Note del corso di Teoria dell'Informazione e Inferenza

Alessandro Verri

Il corso è diviso in tre parti. In dieci lezioni, dalla 1 alla 10, presentiamo gli elementi di teoria della probabilità, ovvero alcuni degli strumenti che consentono di trarre conclusioni certe su fatti governati da incertezza, Il rigore matematico è sacrificato a vantaggio dell'acquisizione di una buona capacità di padroneggiare i concetti principali e di risolvere semplici problemi. L'impostazione è quella del Ross [?].

La seconda parte è dedicata alla teoria dell'informazione. Forti del linguaggio acquisito nella prima parte, le sette lezioni dalla 11 alla 17 discutono le relazioni tra quantità di informazione, entropia, compressione e codifica. Enunciamo i due teoremi di Shannon che aprono la strada ad alcuni algoritmi di compressione in assenza di rumore, ovvero per la codifica di una sorgente, e ad alcuni algoritmi di codifica per la trasmissione di informazione attraverso un canale rumoroso. In questa parte incontriamo un primo esempio di inferenza, intesa come arte di fare un'affermazione sulla distribuzione di probabilità che genera i campioni osservati. La maggior parte del materiale proviene dal Mc Kay [?], con piccoli contributi dal Reza [?] e dal Kinchin [?].

La terza parte è costituita da sei lezioni sul tema dell'inferenza. Dalla lezione 18 alla lezione 23, impariamo alcuni metodi che consentono di simulare al computer il campionamento da una distribuzione di probabilità e i concetti principali alla base dell'inferenza frequentista: test di ipotesi, stime puntuali e stime d'intervallo. Chiude il corso una finestra sulla'inferenza bayesiana che opera aggiornando le ipotesi a priori sulla base delle osservazioni acquisite. Questa terza parte attinge nuovamente al Ross [?], per il campionamento, e al libro di testo digitale Statlect [?], al Rohatgi [?] e alle note del corso 18.05, Introduction to Probability and Statistics [?], tenuto al MIT nel 2014.

Indice

1 Impariamo a contare

In questa lezione introduciamo gli elementi principali del calcolo combinatorio. Nell'ambito del corso questi elementi ci saranno utili per calcolare la probabilità associata a eventi nei casi in cui vale l'ipotesi di equiprobabilità.

Un concetto di base usato nel seguito e' l'idea di esperimento. Con quest'ultimo, si intende intuitivamente un'azione il cui risultato sia uno elemento di un insieme, questi ultimo puó essere visto come l'insieme di risultati possibili di un esperimento. Si assume generalmente che gli elementi dell'insieme siano distinti. Verra' specificato esplicitamente qualora siano presenti ripetizioni.

1.1 Principio base

Se un esperimento fornisce m possibili risultati e se per ciascuno di questi risultati un secondo esperimento fornisce n possibili risultati, allora i due esperimenti forniscono mn possibili risultati.

Esercizio 1.1.1. Se una targa è formata da 4 lettere e 3 cifre, quante sono le targhe possibili?

Assumiamo che le lettere possibili siano 26 e le cifre 10. La scelta di ogni elemento della targa può essere visto come un esperimento. Se le ripetizioni sono ammesse: 26 risultati per la prima, 26 per la seconda, 26 per la terza e 26 per la quarta lettera, 10 risultati per la prima, 10 per la seconda e 10 per la terza cifra: $26^4 \cdot 10^3 = 456,976,000$ targhe in tutto. Se le ripetizioni non sono ammesse: 26 risultati per la prima, 25 per la seconda, 24 per la terza e 23 per la quarta lettera, 10 risultati per la prima, 9 per la seconda e 8 per la terza cifra: $26 \cdot 25 \cdot 24 \cdot 23 \cdot 10 \cdot 9 \cdot 8 = 258,336,000$ targhe in tutto.

1.2 Permutazioni

Una permutazione è un particolare ordinamento di n oggetti. Applicando il principio base per contare le permutazioni possibili, otteniamo $n!$ poichè abbiamo n scelte per il primo oggetto, $n-1$ per il secondo e così via fino alla scelta obbligata dell'n-esimo oggetto, ultimo rimasto.

Esercizio 1.2.1. In quanti modi diversi è possibile ordinare su uno scaffale 2 libri di chimica, 3 di fisica, 4 di matematica e 5 di informatica in modo che i libri di una stessa materia siano in un unico gruppo di libri consecutivi?

Se applichiamo il principio base alle materie otteniamo $4! = 24$ scelte. Se applichiamo il principio base ai libri di ogni materia otteniamo $2! \cdot 3! \cdot 4! \cdot 5! = 34,560$ scelte. Pertanto avremo $24 \cdot 34,560 = 829,440$ permutazioni. In assenza di qualunque vincolo le possibili permutazioni sono invece $14! = 87,178,291,200$.

1.3 Disposizioni

Una disposizione è un particolare ordinamento di i oggetti scelti da n oggetti con $i \leq n$. Se applichiamo il principio base per contare le disposizioni possibili, otteniamo $n(n-1)\dots(n-i+1)$ perchè abbiamo n scelte per il primo, $n-1$ per il secondo e così via fino alla scelta dell'i-esimo oggetto tra gli $n-i+1$ oggetti rimasti.

Osservazione 1.3.1. Dall'identità $n(n-1)\dots(n-i+1) = n!/(n-i)!$ segue che le disposizioni possibili di i oggetti scelti tra n possono essere ottenute anche ragionando in modo diverso. Partiamo dalle permutazioni possibili di n oggetti, $n!$, e consideriamo equivalenti le permutazioni in cui gli i oggetti scelti si trovano nelle stesse posizioni. Queste permutazioni sono $(n-i)!$ coerentemente al fatto che sono tante quanti le possibili permutazioni degli $n-i$ oggetti non scelti.

Esercizio 1.3.1. Gli anagrammi di CINEMA (la maggior parte dei quali non fornisce parole di senso compiuto) sono $6! = 720$. Quanti sono, invece, gli anagrammi di ERRORE?

Dividendo per 3! (per le 3 R) e per 2! (per le 2 E) otteniamo $720/12 = 60$ anagrammi.

1.4 Combinazioni

Una combinazione è una scelta di i oggetti da n oggetti con $i \leq n$. Se applichiamo il principio base per contare le combinazioni possibili e teniamo presente che l'ordine, questa volta, è irrilevante sia per gli i oggetti scelti sia per gli $n - i$ oggetti non scelti, otteniamo

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

Esercizio 1.4.1. Quanti comitati di tre persone possiamo formare partendo da un gruppo di 20 persone?

$20!/(17! \cdot 3!) = 20 \cdot 19 \cdot 18/6 = 1140$.

2 Definizione assiomatica di probabilità

In questa lezione definiamo la probabilità in modo assiomatico discutendo brevemente alcune proprietà nel caso semplice ma importante in cui sia possibile individuare eventi equiprobabili.

2.1 Nozioni fondamentali

Spazio campionario: è l'insieme S dei possibili risultati di un esperimento (testa e croce o i tempi di vita di un componente).

Evento: è un qualunque sottoinsieme E di S che si realizza se il risultato dell'esperimento appartiene a E (testa nel lancio di una moneta o tempo di vita di un componente minore di 5 ore).

Indichiamo con $E \cup F$ l'unione e con EF l'intersezione degli eventi E e F. Due eventi E e F tali che $EF = \emptyset$ sono mutuamente esclusivi. L'evento E^c tale che $E \cup E^c = S$ è il complementare di E.L'uso dei concetti di base della teoria degli insiemi e la loro illustraziione tramite diagrammi di Venn, consentirá di trattare in modo preciso e intuitivo diverse proprietá delle probabilitá.

2.2 Assiomi

Una probabilità $P(\cdot)$ risulta ben definita sugli eventi di uno spazio campionario S se

$A1$: $0 \leq P(E) \leq 1 \ \forall E \subseteq S$

$A2$: $P(S) = 1$

$A3$: Se gli eventi E_i, con $i = 1, 2, \ldots$ sono mutuamente esclusivi, allora $P(\cup E_i) = \sum_i P(E_i)$.

Sono conseguenze immediate di $A1$, $A2$ e $A3$

(i) $\forall E$, $P(E^c) = 1 - P(E)$

(ii) $\forall E$ e F, se $E \subseteq F$ allora $P(E) \leq P(F)$

(iii) $\forall E$ e F, $P(E \cup F) = P(E) + P(F) - P(EF)$

La dimostrazione di queste proprietà costituisce un utile esercizio per impratichirsi con gli assiomi. La prima proprietà segue dal fatto che $E^c E = \emptyset$ e $E^c \cup E = S$, hence $P(E) + P(E^c) = 1$. In modo analogo, la seconda proprieta' si dimostra notando che $F = E^c F \cup EF$ and $F = E^c F \cup EF$. La terza ed ultima proprieta' e' chiara utilizzando i diagrammi di Venn. La dimostrazione si basa sulla scrittura di $E \cup F$ e F come unione di insiemi disgiunti e l'applicazione degli assiomi. Tale dimostrazione e' lasciato come esercizio al lettore.

Esercizio 2.2.1. Quanti sono i componenti di un complesso in cui 3 cantano, 3 suonano la chitarra e 2 entrambe le cose?

Abbiamo 3+3-2=4 componenti.

2.3 Eventi equiprobabili

Supponiamo che S sia costituito da un insieme finito di N risultati che indichiamo con i primi N numeri naturali, ovvero $S = \{1, 2, \dots, N\}$. Se le probabilità $P(\{i\})$ sono tutte uguali allora $P(\{i\}) = 1/N$. La probabilità di E, in questo caso, si calcola come frazione del numero di risultati in E, $\#E$, sul numero di risultati in S, $\#S$.

Esercizio 2.3.1. Calcoliamo la probabilità P di ottenere 7 lanciando 2 dadi.

Le coppie di risultati ottenibili dal lancio di due dadi sono 36. I casi favorevoli sono 6: (1,6), (2,5), (3,4), (4,3), (5,2) e (6,1), per cui $P = 6/36 = 1/6$.

Osservazione 2.3.1. Le coppie sono ordinate, il primo elemento è il risultato del lancio del primo dado, il secondo è il risultato del lancio del secondo dado.

Esercizio 2.3.2. Calcoliamo la probabilità P di estrarre 1 pallina bianca e 2 palline nere da un'urna con 6 palline bianche e 5 nere.

Soluzione con le combinazioni: se consideriamo l'insieme delle palline estratte come non ordinato, i casi possibili sono le combinazioni di 3 palline scelte tra 11, i favorevoli quelle di 1 pallina bianca scelta tra 6 e 2 nere scelte tra 5, o

$$\#S = \binom{11}{3} = 165 \quad \text{e} \quad \#E = \binom{6}{1}\binom{5}{2} = 60$$

da cui segue $P = 60/165 = 4/11$. Soluzione con le disposizioni: consideriamo ora invece come rilevante l'ordine col quale estraiamo le palline. I casi possibili sono le disposizioni di 3 palline scelte tra 11, ovvero $11 \cdot 10 \cdot 9 = 990$. Dividiamo i casi favorevoli in 3 gruppi. Nel primo gruppo la pallina bianca è estratta per prima, $6 \cdot 5 \cdot 4$ casi favorevoli, nel secondo per seconda, $5 \cdot 6 \cdot 4$ casi favorevoli, e nel terzo per terza, $5 \cdot 4 \cdot 6$ casi favorevoli. Nuovamente otteniamo $P = 3 \cdot 120/990 = 4/11$.

Esercizio 2.3.3 (Paradosso del compleanno). Calcoliamo la probabilità P che n individui festeggino il compleanno in n giorni diversi.

Consideriamo l'evento complementare ovvero che nessuno tra n individui sia nato lo stesso giorno. Fissiamo un primo individuo. I casi favorevoli sono 365. Perché un secondo individuo non sia nato nel giorno del primo i casi favorevoli sono (365-1)= 364, per un terzo individuo 365-2=363, ... e perche l'n-esimo non sia nato nei giorni dei primi $(n-1)$ e' $365 - n + 1$. I casi favorevoli sono 365 per ogni individuo. Quindi applicando il principio base nel caso di eventi equiprobabili la probabilitćhe nessuno degli n individui sia nato nello stesso giorno è

$$P = (365-1)(365-2)\dots(365-n+1)/365^{n-1}.$$

Il numeratore cresce molto piú lentamente del denominatore, quindi la probabilitá piccola molto rapidamente al crescere di n. Di conseguenza la probabilitá che almeno due persone siano nati nello stesso giorno tende a uno.

Osservazione 2.3.2. Per $n = 100$ $P < 0.000001$ (esperimento in classe). Al diminuire di n la probabilità aumenta. L'ultimo valore per cui $P < 1/2$ è $n = 23$. Una spiegazione intuitiva di questo risultato apparentemente paradossale è che le coppie possibili di 23 individui sono $23 \cdot 22/2 = 253$ (ben più della metà dei 365 giorni in un anno). Più precisamente un modo per capire questo esempio e' pensare alla probabilità che tra n persone non ce ne sia nessuna con il compleanno lo stesso giorno. La probabilità che 2 persone non abbiano il compleanno lo stesso giorno è $1 - 1/365 = 364/365$, mentre i modi di prendere 2 persone tra n e' $N = \binom{n}{2}$. La probabilità che nessuno tra n persone abbia il compleanno nello stesso giorno è $(364/365)^N$.

2.4 Probabilità soggettiva

Capita di associare il concetto di probabilità a eventi incerti non ripetibili (probabilità di pioggia a Milano o di vittoria in un incontro di scherma). In questi casi parliamo di probabilità soggettiva. Se gli assiomi sono verificati, nulla cambia.

Esercizio 2.4.1. La probabilità dell'evento oggi pioverà è del 40% e che domani pioverà è del 30%. Se la probabilità che oggi o domani pioverà è del 60% mostriamo che non è possibile che la probabilità che oggi e domani pioverà sia del 20%.

Se E è l'evento oggi pioverà e F domani pioverà sappiamo che deve valere la relazione $P(E \cup F) = P(E) + P(F) - P(EF)$. Nel nostro caso, invece, abbiamo che $0.6 \neq (0.4 + 0.3 - 0.2)$.

3 Probabilità condizionata e indipendenza

Il tema di questa lezione, forse il concetto più importante della teoria della probabilità, è la probabilità condizionata, ovvero la probabilità di un evento una volta che si venga a conoscenza della realizzazione di un altro evento.

3.1 Probabilità condizionata

Dati due eventi E e F, siamo interessati a calcolare la probabilità di E quando sappiamo che si è realizzato F. La probabilità di E condizionata a F, indicata con $P(E|F)$, è definita come

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Aggiungiamo due osservazioni di carattere tecnico.

Osservazione 3.1.1. Si noti che, se si realizza anche E dopo che si e' realizzato F significa che il risultato appartiene all'intersezione EF. Inoltre, lo spazio campionario si è ridotto da S a F e misura $P(F)$.

Osservazione 3.1.2. $P(\cdot|F)$ è una probabilità. Che $P(EF) \leq P(F)$ (assioma $A1$) discende dal fatto che $EF \subseteq F$. Che $P(S|F) = 1$ (assioma $A2$) discende dal fatto che $P(SF) = P(F)$. La verifica dell'assioma $A3$, anche se non difficile, è piuttosto laboriosa ed è assegnata per esercizio.

Osservazione 3.1.3. Regola della moltiplicazione: Generalizziamo l'identità $P(EF) = P(F)P(E|F)$ al caso dell'intersezione di n eventi. Abbiamo che

$$P(E_1E_2 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2) \dots P(E_n|E_1E_2 \dots E_{n-1}).$$

Per dimostrare la validità di questa uguaglianza è sufficiente utilizzare la definizione di probabilità condizionata in modo iterativo su ognuno degli $n-1$ fattori. Ad esempio per tre insiemi E, F, G, si ha

$$P(EFG) = P(EF|G)P(G) = P(E|FG)P(F|G)P(G)$$

e dato che

$$P(EF|G) = P(E|FG)P(F|G)$$

concludiamo che

$$P(EFG) = P(E|FG)P(F|G)P(G).$$

Esercizio 3.1.1. Lanciamo una volta un dado. Qual'e' la probabilità di fare 1 se sappiamo che il risultato del lancio e' tra 1 e 3? Qual'e' la probabilità di fare 1 se invece sappiamo che il risultato del lancio e' tra 4 e 6?

L'evento E è il numero 1 mentre l'evento F è $\{1, 2, 3\}$. Si ha $P(F) = 3/6$ e $P(EF) = P(E) + P(F) - P(E \cup F) = 1/6 + 3/6 - 3/6$ e quindi $P(E|F) = 1/3$. Ragionando in modo analogo nell'altro caso si ottiene $P(E|F) = 0$.

Esercizio 3.1.2. Lanciamo due volte una moneta onesta e calcoliamo la probabilità di ottenere 2 volte testa, T, se (i) il risultato del primo lancio è T, o se (ii) il risultato di uno dei due lanci è T.

Nel caso (i) gli eventi sono $E = \{T, T\}$ e $F = \{\{T, C\}, \{T, T\}\}$, e poichè

$$P(E) = P(\{T, T\}) \quad \text{e} \quad P(F) = P(\{\{T, C\}, \{T, T\}\})$$

allora $P(E|F) = P(EF)/P(F) = (1/4)/(2/4) = 1/2$. Nel caso (ii) gli eventi sono $E = \{T, T\}$ e $F = \{\{T, C\}, \{T, T\}\{C, T\}\}$, e poichè

$$P(E) = P(\{T, T\}) \quad \text{e} \quad P(F) = P(\{\{T, C\}, \{T, T\}, \{C, T\}\})$$

allora $P(E|F) = P(EF)/P(F) = (1/4)/(3/4) = 1/3$.

Osservazione 3.1.4. La differenza tra (i) e (ii) nell'Esempio 3.2 è da imputarsi al diverso numero di risultati possibili nei due casi!

3.2 Formula di Bayes

Il seguente risultato e' di fondamentale importanza in un numero di applicazioni della teoria delle probabilitá.

Dati due eventi E e F, applicando due volte la definizione di probabilitá condizionata, si ha.

$$P(EF) = P(E|F)P(F) = P(F|E)P(E).$$

La formula di Bayes e'

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}.$$

Puó essere interpretata come la probabilitá di un'effetto, evento E, a partire dalla conoscenza di possibili cause, evento F.

3.3 Teorema della probabilitá assoluta

Una formulazione alternativa è ottenuta riscrivendo E come l'unione dei due eventi mutuamente esclusivi, EF e EF^c. In questo modo, applicando la definizione di probabilità condizionata, si ha

$$P(E) = P(EF) + P(EF^c) = P(F)P(E|F) + P(F^c)P(E|F^c).$$

Ovvero, la probabilità di E può essere scritta come la media pesata della probabilità di E condizionata alla realizzazione di F e della probabilità di E condizionata alla realizzazione del complementare di F, dove i pesi sono $P(F)$ e $P(F^c) = 1 - P(F)$.

3.4 Teorema di Bayes

Se sostituiamo l'espressione appena ottenuta al posto di $P(E)$ otteniamo una versione alternativa della formula di Bayes

$$P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}.$$

Esercizio 3.4.1 (Paradosso delle tre carte). Viene detto paradosso delle tre carte un classico problema del calcolo delle probabilitá che pur nella sua semplicitá ha una soluzione abbastanza controintuitiva: ci sono tre carte, delle quali la prima (A) è rossa su entrambi i lati, la seconda (B) su un lato è rossa e sull'altro è bianca e la terza (C) è bianca su entrambi i lati. Ponendo su un tavolo una delle tre carte, scelta a caso, ottengo che il lato visibile è di colore rosso. Qual è la probabilitá che anche il lato non visibile sia di colore rosso?

Una prima soluzione e' basata sul calcolo dei casi favorevoli e casi possibili. Estraendo una carta e posandola sul tavolo si possono verificare i seguenti sei casi equiprobabili, che possono capitare in maniera egualmente frequente:
lato visibile = Aa = rosso, lato nascosto = Ab = rosso
lato visibile = Ab = rosso, lato nascosto = Aa = rosso
lato visibile = Ba = rosso, lato nascosto = Bb = bianco
lato visibile = Bb = bianco, lato nascosto = Ba = rosso
lato visibile = Ca = bianco, lato nascosto = Cb = bianco
lato visibile = Cb = bianco, lato nascosto = Ca = bianco
escludendo gli ultimi tre casi in quanto il lato visibile è bianco, rimangono tre casi dove il lato visibile è rosso, due dei quali nascondono un lato anch'esso rosso, dunque la probabilitá è di 2/3.

Lo stesso risultato si trova applicando la formula di Bayes. La probabilitá condizionata cercata è

$$P(\text{lato invisibile rosso}|\text{lato scoperto rosso}) = P(\text{carta con 2 lati rossi}|\text{lato scoperto rosso})$$

che sinteticamente possiamo scrivere $P(A|R)$ dove A è la carta che ha entrambi i lati rossi e $P(A)$ è la probabilitá che essa venga scelta, $P(R)$ è invece la probabilitá che il lato visibile sia rosso.

Utilizzando il teorema di Bayes:

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

essendo $P(R|A) = 1$, ovvero il lato scoperto della carta A è sicuramente rosso, e $P(A) = 1/3$, la probabilitá di scegliere la carta A è $1/3$. Il lato scoperto rosso puó derivare dalla carta A o dalla B, ma mentre per la A la probabilitá è 1, per la B è $1/2$:

$$P(R) = P(A)P(R|A) + P(B)P(R|B) = 1/3 * 1 + 1/3 * 1/2 = 1/2$$

allora

$$P(A|R) = P(R|A)P(A)/P(R) = 1 * 1/3/1/2 = 2/3$$

Esercizio 3.4.2 (Monty Hall). Dietro una delle tre porte a, b e c c'è una macchina, dietro ognuna delle altre due una capra. Scopo del gioco è indovinare la porta dietro la quale si trova la macchina. Scegliamo la porta a e Monty, che vede che cosa si nasconde dietro le tre porte, scopre la porta c mostrando una capra. È meglio mantenere la scelta e chiedere di aprire la porta a o cambiare scelta e chiedere di aprire la porta b?

Sia R_c l'evento Monty sceglie di aprire la porta c. Indichiamo con X l'evento dietro la porta x c'è una macchina. Valutiamo prima di tutto $P(R_c|A)$, $P(R_c|B)$ e $P(R_c|C)$. Se la macchina è dietro la porta a, Monty può aprire le porte b e c con uguale probabilità, per cui $P(R_c|A) = 1/2$. Se la macchina è dietro la porta b, Monty può aprire solo la porta c, per cui $P(R_c|B) = 1$. Se la macchina è dietro la porta c, Monty non può aprire la porta c, per cui $P(R_c|C) = 0$. Pertanto

$$P(A|R_c) = \frac{P(A)P(R_c|A)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3}\frac{1}{2}}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{1}{3}$$

$$P(B|R_c) = \frac{P(B)P(R_c|B)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3}1}{\frac{1}{3}\frac{1}{2} + \frac{1}{3}1 + \frac{1}{3}0} = \frac{2}{3}.$$

Se questo risultato ci lascia perplessi proviamo a pensare di ripetere il gioco con 1000 porte, una macchina e 999 capre e con Monty che, dopo la nostra scelta, apre 998 porte mostrandoci altrettante capre.

3.5 Eventi indipendenti

Due eventi sono indipendenti se la realizzazione di uno non modifica la probabilità dell'altro. Se E e F sono indipendenti allora

$$P(EF) = P(E)P(F),$$

ovvero $P(E|F) = P(E)$ e $P(F|E) = P(F)$.

Esercizio 3.5.1. Lanciamo un dado onesto due volte. Verifichiamo che l'evento E_1 la somma dei due risultati è 7 e l'evento F il primo risultato è 4 sono indipendenti, mentre l'evento E_2 la somma dei due risultati è 6 e F non lo sono.

Chiaramente abbiamo che $P(F) = 1/6$ e $P(E_1F) = P(E_2F) = 1/36$. Per quanto riguarda E_1 abbiamo 6 casi favorevoli su 36 possibili, per cui $P(E_1) = 1/6$. Per E_2, invece, i casi favorevoli sono 5 e quindi $P(E_2) = 5/36$.

Osservazione 3.5.1. Il motivo per cui le cose cambiano è dovuto al fatto che mentre la probabilità di ottenere 7 con due lanci non dipende dal primo risultato, la probabilità di ottenere 6 richiede che il primo risultato non sia 6.

4 Variabili casuali discrete

Con l'introduzione delle variabili casuali nel caso discreto e delle nozioni essenziali di valore atteso e varianza entriamo nel vivo della nostra breve incursione nella teoria della probabilità.

4.1 Variabili casuali

Molto spesso le quantità di interesse in un esperimento non sono i risultati ma una qualche funzione del risultato nota come variabile casuale. Una variabile casuale è, in effetti, una funzione a valori reali definita su un opportuno spazio campionario (il termine variabile per indicare una funzione non è dei più felici). Nel caso di valori discreti, come la somma di due dadi, il numero di teste in n lanci o il genere alla nascita, una variabile casuale X è completamente definita in termini della probabilità con la quale assume ognuno dei suoi possibili valori.

Esercizio 4.1.1. Lanciamo 3 volte una moneta. Il numero X di teste ottenute è una variabile casuale i cui possibili valori sono 0, 1, 2 e 3. Valutiamo la probabilità con la quale X assume i valori 0, 1, 2 e 3 nell'assunzione che la moneta sia onesta.

Poichè tutte le 8 triple sono ugualmente probabili abbiamo

$$P\{X = 0\} = P\{(CCC)\} = 1/8, \quad P\{X = 1\} = P\{(TCC), (CTC), (CCT)\} = 3/8,$$
$$P\{X = 2\} = P\{(CTT), (TCT), (TTC)\} = 3/8 \ \text{ e } \ P\{X = 3\} = P\{(TTT)\} = 1/8$$

con $\sum_i P\{X = i\} = 1$.

Osservazione 4.1.1. Per semplicità, al posto di $P\{X = i\}$ spesso si scrive $p(i)$.

Esercizio 4.1.2. Estraiamo casualmente 3 palline senza reinserimento tra 20 palline numerate da 1 a 20. Il numero estratto più grande X è una variabile casuale i cui possibili valori sono $3, 4, \ldots, 20$. I risultati possibili dell'esperimento, ovvero lo spazio campionario, sono le combinazioni di 3 numeri scelti tra $1, 2, \ldots, 20$. Valutiamo la probabilità con la quale X assume il valore i (con $3 \leq i \leq 20$) nell'assunzione di equiprobabilità.

Se i è il numero estratto più grande, le altre due palline sono numerate con una delle possibili coppie di numeri diversi compresi tra 1 e $i - 1$. Pertanto avremo

$$P\{X = i\} = \frac{\binom{i-1}{2}}{\binom{20}{3}}, \quad i = 3, 4, \ldots, 20.$$

Anche in questo caso abbiamo esaurito i casi possibili per cui $\sum_i P\{X = i\} = 1$.

4.2 Funzione di probabilità di massa

Nel caso di una variabile casuale X a valori discreti la funzione di probabilità di massa $p(\cdot)$ definita sulla retta reale, o pmf o anche solo funzione di probabilità, contiene tutta l'informazione necessaria per descrivere completamente X che può assumere uno tra i valori x_i con $i = 1, 2, \ldots$. Si ha che $p(x_i) = P\{X = x_i\} \geq 0$ (e $p(x) = 0$ altrimenti) con $\sum_i p(x_i) = 1$.

Osservazione 4.2.1. Il vincolo $\sum_i p(x_i) = 1$ deve essere sempre verificato!

4.3 Funzione di probabilità cumulata

Ordiniamo i valori x_i in modo tale che $x_1 < x_2 < \ldots < x_i < \ldots$ e introduciamo la funzione di probabilità cumulata $F(a)$, o cdf, definita come

$$F(a) = \sum_{x_i < a} p(x_i).$$

Osservazione 4.3.1. È facile verificare che la funzione F è continua da destra e crescente da 0 a 1. Una cdf di una pmf è una funzione a gradini. L'i-esimo gradino è localizzato nel punto x_i e il salto corrispondente vale $p(x_i)$.

4.4 Valore atteso

Introduciamo una delle nozioni centrali dell'intera teoria, il valore atteso di una variabile casuale. Dovremo attendere l'ultima lezione, e in particolare la legge dei grandi numeri, per apprezzarne appieno l'importanza.

Il valore atteso μ di una variabile casuale X, indicato con $E[X]$ e che non deve essere confuso con la media empirica, è la media pesata dei valori x_i che può assumere X. Ogni x_i è pesato con la sua probabilità $p(x_i)$ e quindi si ha

$$\mu = E[X] = \sum_i x_i p(x_i).$$

Esercizio 4.4.1. Valutiamo il valore atteso della variabile casuale dell'Esercizio 4.1.

Non sorprendentemente otteniamo $\mu = 0 \cdot 1/8 + 1 \cdot 3/8 + 2 \cdot 3/8 + 3 \cdot 1/8 = 1.5$.

4.5 Valore atteso di una funzione di variabile casuale

Per calcolare il valore atteso di una funzione g di una variabile casuale discreta X possiamo determinare la pmf della variabile casuale discreta $g(X)$, oppure calcolare il valore atteso come media pesata.

Esercizio 4.5.1. Calcoliamo $E[X^2]$ per una variabile casuale X con

$$P\{X = -1\} = 0.2, \quad P\{X = 0\} = 0.5 \quad e \quad P\{X = 1\} = 0.3$$

Abbiamo $P\{Y = 1\} = P\{X = -1\} + P\{X = 1\} = 0.5$ e $P\{Y = 0\} = 0.5$ per cui otteniamo $E[X^2] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5$. Calcoliamo ora il valore atteso come $E[g(X)] = \sum_i g(x_i)p(x_i)$. In questo caso scriviamo $E[X^2] = 1 \cdot 0.2 + 1 \cdot 0.3 = 0.5$.

Esercizio 4.5.2. Valutiamo $E[aX + b]$ con a e $b \in \mathbb{R}$ in funzione di $E[X]$.

Coerentemente con la struttura lineare del valore atteso, abbiamo

$$E[aX + b] = \sum_{x:p(x)>0} (ax + b)p(x) = a \sum_{x:p(x)>0} xp(x) + b = aE[X] + b.$$

4.6 Varianza

Una seconda quantità che cattura proprietà importanti di una variabile casuale X è la varianza $Var(X)$ definita come $Var(X) = E[(X-\mu)^2]$.

Osservazione 4.6.1. Il quadrato nella definizione di varianza è fondamentale per ovviare al fatto che le differenze dal valore atteso hanno valore atteso nullo per qualunque X. Infatti si ha che

$$E[(X-\mu)] = E[X] - E[\mu] = \mu - \mu = 0.$$

Esercizio 4.6.1. Dimostriamo che $E[(X-\mu)^2] = E[X^2] - (E[X])^2$.

$$E[(X-\mu)^2] = E[X^2 + \mu^2 - 2\mu X] = E[X^2] + (E[X])^2 - 2(E[X])^2 = E[X^2] - (E[X])^2.$$

Esercizio 4.6.2. Dimostriamo che $Var(aX+b) = a^2 Var(X)$.

$$Var(aX+b) = E[(aX + b - aE[X] - b)^2] = E[(aX - aE[X])^2] = a^2 Var(X).$$

Osservazione 4.6.2. Una quantità molto usata è la radice quadrata della varianza, nota come deviazione standard o $SD(X) = \sqrt{Var(X)}$.

5 Esemipi di variabili aleatorie discrete

In questa lezione prendiamo in considerazione importanti distribuzioni di probabilità nel caso discreto.

5.1 Bernoulli

Una variabile casuale di Bernoulli X assume due soli valori, 0 e 1 (talvolta associati al fallimento e al successo di un esperimento), con funzione di probabilità di massa $P\{X=0\} = 1-p$ e $P\{X=1\} = p$ con $0 < p < 1$.

Esercizio 5.1.1. Calcoliamo il valore atteso e la varianza di una variabile casuale di Bernoulli.

Da una diretta applicazione delle definizioni di valore atteso e varianza si ha,

$$E[X] = p \cdot 1 + (1-p) \cdot 0 = p$$

e

$$Var(X) = p(1-p)^2 + (1-p)p^2 = p(1-p).$$

5.2 Binomiale

Per introdurre il prossimo esempio di distribuzione e' utile ricordare la seguente equazione

$$\sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k = (a+b)^n$$

detta teorema binomiale (o anche formula di Newton, binomio di Newton e sviluppo binomiale).

La variabile casuale binomiale X conta i successi in una sequenza di n realizzazioni indipendenti di una variabile casuale di Bernoulli. La sua funzione di probabilità di massa si scrive come

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad i = 0, 1, \ldots, n.$$

Esercizio 5.2.1. Verifichiamo che $\sum_i p(i) = 1$.

Scrivendo la somma delle probabilitá in modo esplicito si ha

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

tramite un'applicazione diretta del binomio di Newton ricordato sopra.

Esercizio 5.2.2. Dimostriamo che il valore atteso di una variabile casuale binomiale è np.

L'idea é di ricondursi al binomio di Newton. Iniziamo notando che il primo termine della somma é zero e possiamo semplificare il termine "i" al numeratore,

$$\sum_{i=0}^{n} i\binom{n}{i}p^i(1-p)^{n-i} \quad = \quad \sum_{i=0}^{n} i\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i} = \sum_{i=1}^{n}\frac{n!}{(i-1)!(n-i)!}p^i(1-p)^{n-i}$$

Successivamente raccogliamo un fattore n e p dalla seguente espressione ottenendo

$$np\sum_{i=1}^{n}\frac{n-1!}{(i-1)!(n-i)!}p^{i-1}(1-p)^{n-i}$$

Infine rinominiamo gli indici della sommario considerando $j = i-1$ in modo tale che gli estremi della somma sono 0 e $n-1$ e si ottiene

$$np\sum_{j=0}^{n-1}\frac{n-1!}{j!(n-1-j)!}p^{j}(1-p)^{n-1-j}.$$

L'esercizio é concluso notando che l'espressione di sopra e' uguale a

$$np\sum_{j=0}^{n-1} i\binom{n-1}{j}p^{j}(1-p)^{n-1-j} = np(p+(i-p))^{n-1} = np.$$

Osservazione 5.2.1. Per quanto riguarda $Var(X)$ abbiamo $Var(X) = np(1-p)$. Nell'Esercizio 10.4.1 otterremo questo risultato in modo semplice grazie alla funzione generatrice dei momenti.

5.3 Geometrica

Per introdurre il prossimo esempio di distribuzione e' utile ricordare la seguente equazione

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \qquad 0 < x < 1,$$

detta serie geometrica.

La variabile casuale geometrica X vale n se si ottiene un successo dopo $n-1$ fallimenti in una sequenza di n realizzazioni indipendenti di una variabile casuale di Bernoulli. La pmf è $P(X=n) = (1-p)^{n-1}p$ per $n = 1, 2, \dots$. È immediato verificare che $\sum_{i=1,\dots\infty}(1-p)^{i-1}p = p/(1-(1-p)) = 1$.

Esercizio 5.3.1. Calcoliamo il valore atteso di una variabile casuale binomiale X.

Per ricondursi alla serie geometrica decomponiamo la somma in due addendi

$$E[X] \quad = \quad \sum_{i=1}^{\infty}(i-1+1)(1-p)^{i-1}p = \sum_{i=1}^{\infty}(i-1)(1-p)^{i-1}p + \sum_{i=1}^{\infty}(1-p)^{i-1}p \tag{1}$$

e rinominando gli indizi $j = i-1$ si ha

$$E[X] \quad = \quad \sum_{j=0}^{\infty} j(1-p)^j p + 1 = (1-p)E[X] + 1.$$

Da cui abbiamo che $E[X] = 1/p$.

Osservazione 5.3.1. In modo simile si ottiene $Var(X) = (1-p)/p^2$.

5.4 Poisson

Per introdurre il prossimo esempio di distribuzione e' utile ricordare la seguente equazione

$$\exp(x) \equiv e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

detta serie esponenziale.

Una variabile casuale di Poisson è definita dalla pmf $P\{X = i\} = e^{-\lambda}\lambda^i/i!$ con $i = 0, 1, 2, \ldots$. Ovviamente

$$\sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1.$$

Il numero di errori di stampa per pagina, il numero di ultracentenari di una comunità, il numero di numeri di telefono sbagliati da un centralino, il numero di pacchi di biscotti venduti in un giorno, il numero di clienti in un ufficio postale sono esempi di variabili casuali di Poisson.

Osservazione 5.4.1. Se n grande, p piccolo e $\lambda = np \sim 1$, la funzione di probabilità di massa della binomiale tende a Poisson! Infatti possiamo scrivere

$$\begin{aligned} P\{X = i\} &= \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} = \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\ldots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^i}. \end{aligned}$$

Ma poichè n è grande, abbiamo che $n(n-1)\ldots(n-i+1)/n^i \sim 1$, $(1-\lambda/n)^n \sim e^{-\lambda}$ e $(1-\lambda/n)^i \sim 1$.

Esercizio 5.4.1. Calcoliamo il valore atteso di una variabile casuale di Poisson.

$$E[X] = \sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = \lambda\sum_{i=i}^{\infty} ie^{-\lambda}\frac{\lambda^{i-1}}{(i-1)!} = \lambda\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda.$$

Osservazione 5.4.2. La varianza di una variabile casuale di Poisson è ancora λ.

Osservazione 5.4.3. Se N eventi ciascuno di probabilità p_i piccola sono indipendenti, o solo debolmente dipendenti, il numero di eventi che si realizzano è approssimato da una variabile casuale di Poisson con $\lambda = \sum_{i=1,\ldots,N} p_i$.

6 Variabili casuali continue

In questa lezione estendiamo il nostro studio al caso di variabili casuali che assumono valori nel continuo.

6.1 Funzione densità di probabilità

L'insieme dei valori che può assumere una variabile casuale spesso non è finito o numerabile (pensiamo al tempo di vita di un componente, all'ora d'arrivo di un treno o al tempo di percorrenza di un viaggio in auto). Una variabile casuale X è continua se esiste una funzione $f : \mathbb{R} \to \mathbb{R}^+$ tale che

$$P(X \in B) = \int_B f(x)\mathrm{d}x$$

su ogni sottoinsieme misurabile $B \subset \mathbb{R}$ (la misurabilità è una condizione tecnica che, per i nostri scopi, non ha conseguenze rilevanti in quanto tutti i sottoinsiemi di nostro interesse sono misurabili). La funzione f è la densità di probabilità, o pdf.

Osservazione 6.1.1. La probabilità che una variabile casuale continua X assuma un determinato valore x è 0. Se $B = (x - \epsilon/2, x + \epsilon/2)$, possiamo approssimare l'integrale con un rettangolo di base ϵ e altezza $f(x)$ e scrivere $P\{x - \epsilon/2 < X < x + \epsilon/2\} \sim f(x)\epsilon$.

Esercizio 6.1.1. Sia $f(x) = C(4x - 2x^2)$ per $0 < x < 2$ e 0 altrimenti. Calcoliamo C e $P\{X > 1\}$.

La costante C si ottiene imponendo la condizione di normalizzazione $P(X \in S) = 1$ che, in questo caso, diventa

$$\int_0^2 (4x - 2x^2)\mathrm{d}x = \frac{1}{C}$$

da cui ricaviamo $C = 3/8$. Calcolando l'integrale otteniamo $P\{X > 1\}$=1/2.

Di seguito vediamo come le definizioni introdotte per le variabili aleatorie discrete si estendono al caso continuo, considerando la pdf al posto delle pmf e integrali invece che sommatorie.

6.2 Funzione di distribuzione cumulata

La funzione di distribuzione cumulata $F : \mathbb{R} \to [0, 1]$, o cdf, è definita $\forall a \in \mathbb{R}$ come

$$F(a) = \int_{-\infty}^{a} f(x)\mathrm{d}x.$$

Come nel caso discreto la cdf è una funzione monotona non negativa. In piú utilizzando il teorema fondamentale del calcolo integrale si vede che la pdf e la cdf forniscono due caratterizzazioni equivalenti delle variabili aleatorie continue.

Osservazione 6.2.1. Ricordiamo che la prima parte del teorema del calcolo integrale afferma proprio che se $f : (a, b) \to \mathbb{R}$ e

$$F(x) = \int \mathrm{d}x f(x)$$

allora

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = f(x).$$

Osservazione 6.2.2. La descrizione di una variabile continua in termini di funzione densità di probabilità o funzione di distribuzione cumulata è equivalente.

Osservazione 6.2.3. La nostra trattazione non considera il caso in cui la cdf può presentare discontinuità (come nel caso di variabile casuale discreta).

6.3 Funzione di variabile casuale continua

Vediamo come determinare la densità di probabilità e la distribuzione cumulata di una funzione Y di una variabile casuale X.

Esercizio 6.3.1. Se $f_X(x)$ è la funzione di densità di X e $F_X(x)$ la corrispondente cdf, determiniamo f_Y e F_Y per $Y = 2X$.

Abbiamo

$$F_Y(x) = P\{Y < x\} = P\{2X < x\} = P\left\{X < \frac{x}{2}\right\} = F_X\left(\frac{x}{2}\right).$$

Derivando, si ottiene

$$f_Y(x) = \frac{\mathrm{d}F_Y(x)}{\mathrm{d}x} = \frac{\mathrm{d}F_X(x/2)}{\mathrm{d}x} = \frac{f_X(x/2)}{2}.$$

Si può ottenere lo stesso risultato anche a partire dalla valutazione della probabilità in un intorno di x di semi-ampiezza pari $\epsilon/2$

$$f_Y(x)\epsilon \sim P\{x - \epsilon/2 < Y < x + \epsilon/2\} = P\{x/2 - \epsilon/4 < X < x/2 + \epsilon/4\} \sim \frac{f_X(x/2)}{2}\epsilon.$$

6.4 Distribuzione di una funzione di variabile casuale

Data una variabile casuale X di distribuzione nota vogliamo trovare la distribuzione di $g(X)$ per una funzione g data.

Esempio 6.4.1. Sia X distribuita uniformemente tra 0 e 1. Abbiamo quindi $f_X(x) = 1$ e $F_X(x) = x$ tra 0 e 1. Se $Y = X^n$ allora $F_Y(y) = P\{Y \leq y\} = P\{X^n \leq y\} = P\{X \leq y^{1/n}\} = F_X(y^{1/n}) = y^{1/n}$. Pertanto, per $0 \leq y \leq 1$

$$f_Y(y) = \frac{y^{(1-n)/n}}{n}.$$

Esempio 6.4.2. Data $f_X(x)$, troviamo f_Y per $Y = X^2$. Per $y \geq 0$,

$$F_Y(y) = P\{Y \leq y\} = P\{X^2 \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

da cui

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

per $y \geq 0$.

Osservazione 6.4.1. In generale, se g è strettamente monotona e derivabile allora

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \right|$$

se $y = g(x)$ per un qualche x, 0 altrimenti.

6.5 Valore atteso e varianza

In piena analogia con il caso discrete definiamo il valore atteso μ e la varianza $Var(X)$ di una variabile casuale X continua come

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x \quad \text{e} \quad Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) \mathrm{d}x.$$

Esercizio 6.5.1. Se $f(x) = 2x$ per $0 \leq x \leq 1$ e 0 altrimenti, allora $E[X] = 2/3$.

Innanzitutto, poichè $\int_0^1 2x\mathrm{d}x = 1$, $f(x)$ è una densità di probabilità ben definita. Per il valore atteso abbiamo

$$E[X] = \int_0^1 2x^2 \mathrm{d}x = 2/3.$$

Esercizio 6.5.2. Sia $f(x) = 1$ per $0 \leq x \leq 1$ e 0 altrimenti. Calcoliamo $E[e^X]$.

Per $0 \leq x \leq 1$ abbiamo che $1 \leq e^x \leq e$. Ora, per $1 \leq a \leq e$,

$$F_Y(a) = P\{Y < a\} = P\{e^X < a\} = P\{X < \ln a\} = \int_0^{\ln a} f(x)\mathrm{d}x = \int_0^{\ln a} \mathrm{d}x = \ln a.$$

Segue che

$$f_Y(a) = \frac{\mathrm{d}F_Y(a)}{\mathrm{d}a} = \frac{1}{a}$$

per $1 \leq a \leq e$ e 0 altrimenti. Pertanto,

$$E[e^X] = E[Y] = \int_1^e x\left(\frac{1}{x}\right)\mathrm{d}x = e - 1.$$

Otteniamo la stessa soluzione con una strategia alternativa. In analogia col caso discreto possiamo scrivere $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x$. Quindi,

$$E[e^X] = \int_{-\infty}^{+\infty} e^x f(x)\mathrm{d}x = \int_0^1 e^x \mathrm{d}x = e - 1$$

.

6.6 Distribuzione uniforme

Una variabile casuale è distribuita uniformemente in un intervallo I se ha densità di probabilità costante su I pari a $1/I$.

Esercizio 6.6.1. Calcoliamo $P\{X > 6\}$ e $P\{4 < X < 9\}$ per una distribuzione uniforme tra 0 e 10.

$$P\{X > 6\} = \frac{1}{10}\int_6^{10} \mathrm{d}x = \frac{2}{5} \quad \text{e} \quad P\{4 < X < 9\} = \frac{1}{10}\int_4^9 \mathrm{d}x = \frac{1}{2}.$$

7 Distribuzioni importanti

Anche nel caso continuo vale la pena soffermarsi su alcune funzioni di distribuzioni importanti.

7.1 Digressione sull'integrazione per parti

Siano f e g due funzioni continue e derivabili. La derivata del prodotto delle due funzioni é pari a:

$$\frac{\mathrm{d}}{\mathrm{d}x}[f(x)g(x)] = \frac{\mathrm{d}f(x)}{\mathrm{d}x}g(x) + f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x} = f'(x)g(x) + f(x)g'(x)$$

Considerando l'integrale di entrambi i membri dell'equazione si ottiene:

$$\int \frac{\mathrm{d}}{\mathrm{d}x}[f(x)g(x)]\mathrm{d}x = \int [f'(x)g(x) + f(x)g'(x)]\mathrm{d}x = \int [f'(x)g(x)]\mathrm{d}x + \int [f(x)g'(x)]\mathrm{d}x$$

dove abbiamo tacitamente assunto che gli integrali al secondo membro dell'equazione esistano. Per il teorema fondamentale del calcolo integrale si ha che:

$$f(x)g(x) = \int [f'(x)g(x)]\mathrm{d}x + \int [f(x)g'(x)]\mathrm{d}x$$

quindi per risolvere un integrale possiamo considerare

$$\int [f'(x)g(x)]\mathrm{d}x = f(x)g(x) - \int [f(x)g'(x)]\mathrm{d}x$$

La forza di questo metodo risiede nella capacitá di individuare, fra le due funzioni quella piú facilmente derivabile/integrabile in maniera da poterla utilizzare per eliminare la difficoltdi integrazione insorta. Volendo applicare il procedimento appena eseguito su un intervallo di integrazione (a, b) si ottiene

$$f(x)g(x)|_a^b = \int_a^b [f'(x)g(x)]\mathrm{d}x + \int_a^b [f(x)g'(x)]\mathrm{d}x$$

cioé:

$$\int_a^b [f'(x)g(x)]\mathrm{d}x = f(x)g(x)|_a^b - \int_a^b [f(x)g'(x)]\mathrm{d}x$$

Esempio 7.1.1. Vogliamo risolvere per parti:

$$\int xe^x \mathrm{d}x$$

Poniamo

$$g'(x) = e^x$$

nell'espressione, come in precedenza:

$$\int [f(x)g'(x)]\mathrm{d}x = f(x)g(x) - \int [f'(x)g(x)]\mathrm{d}x$$

cioé:

$$\int xe^x \mathrm{d}x = xe^x - \int e^x \mathrm{d}x$$

$$\int xe^x \mathrm{d}x = xe^x - e^x + C$$

$$\int xe^x \mathrm{d}x = e^x(x-1) + C$$

7.2 Distribuzione normale ovvero Gaussiana

La pdf di una variabile casuale normale X è

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

Esercizio 7.2.1. E' possibile verificare che l'integrale di f vale 1, ma tale esercizio richiede l'uso degli integrali doppi ed é omesso.

Osservazione 7.2.1. Se indichiamo una variabile casuale distribuita normalmente con $X = \mathcal{N}(\mu, \sigma^2)$, e $Y = aX + b$. Abbiamo che $Y = \mathcal{N}(a\mu + b, a^2\sigma^2)$. Infatti,

$$F_Y(x) = P\{Y \leq x\} = P\{aX + b \leq x\} = P\left\{X \leq \frac{x-b}{a}\right\} = F_X\left(\frac{x-b}{a}\right) \quad e$$

$$f_Y(x) = \frac{dF_Y(x)}{dx} = \frac{f_X\left(\frac{x-b}{a}\right)}{a} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-((x-b)/a-\mu)^2/2\sigma^2} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-(x-a\mu-b)^2/2a^2\sigma^2}.$$

In particolare, ponendo $a = 1/\sigma$ e $b = -\mu/\sigma$ otteniamo $Z = (X-\mu)/\sigma = \mathcal{N}(0,1)$. La variabile $Z = \mathcal{N}(0,1)$ è detta normale standard.

Esercizio 7.2.2. Determiniamo $E[X]$ e $Var(X)$ per $X = \mathcal{N}(\mu, \sigma^2)$.

Valutiamo prima il valore atteso e la varianza della normale standard,

$$E[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-x^2/2} \mathrm{d}x = 0,$$

dove il risultato segue dall'osservazione che la funzione integranda e' dispari. Per la varianza integrando per parti, otteniamo

$$Var(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} \mathrm{d}x = 1.$$

Piú nel dettaglio, si ha

$$\int \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \int \frac{x\left(xe^{-\frac{x^2}{2}}\right)}{\sqrt{2\pi}} dx =$$

A questo punto scegliendo

$$f(x) = \frac{x}{\sqrt{2\pi}} \implies f'(x) = \frac{1}{\sqrt{2\pi}}$$

e

$$g'(x) = xe^{-\frac{x^2}{2}} \implies g(x) = -e^{-\frac{x^2}{2}}$$

utilizzando la formula di integrazione per parti si ha

$$\int \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = -\frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}} - \int -\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \int \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = -\frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + \int \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$$

e il risultato segue prendendo gli opportuni estremi di integrazioni, ovvero $\pm\infty$. Pertanto, poichè $X = \sigma Z + \mu$, abbiamo che $E[X] = \mu$ e $Var(X) = \sigma^2$.

Osservazione 7.2.2. La funzione di probabilità cumulata per $Z = \mathcal{N}(0,1)$ è solitamente indicata con la lettera Φ e data da,

$$\Phi(a) = (1/\sqrt{2\pi}) \int_{-\infty}^{a} e^{-x^2/2} \mathrm{d}x$$

per $a > 0$ e $1 - \Phi(-a)$ per $a < 0$.

Esercizio 7.2.3. Calcoliamo $P\{2 < X < 5\}$ per $\mathcal{N}(3,9)$.

$$P\{2 < X < 5\} = P\left\{-\frac{1}{3} < \frac{X-3}{3} < \frac{2}{3}\right\} = \Phi(2/3) - \Phi(0) - (1 - \Phi(1/3) + \Phi(0)) \sim 0.38.$$

7.3 Distribuzione esponenziale

La densità di probabilità di una variabile casuale esponenziale è $f(x) = \lambda e^{-\lambda x}$ per $x \geq 0$ e 0 altrimenti.

Osservazione 7.3.1. Si verifica facilmente che $\int_0^{+\infty} \lambda e^{-\lambda x} \mathrm{d}x = 1$, dato che

$$\int_0^{+\infty} e^{-\lambda x} \mathrm{d}x = -\frac{1}{\lambda} e^{-\lambda x}|_0^{+\infty} = 0 - (-\frac{1}{\lambda})$$

Osservazione 7.3.2. La funzione di probabilità cumulata è

$$F(x) = P\{X \leq x\} = -\lambda \frac{1}{\lambda} e^{-\lambda x}|_0^x = 1 - e^{-\lambda x}$$

per $x \geq 0$.

Esercizio 7.3.1. Facciamo vedere che $E[X] = 1/\lambda$.

Integrando per parti si ha

$$E[X] = \int_0^{+\infty} x \lambda e^{-\lambda x} \mathrm{d}x = -x e^{-\lambda x}|_0^{+\infty} - (-\int_0^{+\infty} e^{-\lambda x} \mathrm{d}x) = 1/\lambda.$$

Esercizio 7.3.2. Per la varianza dell'esponenziale abbiamo $Var(X) = 1/\lambda^2$.

É utile ricordare che $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Calcoliamo quindi

$$\mathbb{E}[X^2] = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} \mathrm{d}x.$$

Integrando per parti si ha

$$\mathbb{E}[X^2] = -x^2 e^{-\lambda x}|_0^{+\infty} - (-\int_0^{+\infty} 2x e^{-\lambda x} \mathrm{d}x) = 2 \int_0^{+\infty} x e^{-\lambda x} \mathrm{d}x$$

Moltiplicando e dividendo per λ si ottiene

$$\mathbb{E}[X^2] = \frac{2}{\lambda} \mathbb{E}[X] = \frac{2}{\lambda^2}.$$

Infine, ne segue che

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Osservazione 7.3.3. Se $P\{X > s + t | X > t\} = P\{X > s\}$ per tutti gli $s, t \geq 0$, la distribuzione è senza memoria. L'esponenziale è senza memoria: infatti

$$P\{X > s + t | X > t\} = \frac{P\{X > s + t, X > t\}}{P\{X > t\}} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P\{X > s\}.$$

8 Distribuzioni congiunte e indipendenza

In questa lezione estendiamo quanto visto nelle precedenti lezioni al caso di più variabili casuali. La distribuzione di una coppia di variabili congiunte (X, Y) con valori in $\{x_1, \dots, x_N\}$ e $\{y_1, \dots, y_M\}$ e' data da una pmf $p(i, j)$, $i = 1, \dots, N$, $j = 1, \dots, M$.

La cdf congiunta é data da

$$F(a,b) = \sum_{i\ : x_i \le a} \sum_{j\ : y_j \le b} p(i,j).$$

8.1 Distribuzioni marginali

A partire dalla cdf congiunta di due variabili casuali, $F(a, b) = P\{X \le a, Y \le b\}$, è immediato definire una cdf per le singole variabili, note come distribuzioni marginali

$$F_X(a) = P\{X \le a\} = P\{X \le a, Y < +\infty\} = F(a, +\infty)$$

$$F_Y(b) = P\{Y \le b\} = P\{X < +\infty, Y \le b\} = F(+\infty, b).$$

Esercizio 8.1.1. Dimostriamo che $P\{X > a, Y > b\} = 1 + F(a, b) - F_X(a) - F_Y(b)$.

Abbiamo

$$\begin{aligned} P\{X > a, Y > b\} &= 1 - P(\{X > a, Y > b\}^c) = 1 - P(\{X > a\}^c \cup \{Y > b\}^c) \\ &= 1 - P(\{X \le a\} \cup \{Y \le b\}) \\ &= 1 - (P\{X \le a\} + P\{Y \le b\} - P\{X \le a, Y \le b\}) \\ &= 1 + F(a,b) - F_X(a) - F_Y(b). \end{aligned}$$

Osservazione 8.1.1. Nel caso di variabili discrete X e Y $p(x, y) = P\{X = x, Y = y\}$, allora

$$p_X(x) = P\{X = x\} = \sum_{y : p(x,y) > 0} p(x,y) \quad \text{e} \quad p_Y(y) = P\{Y = y\} = \sum_{x : p(x,y) > 0} p(x,y).$$

Osservazione 8.1.2. Il caso continuo é in completa analogia col discreto, ma richiede l'uso degli integrali doppi,

$$P\{(X,Y) \in C\} = \int \int_{(x,y) \in C} f(x,y) \mathrm{d}x \mathrm{d}y = P\{X \in A, Y \in B\} = \int_B \left(\int_A f(x,y) \mathrm{d}x \right) \mathrm{d}y$$

con

$$F(a,b) = \int_{-\infty}^{b} \left(\int_{-\infty}^{a} f(x,y) \mathrm{d}x \right) \mathrm{d}y \quad \text{e} \quad f(a,b) = \frac{\partial^2 F(a,b)}{\partial a \partial b}.$$

Inoltre, per le probabilità marginali, se $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y$ e $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x$ abbiamo

$$P\{X \in A\} = P\{X \in A, Y < +\infty\} = \int_A \left(\int_{-\infty}^{+\infty} f(x,y) \mathrm{d}y \right) \mathrm{d}x = \int_A f_X(x) dx$$

$$P\{X \in B\} = P\{X < +\infty, Y \in B\} = \int_B \left(\int_{-\infty}^{+\infty} f(x,y) \mathrm{d}x \right) \mathrm{d}y = \int_B f_Y(y) dy.$$

Esercizio 8.1.2. Estraiamo 3 palline a caso da un'urna che ne contiene 3 rosse, 4 bianche e 5 blu. Se X conta le rosse estratte e Y le bianche, calcoliamo la distribuzione congiunta di probabilità di massa $P\{X = x, Y = y\}$.

Abbiamo 220 casi possibili e

$p(0,0) = 10/220$ $p(0,1) = 40/220$ $p(0,2) = 30/220$ $p(0,3) = 4/220$
$p(1,0) = 30/220$ $p(1,1) = 60/220$ $p(1,2) = 18/220$
$p(2,0) = 15/220$ $p(2,1) = 12/220$
$p(3,0) = 1/220$

Osservazione 8.1.3. Sommando le probabilità per righe o per colonne otteniamo le probabilità marginali (chiamate in questo modo perchè storicamente scritte a margine della tabella).

8.2 Variabili casuali indipendenti

Due variabili casuali sono indipendenti se per ogni coppia di insiemi A e B

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

Una definizione equivalente richiede che per ogni coppia a e b di numeri reali

$$P\{X \leq a, Y \leq b\} = P\{X \leq a\}P\{Y \leq b\}.$$

Se X e Y sono indipendenti, $F(a,b) = F_X(a)F_Y(b)$. Inoltre, abbiamo $p(x,y) = p_X(x)p_Y(y)$ nel caso discreto.

9 Proprietà dei valori attesi

La capitale importanza della nozione di valore atteso merita una lezione a parte sia per il ruolo giocato nel caso di somme di variabili casuali, sia per alcuni concetti fondamentali nell'analisi di dati quali covarianza e correlazione.

9.1 Somme di variabili casuali

Se X e Y sono due variabili casuali, il valore atteso di una funzione $g(X,Y)$ può essere calcolato nel caso discreto come
$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p(x,y).$$

Se $g(X,Y) = X + Y$ avremo

$$E[X+Y] = \sum_x \sum_y (x+y)p(x,y) = \sum_x \sum_y xp(x,y) + \sum_y \sum_x yp(x,y) = E[X] + E[Y],$$

ricordando la definizione di distribuzione marginale. Inoltre, se $X \geq Y$, $E[X] \geq E[Y]$.

Esercizio 9.1.1 (Passeggiata dell'ubriaco). Un ubriaco si muove con passi di lunghezza unitaria. Ogni passo è indipendente dai precedenti e in una direzione Θ uniformemente distribuita tra 0 e 2π. Dopo n passi a quale distanza D si troverà dal punto di partenza?

Se θ_i è la direzione del passo i-esimo, dopo i passi l'ubriaco si troverà nella posizione $(\sum_i X_i, \sum_i Y_i)$ con $(X_i, Y_i) = (\cos\theta_i, \sin\theta_i)$. Pertanto, dopo n passi avremo

$$D^2 = (\sum_i X_i)^2 + (\sum_i Y_i)^2 = \sum_i (\cos^2\theta_i + \sin^2\theta_i) + (\sum_i \cos\theta_i)(\sum_{j \neq i} \cos\theta_j) + (\sum_i \sin\theta_i)(\sum_{j \neq i} \sin\theta_j).$$

Poichè $E[\cos\Theta] = E[\sin\Theta] = 0$ per n grande avremo $\sum_i \cos\theta_i \approx \sum_i \sin\theta_i \approx 0$ e, pertanto, $D \approx \sqrt{n}$.

Esercizio 9.1.2. L'algoritmo quick-sort ordina gli interi da 1 a n disposti casualmente in un vettore di dimensione n nel seguente modo. Se $n = 2$ ordina i due valori dopo un confronto. Per $n > 2$ estrae un valore x in una posizione a caso del vettore e divide il vettore in due parti: una costituita da tutti i valori minori di x e una da tutti i valori maggiori di x. L'algoritmo si ripete uguale a se stesso sulle due parti terminando quando una parte consiste al più di due valori. Valutiamo la complessità computazionale di quick-sort.

Poniamo $I(i, j) = 1$ se i e j saranno confrontati, 0 altrimenti. Indichiamo il numero di confronti effettuati con

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(i, j).$$

Se p_{ij} è la probabilità che i e j saranno confrontati, il valore atteso dei confronti è

$$E[X] = E\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(i, j)\right] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij}.$$

Valutiamo ora p_{ij}. Inizialmente c'è una sola parte per tutti e $i, i+1, \ldots, j$ sono dalla stessa parte. A ogni passo, se il numero x scelto è minore di i o maggiore di j, i valori $i, i+1, \ldots, j$ continueranno a essere tutti dalla stessa parte. Quindi non cambia nulla fino a che non è scelto un valore compreso tra i e j. Se è scelto un valore compreso tra i e j, ma diverso da i o da j, i valori i e j finiscono in parti diverse e non saranno mai confrontati. Nel caso, invece, che il valore scelto sia i o j, i e j sono confrontati. Con 2 casi favorevoli su $j - i + 1$ casi possibili si ha $p_{ij} = 2/(j - i + 1)$ e, quindi,

$$\sum_{j=i+1}^{n} \frac{2}{j - i + 1} \approx \int_{i+1}^{n} \frac{2}{x - i + 1} dx = 2\ln(n - i + 1) - 2\ln(2) \approx 2\ln(n - i + 1) \quad \text{e}$$

$$E[X] \approx \sum_{i=1}^{n-1} 2\ln(n - i + 1) \approx \int_{1}^{n-1} 2\ln(n - x + 1) dx \approx 2\int_{2}^{n} \ln y dy \approx 2n\ln n.$$

9.2 Covarianza e varianza di somme

Se X e Y sono indipendenti, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$. Infatti

$$\begin{aligned} E[g(X)h(Y)] &= \sum_x \sum_y g(x)h(y)p(x, y) = \sum_x \sum_y g(x)h(y)p(x)p(y) \\ &= \sum_x g(x)p(x) \sum_y h(y)p(y) = E[g(X)]E[h(Y)]. \end{aligned} \tag{2}$$

La covarianza di due variabili casuali X e Y è definita come

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

ed é facile vedere che $Cov(X, Y) = E[XY] - E[X]E[Y]$.

Esercizio 9.2.1. Esprimiamo la varianza di una somma di due variabili aleatorie in termini delle varianze delle singole variabili e della loro covarianza.

$$\begin{aligned} Var(X + Y) &= E[(X + Y - (E[X + Y])^2] = E[(X - E[X] + Y - E[Y])^2] \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])] \\ &= Var(X) + Var(Y) + 2Cov(X, Y) \end{aligned}$$

Osservazione 9.2.1. Se X e Y sono indipendenti, allora $Cov(X, Y) = 0$ (ma non viceversa).

Esercizio 9.2.2. Dimostriamo che se a è una costante, allora $Cov(aX, Y) = aCov(X, Y)$.

Abbiamo
$$Cov(aX, Y) = E[(aX - E[aX])(Y - E[Y])] = aE[XY] - aE[X]E[Y].$$

Esercizio 9.2.3. Se X_i per $i = 1, \ldots, n$ sono *i.i.d.* con valore atteso μ, varianza σ^2 definiamo le variabili casuali media campionaria, X^*, e varianza campionaria, S^2, come

$$\mu_n = \frac{\sum_i X_i}{n} \qquad \text{e} \qquad \sigma_n^2 = \frac{\sum_i (X_i - \mu_n)^2}{n-1}.$$

Calcoliamo $E[X^*]$, $Var(X^*)$ e $E[S^2]$.

$$\begin{aligned}
E[\mu_n] &= E\left[\frac{\sum_i X_i}{n}\right] = \frac{E[\sum_i X_i]}{n} = \frac{\sum_i E[X_i]}{n} = \mu \\
Var(\mu_n) &= E[(\frac{\sum_i X_i}{n} - \mu)^2]) = \frac{1}{n^2}E[(\sum_i (X_i - \mu))^2] = \qquad (3)\\
&= \frac{1}{n^2}E[\sum_i (X_i - \mu)^2] + \frac{1}{n^2}E[\sum_i \sum_{j \neq i} (X_i - \mu)(X_j - \mu)] = \frac{\sum_i \sigma^2}{n^2} = \frac{\sigma^2}{n} \\
E[\sigma_n^2] &= E\left[\frac{\sum_i (X_i - \mu_n)^2}{n-1}\right] = \frac{E[\sum_i (X_i - \mu + \mu - \mu_n)^2]}{n-1} \\
&= \frac{E[\sum_i (X_i - \mu)^2] + E[\sum_i (\mu - \mu_n)^2] - 2E[\sum_i (\mu - X_i) \sum_i (\mu - \mu_n)]}{n-1} \\
&= \frac{n\sigma^2 + nVar(\mu_n) - 2nVar(\mu_n)}{n-1} = \frac{n\sigma^2 - nVar(\mu_n)}{n-1} = \frac{(n-1)\sigma^2}{n-1} = \sigma^2.
\end{aligned}$$

Osservazione 9.2.2. Torneremo in seguito su questo esempio quando introdurremo la nozione di distorsione e stimatori corretti e distorti in statistica. Il fatto che al denominatore di S^2 compaia $n-1$ anzichè n rende la varianza campionaria uno stimatore corretto della varianza.

9.3 Correlazione

La correlazione $\rho(X, Y)$ di due variabili casuali X e Y è definita come

$$\rho(X, Y) = \frac{Cov(X.Y)}{\sqrt{Var(X)Var(Y)}}.$$

Esercizio 9.3.1. Dimostriamo che $-1 \leq \rho(X, Y) \leq 1$. Sia $Var(X) = \sigma^2$ e $Var(Y) = \tau^2$.

Ragionando in modo analogo all'esercizio 9.2.2, abbiamo che $-1 \leq \rho(X, Y)$ poichè

$$0 \leq Var\left(\frac{X}{\sigma} + \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} + 2\frac{Cov(X)}{\sigma\tau} = 2(1 + \rho(X, Y)).$$

Inoltre $\rho(X, Y) \leq 1$ poichè

$$0 \leq Var\left(\frac{X}{\sigma} - \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} - 2\frac{Cov(X)}{\sigma\tau} = 2(1 - \rho(X, Y)).$$

Osservazione 9.3.1. Per $\rho(X, Y) \approx 1$ la relazione tra X e Y è ben approssimata da una retta con coefficiente angolare positivo (negativo per $\rho(X, Y) \approx -1$).

10 Risultati asintotici

In questa ultima lezione enunciamo i risultati fondamentali della Teoria della Probabilità: la legge dei grandi numeri e il teorema centrale del limite.

10.1 Disuguaglianze fondamentali

Sia X una variabile casuale che assume valori non negativi. Allora vale la disuguaglianza di Markov

$$\forall a > 0, P\{X \geq a\} \leq \frac{E[X]}{a}.$$

Infatti

$$E[X] = \int_0^{+\infty} x f(x) \mathrm{d}x \geq \int_a^{+\infty} x f(x) \mathrm{d}x$$

dato che $xf(x) \geq 0$. Inoltre moltiplicando e dividendo per a si

$$E[X] \geq a \int_a^{+\infty} \frac{x}{a} f(x) \mathrm{d}x \geq a \int_a^{+\infty} f(x) \mathrm{d}x,$$

notando che $x > a$ e quindi $x/a > 1$. Infine ricordando la definizione di cumulata si ha

$$E[X] \geq aP\{X \geq a\}.$$

Sia X una variabile casuale con valore atteso μ e varianza σ^2 finiti. Allora vale la disuguaglianza di Chebyshev

$$\forall \epsilon > 0, P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

Basta applicare la disuguaglianza di Markov a $(X - \mu)^2$ con $a = \epsilon^2$ e notare che

$$\forall \epsilon > 0, P\{|X - \mu| \geq \epsilon\} = P\{|X - \mu|^2 \geq \epsilon^2\}.$$

Osservazione 10.1.1. La portata fondamentale delle disuguaglianze di Markov e Chebyshev è la loro validità per qualunque distribuzione di probabilità.

Esercizio 10.1.1. Sia X uniformemente distribuita tra 0 e 10. Quanto vale la probabilita' che X si scosti di 4 dal suo valore medio? Confrontare il risultato ottenuto utilizzando la disuguaglianza di Chebyshev con quello ottenuto sfruttando il fatto che la distribuzione e ' uniforme.

Abbiamo che $E[X] = 5$ e $\sigma^2 = 25/3$. Se applichiamo la disuguaglianza di Chebyshev con $\epsilon = 4$ otteniamo $P\{|X - 5| \geq 4\} \leq 25/48$, mentre, utilizzando l'informazione sulla forma della distribuzione di X otteniamo un valore più piccolo, ovvero $P\{|X - 5| \geq 4\} = (2/10) \int_0^1 \mathrm{d}x = 2/10$.

Esempio 10.1.1. Sia $X = \mathcal{N}(\mu, \sigma^2)$. Se applichiamo la disuguaglianza di Chebyshev con $\epsilon = 2\sigma$ otteniamo

$$P\{|X - \mu| \geq 2\sigma\} \leq \frac{1}{4}.$$

Anche in questo caso, invece, possiamo l'informazione sulla forma della distribuzione sappiamo che $P\{|X - \mu|/\sigma > 2\} = 2(1 - \Phi(2)) \approx 0.05$.

10.2 Legge dei grandi numeri

Questo risultato e' alla base di un gran numero di metodi per la stima di una quantitá ignota a partire da un numero finito di osservazioni.

Teorema 10.2.1 (Legge debole dei grandi numeri). Siano X_i con $i = 1, 2, \ldots, n$ variabili casuali indipendenti e identicamente distribuite con $E[X_i] = \mu$. Allora

$$\forall \epsilon > 0 \lim_{n\to\infty} P\left\{\left|\frac{\sum X_i}{n} - \mu\right| \geq \epsilon\right\} = 0.$$

Con l'ipotesi aggiuntiva (e non necessaria) che la varianza σ^2 sia finita, poichè $E[\sum_i X_i/n] = \mu$ e $Var(\sum_i X_i/n) = \sigma^2/n$, applicando la disuguaglianza di Chebyshev per $k = \epsilon$ otteniamo

$$P\left\{\left|\sum_i \frac{X_i}{n} - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

Osservazione 10.2.1. La legge dei grandi numeri in forma debole afferma che al crescere di n la probabilità che la media empirica differisca dal valore atteso tende a 0. Differenze significative per n sufficientemente grande possono essere rilevate anche se non frequentemente.

Esercizio 10.2.1. Le misure X_i con le quali un astronomo stima una distanza d sono variabili casuali indipendenti e identicamente distribuite con $E[X_i] = d$ e $Var(X_i) = \sigma^2 = 4$ anni luce. Calcoliamo il numero di misure necessarie per determinare d con una precisione di 0.5 anni luce con confidenza pari al 95%.

Con la disuguaglianza di Chebyshev: poichè $E[\sum_i X_i/n] = d$ e $Var(\sum_i X_i/n) = 4/n$, segue che

$$P\left\{\left|\frac{\sum_i X_i}{n} - d\right| \geq 0.5\right\} \leq \frac{4}{0.25n} = \frac{16}{n}.$$

Ponendo $16/n = 5\%$ otteniamo che per raggiungere la confidenza del 95% sono necessarie 320 misure. Sono molte più del caso precedente ma non abbiamo dovuto fare ipotesi circa la bontà dell'approssimazione (come invece serve fare nel caso del teorema del limite centrale).

10.3 Teorema del limite centrale

Il teorema del limite centrale e' uno dei risultati più importanti della matematica.

Teorema 10.3.1 (Teorema centrale del limite). Siano X_i con $i = 1, 2, \ldots, n$ variabili casuali indipendenti e identicamente distribuite con $E[X_i] = \mu$ e $Var(X_i) = \sigma^2$. Allora

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

Osservazione 10.3.1. Questo teorema garantisce che le frequenze empiriche, al crescere di n, si distribuiscono come una gaussiana centrata sul valore atteso indipendentemente dalla distribuzione sottostante. La varianza della gaussiana è la varianza della distribuzione sottostante divisa per n.

Esercizio 10.3.1. Come cambia la soluzione dell'esercizio 10.2.1 utilizzando il teorema del limite centrale?

Con il teorema del limite centrale: sappiamo che $Z_n = (\sum_i X_i - nd)/(2\sqrt{n})$, al crescere di n, tende alla distribuzione normale standard. Poichè

$$P\left\{-0.5 \leq \left(\frac{\sum_i X_i}{n} - d\right) \leq 0.5\right\} = P\left\{-0.5\frac{\sqrt{n}}{2} \leq Z_n \leq 0.5\frac{\sqrt{n}}{2}\right\}$$

abbiamo che

$$P\left\{-0.5\frac{\sqrt{n}}{2} \leq Z_n \leq 0.5\frac{\sqrt{n}}{2}\right\} \approx \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1.$$

Ponendo $2\Phi(\sqrt{n}/4) - 1 = 95\%$, ricaviamo che per raggiungere la confidenza del 95% è necessario effettuare almeno 63 misure.

10.4 Funzione generatrice dei momenti

La funzione generatrice dei momenti di una variabile casuale X è definita $\forall t \in \mathbb{R}$ come $M(t) = E[e^{tX}]$. Nel discreto abbiamo $M(t) = \sum_x e^{tx} p(x)$ mentre, nel continuo, $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\mathrm{d}x$.

Osservazione 10.4.1. Nell'ipotesi che sia sempre possibile scambiare la derivata rispetto a t con il calcolo del valore atteso, la derivata n-esima di $M(t)$ calcolata in $t = 0$ è uguale al momento n-esimo della distribuzione di X. Infatti, valutando

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = \frac{\mathrm{d}E[e^{tX}]}{\mathrm{d}t} = E[Xe^{tX}]$$

in $t = 0$ otteniamo $M'(0) = E[X]$ e, continuando a derivare, $M^n(0) = E[X^n]$.

Esercizio 10.4.1. Calcoliamo il valore atteso e la varianza per una binomiale con parametri n e p mediante la funzione generatrice.

Per la funzione generatrice abbiamo

$$M(t) = E[e^{tX}] = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{n}\binom{n}{k}(e^t p)^k (1-p)^{n-k} = (pe^t + 1 - p)^n.$$

Derivando una prima volta otteniamo $M'(t) = n(pe^t + 1 - p)^{n-1}pe^t$, da cui segue $M'(0) = E[X] = np$. Derivando una seconda volta otteniamo

$$M''(t) = n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + n(pe^t + 1 - p)^{n-1}pe^t$$

da cui segue $M''(0) = E[X^2] = n(n-1)p^2 + np$ e, quindi, $Var(X) = E[X^2] - (E[X])^2 = np(1-p)$.

Esercizio 10.4.2. Calcoliamo la funzione generatrice per $Z = \mathcal{N}(0,1)$. Abbiamo

$$M(t) = \int_{-\infty}^{+\infty} e^{tx}e^{-x^2/2}\mathrm{d}x = \int_{-\infty}^{+\infty} e^{t^2/2}e^{-(x-t)^2/2}\mathrm{d}x = e^{t^2/2}.$$

10.5 Disuguaglianze di Chernoff

Applicando la disuguaglianza di Markov a e^{tX}, otteniamo le disuguaglianze di Chernoff

$$\forall t > 0, \quad P\{X \geq a\} = P\{e^{tX} \geq e^{ta}\} \leq E[e^{tX}]e^{-ta} = M(t)e^{-ta}$$

$$\forall t < 0, \quad P\{X \leq a\} = P\{e^{tX} \geq e^{ta}\} \leq E[e^{tX}]e^{-ta} = M(t)e^{-ta}.$$

Osservazione 10.5.1. Entrambe le disuguaglianze di Chernoff sono valide per tutti i valori di t. Pertanto, per una distribuzione fissata che abbia $M(t)$ come funzione generatrice dei momenti, è possibile ottenere disuguaglianze più stringenti di Chebyshev scegliendo il valore di t che minimizza $M(t)e^{-ta}$.

Esempio 10.5.1. Per $Z = \mathcal{N}(0,1)$ abbiamo $M(t) = e^{t^2/2}$ e, quindi, $\forall t > 0$

$$P\{Z \geq a\} \leq M(t)e^{-ta} = e^{t^2/2}e^{-ta}.$$

Uguagliando a 0 la derivata otteniamo $t = a$. Pertanto per $a > 0$

$$P\{Z \geq a\} \leq e^{-a^2/2}.$$

Ripetendo il tutto per $t < 0$ otteniamo che per $a < 0$

$$P\{Z \leq a\} \leq e^{-a^2/2}.$$

Pertanto $\forall a > 0$ possiamo scrivere

$$P\{|Z| \geq a\} \leq e^{-a^2/2} \ \text{ (Chernoff) } \ \text{invece di} \ \ P\{|Z| \geq a\} \leq \frac{1}{a^2} \ \text{ (Chebyshev)}.$$