La grandezza che si rileva su ogni unità si chiama **carattere** (o variabile statistica) e i valori che assume sono le **modalità**.

Un carattere è **qualitativo** se le modalità sono categorie (il sistema operativo usato, il corso di laurea) e **quantitativo** se sono numeri. I caratteri quantitativi sono **discreti** se assumono valori isolati, tipicamente conteggi (il numero di bug in un modulo), e **continui** se possono assumere qualunque valore in un intervallo (un tempo, un peso). Il tipo di carattere decide come raggruppare i dati e quali grafici usare.

Raggruppamento per valori singoli

Quando le modalità sono poche, come per un carattere qualitativo o per un conteggio con pochi valori, si costruisce la **tabella delle frequenze** elencando ogni modalità con il numero di volte in cui compare.

DEFINIZIONE

Dati N valori osservati, per ogni modalità x_i si definiscono: la **frequenza assoluta** f_i, numero di dati uguali a x_i; la **frequenza relativa** $r_i = f_i/N$; la **frequenza percentuale** $100 \cdot r_i$; la **frequenza cumulata** F_i, somma delle frequenze delle modalità minori o uguali a x_i (definita per caratteri ordinabili).

Valgono sempre $\Sigma f_i = N$ e $\Sigma r_i = 1$.

Esempio. In un progetto software si conta il numero di bug trovati in ciascuno dei 20 moduli. I dati raggruppati per valori singoli danno la tabella seguente.

BUG PER MODULO	FREQ. ASSOLUTA	FREQ. RELATIVA	FREQ. PERCENTUALE	FREQ. CUMULATA
0	4	0,20	20%	4
1	7	0,35	35%	11
2	5	0,25	25%	16
3	3	0,15	15%	19
4	1	0,05	5%	20
Totale	20	1	100%	

Raggruppamento in classi

Per un carattere continuo, o discreto con moltissimi valori, i dati quasi mai si ripetono e la tabella per valori singoli sarebbe lunga quanto l'elenco dei dati. Si suddivide allora l'intervallo dei valori in **classi**, cioè intervalli contigui e non sovrapposti, e si conta quanti dati cadono in ciascuna.

Per ogni classe [a, b) si definiscono l'**ampiezza** b − a, il **valore centrale** (a + b)/2, usato come rappresentante della classe nei calcoli, e la **densità di frequenza**, rapporto tra frequenza e ampiezza. Le classi possono avere ampiezze diverse; in quel caso è la densità, e non la frequenza, a dire dove i dati sono più concentrati. Il numero di classi non deve essere né troppo piccolo, perché si perde informazione, né troppo grande, perché le classi restano quasi vuote: una regola empirica diffusa suggerisce circa √N classi.

Esempio. Si registra il tempo di consegna, in giorni, di 50 ordini e si raggruppano i dati in quattro classi, l'ultima più ampia delle altre.

CLASSE (GIORNI)	VALORE CENTRALE	FREQUENZA	FREQ. RELATIVA	AMPIEZZA	DENSITÀ
da 0 a 2	1	8	0,16	2	4
da 2 a 4	3	18	0,36	2	9
da 4 a 6	5	14	0,28	2	7
da 6 a 10	8	10	0,20	4	2,5
Totale		50	1		

Rappresentazioni grafiche

Un buon grafico mostra a colpo d'occhio la forma dei dati. Il **diagramma a barre** si usa per caratteri qualitativi o discreti: una barra per modalità, di altezza pari alla frequenza, con barre separate. Il **diagramma a torta** rappresenta le frequenze relative di un carattere qualitativo come settori di un cerchio con angolo proporzionale alla frequenza, ed è utile solo con poche modalità.

L'**istogramma** è il grafico dei dati raggruppati in classi: rettangoli adiacenti con base sulla classe e **area** proporzionale alla frequenza. Se le classi hanno la stessa ampiezza l'altezza può essere la frequenza; se le ampiezze sono diverse l'altezza deve essere la densità di frequenza, altrimenti le classi larghe sembrerebbero più importanti di quanto sono. Nell'esempio, la classe da 6 a 10 giorni contiene 10 ordini ma ha altezza 2,5, più bassa delle altre.

Il **poligono delle frequenze** unisce i punti medi delle sommità dei rettangoli; l'**ogiva** rappresenta le frequenze cumulate ed è la versione empirica della funzione di ripartizione. Il **box plot**, infine, riassume i dati con cinque numeri: una scatola che va dal primo al terzo quartile, una linea interna alla mediana e due baffi che arrivano fino ai valori estremi non anomali; i valori che distano dalla scatola più di 1,5 volte la sua lunghezza vengono disegnati come punti isolati e segnalati come possibili valori anomali.

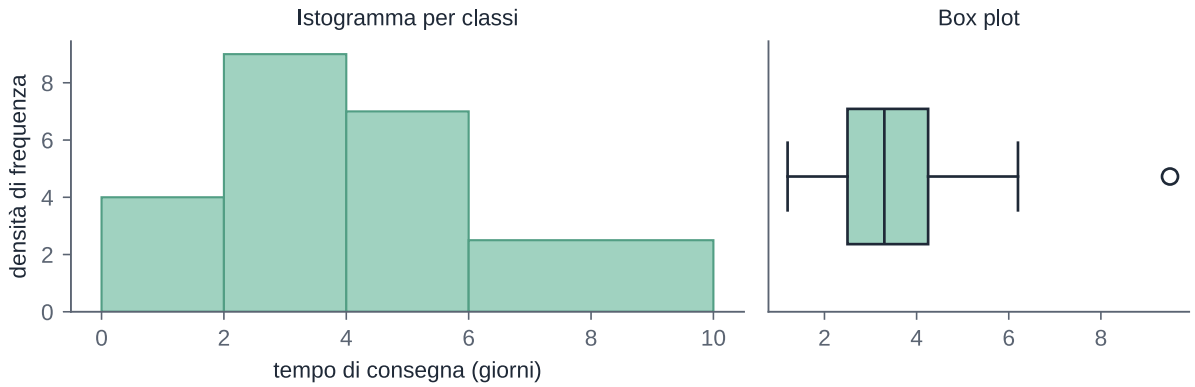

Figura 4.1. A sinistra l'istogramma della tabella per classi; a destra un box plot con un valore anomalo (cerchietto).

Indici di posizione

Gli indici di posizione indicano attorno a quale valore si collocano i dati.

DEFINIZIONE

La **media aritmetica** di N dati $x_1, \dots, x_N$ è la loro somma divisa per N. Se i dati sono raggruppati per valori singoli, con frequenze f_i, ogni modalità si pesa con la sua frequenza; se sono raggruppati in classi, si usano i valori centrali c_i, ottenendo un valore approssimato:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \bar{x} = \frac{1}{N}\sum_i x_i f_i, \qquad \bar{x} \approx \frac{1}{N}\sum_i c_i f_i$$

La **mediana** è il valore che divide i dati ordinati in due metà di uguale numerosità: il valore centrale se N è dispari, la media dei due valori centrali se N è pari.

La **moda** è la modalità con frequenza più alta; per dati in classi si parla di classe modale, quella con densità maggiore.

Per i bug per modulo la media è (0·4 + 1·7 + 2·5 + 3·3 + 4·1)/20 = 30/20 = 1,5. I dati ordinati hanno N = 20: la mediana è la media del decimo e dell'undicesimo valore, entrambi uguali a 1 (i valori 1 occupano le posizioni dalla quinta all'undicesima, come mostra la frequenza cumulata), quindi la mediana è 1. Anche la moda è 1. Per i tempi di consegna la media approssimata è (1·8 + 3·18 + 5·14 + 8·10)/50 = 212/50 = 4,24 giorni.

DEFINIZIONE

Il **quantile di ordine p** (con $0 < p < 1$) è un valore che lascia alla sua sinistra una frazione p dei dati ordinati e alla sua destra una frazione $1 - p$. La mediana è il quantile di ordine 0,5; il **primo quartile** Q_1 e il **terzo quartile** Q_3 sono i quantili di ordine 0,25 e 0,75; i **percentili** sono i quantili di ordine 0,01, 0,02 e così via.

Con pochi dati il quantile non è unico e i diversi testi e programmi adottano convenzioni leggermente diverse per interpolare tra due dati consecutivi; per campioni grandi le differenze diventano trascurabili. Per dati in classi la mediana si stima per interpolazione lineare dentro la classe che la contiene: nell'esempio dei tempi di consegna il venticinquesimo ordine cade nella classe da 2 a 4 (prima di essa ci sono 8 ordini, dentro 18), quindi la mediana è circa $2 + (25 - 8)/18 \cdot 2 \approx 3{,}89$ giorni.

DA RICORDARE

La media è sensibile ai valori anomali, la mediana no: un solo dato molto grande sposta la media ma lascia la mediana quasi invariata. Per questo per i redditi o per i tempi di risposta di un sistema si riporta spesso la mediana. Nella figura del box plot il valore anomalo 9,5 alza la media ma non la mediana.

Indici di dispersione

Due insiemi di dati possono avere la stessa media ed essere molto diversi: tempi di consegna tutti vicini a 4 giorni o sparsi tra 1 e 10. Gli indici di dispersione misurano quanto i dati si allontanano dal centro.

DEFINIZIONE

Il **campo di variazione** (range) è la differenza tra il valore massimo e il minimo. La **differenza interquartile** è IQR = $Q_3 - Q_1$, ampiezza dell'intervallo che contiene la metà centrale dei dati.

La **varianza empirica** è la media dei quadrati degli scarti dalla media; la **varianza campionaria** divide invece per N − 1:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

La **deviazione standard** (o scarto quadratico medio) è la radice quadrata della varianza. Il **coefficiente di variazione** CV = s / $\bar{x}$ è un indice adimensionale, utile per confrontare la variabilità di fenomeni con unità o scale diverse.

La varianza empirica è un indice puramente descrittivo dei dati osservati; la varianza campionaria, con N − 1, è quella che si usa quando i dati sono un campione e si vuole stimare la varianza della popolazione, perché, come si vedrà nel prossimo capitolo, è uno stimatore non distorto. Le due differiscono per il fattore (N − 1)/N, trascurabile solo per N grande. Per i calcoli a mano è comoda la **formula computazionale**, che evita di calcolare tutti gli scarti:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

ESEMPIO: VARIANZA DEI BUG PER MODULO

Passo 1. Media dei quadrati, pesando con le frequenze: (0·4 + 1·7 + 4·5 + 9·3 + 16·1)/20 = 70/20 = 3,5.

Passo 2. Varianza empirica con la formula computazionale: $3{,}5 - 1{,}5^2 = 3{,}5 - 2{,}25 = 1{,}25$.

Passo 3. Varianza campionaria: 1,25 · 20/19 ≈ 1,316, da cui la deviazione standard s ≈ 1,147 bug.

Passo 4. Coefficiente di variazione: CV ≈ 1,147/1,5 ≈ 0,76; la variabilità è grande rispetto alla media.

Indici di forma

Gli indici di forma descrivono la simmetria della distribuzione dei dati e il peso delle sue code, e sono le versioni empiriche degli indici visti per le variabili aleatorie, calcolati con i momenti centrati dei dati:

$$\gamma_1 = \frac{\frac{1}{N}\sum(x_i - \bar{x})^3}{\hat{\sigma}^3}, \qquad \gamma_2 = \frac{\frac{1}{N}\sum(x_i - \bar{x})^4}{\hat{\sigma}^4} - 3$$

Un'**asimmetria** γ_1 positiva indica una coda lunga verso destra, con la media spostata a destra della mediana, come accade per i tempi di consegna; una negativa indica una coda verso sinistra; un valore vicino a zero indica dati simmetrici. Una **curtosi** γ_2 positiva indica una distribuzione più appuntita e con code più pesanti della normale, una negativa una distribuzione più piatta. Per i bug per modulo i momenti centrati di ordine 3 e 4 valgono 0,6 e 3,7625, da cui $\gamma_1 \approx 0{,}6/1{,}398 \approx 0{,}43$ (lieve asimmetria a destra) e $\gamma_2 \approx 3{,}7625/1{,}5625 - 3 \approx -0{,}59$ (distribuzione un po' più piatta della normale).

CAPITOLO 5

Statistica inferenziale: stimatori e intervalli di confidenza

Come usare un campione per stimare i parametri di una popolazione: stimatori puntuali di media, varianza e proporzione, loro proprietà, e intervalli di confidenza.

Campione, stimatore e stima

La popolazione è descritta da una variabile aleatoria X la cui distribuzione dipende da un parametro incognito θ (per esempio la media μ, la varianza σ^2 o una proporzione p). Per conoscere θ si osserva un campione e si combina l'informazione dei dati in un unico numero.

DEFINIZIONE

Un **campione casuale** di numerosità n è una sequenza $X_1, \ldots, X_n$ di variabili aleatorie indipendenti, tutte con la stessa distribuzione di X.

Uno **stimatore** del parametro θ è una funzione del campione $T = T(X_1, \ldots, X_n)$ che **non dipende da θ**. Essendo funzione di variabili aleatorie, uno stimatore è a sua volta una variabile aleatoria.

Il valore numerico $t = T(x_1, \ldots, x_n)$ che lo stimatore assume sui dati effettivamente osservati si chiama **stima**.

La distinzione tra stimatore e stima è la stessa che c'è tra una ricetta e il piatto cucinato: lo stimatore è la regola di calcolo, valida per qualunque campione, e ha una distribuzione; la stima è il numero ottenuto applicando la regola a un campione concreto. Per esempio la media campionaria è lo stimatore naturale di μ, e 12,4 ore è la stima ottenuta da un particolare campione di batterie. La condizione che T non dipenda da θ è essenziale: uno stimatore che richiedesse di conoscere θ non servirebbe a stimarlo.

Proprietà degli stimatori

Per lo stesso parametro si possono costruire molti stimatori diversi; servono criteri per dire quali sono buoni.

DEFINIZIONE

Uno stimatore T di θ è **non distorto** (o corretto) se $E[T] = \theta$ per ogni valore possibile di θ. La differenza $E[T] - \theta$ si chiama **distorsione** (bias).

L'**errore quadratico medio** di T misura la distanza media dal valore vero e si scompone in varianza più distorsione al quadrato:

$$\mathrm{MSE}(T) = E\left[(T-\theta)^2\right] = \mathrm{Var}(T) + (E[T]-\theta)^2$$

Uno stimatore è **consistente** se converge in probabilità a θ al crescere di n.

La scomposizione dell'errore quadratico medio separa due tipi di errore. La distorsione è un **errore sistematico**, come una bilancia tarata male, che non si riduce aumentando i dati; la varianza è un **errore casuale**, che si riduce aumentando n. Per uno stimatore non distorto l'errore quadratico medio coincide con la varianza. Uno stimatore non distorto la cui varianza tende a zero è consistente: lo garantisce la disuguaglianza di Chebyshev, esattamente come nella dimostrazione della legge dei grandi numeri.

Stimatore della media

PROPRIETÀ

La media campionaria $\bar{X}_n = (1/n)\,\Sigma\, X_i$ è uno stimatore **non distorto** e **consistente** della media μ della popolazione, con $E[\bar{X}_n] = \mu$ e $Var(\bar{X}_n) = \sigma^2/n$.

La dimostrazione è quella del capitolo 3: la linearità del valore atteso dà $E[\bar{X}_n] = \mu$, cioè la non distorsione; l'indipendenza dà $Var(\bar{X}_n) = \sigma^2/n$, che tende a zero, e quindi per Chebyshev $\bar{X}_n$ converge in probabilità a μ. Lo stesso ragionamento si applica a una **proporzione**: se X è il numero di successi in n prove, la frequenza relativa $\hat{p} = X/n$ è la media di n variabili di Bernoulli e quindi è uno stimatore non distorto e consistente di p, con $Var(\hat{p}) = p(1 - p)/n$.

Stimatore della varianza

Per la varianza bisogna distinguere due situazioni, perché lo stimatore cambia a seconda che la media μ della popolazione sia nota oppure no.

CASO 1: MEDIA NOTA

PROPRIETÀ

Se μ è nota, lo stimatore

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

è non distorto per σ^2. Infatti, per la linearità e perché $E[(X_i - \mu)^2] = \sigma^2$ per definizione di varianza,

$$E[S_\mu^2] = \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] = \frac{1}{n}\cdot n\sigma^2 = \sigma^2$$

CASO 2: MEDIA NON NOTA

Di solito μ non è nota e va sostituita con la sua stima $\bar{X}_n$. La sostituzione però ha un costo: lo stimatore "naturale" che divide per n diventa distorto.

ATTENZIONE

Lo stimatore $\hat{\sigma}^2 = (1/n)\,\Sigma\,(X_i - \bar{X}_n)^2$ è **distorto**: $E[\hat{\sigma}^2] = (n - 1)\sigma^2/n$, quindi sottostima sistematicamente la varianza. Gli scarti dalla media campionaria sono in media più piccoli degli scarti dalla media vera, perché $\bar{X}_n$ è per costruzione il valore più vicino ai dati.

DIMOSTRAZIONE

Passo 1. Si scrive ogni scarto come $X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu)$ e si sviluppa il quadrato, sommando su i:

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_i (X_i - \mu) + n(\bar{X} - \mu)^2$$

Passo 2. Poiché $\Sigma\,(X_i - \mu) = n(\bar{X} - \mu)$, il termine centrale vale $-2n(\bar{X} - \mu)^2$ e si somma all'ultimo:

$$\sum_i (X_i - \bar{X})^2 = \sum_i (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Passo 3. Si prende il valore atteso, usando $E[(X_i - \mu)^2] = \sigma^2$ e $E[(\bar{X} - \mu)^2] = \mathrm{Var}(\bar{X}) = \sigma^2/n$:

$$E\left[\sum_i (X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\,\sigma^2$$

Dividendo per n si ottiene la distorsione di $\hat{\sigma}^2$; dividendo per n − 1 la distorsione scompare.

PROPRIETÀ

Se μ non è nota, lo stimatore non distorto di σ^2 è la **varianza campionaria**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \qquad E[S^2] = \sigma^2$$

La divisione per n − 1 invece che per n si chiama **correzione di Bessel**.

Il numero n − 1 si chiama **gradi di libertà**. Dopo aver calcolato $\bar{X}_n$, gli n scarti $X_i - \bar{X}_n$ non sono più liberi, perché la loro somma è sempre zero: noti n − 1 scarti, l'ultimo è determinato. Un'unità di informazione è stata spesa per stimare μ, e dividere per n − 1 compensa esattamente questa perdita. Per lo stesso motivo la t di Student che compare negli intervalli per la media ha n − 1 gradi di libertà.

SIMBOLO	FORMULA	RUOLO
σ^2	$E[(X - \mu)^2]$	parametro della popolazione, incognito
S_μ^2	$(1/n)\ \Sigma\ (X_i - \mu)^2$	stimatore con μ nota, non distorto
$\hat{\sigma}^2$	$(1/n)\ \Sigma\ (X_i - \bar{X})^2$	stimatore con μ non nota, distorto (varianza empirica)
S^2	$(1/(n-1))\ \Sigma\ (X_i - \bar{X})^2$	stimatore con μ non nota, non distorto (varianza campionaria)

Se la popolazione è normale vale inoltre un risultato che sarà essenziale per intervalli e test: la variabile $(n-1)S^2/\sigma^2$ ha distribuzione chi quadro con n − 1 gradi di libertà, ed è indipendente da $\bar{X}_n$. Ne segue che, sostituendo σ con S nella media standardizzata, si ottiene una t di Student:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \qquad T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Intervalli di confidenza

Una stima puntuale, da sola, non dice quanto ci si possa fidare. Un intervallo di confidenza accompagna la stima con un margine di errore, costruito in modo da contenere il parametro vero con una probabilità fissata in anticipo.

DEFINIZIONE

Un **intervallo di confidenza** di livello 1 − α per il parametro θ è un intervallo $[T_1, T_2]$, i cui estremi T_1 e T_2 sono funzioni del campione, tale che

$$P(T_1 \le \theta \le T_2) = 1 - \alpha$$

Il numero 1 − α (per esempio 0,95) si chiama **livello di confidenza**.

ATTENZIONE

Casuali sono gli estremi dell'intervallo, non θ, che è un numero fisso. Dopo aver calcolato l'intervallo su un campione concreto non si può dire che θ vi cada con probabilità 0,95: o lo contiene o non lo contiene. L'interpretazione corretta riguarda il procedimento: ripetendo il campionamento molte volte, il 95% degli intervalli così costruiti conterrebbe il valore vero.

MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA NOTA

Passo 1. Se la popolazione è normale con σ nota, la media campionaria standardizzata è esattamente normale standard: $Z = (\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$. Per popolazioni non normali vale lo stesso in modo approssimato quando n è grande, grazie al teorema del limite centrale.

Passo 2. Sia $z_{1-\alpha/2}$ il quantile di ordine $1 - \alpha/2$ della normale standard. Per la simmetria della normale

$$P\left(-z_{1-\alpha/2} \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le z_{1-\alpha/2}\right) = 1 - \alpha$$

Passo 3. Si isola μ al centro della doppia disuguaglianza, moltiplicando per $\sigma/\sqrt{n}$ e riordinando:

$$\left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

La quantità $z_{1-\alpha/2}\,\sigma/\sqrt{n}$ si chiama **margine di errore**. Con α = 0,05 il quantile è 1,96.

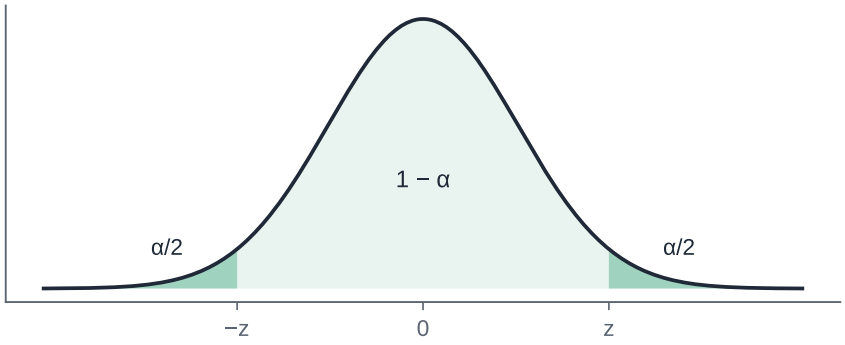

Figura 5.1. Il quantile z lascia al centro un'area 1 − α e α/2 in ciascuna coda (zone più scure). Con α = 0,05 si ha z = 1,96.

Esempio. La durata di 36 batterie ha media 12,4 ore; da prove precedenti si sa che σ = 1,8 ore. Il margine di errore al 95% è 1,96 · 1,8/√36 = 1,96 · 0,3 = 0,588 ore, quindi l'intervallo è [11,81 ; 12,99] ore.

MEDIA DI UNA POPOLAZIONE NORMALE CON VARIANZA NON NOTA

Se σ non è nota la si sostituisce con la deviazione standard campionaria S. La quantità standardizzata non è più normale ma, come visto sopra, una t di Student con n − 1 gradi di libertà; il ragionamento è identico con il quantile della t al posto di quello della normale:

$$\left[\bar{X}_n - t_{1-\alpha/2,\,n-1}\frac{S}{\sqrt{n}},\ \bar{X}_n + t_{1-\alpha/2,\,n-1}\frac{S}{\sqrt{n}}\right]$$

Esempio. Dieci misure del tempo di risposta di un servizio danno media 230 ms e deviazione standard campionaria s = 18 ms. Con 9 gradi di libertà il quantile è $t_{0{,}975;\,9} \approx 2{,}262$, il margine è $2{,}262 \cdot 18/\sqrt{10} \approx 12{,}88$ ms e l'intervallo al 95% è [217,1 ; 242,9] ms. Con la normale si sarebbe usato 1,96 e l'intervallo sarebbe risultato più stretto, ma ingiustificatamente: la t tiene conto dell'incertezza in più dovuta alla stima di σ, e la differenza conta soprattutto per campioni piccoli.

PROPORZIONE

Per una proporzione p e un campione grande, $\hat{p}$ è approssimativamente normale con varianza $p(1-p)/n$; sostituendo p con $\hat{p}$ nella varianza si ottiene l'intervallo approssimato

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Esempio. Su 400 utenti intervistati, 92 usano l'applicazione in modalità scura: $\hat{p} = 0{,}23$. Il margine al 95% è $1{,}96 \cdot \sqrt{(0{,}23 \cdot 0{,}77/400)} \approx 0{,}041$, quindi l'intervallo è circa [0,189 ; 0,271].

VARIANZA DI UNA POPOLAZIONE NORMALE

Per la varianza si sfrutta il fatto che $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. La chi quadro non è simmetrica, quindi servono due quantili diversi, uno per coda, che indichiamo con $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$. Isolando σ^2 nella doppia disuguaglianza (dividere per una quantità più grande dà un estremo più piccolo, per questo i quantili si scambiano):

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}\ ,\ \frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}}\right]$$

Esempio. Con i dieci tempi di risposta di prima, $s^2 = 324$. I quantili della chi quadro con 9 gradi di libertà sono $\chi^2_{0{,}975} \approx 19{,}02$ e $\chi^2_{0{,}025} \approx 2{,}70$. L'intervallo al 95% per σ^2 è $[9 \cdot 324/19{,}02\ ;\ 9 \cdot 324/2{,}70] \approx [153{,}3\ ;\ 1079{,}8]$, e prendendo le radici quello per σ è circa [12,4 ; 32,9] ms. L'intervallo non è centrato sulla stima ed è molto ampio: con dieci dati la varianza si stima con poca precisione.

DA CHE COSA DIPENDE L'AMPIEZZA

Per la media con σ nota l'ampiezza dell'intervallo è $2\,z_{1-\alpha/2}\,\sigma/\sqrt{n}$, e da questa formula si leggono le tre dipendenze. Al crescere di **n** l'ampiezza diminuisce come $1/\sqrt{n}$: per dimezzarla serve un campione quattro volte più grande. Al crescere di **σ** l'ampiezza aumenta in proporzione: una popolazione più variabile si conosce con meno precisione. Al crescere del **livello di confidenza** $1-\alpha$ (cioè al diminuire di α) il quantile aumenta, da 1,645 al 90% a 1,96 al 95% e 2,576 al 99%, e l'intervallo si allarga: per essere più sicuri di contenere il valore vero bisogna accettare un intervallo meno preciso.

DA RICORDARE

Per ottenere un margine di errore non superiore a E con livello $1-\alpha$, la numerosità del campione deve essere

$$n \geq \left(\frac{z_{1-\alpha/2}\,\sigma}{E}\right)^2$$

Con le batterie (σ = 1,8 ore), per un margine di 0,3 ore al 95% servono almeno $(1{,}96 \cdot 1{,}8/0{,}3)^2 \approx 138{,}3$, cioè 139 batterie.

CAPITOLO 6

Verifica di ipotesi

Come decidere, sulla base di un campione, se un'affermazione sulla popolazione è compatibile con i dati: caratteristiche generali dei test, test sulla media, sulla proporzione e sulla varianza, test chi quadro.

Caratteristiche generali dei test

Un produttore dichiara che i suoi bulloni sono lunghi in media 50 mm; un algoritmo nuovo dovrebbe essere più veloce del precedente; un dado potrebbe essere truccato. In tutti questi casi c'è un'affermazione sulla popolazione e un campione di dati, e bisogna decidere se i dati sono compatibili con l'affermazione oppure la smentiscono. Il test di ipotesi formalizza questa decisione, in modo simile a un processo: l'imputato è considerato innocente finché le prove non dimostrano il contrario oltre ogni ragionevole dubbio.

DEFINIZIONE

Un **test di ipotesi** mette a confronto due affermazioni sul parametro θ della popolazione: l'**ipotesi nulla** H_0, che rappresenta la situazione di riferimento e viene mantenuta salvo forte evidenza contraria, e l'**ipotesi alternativa** H_1, che è ciò che si cerca di dimostrare.

La decisione si basa su una **statistica test**, funzione del campione la cui distribuzione è nota quando H_0 è vera. L'insieme dei valori della statistica che portano a rifiutare H_0 si chiama **regione di rifiuto** (o regione critica); il suo confine è il **valore critico**.

Il test può sbagliare in due modi, che hanno conseguenze diverse e vanno tenuti distinti.

	H_0 È VERA	H_0 È FALSA
Si rifiuta H_0	errore di primo tipo, probabilità α	decisione corretta, probabilità $1-\beta$ (potenza)
Non si rifiuta H_0	decisione corretta	errore di secondo tipo, probabilità β

DEFINIZIONE

Il **livello di significatività** α è la probabilità di commettere un errore di primo tipo, cioè di rifiutare H_0 quando è vera; si fissa prima di guardare i dati, di solito 0,05 o 0,01.

β è la probabilità dell'errore di secondo tipo, cioè di non rifiutare H_0 quando è falsa; $1-\beta$ si chiama **potenza** del test.

Il **p value** (o livello di significatività osservato) è la probabilità, calcolata supponendo vera H_0, di ottenere un valore della statistica test almeno altrettanto estremo di quello osservato. Si rifiuta H_0 se il p value è minore di α.

Nel paragone con il processo, l'errore di primo tipo è condannare un innocente e quello di secondo tipo è assolvere un colpevole. α è sotto il nostro controllo; β invece dipende da quanto H_1 è lontana da H_0 e diminuisce se aumenta n. A parità di dati non si possono ridurre entrambi: abbassare α rende il test più prudente e quindi aumenta β. Il p value è comodo perché non richiede di scegliere un valore critico e dice anche **quanto** forte è l'evidenza contro H_0: un p value di 0,001 è un'evidenza molto più forte di uno di 0,04.

ATTENZIONE

Il p value **non** è la probabilità che H_0 sia vera: è la probabilità dei dati, supponendo vera H_0.

Se non si rifiuta H_0 non si è dimostrato che H_0 sia vera. La conclusione corretta è "i dati non forniscono evidenza sufficiente, al livello α, per rifiutare H_0": assenza di prove non è prova di assenza.

TEST BILATERALI E UNILATERALI

La forma di H_1 decide dove si trova la regione di rifiuto. Se si vuole verificare se il parametro è **diverso** da un valore θ_0 (H_1: $\theta \neq \theta_0$) il test è **bilaterale** e la regione di rifiuto è divisa tra le due code, con α/2 in ciascuna. Se si vuole verificare se è **maggiore** (H_1: $\theta > \theta_0$) il test è **unilaterale destro** e tutta la probabilità α sta nella coda destra; se si vuole verificare se è **minore** il test è **unilaterale sinistro**. La scelta va fatta leggendo la domanda, prima di guardare i dati.

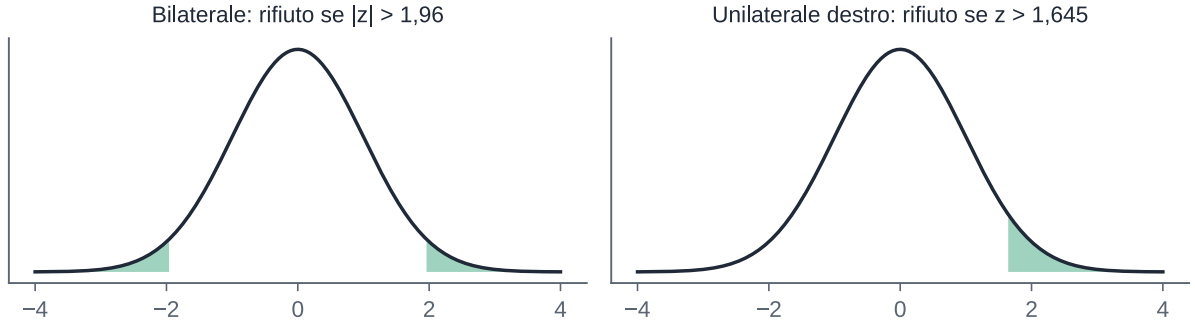

Figura 6.1. Regioni di rifiuto (zone più scure) per un test al livello α = 0,05.

DA RICORDARE

Ogni test si svolge in cinque passi. **1. Ipotesi:** scrivere H_0 e H_1 e dire se il test è bilaterale o unilaterale. **2. Livello:** fissare α e trovare il valore critico. **3. Statistica test:** scegliere la statistica, dire qual è la sua distribuzione sotto H_0 e calcolarne il valore. **4. Decisione:** confrontare la statistica con il valore critico, oppure il p value con α. **5. Conclusione:** scrivere che cosa significa la decisione nel contesto del problema.

Test sulla media

VARIANZA NOTA: IL TEST Z

Se la popolazione è normale con σ nota (o se n è grande, per il teorema del limite centrale), sotto H_0: $\mu = \mu_0$ la statistica

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Si rifiuta H_0 se $|Z| > z_{1-\alpha/2}$ nel test bilaterale, se $Z > z_{1-\alpha}$ nel test unilaterale destro, se $Z < -z_{1-\alpha}$ in quello sinistro.

Esempio. I bulloni devono avere lunghezza media 50 mm e il processo ha σ = 0,4 mm. Su 25 bulloni si misura una media di 50,18 mm. Il processo è fuori taratura?

1. H_0: $\mu = 50$ contro H_1: $\mu \neq 50$; il test è bilaterale perché interessa uno scostamento in entrambe le direzioni. **2.** Con α = 0,05 il valore critico è 1,96. **3.** La statistica vale

$$z = \frac{50{,}18 - 50}{0{,}4/\sqrt{25}} = \frac{0{,}18}{0{,}08} = 2{,}25$$

4. Poiché $|z| = 2{,}25 > 1{,}96$ si rifiuta H_0; equivalentemente il p value è $2(1 - \Phi(2{,}25)) \approx 0{,}024$, minore di 0,05. **5.** Al livello del 5% i dati indicano che la lunghezza media si è spostata da 50 mm: il processo va ricalibrato.

VARIANZA NON NOTA: IL TEST T

Se σ non è nota la si stima con la deviazione standard campionaria S. Per una popolazione normale, sotto H_0 la statistica ha distribuzione t di Student con $n - 1$ gradi di libertà, e i valori critici si leggono da quella distribuzione:

$$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

Esempio. Il vecchio algoritmo impiega in media 120 ms. Su 16 esecuzioni il nuovo algoritmo ha media 114 ms e s = 10 ms. Il nuovo algoritmo è più veloce?

1. La domanda è direzionale: H_0: μ = 120 contro H_1: μ < 120, test unilaterale sinistro. **2.** σ non è nota, si usa la t con 15 gradi di libertà; con α = 0,05 il valore critico è $-t_{0{,}95;\,15} \approx -1{,}753$. **3.** La statistica vale

$$t = \frac{114 - 120}{10/\sqrt{16}} = \frac{-6}{2{,}5} = -2{,}4$$

4. Poiché $-2{,}4 < -1{,}753$ si rifiuta H_0; il p value è $P(T_{15} < -2{,}4) \approx 0{,}015$. **5.** Al livello del 5% c'è evidenza che il nuovo algoritmo sia più veloce.

DA RICORDARE

Varianza della popolazione nota: statistica Z e normale standard. Varianza non nota e stimata con S: statistica T e t di Student con $n - 1$ gradi di libertà. Per n grande le due distribuzioni quasi coincidono.

Test su una proporzione

Per verificare un'ipotesi H_0: $p = p_0$ su una proporzione, con n grande (np_0 e $n(1 - p_0)$ almeno 5) si usa l'approssimazione normale; nella varianza si mette p_0, perché il calcolo si fa supponendo vera H_0:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1) \text{ (approssimativamente)}$$

Con n piccolo si usa invece direttamente la distribuzione binomiale del numero di successi, calcolando il p value come probabilità binomiale delle code.

Esempio. Un fornitore garantisce al massimo il 3% di pezzi difettosi. In un lotto di 500 pezzi se ne trovano 24 difettosi, cioè $\hat{p} = 0{,}048$. H_0: $p = 0{,}03$ contro H_1: $p > 0{,}03$, unilaterale destro con valore critico 1,645 al 5%. La statistica vale $z = (0{,}048 - 0{,}03)/\sqrt{(0{,}03 \cdot 0{,}97/500)} \approx 0{,}018/0{,}00763 \approx 2{,}36 > 1{,}645$, con p value circa 0,009: si rifiuta H_0, il lotto non rispetta la garanzia.

Confronto tra le medie di due popolazioni

Spesso si vogliono confrontare due gruppi indipendenti, per esempio i tempi di risposta di due server. Con campioni di numerosità n e m, medie $\bar{X}$ e $\bar{Y}$ e ipotesi nulla H_0: $\mu_X = \mu_Y$, la statistica si costruisce sulla differenza delle medie, che ha varianza $\sigma_X^2/n + \sigma_Y^2/m$.

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

Se le varianze non sono note ma si possono supporre uguali, si stimano insieme con la **varianza combinata** (pooled) e si usa una t con n + m − 2 gradi di libertà:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}, \qquad T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/n + 1/m}} \sim t(n+m-2)$$

Esempio. Il server A, su 40 richieste, ha tempo medio 210 ms con σ_A = 30 ms; il server B, su 50 richieste, 196 ms con σ_B = 25 ms, varianze note da lunghe serie storiche. Per H_1: $\mu_A \neq \mu_B$ la statistica vale z = 14/√(900/40 + 625/50) = 14/√35 ≈ 2,37, maggiore di 1,96: al 5% la differenza tra i due server è significativa (p value circa 0,018).

Test sulla varianza

Nei processi produttivi conta spesso più la regolarità che la media: una macchina che produce pezzi della lunghezza giusta in media, ma molto variabili, è comunque da regolare. Per una popolazione normale l'ipotesi H_0: $\sigma^2 = \sigma_0^2$ si verifica con la statistica

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

che sotto H_0 ha distribuzione chi quadro con n − 1 gradi di libertà. Se H_1: $\sigma^2 > \sigma_0^2$ si rifiuta quando χ^2 supera il quantile $\chi^2_{1-\alpha}$; se H_1: $\sigma^2 < \sigma_0^2$ quando è minore di χ^2_{α}; nel caso bilaterale quando cade fuori dall'intervallo tra $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$. A differenza della normale, la chi quadro non è simmetrica, quindi i due valori critici del caso bilaterale vanno calcolati separatamente.

Esempio. Una macchina è considerata regolare se la deviazione standard del diametro dei pezzi non supera 0,5 mm, cioè σ_0^2 = 0,25. Su 20 pezzi si osserva s^2 = 0,42. Con H_1: $\sigma^2 > 0{,}25$ e α = 0,05, il valore critico è $\chi^2_{0{,}95;\,19} \approx 30{,}14$. La statistica vale 19 · 0,42/0,25 = 31,92, che supera il valore critico (p value circa 0,032): la variabilità è eccessiva e la macchina va regolata.

Test chi quadro di adattamento

Il test chi quadro di **bontà di adattamento** verifica se i dati osservati sono compatibili con una distribuzione ipotizzata. I dati sono divisi in k categorie; per ciascuna si confronta la frequenza osservata O_i con la frequenza attesa $E_i = n \cdot p_i$, dove p_i è la probabilità della categoria secondo l'ipotesi nulla.

DEFINIZIONE

Ipotesi nulla H_0: i dati seguono la distribuzione ipotizzata. Statistica test:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Sotto H_0, per n grande, χ^2 ha approssimativamente distribuzione chi quadro con k − 1 − r gradi di libertà, dove r è il numero di parametri della distribuzione stimati dai dati (r = 0 se la distribuzione è completamente specificata). Si rifiuta H_0 se χ^2 supera il quantile $\chi^2_{1-\alpha}$.

L'idea è semplice: ogni termine misura quanto una categoria si discosta da ciò che ci si aspetta, relativamente a quanto ci si aspetta. Se l'ipotesi è vera gli scarti sono solo fluttuazioni casuali e la somma resta piccola; se è falsa almeno una categoria si discosta molto e la somma cresce. Per questo il test è sempre unilaterale destro. L'approssimazione è considerata affidabile se tutte le frequenze attese sono almeno 5; altrimenti si accorpano le categorie con frequenze attese troppo piccole. Si perde un grado di libertà perché le frequenze sommano a n, e uno per ogni parametro stimato.

ESEMPIO: IL DADO È EQUILIBRATO?

Un dado viene lanciato 120 volte e si ottengono le frequenze riportate nella tabella. Se il dado è equilibrato ogni faccia ha probabilità 1/6 e frequenza attesa 120/6 = 20.

FACCIA	1	2	3	4	5	6	TOTALE
Osservate O_i	14	22	18	25	16	25	120
Attese E_i	20	20	20	20	20	20	120
$(O_i - E_i)^2/E_i$	1,80	0,20	0,20	1,25	0,80	1,25	5,50

Passo 1. H_0: il dado è equilibrato. **Passo 2.** Le categorie sono k = 6 e non si stima nessun parametro, quindi i gradi di libertà sono 5; con α = 0,05 il valore critico è $\chi^2_{0,95;\,5} \approx 11,07$. **Passo 3.** La statistica vale 5,50. **Passo 4.** 5,50 < 11,07 (p value circa 0,36), quindi non si rifiuta H_0. **Passo 5.** Le differenze osservate sono compatibili con le fluttuazioni casuali: i dati non forniscono evidenza che il dado sia truccato.

Test chi quadro di indipendenza

Lo stesso principio permette di verificare se due caratteri qualitativi sono indipendenti. I dati si organizzano in una **tabella di contingenza** con r righe e c colonne, dove la cella (i, j) contiene il numero O_{ij} di unità che presentano la modalità i del primo carattere e la modalità j del secondo. Se i caratteri fossero indipendenti, la probabilità di una cella sarebbe il prodotto delle probabilità marginali, e la frequenza attesa si stima con i totali di riga e di colonna:

$$E_{ij} = \frac{(\text{totale riga i}) \cdot (\text{totale colonna j})}{n}, \qquad \chi^2 = \sum_{i,\,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2[(r-1)(c-1)]$$

Esempio. A 160 studenti si chiede quale sistema operativo usano; 80 frequentano un corso di informatica e 80 un corso di economia. Le frequenze attese, calcolate con i totali, sono riportate tra parentesi.

CORSO	WINDOWS	MACOS	LINUX	TOTALE
Informatica	30 (40)	20 (22,5)	30 (17,5)	80
Economia	50 (40)	25 (22,5)	5 (17,5)	80
Totale	80	45	35	160

La statistica vale 2,5 + 0,28 + 8,93 + 2,5 + 0,28 + 8,93 ≈ 23,41, con (2 − 1)(3 − 1) = 2 gradi di libertà. Il valore critico al 5% è $\chi^2_{0,95;\,2} \approx 5,99$, molto più piccolo: si rifiuta l'ipotesi di indipendenza. La scelta del sistema operativo dipende dal corso frequentato, e il contributo maggiore viene dalla colonna Linux.

Test e intervalli di confidenza

Test bilaterali e intervalli di confidenza sono due facce della stessa medaglia. Il test bilaterale di livello α per H_0: $\mu = \mu_0$ rifiuta H_0 esattamente quando μ_0 non appartiene all'intervallo di confidenza di livello 1 − α per μ. Nell'esempio dei bulloni l'intervallo al 95% è 50,18 ± 1,96 · 0,08, cioè [50,02 ; 50,34], che non contiene 50: coerentemente il test ha rifiutato H_0. L'intervallo dà però un'informazione in più, perché dice anche di quanto la media si è spostata.

CAPITOLO 7

Regressione lineare

Come misurare il legame tra due grandezze e descriverlo con un'equazione: covarianza, coefficiente di correlazione, metodo dei minimi quadrati, coefficiente di determinazione e regressione multipla.

Il problema

Molte grandezze variano insieme: il tempo di esecuzione di un programma e la dimensione dell'input, le ore di studio e il punteggio a un test, la temperatura e il consumo di energia. Dati n punti $(x_1, y_1), \ldots, (x_n, y_n)$ ci si pongono due domande diverse. La prima è se esiste un legame e quanto è forte: risponde la **correlazione**, con un numero. La seconda è quale sia il legame, per poter prevedere y conoscendo x: risponde la **regressione**, con un'equazione.

Le due analisi non sono simmetriche allo stesso modo. La correlazione tratta x e y alla pari; la regressione distingue una **variabile esplicativa** x (o indipendente) da una **variabile risposta** y (o dipendente), e scambiando i ruoli si ottiene una retta diversa.

ATTENZIONE

Correlazione non significa causalità. Due grandezze possono essere fortemente correlate senza che una causi l'altra, perché entrambe dipendono da una terza (le vendite di gelati e il numero di scottature crescono insieme, ma la causa comune è l'estate). Si parla in questi casi di **correlazione spuria**.

Covarianza e coefficiente di correlazione

DEFINIZIONE

La **covarianza empirica** dei dati misura quanto x e y tendono a scostarsi dalle rispettive medie nello stesso verso:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

Il **coefficiente di correlazione lineare** si ottiene dividendo la covarianza per il prodotto delle deviazioni standard, con σ_x^2 e σ_y^2 varianze empiriche:

$$\rho = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}, \qquad -1 \le \rho \le 1$$

Quando un punto ha sia x sia y sopra la media, o entrambi sotto, il prodotto degli scarti è positivo; quando sono da parti opposte è negativo. La covarianza è positiva se prevalgono i primi casi, negativa se prevalgono i secondi. Il suo difetto è che dipende dalle unità di misura: esprimendo x in millisecondi invece che in secondi la covarianza si moltiplica per mille. Il coefficiente ρ elimina questo problema ed è un numero puro.

Il segno di ρ indica la direzione del legame: positivo se y tende a crescere con x, negativo se tende a diminuire. Il valore assoluto indica la forza: ρ = ±1 significa che tutti i punti stanno esattamente su una retta, ρ vicino a 0 che non c'è legame lineare. Come guida si parla di correlazione **debole** per |ρ| fino a 0,3, **moderata** per |ρ| tra 0,3 e 0,7 e **forte** per |ρ| superiore a 0,7.

ATTENZIONE

ρ misura solo il legame **lineare**. Dati disposti lungo una parabola o una curva a campana possono avere ρ vicino a zero pur essendo legati in modo preciso. Prima di commentare ρ bisogna sempre guardare il diagramma di dispersione: se $\rho \approx 0$ la conclusione corretta è che non c'è una relazione lineare, non che non c'è relazione.

Il metodo dei minimi quadrati

Si cerca la retta $y = \beta_0 + \beta_1 x$ che meglio si adatta ai punti. Per ogni x_i la retta prevede il valore $\hat{y}_i = \beta_0 + \beta_1 x_i$; la differenza $e_i = y_i - \hat{y}_i$ tra valore osservato e previsto si chiama **residuo**. Il metodo dei minimi quadrati sceglie β_0 e β_1 in modo da rendere minima la somma dei quadrati dei residui:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Si usano i quadrati perché i residui positivi e negativi non si compensino e perché la funzione risultante si minimizza facilmente con le derivate.

TEOREMA

La retta dei minimi quadrati, o **retta di regressione**, ha coefficienti

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

e passa sempre per il baricentro dei dati $(\bar{x}, \bar{y})$.

DERIVAZIONE

Passo 1. Si annulla la derivata parziale rispetto a β_0, dividendo poi per meno 2n:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \implies \bar{y} - \beta_0 - \beta_1 \bar{x} = 0 \implies \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Passo 2. Si annulla la derivata parziale rispetto a β_1:

$$\frac{\partial S}{\partial \beta_1} = -2\sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \implies \sum_i x_i y_i - n\bar{x}\beta_0 - \beta_1 \sum_i x_i^2 = 0$$

Passo 3. Si sostituisce β_0 trovato al passo 1 e si raccoglie β_1:

$$\sum_i x_i y_i - n\bar{x}\bar{y} = \beta_1 \left(\sum_i x_i^2 - n\bar{x}^2 \right) \implies n\,\sigma_{xy} = \beta_1\, n\,\sigma_x^2$$

da cui $\beta_1 = \sigma_{xy}/\sigma_x^2$. Annullare le derivate garantisce un punto stazionario; che si tratti di un minimo si verifica con la matrice hessiana, che qui è definita positiva (S è una somma di quadrati, una funzione convessa).

Per i calcoli a mano si usano spesso le somme $S_{xx} = \Sigma\, x_i^2 - n\,\bar{x}^2$, $S_{yy} = \Sigma\, y_i^2 - n\,\bar{y}^2$ e $S_{xy} = \Sigma\, x_i y_i - n\,\bar{x}\bar{y}$, che sono le stesse quantità moltiplicate per n: $\beta_1 = S_{xy}/S_{xx}$ e $\rho = S_{xy}/\sqrt{(S_{xx}S_{yy})}$, perché il fattore n si semplifica.

Bontà dell'adattamento

La variabilità totale di y attorno alla sua media si scompone in una parte spiegata dalla retta e in una parte residua, che la retta non riesce a spiegare:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

cioè devianza totale uguale a devianza spiegata più devianza residua.

DEFINIZIONE

Il **coefficiente di determinazione** è la frazione della variabilità di y spiegata dalla regressione:

$$R = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \rho^2$$

Vale tra 0 e 1, e il modello è tanto migliore quanto più R è vicino a 1. Molti testi e programmi indicano la stessa quantità con R^2; ciò che conta è dichiarare la notazione e usarla in modo coerente.

ESEMPIO: ORE DI STUDIO E PUNTEGGIO

Otto studenti riportano le ore di studio x e il punteggio y ottenuto in una prova. Calcoliamo correlazione, retta e coefficiente di determinazione.

X (ORE)	2	3	5	6	8	9	11	12
y (punti)	52	55	63	61	72	75	82	84

Passo 1, medie. $\bar{x} = 56/8 = 7$ e $\bar{y} = 544/8 = 68$.

Passo 2, varianze e covarianza con le formule computazionali: $\Sigma x_i^2 = 484$, $\Sigma y_i^2 = 38008$, $\Sigma x_i y_i = 4111$, quindi

$$\sigma_x^2 = \frac{484}{8} - 7^2 = 11{,}5, \qquad \sigma_y^2 = \frac{38008}{8} - 68^2 = 127, \qquad \sigma_{xy} = \frac{4111}{8} - 7 \cdot 68 = 37{,}875$$

Passo 3, correlazione. $\rho = 37{,}875/\sqrt{(11{,}5 \cdot 127)} \approx 37{,}875/38{,}22 \approx 0{,}991$: correlazione positiva e forte.

Passo 4, retta. $\beta_1 = 37{,}875/11{,}5 \approx 3{,}293$ e $\beta_0 = 68 - 3{,}293 \cdot 7 \approx 44{,}95$. La retta è $\hat{y} = 44{,}95 + 3{,}293x$: ogni ora di studio in più vale in media circa 3,3 punti.

Passo 5, bontà e previsione. $R = \rho^2 \approx 0{,}982$: la retta spiega circa il 98% della variabilità dei punteggi. Per 10 ore di studio la retta prevede circa $44{,}95 + 32{,}93 \approx 77{,}9$ punti. Le previsioni sono affidabili solo all'interno dell'intervallo dei dati osservati: estrapolare a 40 ore darebbe un punteggio superiore al massimo possibile.

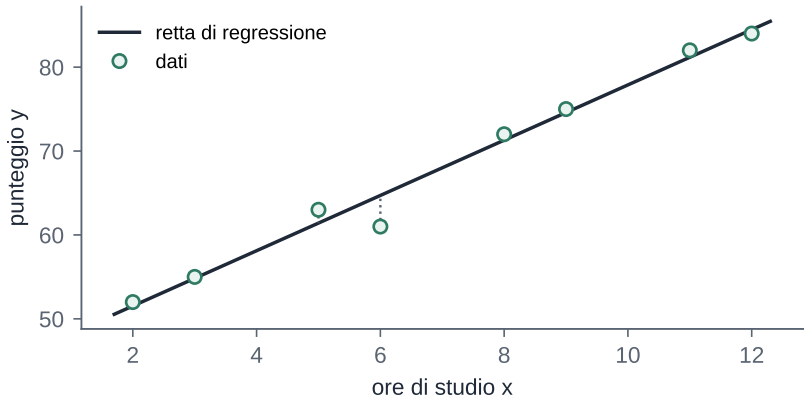

Figura 7.1. Dati, retta dei minimi quadrati e residui (segmenti punteggiati): la retta rende minima la somma dei loro quadrati.

MODELLI LINEARIZZABILI

Il metodo si estende a molte relazioni non lineari con un cambio di variabile. Se i dati seguono una parabola $y = \beta_0 + \beta_1 x^2$, ponendo $z = x^2$ si ottiene $y = \beta_0 + \beta_1 z$, una retta in z. Se seguono un andamento esponenziale $y = a \cdot e^{bx}$, passando ai logaritmi si ha $\ln y = \ln a + bx$, una retta tra x e ln y.

Regressione lineare multipla

Spesso la risposta dipende da più variabili esplicative insieme: il tempo di esecuzione di un programma dipende dalla dimensione dell'input ma anche dal numero di processori, il prezzo di una casa dalla superficie, dalla zona e dall'età. La **regressione multipla** estende il modello a p variabili esplicative $x_1, \ldots, x_p$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

dove ε è l'errore, la parte di y non spiegata dal modello. Il coefficiente β_j si interpreta come la variazione media di y quando x_j aumenta di un'unità **a parità di tutte le altre variabili**. Il modello si chiama lineare perché è lineare nei coefficienti β, anche se le x possono essere trasformazioni di altre grandezze (per esempio $x_2 = x_1^2$).

FORMA MATRICIALE ED EQUAZIONI NORMALI

Con n osservazioni conviene usare le matrici. Si raccolgono le risposte nel vettore colonna y, i coefficienti nel vettore β e i valori delle variabili esplicative nella **matrice del disegno** X, che ha n righe e p + 1 colonne: la prima colonna è fatta di 1 (corrisponde all'intercetta β_0) e la colonna j + 1 contiene i valori di x_j. Il modello diventa $y = X\beta + \varepsilon$ e la somma dei quadrati dei residui è $(y - X\beta)^T(y - X\beta)$.

TEOREMA

Se le colonne di X sono linearmente indipendenti, i coefficienti dei minimi quadrati sono la soluzione delle **equazioni normali**

$$X^T X \beta = X^T y \quad \Longrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

Per p = 1 queste formule si riducono esattamente a quelle della regressione semplice.

Le equazioni normali si ottengono, come nel caso semplice, annullando le derivate parziali della somma dei quadrati rispetto a ciascun coefficiente: si ottiene un sistema lineare di p + 1 equazioni in p + 1 incognite. La condizione sulle colonne di X esclude che una variabile esplicativa sia combinazione lineare delle altre: in quel caso, detto **multicollinearità**, il contributo delle singole variabili non è separabile e $X^T X$ non è invertibile.

La bontà dell'adattamento si misura ancora con il coefficiente di determinazione, definito come 1 − (devianza residua)/(devianza totale). Aggiungere una variabile non può mai peggiorarlo, anche se la variabile è inutile; per confrontare modelli con un numero diverso di variabili si usa allora il **coefficiente di determinazione corretto**, che penalizza le variabili in più:

$$\bar{R} = 1 - (1 - R)\frac{n - 1}{n - p - 1}$$

ESEMPIO: TEMPO DI ESECUZIONE

Si misura il tempo di esecuzione y (in secondi) di un programma al variare della dimensione dell'input x_1 (in migliaia di righe) e del numero di processori x_2.

	1	2	3	4	5	6
x_1 (migliaia di righe)	1	2	3	4	5	6
x_2 (processori)	1	2	1	2	1	2
y (secondi)	2,1	2,9	6,2	6,8	10,1	10,9

Passo 1, le matrici. X ha una colonna di 1, la colonna di x_1 e quella di x_2. I prodotti che servono si ottengono sommando: X^TX contiene n e le somme Σx_1, Σx_2, Σx_1^2, Σx_1x_2, Σx_2^2; X^Ty contiene Σy, Σx_1y, Σx_2y.

$$X^TX = \begin{bmatrix} 6 & 21 & 9 \\ 21 & 91 & 33 \\ 9 & 33 & 15 \end{bmatrix} \qquad X^Ty = \begin{bmatrix} 39 \\ 169{,}6 \\ 59{,}6 \end{bmatrix}$$

Passo 2, le equazioni normali. Il sistema $X^TX\beta = X^Ty$, risolto per esempio per sostituzione, dà $\beta_0 = 1{,}4$, $\beta_1 = 2$ e $\beta_2 \approx -1{,}267$. Il modello stimato è

$$\hat{y} = 1{,}4 + 2\,x_1 - 1{,}267\,x_2$$

Passo 3, interpretazione. A parità di processori, ogni mille righe in più costano in media 2 secondi; a parità di dimensione dell'input, ogni processore in più fa risparmiare in media circa 1,27 secondi.

Passo 4, bontà. I valori previsti sono molto vicini a quelli osservati e il coefficiente di determinazione vale circa 0,9998 (corretto circa 0,9997). Con una sola variabile esplicativa, la dimensione dell'input, il modello spiegherebbe meno, perché attribuirebbe al caso l'effetto del numero di processori.

CAPITOLO 8

Numeri pseudocasuali e metodo Monte Carlo

Come un calcolatore, che è deterministico, produce numeri che si comportano come casuali; come trasformarli in campioni di una distribuzione assegnata; come usarli per calcolare integrali e probabilità.

Numeri pseudocasuali

Le simulazioni richiedono grandi quantità di numeri casuali, tipicamente uniformi sull'intervallo [0, 1). Generarli da un processo fisico (rumore elettronico, decadimento radioattivo) è lento e non riproducibile; per questo si usano algoritmi che producono sequenze che **sembrano** casuali.

DEFINIZIONE

Un numero **pseudocasuale** è un numero prodotto da un algoritmo **deterministico** a partire da un valore iniziale detto **seme** (seed), in modo che la sequenza generata abbia le proprietà statistiche di una sequenza di variabili aleatorie indipendenti e uniformi su [0, 1).

La definizione contiene tre idee. La sequenza è **deterministica**: nota la regola e noto il seme, ogni numero è prevedibile. È **riproducibile**: lo stesso seme produce sempre la stessa sequenza, il che non è un difetto ma una proprietà preziosa, perché permette di ripetere esattamente una simulazione per verificarla o correggerla; semi diversi producono sequenze diverse. Infine deve **imitare il caso**: i numeri devono essere distribuiti uniformemente, senza correlazioni apprezzabili tra numeri successivi, e la sequenza deve ripetersi solo dopo moltissimi valori.

Il generatore congruenziale lineare

DEFINIZIONE

Fissati quattro interi non negativi, il **modulo** $m > 0$, il **moltiplicatore** $a < m$, l'**incremento** $c < m$ e il **seme** X_0, il **generatore congruenziale lineare** produce la successione

$$X_{n+1} = (a\,X_n + c) \bmod m, \qquad U_n = \frac{X_n}{m} \in [0, 1)$$

dove "mod m" indica il resto della divisione per m. Se $c = 0$ il generatore si dice **moltiplicativo**, altrimenti **misto**.

Poiché X_n può assumere solo i valori da 0 a $m - 1$, prima o poi un valore si ripete e da quel momento la sequenza ricomincia identica: la successione è **periodica**. Il **periodo** è la lunghezza del ciclo e non può superare m; un generatore ha **periodo pieno** se il periodo è proprio m, cioè se produce tutti i valori possibili prima di ripetersi.

TEOREMA

Hull e Dobell. Il generatore misto ($c \neq 0$) ha periodo pieno m se e solo se valgono tre condizioni: c e m sono primi tra loro; $a - 1$ è divisibile per ogni numero primo che divide m; $a - 1$ è divisibile per 4 se m è divisibile per 4.

ESEMPIO: UN GENERATORE A MANO

Consideriamo m = 16, a = 9, c = 7 e seme $X_0 = 1$.

Passo 1, verifica delle condizioni. 7 e 16 sono primi tra loro. L'unico primo che divide 16 è 2, e a − 1 = 8 è divisibile per 2. Infine 16 è divisibile per 4 e anche 8 lo è. Il periodo è pieno, uguale a 16.

Passo 2, primi valori. $X_1 = (9 \cdot 1 + 7) \bmod 16 = 16 \bmod 16 = 0$; $X_2 = (0 + 7) \bmod 16 = 7$; $X_3 = (63 + 7) \bmod 16 = 70 \bmod 16 = 6$, perché $70 = 4 \cdot 16 + 6$; $X_4 = (54 + 7) \bmod 16 = 61 \bmod 16 = 13$. I corrispondenti numeri in [0, 1) sono $U_1 = 0$, $U_2 = 7/16 = 0{,}4375$, $U_3 = 6/16 = 0{,}375$, $U_4 = 13/16 = 0{,}8125$.

Passo 3, il ciclo completo. Proseguendo si ottiene 0, 7, 6, 13, 12, 3, 2, 9, 8, 15, 14, 5, 4, 11, 10, 1 e poi di nuovo 0: compaiono tutti i 16 valori. Se invece si sceglie c = 4, che non è primo con 16, la prima condizione cade e dal seme 1 si ottiene il ciclo 13, 9, 5, 1, di periodo appena 4.

I generatori congruenziali sono semplici e veloci, ma hanno difetti noti: il periodo è al più m, quindi servono moduli enormi (dell'ordine di 2^{31} o 2^{64}); le coppie di valori successivi (U_n, U_{n+1}) non riempiono uniformemente il quadrato ma si dispongono su un numero limitato di rette parallele; le cifre meno significative sono poco casuali. Per questo le librerie moderne usano generatori più sofisticati, come il Mersenne Twister, con periodo $2^{19937} - 1$; il generatore congruenziale resta però il modello per capire il principio.

Generazione con densità assegnata

Un generatore produce numeri uniformi su [0, 1). Per simulare un tempo di attesa esponenziale, un lancio di dado o una qualunque altra distribuzione bisogna trasformarli. Il metodo più generale è il metodo dell'inversione.

TEOREMA

Metodo dell'inversione. Sia U uniforme su (0, 1) e sia F una funzione di ripartizione continua e strettamente crescente. Allora la variabile $X = F^{-1}(U)$ ha funzione di ripartizione F.

DIMOSTRAZIONE

Poiché F è strettamente crescente, la disuguaglianza $F^{-1}(U) \le x$ equivale a $U \le F(x)$. Per una variabile uniforme su (0, 1) si ha $P(U \le t) = t$ per ogni t tra 0 e 1, quindi

$$P(X \le x) = P\left(F^{-1}(U) \le x\right) = P(U \le F(x)) = F(x)$$

L'idea si visualizza sul grafico di F: si estrae un'altezza uniforme sull'asse verticale e si legge il valore corrispondente sull'asse orizzontale. Dove F è ripida, cioè dove la densità è alta, un ampio intervallo di altezze corrisponde a un intervallo stretto di valori, che quindi vengono generati spesso.

ESEMPIO: LA DISTRIBUZIONE ESPONENZIALE

Passo 1. La ripartizione dell'esponenziale è $F(x) = 1 - e^{-\lambda x}$. Si pone $F(x) = u$.

Passo 2. Si risolve rispetto a x: $e^{-\lambda x} = 1 - u$, quindi $-\lambda x = \ln(1 - u)$ e

$$x = F^{-1}(u) = -\frac{1}{\lambda}\ln(1 - u)$$

Passo 3. Poiché anche 1 − U è uniforme su (0, 1), si può usare la forma più semplice $X = -\ln(U)/\lambda$. Con $\lambda = 2$ e $u = 0{,}25$ nella prima forma si ottiene $x = -\ln(0{,}75)/2 \approx 0{,}144$.

L'esponenziale è uno dei pochi casi in cui F si inverte con una formula esplicita; per la normale l'inversa non ha forma chiusa e le librerie usano approssimazioni numeriche o metodi specifici.

VARIABILI DISCRETE

Per una variabile discreta con valori $x_1 < x_2 < \dots$ e probabilità $p_1, p_2, \dots$ l'inversione diventa una ricerca: si divide $[0, 1)$ in intervalli consecutivi di lunghezza $p_1, p_2, \dots$ e si restituisce x_k se U cade nel k esimo intervallo, cioè se $F(x_{k-1}) \le U < F(x_k)$. Per esempio, per simulare un server che è attivo con probabilità 0,7, carico con probabilità 0,2 e in manutenzione con probabilità 0,1, si restituisce "attivo" se $U < 0{,}7$, "carico" se $0{,}7 \le U < 0{,}9$ e "manutenzione" altrimenti.

METODO DEL RIGETTO

Quando F non si sa invertire si può usare il metodo del rigetto. Si sceglie una densità g facile da simulare (spesso l'uniforme) e una costante $M \ge 1$ tale che $f(x) \le M \cdot g(x)$ per ogni x; poi si ripete il procedimento seguente.

1. Si genera Y con densità g e, indipendentemente, U uniforme su $(0, 1)$.
2. Se $U \le f(Y)/(M \cdot g(Y))$ si accetta e si pone $X = Y$.
3. Altrimenti si scarta Y e si ricomincia dal primo passo.

I valori accettati hanno densità f. Geometricamente si generano punti uniformi sotto la curva $M \cdot g$ e si tengono quelli che cadono sotto f; la probabilità di accettazione è $1/M$, quindi il metodo è tanto più efficiente quanto più $M \cdot g$ aderisce a f.

Il metodo Monte Carlo

DEFINIZIONE

Si chiama **metodo Monte Carlo** ogni tecnica che stima una quantità deterministica (una probabilità, un integrale, un valore atteso) come media di un gran numero di esperimenti simulati con numeri casuali. La legge dei grandi numeri garantisce che la stima converga al valore vero, e il teorema del limite centrale permette di valutarne l'errore.

METODO HIT OR MISS

Si vuole calcolare l'integrale I di una funzione f non negativa su $[a, b]$, con $f(x) \le c$, cioè l'area della regione sotto il grafico. Si racchiude la regione nel rettangolo $[a, b] \times [0, c]$ e si procede come lanciando freccette a caso sul rettangolo.

1. Si generano N punti (x_i, y_i) uniformi nel rettangolo: x_i uniforme su $[a, b]$ e y_i uniforme su $[0, c]$.
2. Si conta il numero H di punti "colpiti", cioè con $y_i \le f(x_i)$.
3. Si stima l'integrale come area del rettangolo per frazione di colpi: $\hat{I} = c(b - a) \cdot H/N$.

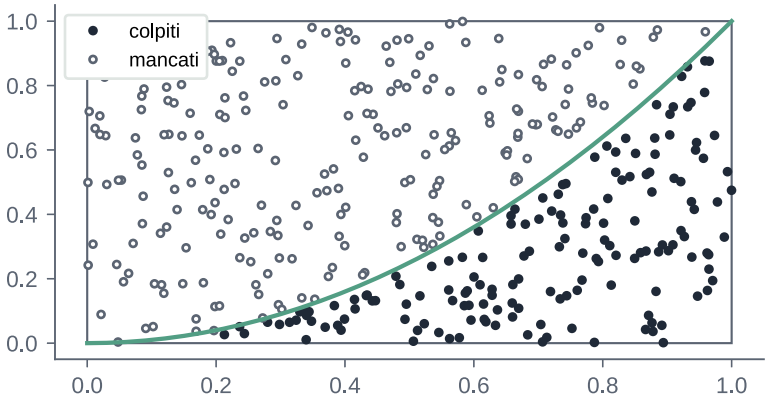

Figura 8.1. Hit or miss per l'area sotto $f(x) = x^2$ su $[0, 1]$: la frazione di punti pieni stima l'area (valore esatto 1/3).

Ogni punto è un esperimento di Bernoulli con probabilità di successo p = I/(c(b − a)), il rapporto tra l'area cercata e quella del rettangolo. La frazione H/N è quindi una media di N Bernoulli: è uno stimatore non distorto di p, con varianza p(1 − p)/N. Ne segue che l'errore standard della stima è

$$\text{errore standard}(\hat{I}) = c\,(b-a)\sqrt{\frac{p(1-p)}{N}}$$

Esempio. Per $I = \int x^2\,dx$ su [0, 1], il cui valore esatto è 1/3, si usa il quadrato unitario (c = 1). Con N = 10 000 punti l'errore standard è √((1/3)(2/3)/10 000) ≈ 0,0047: una stima tipica sarà 0,333 più o meno un paio di millesimi, e per guadagnare una cifra decimale servirebbero 100 volte più punti.

METODO DELLA MEDIA CAMPIONARIA

Un modo più efficiente sfrutta direttamente la definizione di valore atteso. Se X è uniforme su [a, b], allora E[f(X)] è l'integrale di f diviso per b − a; per la legge dei grandi numeri

$$\hat{I} = (b-a)\frac{1}{N}\sum_{i=1}^{N} f(X_i), \qquad \text{errore standard}(\hat{I}) = (b-a)\frac{\sigma_f}{\sqrt{N}}$$

dove $\sigma_f^2 = \text{Var}(f(X))$. Per x^2 su [0, 1] si ha $\text{Var}(U^2) = E[U^4] - (1/3)^2 = 1/5 - 1/9 = 4/45$, e con N = 10 000 l'errore standard è circa 0,0030, più piccolo di quello dell'hit or miss a parità di punti. Questo metodo non richiede di conoscere un limite superiore c della funzione.

VANTAGGI E SVANTAGGI

Il metodo Monte Carlo ha il grande pregio della **generalità**: funziona per funzioni prive di primitiva, per regioni di forma complicata e per problemi in molte dimensioni, ed è semplice da programmare. Il suo errore decresce come 1/√N **indipendentemente dalla dimensione** del problema, mentre i metodi numerici classici peggiorano rapidamente all'aumentare delle dimensioni: per integrali in molte variabili è spesso l'unico metodo praticabile. Inoltre l'errore si stima dai dati stessi, con l'errore standard.

D'altra parte la convergenza è **lenta**: per dimezzare l'errore servono quattro volte più simulazioni, per una cifra decimale in più cento volte di più. Il risultato è una stima con un margine di errore, mai un valore esatto, e la sua qualità dipende da quella del generatore di numeri pseudocasuali: correlazioni nascoste nella sequenza possono distorcere la stima senza che ce ne si accorga.

DA RICORDARE

Una probabilità stimata per simulazione si calcola sempre come numero di esperimenti favorevoli diviso numero di esperimenti simulati, e deve stare tra 0 e 1. L'errore standard di una frequenza stimata su N esperimenti è √(p(1 − p)/N).

CAPITOLO 9

Catene di Markov

Sistemi che evolvono nel tempo in modo casuale e senza memoria del passato: processi stocastici, matrice di transizione, classificazione degli stati, distribuzione stazionaria e simulazione.

Processi stocastici

Finora abbiamo studiato variabili aleatorie isolate o campioni di variabili indipendenti. Molti fenomeni però evolvono nel tempo e il loro stato a un certo istante dipende dagli stati precedenti: il numero di utenti collegati a un servizio minuto per minuto, il meteo giorno per giorno, la posizione di un robot che si sposta su una griglia.

DEFINIZIONE

Un **processo stocastico** è una famiglia di variabili aleatorie $\{X_t : t \in T\}$, tutte definite sullo stesso spazio di probabilità e indicizzate da un parametro t che rappresenta il tempo.

Se T è un insieme discreto, per esempio {0, 1, 2, …}, il processo è **a tempo discreto**; se è un intervallo, è **a tempo continuo**. L'insieme S dei valori che le X_t possono assumere si chiama **spazio degli stati**, e anch'esso può essere discreto o continuo.

Un processo stocastico non è dunque una singola variabile aleatoria, ma un'intera successione di variabili, una per ogni istante. Fissato un esito dell'esperimento, il processo produce una **traiettoria**, cioè una sequenza di stati nel tempo, come la sequenza dei prezzi di un titolo in borsa giorno dopo giorno. Combinando tempo e stati discreti o continui si ottengono quattro famiglie di processi: le catene di Markov, oggetto di questo capitolo, hanno tempo e stati discreti; il processo di Poisson ha tempo continuo e stati discreti.

La catena di Markov

DEFINIZIONE

Un processo stocastico a tempo discreto $\{X_n : n = 0, 1, 2, \ldots\}$ con spazio degli stati discreto S è una **catena di Markov** se vale la **proprietà di Markov**: per ogni n e per ogni scelta degli stati

$$P(X_{n+1} = j \mid X_n = i,\ X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i)$$

La catena è **omogenea** se queste probabilità non dipendono da n; in tal caso si indicano con p_{ij} e si chiamano **probabilità di transizione** dallo stato i allo stato j.

La proprietà di Markov dice che il futuro dipende dal passato solo attraverso il presente: noto lo stato attuale, sapere come ci si è arrivati non aggiunge nessuna informazione sul passo successivo. È la stessa assenza di memoria già incontrata per la geometrica e l'esponenziale. Un gioco da tavolo in cui ci si sposta lanciando un dado è una catena di Markov: la casella in cui si arriverà dipende solo da quella in cui ci si trova ora, non dal percorso fatto. Nel seguito tutte le catene sono omogenee e con un numero finito di stati.

Matrice di transizione e grafo

DEFINIZIONE

La **matrice di transizione** di una catena con stati 1, …, k è la matrice quadrata $P = (p_{ij})$ il cui elemento di riga i e colonna j è la probabilità di passare da i a j in un passo. È una **matrice stocastica**, cioè soddisfa

$$p_{ij} \geq 0 \quad \text{per ogni } i, j, \qquad \sum_{j=1}^{k} p_{ij} = 1 \quad \text{per ogni } i$$

ATTENZIONE

Sono le **righe** a sommare a 1, non le colonne. La riga i elenca le probabilità di tutte le destinazioni possibili partendo da i: da qualche parte bisogna andare, eventualmente restando in i. Per verificare che una matrice sia di transizione si controllano quindi le righe.

La catena si rappresenta anche con un **grafo orientato**: un nodo per ogni stato e una freccia da i a j, etichettata con p_{ij}, per ogni transizione con probabilità positiva. Le frecce che escono da ciascun nodo hanno etichette che sommano a 1.

ESEMPIO: LO STATO DI UN SERVER

Ogni minuto un server può essere attivo (A), sotto carico (C) o in manutenzione (M). Da A resta attivo con probabilità 0,7, va sotto carico con probabilità 0,2 e in manutenzione con probabilità 0,1; da C torna attivo con probabilità 0,5, resta sotto carico con 0,3 e va in manutenzione con 0,2; da M torna attivo con probabilità 0,6 e resta in manutenzione con 0,4. Con gli stati nell'ordine A, C, M:

$$P = \begin{bmatrix} 0{,}7 & 0{,}2 & 0{,}1 \\ 0{,}5 & 0{,}3 & 0{,}2 \\ 0{,}6 & 0 & 0{,}4 \end{bmatrix}$$

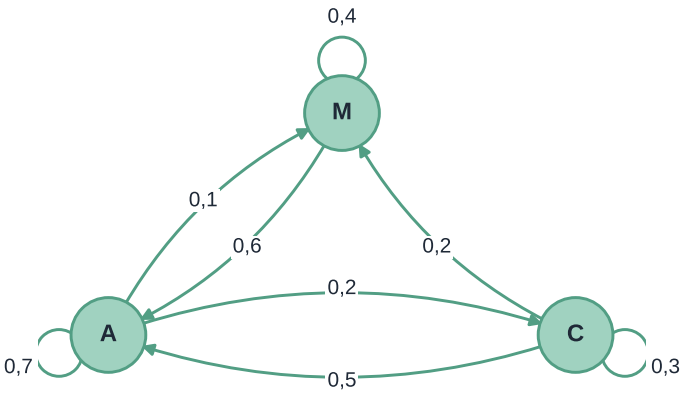

Figura 9.1. Grafo della catena del server: A attivo, C carico, M manutenzione. Gli anelli sono le probabilità di restare nello stesso stato.

TRANSIZIONI IN PIÙ PASSI

La probabilità di passare da i a j in esattamente n passi si indica con $p_{ij}^{(n)}$. Per calcolarla si condiziona sullo stato intermedio: per andare da i a j in m + n passi bisogna trovarsi in qualche stato l dopo m passi, e poi andare da l a j negli n passi restanti.

TEOREMA

Equazioni di Chapman e Kolmogorov. Per ogni m, n ≥ 0

$$p_{ij}^{(m+n)} = \sum_{l \in S} p_{il}^{(m)}\, p_{lj}^{(n)}$$

In forma matriciale $P^{(m+n)} = P^{(m)}P^{(n)}$, e quindi la matrice di transizione in n passi è la potenza n esima di P: $P^{(n)} = P^n$.

La dimostrazione usa il teorema delle probabilità totali con la partizione data dagli stati possibili al tempo m, e la proprietà di Markov per affermare che, una volta in l, il seguito non dipende da come ci si è arrivati.

PROPRIETÀ

Sia $\pi^{(0)}$ il vettore riga della **distribuzione iniziale**, con componenti $\pi_i^{(0)} = P(X_0 = i)$. La distribuzione al tempo n si ottiene moltiplicando a destra per la potenza della matrice:

$$\pi^{(n)} = \pi^{(0)} P^n, \qquad P(X_n = j) = \sum_i \pi_i^{(0)}\, p_{ij}^{(n)}$$

Esempio. Se il server parte in manutenzione, $\pi^{(0)} = (0, 0, 1)$. Dopo un minuto la distribuzione è la terza riga di P, $\pi^{(1)} = (0{,}6\ ;\ 0\ ;\ 0{,}4)$. Dopo due minuti

$$\pi^{(2)} = \pi^{(1)}P = (0{,}6 \cdot 0{,}7 + 0{,}4 \cdot 0{,}6\ ;\ 0{,}6 \cdot 0{,}2 + 0\ ;\ 0{,}6 \cdot 0{,}1 + 0{,}4 \cdot 0{,}4) = (0{,}66\ ;\ 0{,}12\ ;\ 0{,}22)$$

Classificazione degli stati

DEFINIZIONE

Lo stato j è **accessibile** da i (si scrive $i \to j$) se esiste $n \geq 0$ con $p_{ij}^{(n)} > 0$, cioè se nel grafo c'è un cammino da i a j. Due stati **comunicano** ($i \leftrightarrow j$) se ciascuno è accessibile dall'altro.

La comunicazione è una relazione di equivalenza e divide gli stati in **classi di comunicazione**: all'interno di una classe tutti gli stati comunicano tra loro.

DEFINIZIONE

Un insieme di stati C è una **classe chiusa** se dalla classe non si può uscire: $p_{ij} = 0$ per ogni i in C e ogni j fuori da C. Uno stato i con $p_{ii} = 1$ si dice **assorbente**: da solo forma una classe chiusa.

Una catena è **irriducibile** se tutti gli stati comunicano tra loro, cioè se c'è un'unica classe di comunicazione. Equivalentemente, l'unica classe chiusa è l'intero spazio degli stati.

La catena del server è irriducibile: da ogni stato si raggiunge ogni altro (per esempio A → C → M → A). Una catena con uno stato assorbente e almeno un altro stato non è mai irriducibile, perché lo stato assorbente forma una classe chiusa più piccola dell'intero spazio.

DEFINIZIONE

Uno stato i è **ricorrente** se, partendo da i, la catena vi ritorna prima o poi con probabilità 1; è **transitorio** se c'è una probabilità positiva di non tornarci mai più.

Il **periodo** di uno stato i è il massimo comun divisore dei numeri $n \geq 1$ per cui $p_{ii}^{(n)} > 0$. Lo stato è **aperiodico** se il periodo è 1.

Ricorrenza, transitorietà e periodo sono proprietà di classe: tutti gli stati di una stessa classe le condividono. In una catena finita le classi chiuse sono formate da stati ricorrenti e le classi non chiuse da stati transitori, perché prima o poi la catena ne esce e non vi rientra più. Basta che un solo stato di una classe abbia $p_{ii} > 0$ perché la classe sia aperiodica, dato che allora si può tornare in i in un passo. Un esempio di periodicità è la catena che alterna deterministicamente due stati, con $p_{12} = p_{21} = 1$: si torna nello stato di partenza solo dopo un numero pari di passi, e il periodo è 2.

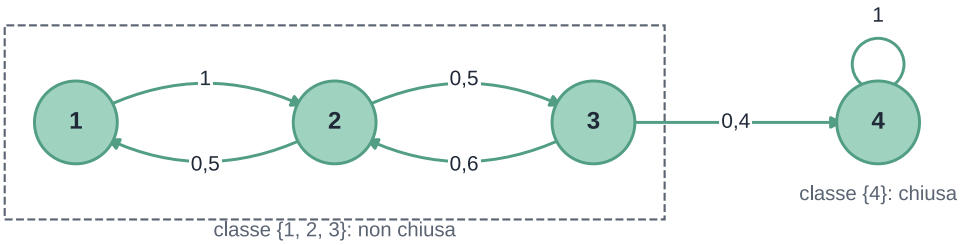

Figura 9.2. Una catena con due classi di comunicazione. Lo stato 4 è assorbente; gli stati 1, 2 e 3 sono transitori.

Nella catena della figura le classi di comunicazione sono {1, 2, 3} e {4}. La classe {4} è chiusa, perché lo stato 4 è assorbente; la classe {1, 2, 3} non è chiusa, perché da 3 si va in 4 con probabilità 0,4. La catena non è irriducibile. Gli stati 1, 2 e 3 sono transitori, perché prima o poi la catena finisce in 4 e non torna più indietro; lo stato 4 è ricorrente.

DA RICORDARE

Per leggere una catena dal grafo: controllare che le frecce uscenti da ogni nodo sommino a 1; raggruppare gli stati che si raggiungono a vicenda nelle classi di comunicazione; per ogni classe verificare se ci sono frecce che escono (non chiusa) oppure no (chiusa); la catena è irriducibile se c'è una sola classe.

Distribuzione stazionaria

Che cosa succede alla catena dopo molto tempo? In molti casi la distribuzione $\pi^{(n)}$ si stabilizza su un vettore che non dipende più dallo stato di partenza.

DEFINIZIONE

Un vettore di probabilità $\pi = (\pi_1, \ldots, \pi_k)$, con $\pi_i \geq 0$ e $\Sigma\, \pi_i = 1$, è una **distribuzione stazionaria** (o invariante) se

$$\pi = \pi P, \qquad \text{cioè} \qquad \pi_j = \sum_i \pi_i\, p_{ij} \quad \text{per ogni } j$$

Il nome si spiega subito: se la catena parte con distribuzione π allora $\pi^{(1)} = \pi P = \pi$ e, ripetendo, $\pi^{(n)} = \pi$ per ogni n. La distribuzione non cambia nel tempo, pur continuando la catena a muoversi da uno stato all'altro: è un equilibrio statistico, non un arresto.

TEOREMA

Una catena finita e **irriducibile** ha una e una sola distribuzione stazionaria π, con tutte le componenti positive. Inoltre $\pi_i = 1/\mu_i$, dove μ_i è il tempo medio di ritorno in i partendo da i.

Se la catena è anche **aperiodica** (in tal caso si dice **ergodica**), qualunque sia la distribuzione iniziale $\pi^{(n)}$ converge a π, e tutte le righe di P^n convergono a π.

ATTENZIONE

L'esistenza della distribuzione stazionaria non garantisce la convergenza. Nella catena che alterna due stati $\pi = (1/2, 1/2)$ è stazionaria, ma partendo dallo stato 1 la distribuzione oscilla per sempre tra (1, 0) e (0, 1): per la convergenza serve l'aperiodicità. In una catena non irriducibile, invece, possono esistere più distribuzioni stazionarie, una per ogni classe chiusa.

COME SI CALCOLA

Si risolve il sistema lineare $\pi = \pi P$ insieme alla condizione di normalizzazione $\Sigma\, \pi_i = 1$. Le k equazioni di $\pi = \pi P$ non sono indipendenti (la loro somma è un'identità), quindi se ne scarta una e la si sostituisce con la normalizzazione.

Esempio. Per la catena del server le equazioni sono

$$\pi_A = 0{,}7\,\pi_A + 0{,}5\,\pi_C + 0{,}6\,\pi_M, \quad \pi_C = 0{,}2\,\pi_A + 0{,}3\,\pi_C, \quad \pi_M = 0{,}1\,\pi_A + 0{,}2\,\pi_C + 0{,}4\,\pi_M$$

Passo 1. Si scarta la prima equazione. Dalla seconda, $0{,}7\pi_C = 0{,}2\pi_A$, quindi $\pi_C = (2/7)\pi_A$.

Passo 2. Dalla terza, $0{,}6\pi_M = 0{,}1\pi_A + 0{,}2\pi_C = 0{,}1\pi_A + (0{,}4/7)\pi_A = (1{,}1/7)\pi_A$, quindi $\pi_M = (11/42)\pi_A$.

Passo 3. Si impone la normalizzazione: $\pi_A(1 + 2/7 + 11/42) = \pi_A \cdot 65/42 = 1$, da cui

$$\pi = \left(\frac{42}{65},\ \frac{12}{65},\ \frac{11}{65}\right) \approx (0{,}646\ ;\ 0{,}185\ ;\ 0{,}169)$$

Passo 4, interpretazione. Nel lungo periodo il server è attivo circa il 64,6% del tempo, sotto carico il 18,5% e in manutenzione il 16,9%. La catena è irriducibile e aperiodica ($p_{AA} > 0$), quindi è ergodica: partendo da qualunque stato la distribuzione converge a π, e già $\pi^{(2)} = (0{,}66\ ;\ 0{,}12\ ;\ 0{,}22)$ calcolata prima è vicina al limite. Il tempo medio di ritorno nello stato M, cioè il numero medio di minuti tra un minuto di manutenzione e il successivo, è $1/\pi_M = 65/11 \approx 5{,}9$.

Al calcolatore π si trova anche elevando P a una potenza alta, perché per una catena ergodica tutte le righe di P^n tendono a π, oppure come autovettore sinistro di P relativo all'autovalore 1, cioè come soluzione di $\pi(P - I) = 0$.

Simulazione di una catena

Simulare una catena significa generarne una traiettoria. Si parte da uno stato iniziale X_0, scelto o estratto dalla distribuzione iniziale; a ogni passo, trovandosi nello stato i, si estrae lo stato successivo dalla distribuzione discreta della riga i di P, con il metodo dell'inversione visto nel capitolo precedente: si genera U uniforme e si sceglie il primo stato j per cui la somma $p_{i1} + \dots + p_{ij}$ supera U. Ripetendo per molti passi e contando la frazione di tempo trascorsa in ciascuno stato si ottiene, per la legge dei grandi numeri applicata alle catene ergodiche, una stima della distribuzione stazionaria. È un modo concreto di verificare i calcoli teorici e l'unico disponibile quando la catena è troppo grande per risolvere il sistema.

CAPITOLO 10

Formulario

Le formule principali della dispensa raccolte in tabelle, per il ripasso finale.

Probabilità

RISULTATO	FORMULA
Contrario	$P(A^c) = 1 - P(A)$
Unione	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Probabilità condizionata	$P(A \mid B) = P(A \cap B) / P(B)$
Indipendenza	$P(A \cap B) = P(A)\,P(B)$
Probabilità totali	$P(B) = \Sigma_i\, P(B \mid A_i)\, P(A_i)$
Bayes	$P(A_k \mid B) = P(B \mid A_k)\, P(A_k) / \Sigma_i\, P(B \mid A_i)\, P(A_i)$
Coefficiente binomiale	$C(n, r) = n! / (r!\,(n - r)!)$

Variabili aleatorie e teoremi limite

RISULTATO	FORMULA
Varianza	$\mathrm{Var}(X) = E[X^2] - E[X]^2$
Trasformazione lineare	$E[aX + b] = aE[X] + b$, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
Somma	$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
Standardizzazione	$Z = (X - \mu)/\sigma \sim N(0, 1)$ se $X \sim N(\mu, \sigma^2)$
Media campionaria	$E[\bar{X}_n] = \mu$, $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$
Markov	$P(X \geq \varepsilon) \leq E[X]/\varepsilon$ per $X \geq 0$
Chebyshev	$P(\lvert X - \mu\rvert \geq \varepsilon) \leq \sigma^2/\varepsilon^2$
Limite centrale	$(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ tende in legge a $N(0, 1)$
Quantili normali	$z_{0,95} = 1{,}645$, $z_{0,975} = 1{,}96$, $z_{0,995} = 2{,}576$

Statistica descrittiva e stimatori

INDICE O STIMATORE	FORMULA
Media	$\bar{x} = (1/N)\,\Sigma\, x_i = (1/N)\,\Sigma\, x_i f_i$
Varianza empirica	$\hat{\sigma}^2 = (1/N)\,\Sigma\,(x_i - \bar{x})^2 = (1/N)\,\Sigma\, x_i^2 - \bar{x}^2$
Varianza campionaria	$s^2 = (1/(N-1))\,\Sigma\,(x_i - \bar{x})^2$
Coefficiente di variazione	$CV = s / \bar{x}$
Differenza interquartile	$IQR = Q_3 - Q_1$
Varianza con μ nota	$S_\mu^{\,2} = (1/n)\,\Sigma\,(X_i - \mu)^2$, non distorto
Errore quadratico medio	$MSE(T) = Var(T) + (E[T] - \theta)^2$
Proporzione	$\hat{p} = X/n$, $Var(\hat{p}) = p(1-p)/n$

Intervalli di confidenza di livello 1 − α

PARAMETRO	INTERVALLO
Media, σ nota	$\bar{x} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$
Media, σ non nota	$\bar{x} \pm t_{1-\alpha/2;\,n-1}\,s/\sqrt{n}$
Proporzione (n grande)	$\hat{p} \pm z_{1-\alpha/2}\,\sqrt{\hat{p}(1-\hat{p})/n}$
Varianza (popolazione normale)	$[(n-1)s^2/\chi^2_{1-\alpha/2}\,;\,(n-1)s^2/\chi^2_{\alpha/2}]$, con n − 1 gradi di libertà
Numerosità per margine E	$n \geq (z_{1-\alpha/2}\,\sigma / E)^2$

Statistiche test

IPOTESI NULLA	STATISTICA	DISTRIBUZIONE SOTTO H_0
$\mu = \mu_0$, σ nota	$(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$	$N(0, 1)$
$\mu = \mu_0$, σ non nota	$(\bar{X} - \mu_0)/(S/\sqrt{n})$	$t(n-1)$
$p = p_0$	$(\hat{p} - p_0)/\sqrt{(p_0(1-p_0)/n)}$	$N(0, 1)$ per n grande
$\mu_X = \mu_Y$, varianze note	$(\bar{X} - \bar{Y})/\sqrt{(\sigma_X^{\,2}/n + \sigma_Y^{\,2}/m)}$	$N(0, 1)$
$\mu_X = \mu_Y$, varianze uguali non note	$(\bar{X} - \bar{Y})/(S_p\sqrt{(1/n + 1/m)})$	$t(n+m-2)$
$\sigma^2 = \sigma_0^{\,2}$	$(n-1)S^2/\sigma_0^{\,2}$	$\chi^2(n-1)$
Adattamento	$\Sigma\,(O_i - E_i)^2/E_i$	$\chi^2(k-1-r)$
Indipendenza	$\Sigma\,(O_{ij} - E_{ij})^2/E_{ij}$	$\chi^2((r-1)(c-1))$

Regressione, Monte Carlo, catene di Markov

RISULTATO	FORMULA
Covarianza empirica	$\sigma_{xy} = (1/n)\ \Sigma\ x_i y_i - \bar{x}\bar{y}$
Correlazione	$\rho = \sigma_{xy}/(\sigma_x \sigma_y)$
Retta di regressione	$\beta_1 = \sigma_{xy}/\sigma_x^2$, $\beta_0 = \bar{y} - \beta_1 \bar{x}$
Coefficiente di determinazione	$R = \rho^2 = 1 -$ devianza residua / devianza totale
Regressione multipla	$\beta = (X^T X)^{-1} X^T y$
Generatore congruenziale	$X_{n+1} = (aX_n + c) \bmod m$, $U_n = X_n/m$
Inversione esponenziale	$X = -\ln(U)/\lambda$
Hit or miss	$\hat{I} = c(b-a)H/N$, errore standard $c(b-a)\sqrt{(p(1-p)/N)}$
Media campionaria	$\hat{I} = (b-a)(1/N)\ \Sigma\ f(X_i)$, errore standard $(b-a)\sigma_f/\sqrt{N}$
Distribuzione al tempo n	$\pi^{(n)} = \pi^{(0)} P^n$
Distribuzione stazionaria	$\pi = \pi P$, $\Sigma\ \pi_i = 1$, tempo medio di ritorno $1/\pi_i$