Elementi di Probabilità, Statistica e Processi Stocastici

Franco Flandoli

23 ottobre 2011

Indice

Prefazione

Il materiale qui raccolto ha la forma di appunti più che di libro organico. Il testo è pensato per le lauree magistrali in Ingegneria e raccoglie materiale utilizzato in numerosi corsi in anni recenti. Alcune parti devono molto al contributo di alcuni collaboratori e di numerosi studenti; in particolare merita di essere ricordato il contributo di Michele Barsanti alle due sezioni sull'analisi di Fourier dei processi stocastici, oltre che a vari altri punti ed esercizi, di Michele Tocchet alla sezione sul metodo PCA, di Giuseppe Matisi e Lorenzo Doccini ad alcuni esercizi di statistica multivariata (4 e 5).

Capitolo 1

Elementi di Calcolo delle Probabilità

Questo capitolo è dedicato ad un riassunto degli elementi di Calcolo delle Probabilità che verranno utilizzati nel seguito. L'esposizione di questi elementi è sommaria per cui, chi sentisse la necessità di approfondimenti, può leggere il testo di S. Ross, Probabilità e Statistica, Apogeo 2008 (per un'esposizione adatta ad un triennio di Ingegneria) o di P. Baldi, Calcolo delle Probabilità, McGraw-Hill 2007 (più adatto per le lauree magistrali in Ingegneria), così come molti altri.

La prima sezione è dedicata all'illustrazione di alcuni primi "oggetti" del calcolo delle probabilità:

- *gli eventi*; in parole povere sono *affermazioni*, più formalmente saranno *insiemi*; su di essi si opera con operazioni logiche, o insiemistiche, a seconda del punto di vista;
- *la probabilità*; si calcola la probabilità di eventi; ad ogni evento è associato un numero dell'intervallo $[0, 1]$, la sua probabilità; la probabilità sarà quindi un'*applicazione* che ad ogni evento associa un numero, con certe regole.

Nella sezione successiva vedremo poi:

- *le variabili aleatorie*; a livello intuitivo sono grandezze (numeriche o di altro tipo) con un qualche grado di imprevedibilità, quantificato da nozioni probabilistiche; nella formalizzazione matematica saranno funzioni;
- *i valori medi*; indicatori numerici associati a variabili aleatorie che ne riassumono alcune caratteristiche.

Segue poi una sezione di esempi, una sui teoremi limite ed una più specifica sui vettori aleatori, soprattutto gaussiani.

1.1 Eventi e loro probabilità

1.1.1 Universo ed eventi elementari

Nella costruzione dello schema matematico fondamentale della probabilità, lo *spazio probabilizzato* $(\Omega, \mathcal{F}, P)$ che verrà introdotto un po' per volta, si parte da un insieme ambiente, di solito indicato con Ω, o S, spesso detto *universo*, o *insieme degli eventi elementari* (o *insieme degli esiti*). I suoi elementi $\omega \in \Omega$ si dicono *eventi elementari* (o *esiti*). Intuitivamente, di fronte ad una situazione casuale, come ad esempio un esperimento, il risultato dell'esperimento è un esito, quindi Ω è l'insieme dei risultati possibili dell'esperimento.

Ad esempio, se osserviamo il simbolo, 0 o 1, che entra in un canale di trasmissione (che trasmette un simbolo alla volta), ed il simbolo, 0 o 1, che ne esce, un evento elementare è una coppia (a, b) dove a (simbolo in entrata) e b (simbolo in uscita) possono valere 0 o 1. Quindi i possibili eventi elementari sono

$$(0,0) \quad (0,1) \quad (1,0) \quad (1,1)\,.$$

Lo *spazio* Ω in questo caso è l'insieme di questi oggetti, quindi semplicemente

$$\Omega = \{(0,0)\,,\ (0,1)\,,\ (1,0)\,,\ (1,1)\}\,.$$

Un esempio di evento elementare è

$$\omega = (0,1)\,.$$

Va notato che un singolo evento elementare ω contiene l'informazione più dettagliata possibile relativamente al problema che si sta studiando. Nell'esempio appena visto, uno specifico valore del simbolo in uscita non è un evento elementare: l'affermazione

"il simbolo in uscita è 1"

non corrisponde ad un evento elementare. Invece l'affermazione "il simbolo in entrata è 0 ed il simbolo in uscita è 1" corrisponde all'evento elementare $\omega = (0,1)$.

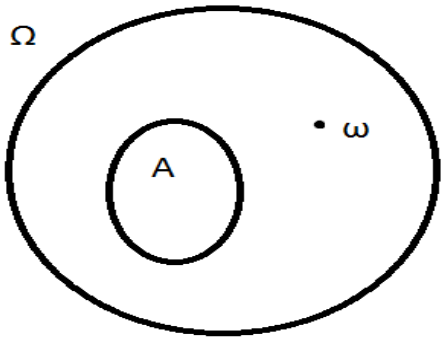
Analogamente, se si sta esaminando un gioco che consiste nel lancio di 5 dadi, il risultato del primo lancio non è un evento elementare, mentre una qualsiasi stringa $(n_1, ..., n_5)$ dei risultati dei cinque lanci è un evento elementare.

Se si osserva un fluido turbolento e lo si descrive con grandezze aleatorie, un evento elementare è una possibile configurazione complessiva del fluido (nel senso della specifica di velocità, pressione ecc. in ogni punto della regione occupata dal fluido). Invece, l'osservazione del valore della velocità in un certo punto fissato non è un evento elementare.

1.1.2 Eventi

Gli esempi precedenti mostrano che il dettaglio di conoscenza insito nel concetto di evento elementare è spesso sovrabbondante. E' perfettamente sensato porsi domande relative a grandezze meno dettagliate, come il valore del simbolo ricevuto da un canale di comunicazione o il valore della velocità di un fluido turbolento in un punto specifico. Si introducono allora gli *eventi* (non più necessariamente elementari).

In prima approssimazione, possiamo dire che un *evento* è un *sottoinsieme* di Ω.



Universo Ω, un evento elementare ω ed un evento A

Riprendendo il primo esempio fatto sopra dei simboli 0 e 1 in entrata ed uscita, l'insieme

$$A = \{(0,1)\,,\ (1,1)\}$$

corrisponde all'affermazione "il simbolo in uscita è 1". A è l'insieme di tutti gli eventi elementari che corrispondono a tale affermazione. Questo è un esempio di *evento*.

In prima approssimazione, ogni sottoinsieme $A \subset \Omega$ è un possibile evento di interesse. Ci sono però due ragioni per restringere l'attenzione, in alcuni casi, ad una famiglia più ristretta di eventi, che non comprenda necessariamente tutti i sottoinsiemi $A \subset \Omega$ ma solo alcuni. Una ragione è meramente tecnica nel senso matematico del termine: in certi esempi non è possibile definire la probabilità (di cui parleremo tra un attimo) di *ogni* sottoinsieme di Ω, in modo coerente secondo certe regole; per cui è necessario sacrificare certi sottoinsiemi troppo strani. Purtroppo questa ragione, assai noiosa, si apprezza solo dopo lunghe premesse di teoria della misura e teoria degli insiemi (ad esempio, per costruire sottoinsiemi strani che creino problemi si deve usare l'assioma della scelta). Per scopi pratici questa restrizione, o patologia, è irrilevante: tutti gli insiemi che introdurremo nel corso sono accettabili come eventi.

La seconda ragione invece è molto più interessante per le applicazioni: essa corrisponde al concetto di maggior o minor informazione che abbiamo su un problema. Premettiamo quindi una breve introduzione al concetto di informazione.

Circa la distinzione tra evento ed evento elementare si osservi il seguente fatto: quando l'esperimento, o osservazione, si è conclusa, osserviamo il verificarsi di un evento elementare ω. Molti eventi A si sono verificati, allora: tutti gli eventi A che contengono l'elemento ω. Se ad esempio dal lancio di un dato è uscito il numero 2 (evento elementare), si è verificato l'evento "è uscito un numero pari", ed anche "è uscito un numero inferiore a 4", e così via.

1.1.3 Informazione contenuta in una famiglia di eventi

Non esiste alcuna definizione univoca del concetto di informazione, che ha molte facce suscettibili di varie descrizioni rigorose. Una di queste è data da certi indicatori numerici chiamati *entropia* (ce ne sono di vario tipo) che vengono introdotti per descrivere l'informazione contenuta ad esempio in sequenze numeriche o in distribuzioni di probabilità.

Qui invece ci indirizziamo in un'altra direzione. Pensiamo per fissare le idee ad un esperimento eseguito per misurare il valore di una grandezza fisica. Supponiamo ad esempio

che lo strumento di misura abbia una sua incertezza intrinseca. Un modo per tenerne conto può essere il seguente: invece che sperare di ottenere un ben preciso valore x come risultato dell'esperimento, immaginiamo che il risultato consista in un intervallo, preso in una famiglia prefissata di intervalli possibili $(-\infty, x_1]$, $(x_1, x_2]$, ... , $(x_{n-1}, x_n]$, (x_n, ∞). Ad esempio, immaginiamo a priori di non poterci fidare della misura dello strumento oltre la prima cifra decimale, e che i valori inferiori a -10 o superiori a 10 non siano distinguibili. Allora il risultato dell'esperimento può essere solo uno dei seguenti intervalli: $(-\infty, -10]$, $(-10, -9.9]$, $(-9.9, -9.8]$, ... , $(9.8, 9.9]$, $(9.9, 10]$. (Esempio: quando si usano le tavole gaussiane dei quantili, ci si deve accontentare dei numeri riportati sulle tavole, che non sono tutti i numeri reali, e ci si deve accontatare della precisione del risultato espressa con un numero finito e basso di cifre, secondo la disponibilità di quelle tavole.)

Questa famiglia di intervalli descrive il nostro grado di infomazione (o se si vuole il grado di informazione raggiungibile con l'esperimento).

Se in un momento successivo si riesce a migliorare lo strumento di misura in modo da poterci fidare di due cifre decimali e magari di allargare lo spettro dei valori da -20 a 20, la famiglia che descrive la nostra informazione diventa $(-\infty, -20]$, $(-20, -19.99]$, $(-19.99, -19.98]$, ... , $(19.98, 19.99]$, $(19.99, 20]$.

In questo esempio l'insieme universo Ω naturale da introdurre è l'insieme $\mathbb{R}$ dei numeri reali, ma gli unici sottoinsiemi che ci interessano per la descrizione dell'esperimento sono gli intervalli scritti sopra. Oppure possiamo adottare un'altro punto di vista: in teoria ci interesserebbero tutti i sottoinsiemi, in particolare quelli composti dai singoli numeri reali (che darebbero il risultato con precisione infinita), ma in pratica evidenziamo che il grado di informazione contenuto nel nostro esperimento è descritto dalla famiglia più ristretta degli intervalli detti sopra.

Vediamo un'altro esempio.

Esempio 1 *In un capitolo successivo studieremo i processi stocastici. Per lo scopo di questo esampio, basti pensare intuitivamente che un processo stocastico è la descrizione matematica di una grandezza (fisica, economica ecc.) che varia nel tempo ed è aleatoria. Indichiamo con X_t questa grandezza al tempo t. Supponiamo di studiare il fenomeno per tutti i tempi $t \geq 0$. Prendiamo come Ω l'insieme di tutte le "storie" possibili di questa grandezza, tutte le funzioni $t \mapsto x_t$ che possono realizzarsi. Gli eventi sono sottoinsiemi di Ω, cioè famiglie di tali "storie", "realizzazioni". Un esempio è l'evento $A =$ "al tempo $t = t_1$ il valore di X_t è positivo", evento che possiamo riassumere con la scrittura*

$$A = \{X_{t_1} > 0\}.$$

Un altro è $B = \{X_{t_2} \in I\}$ con I un certo intervallo. Intersecando eventi di questo tipo troviamo eventi della forma

$$\{X_{t_1} \in I_1, ..., X_{t_n} \in I_n\}$$

cioè eventi che affermano che X_t, in certi istanti assume certi possibili valori. Fatte queste premesse, fissiamo un tempo $T > 0$ e consideriamo la famiglia $\mathcal{F}_T^0$ di tutti gli eventi del tipo $\{X_{t_1} \in I_1, ..., X_{t_n} \in I_n\}$ con

$$0 \leq t_1 \leq ... \leq t_n \leq T.$$

Sono eventi che affermano qualcosa del processo X_t solo entro il tempo T, solo relativamente all'intervallo $[0,T]$. La famiglia $\mathcal{F}_T^0$ di tutti questi eventi descrive un certo grado di informazione, l'informazione di cosa può accadere nell'intervallo $[0,T]$.

Al crescere di T questa famiglia cresce, cioè $\mathcal{F}_T^0 \subset \mathcal{F}_S^0$ se $T < S$. Si acquisisce nuova informazione, su un periodo di tempo maggiore.

1.1.4 Algebre di eventi

Ricordiamo che la famiglia di tutti i sottoinsiemi di Ω, detta famiglia delle *parti* di Ω, si usa indicare con $\mathcal{P}(\Omega)$.

Definizione 1 *Chiamiamo algebra di insiemi di Ω una famiglia $\mathcal{F} \subset \mathcal{P}(\Omega)$ che sia chiusa per tutte le operazioni insiemistiche finite e tale che $\Omega \in \mathcal{F}$.*

Chiusa per tutte le operazioni insiemistiche *finite* significa che se $A, B \in \mathcal{F}$ allora

$$A \cup B \in \mathcal{F},\ A \cap B \in \mathcal{F},\ A^c \in \mathcal{F},$$

(il complementare A^c è inteso rispetto allo spazio ambiente Ω) e di conseguenza anche $A \backslash B \in \mathcal{F}$, $A \triangle B \in \mathcal{F}$, dove $A \backslash B$ è l'insieme dei punti di A che non stanno in B, e la differenza simmetrica $A \triangle B$ è l'unione di $A \backslash B$ più $B \backslash A$. Dal fatto che $\Omega \in \mathcal{F}$ e $A^c \in \mathcal{F}$ discende che $\emptyset \in \mathcal{F}$. Si ricordino le formule di De Morgan

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c \end{aligned}$$

che si usano spesso quando si eseguono nei dettagi certe verifiche.Valgono inoltre proprietà distributive tra unione e intersezione, su cui non insistiamo.

Due esempi semplici di algebre di insiemi sono $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathcal{F} = \{\emptyset, \Omega\}$. La verifica che $\mathcal{P}(\Omega)$ è un'algebra è ovvia. Nel caso di $\{\emptyset, \Omega\}$, si osservi ad esempio che $\emptyset \cup \Omega = \Omega$, $\emptyset \cap \Omega = \emptyset$, $\emptyset^c = \Omega$; non si esce dalla famiglia $\{\emptyset, \Omega\}$ effettuando operazioni insiemistiche sui suoi elementi.

Un altro esempio, per così dire intermedio tra i due, è

$$\mathcal{F} = \{\emptyset, \Omega, A, A^c\}.$$

La verifica che sia un'algebra è identica al caso di $\{\emptyset, \Omega\}$. L'algebra $\{\emptyset, \Omega\}$ non contiene alcuna informazione, $\{\emptyset, \Omega, A, A^c\}$ contiene l'informazione relativa al solo evento A, $\mathcal{P}(\Omega)$ contiene tutte le infomazioni possibili.

Un esempio importante, nello spazio $\Omega = \mathbb{R}$, è la famiglia $\mathcal{F}$ dei pluri-intervalli, composta da tutti i seguenti insiemi, che elenchiamo:

- $\emptyset$ ed $\mathbb{R}$ stesso
- gli intervalli (chiusi, aperti, semi-aperti) di estremi $a < b$ (anche infiniti)
- tutti gli insiemi che si ottengono per unione *finita* dei precedenti.

Detto un po' sommariamente, gli elementi di $\mathcal{F}$ sono tutte le unioni finite di intervalli. E' immediato che questa famiglia, oltre a contenere Ω, sia chiusa per unione finita; siccome l'intersezione di due intervalli è un intervallo o l'insieme $\emptyset$, la famiglia è anche chiusa per intersezione finita (grazie alle proprietà distributive); ed infine, il complementare di un intervallo è unione finita di intervalli, quindi (per le formule di De Morgan) la famiglia è chiusa anche per complementare.

Invece la famiglia degli intervalli non è un'algebra, perché non è chiusa per unione finita e per complementare.

Esempio 2 *Riprendendo l'esempio del paragrafo precedente, la famiglia $\mathcal{F}_T^0$ non è un'algebra, per colpa di due fatti. Da un lato, ci siamo ristretti a prendere intervalli I_j e questo pone gli stessi problemi appena visti su $\Omega = \mathbb{R}$; decidiamo allora che nella definizione di $\mathcal{F}_T^0$ usiamo pluri-intervalli I_j. Dall'altro, se ad esempio uniamo gli eventi $\{X_{t_1} > 0\}$ e $\{X_{t_2} > 0\}$, non riusciamo a scrivere questo insieme nella forma $\{X_{t_1} \in I_1, X_{t_2} \in I_2\}$. Allora chiamiamo $\mathcal{F}_T$ la famiglia formata da tutte le unioni finite di insiemi di $\mathcal{F}_T^0$. Questa è un'algebra.*

1.1.5 σ-algebre di eventi

Quasi tutta la matematica di un certo livello è basata su operazioni limite (derivate, integrali, e così via). Anche in probabilità dobbiamo poter effettuare operazioni limite per raggiungere una certa ricchezza di risultati. A livello di eventi, questa richiesta si traduce nel concetto di *σ-algebra* di insiemi: con questo nome si intendono le algebre $\mathcal{F}$ che siano chiuse anche per unione (ed automaticamente interesezione) *numerabile*.

Definizione 2 *Una σ-algebra di insiemi di Ω una famiglia $\mathcal{F} \subset \mathcal{P}(\Omega)$ che abbia le proprietà di un'algebra e tale che, se $A_1, ..., A_n, ...$ sono eventi appartenenti ad $\mathcal{F}$, allora*

$$\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}.$$

Le algebre $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathcal{F} = \{\emptyset, \Omega\}$, $\{\emptyset, \Omega, A, A^c\}$ sono anche σ-algebre*:* $\mathcal{P}(\Omega)$ lo è sempre e lo sono anche le algebre composte da un numero finito di eventi. Invece l'algebra dei pluri-intervalli non è una σ-algebra: l'unione di una quantità numerabile di intervalli non è riscrivibile necessariamente come unione finita di intervalli, si pensi a $\bigcup_{i=1}^{\infty} [2i, 2i+1]$.

La σ-algebra più comunemente usata su $\Omega = \mathbb{R}$, è quella degli insiemi *boreliani*. Essa è definita come la più piccola σ-algebra a cui appartengono tutti gli intervalli (e quindi tutti i pluri-intervalli). Premettiamo un fatto di carattere generale. Sia $\mathcal{F}^0$ una famiglia di sottoinsiemi di Ω. Con ragionamenti insiemistici si può verificare che il concetto di "più piccola σ-algebra" contenente $\mathcal{F}^0$ è un concetto bene definito ed univoco: definisce una ben precisa σ-algebra; purtroppo non costruttiva, cioè non immediatamente esprimibile con operazioni fatte a partire da $\mathcal{F}^0$. Detto questo, se prendiamo l'algebra $\mathcal{F}^0$ dei pluri-intervalli di $\Omega = \mathbb{R}$, o anche solo la famiglia $\mathcal{F}^0$ degli intervalli, allora esiste la più piccola σ-algebra che contiene $\mathcal{F}^0$ ed è, per definizione la σ-algebra dei boreliani. Essa contiene insiemi anche piuttosto complessi, come $\mathbb{Q}$ o l'insieme dei numeri irrazionali.

Se si conosce la definizione di insieme aperto, si può osservare che i boreliani sono anche la più piccola σ-algebra che contiene la famiglia degli insiemi aperti. Queste definizioni si estendono a $\mathbb{R}^n$, usando di nuovo gli aperti oppure altre famiglie $\mathcal{F}^0$ come le sfere $B(x_0, r) = \{x \in \mathbb{R}^n : \|x - x_0\| < r\}$, o rettangoli o altro.

Pur essendo vastissima, la σ-algebra dei boreliani, non coincide con $\mathcal{P}(\mathbb{R})$. Però, parlando in pratica, ogni insieme che si costruisca con operazioni usuali (in cui non includiamo l'uso dell'assioma della scelta), risulta essere un boreliano. Abbiamo detto che la σ-algebra dei boreliani è la più usata. Ci si chiederà perché non si usi più semplicemente $\mathcal{P}(\mathbb{R})$. la ragione è molto tecnica. La teoria dell'integrazione secondo Lebesgue, che permette di estendere il concetto di integrale $\int_A f(x)\,dx$ dal caso di funzioni f "facili" (es. continue a tratti) su insiemi A "facili" (es. pluri-intervalli), al caso di funzioni assai più irregolari su insiemi assai più complessi, non permette però di prendere qualsiasi insieme $A \subset \mathbb{R}$. Permette di considerare boreliani A ed anche qualcosa in più (gli insiemi misurabili secondo Lebesgue), ma non tutti gli insiemi. Quindi $\mathcal{P}(\mathbb{R})$ non risulta opportuna per poi sviluppare calcoli basati su integrali.

Nel discreto invece la famiglia delle parti va benissimo. Se ad esempio si considera $\Omega = \mathbb{N}$, si può tranquillamente prendere $\mathcal{F} = \mathcal{P}(\Omega)$, senza incorrere in problemi tecnici di tipo matematico.

1.1.6 Spazio probabilizzabile

La prima parte dello schema matematico è stata definita: un insieme (o "spazio") Ω ed una σ-algebra $\mathcal{F}$ di sottoinsiemi di Ω. In questo schema chiameremo *eventi* tutti gli elementi di $\mathcal{F}$.

Definizione 3 *Una coppia $(\Omega, \mathcal{F})$, dove Ω è un insieme ed $\mathcal{F}$ è una σ-algebra di suoi sottoinsiemi, si chiama spazio probabilizzabile.*

1.1.7 Probabilità

Definizione 4 *Su uno spazio probabilizzabile $(\Omega, \mathcal{F})$, si chiama probabilità (o distribuzione di probabilità, o misura di probabilità) ogni funzione*

$$P : \mathcal{F} \to [0, 1]$$

che soddisfa le seguenti due proprietà:

i) $P(\Omega) = 1$

ii) se $A_1, ..., A_n, ...$ è una famiglia finita, o una successione infinita, di eventi, a due a due disgiunti, allora

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n).$$

Scrivendo che P è una funzione da $\mathcal{F}$ in $[0, 1]$ intendiamo dire che calcoleremo P su ogni elemento del suo dominio $\mathcal{F}$, ottenendo come risultato un numero del codominio $[0, 1]$. Quindi, preso un qualsiasi evento $A \in \mathcal{F}$, calcoleremo la sua probabilità

$$P(A) \in [0, 1].$$

Per quanto rigiarda e due proprietà che deve soddisfare P, la prima è una convenzione di normalizzazione; osserviamo solo che la scrittura $P(\Omega)$ ha senso, in quanto abbiamo presupposto che $\Omega \in \mathcal{F}$. La seconda è la proprietà essenziale, che distingue il concetto di probabilità dagli altri (comune però ad alcuni altri concetti simili, come quello di misura o, fisicamente, di massa). Osserviamo che la scrittura $P\left(\bigcup_n A_n\right)$ ha senso, in quanto $\bigcup_n A_n \in \mathcal{F}$ per la proprietà di σ-algebra. Disgiunti a due a due significa $A_i \cap A_j = \varnothing$ per ogni $i \neq j$. La proprietà (ii) si chiama σ-*additività* (e semplicemente *additività* nel caso di un numero finito di insiemi).

Per avere un modello intuitivo di grande aiuto, si può pensare ad una distribuzione di massa su una regione Ω, normalizzata in modo che la massa totale sia uno. Se prendiamo sottoinsiemi disgiunti di Ω, la massa della loro unione è la somma delle masse.

Per inciso, esistono varie generalizzazioni del concetto di probabilità, che abbandonano la richiesta $P(A) \in [0,1]$, ma in genere mantengono la σ-additività. La generalizzazione più nota è quella in cui si richiede solo $P(A) \geq 0$ (eventualmente infinito), nel qual caso si parla di misura; l'esempio a tutti noto è la misura euclidea sulla retta, o sul piano, o nello spazio, ecc. (detta misura di Lebesgue, nella sua accezione σ-additiva su un'opportuna σ-algebra $\mathcal{F}$ molto ampia, detta degli insiemi misurabili secondo Lebesgue). Ma con l'ispirazione della carica elettrica al posto della massa si può costruire la nozione di misura con segno, in cui $P(A)$ può avere anche segno negativo, ed infine anche il caso vettoriale in cui $P(A)$ è un vettore di un certo spazio, sempre σ-additivo rispetto ad A. Non tratteremo queste generalizzazioni, ma può essere utile sapere che si possono sviluppare.

Per esercizio si può cercare di dimostrare che:

- $A \subset B$ implica $P(A) \leq P(B)$
- $P\left(A^c\right) = 1 - P(A)$
- $P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) - P\left(A \cap B\right)$.

Ad esempio la seconda segue subito dal fatto che $\Omega = A \cup A^c$, che sono disgiunti, quindi

$$1 = P\left(\Omega\right) = P\left(A\right) + P\left(A^c\right)$$

da cui $P\left(A^c\right) = 1 - P\left(A\right)$.

Concludiamo questo paragrafo osservando che abbiamo definito la struttura fondamentale del calcolo delle probabilità, ovvero:

Definizione 5 *Si chiama spazio probabilizzato una terna $(\Omega, \mathcal{F}, P)$, dove Ω è un insieme, $\mathcal{F}$ è una σ-algebra di sottoinsiemi di Ω e P è una probabilità.*

Naturalmente in ogni esempio dovremo (o dovremmo) specificare chi sono esattamente questi tre oggetti; indipendentemente dall'esempio specifico, essi devono però soddisfare i requisiti elencati sopra ($\mathcal{F}$ chiusa per operazion i numerabili, P che sia σ-additiva), dai quali derivano i vari teoremi del calcolo delle probabilità, validi in ogni esempio. Sottolineiamo che la specifica quantitativa di P nei singoli esempi può essere assai laboriosa e non discende assolutamente in modo automatico dalle regole (i) e (ii), quindi lo schema descritto fino ad

ora è solo un vago contenitore di idee astratte. Queste regole generali (insieme ad altre che vedremo relative al concetto di probabilità condizionale e indipendenza) servono di solito a calcolare la probbilità di certi eventi a partire da quella di altri; ma da qualche parte bisogna introdurre informazioni specifiche di ciascun esempio, da cui partire.

Osservazione 1 *Se $\mathcal{F}^0$ è solo un'algebra e $P : \mathcal{F}^0 \to [0,1]$ soddisfa $P(\Omega) = 1$ e*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right) \tag{1.1}$$

quando gli insiemi $A_i \in \mathcal{F}^0$ sono a due a due disgiunti, allora diciamo che P è una probabilità finitamente additiva.

1.1.8 Probabilità associata ad una densità

Definizione 6 *Si chiama densità di probabilità (che abbrevieremo con pdf, dall'inglese) ogni funzione $f : \mathbb{R} \to \mathbb{R}$ avente le seguenti due proprietà:*

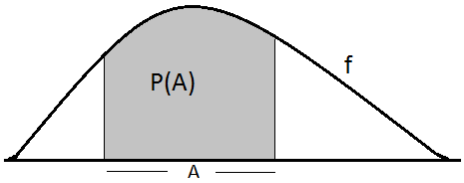
$$f\left(x\right) \geq 0 \quad \textit{per ogni } x$$

$$\int_{-\infty}^{+\infty} f\left(x\right) dx = 1.$$

L'integrale ora scritto è un integrale improprio. Ricordiamo che nel caso di funzioni positive, un integrale improprio può solo convergere o divergere a $+\infty$. Per una densità esso deve convergere (cosa non ovvia) ed avere convenzionalmente valore 1. Supponiamo per semplicità che f sia Riemann integrabile su ogni intervallo limitato, essendo questa la teoria usualmente appresa nei corsi di Ingegnera. Si ricordi che le funzioni continue, o anche solo continue a tratti, sono Riemann integrabili, e tali sarano tutti i nostri esempi.

Definizione 7 *Data una pdf f, dato un intervallo A o più in generale un insieme A che sia unione finita di intervalli, poniamo*

$$P\left(A\right) = \int_A f\left(x\right) dx.$$



Ad esempio

$$P\left([10,\infty)\right) = \int_{10}^{+\infty} f\left(x\right) dx.$$

Abbiamo così definito una funzione P : dall'algebra $\mathcal{F}^0$ dei pluri-intervalli a valori reali. Siccome $f \geq 0$, vale $P(A) \geq 0$. Siccome $A \subset \mathbb{R}$, vale $\int_A f(x)\,dx \leq \int_{\mathbb{R}} f(x)\,dx$ e quindi $P(A) \leq 1$. Pertanto è vero che

$$P : \mathcal{F}^0 \to [0,1]\,.$$

Ovviamente vale $P(\Omega) = 1$. Con ragionamenti elementari ma noiosi da scrivere, sei verifica che P è finitamente additiva, cioè vale (1.1) se i pluri-intervalli A_i sono a due a due disgiunti. In questo modo abbiamo definito una $P : \mathcal{F}^0 \to [0,1]$ finitamente additiva. Usando la teoria dell'integrazione secondo Lebesgue, si può estendere P ad una probabilità sulla σ-algebra $\mathcal{F}$ dei boreliani. Questa estensione però esula dal nostro corso (qualche elemento si può vedere alla sezione 1.2.22).

Esempio 3 *Dati due numeri reali C, λ, chiediamoci quando la funzione*

$$f(x) = \begin{cases} Ce^{-\lambda x} & per \quad x \geq 0 \\ 0 & per \quad x < 0 \end{cases}$$

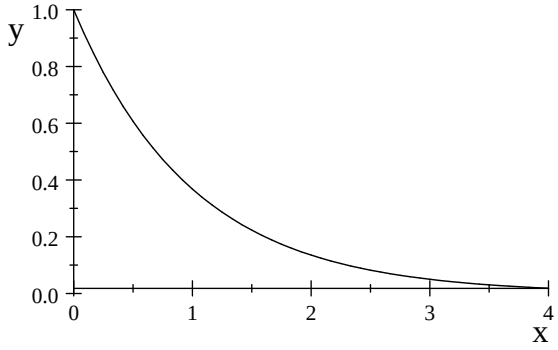
è una densità di probabilità. Dovendo essere $f \geq 0$, dev'essere $C \geq 0$. L'integrale non può essere finito se $\lambda = 0$ o ancor perggio se $\lambda < 0$ (in entrambi i casi la funzione non tende a zero per $x \to \infty$ ed anzi è maggiore di una costante positiva, se $C > 0$) quindi esaminiamo solo il caso $\lambda > 0$. Vale

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_0^{\infty} Ce^{-\lambda x}dx = -C\int_0^{\infty} \frac{d}{dx}\frac{e^{-\lambda x}}{\lambda}dx = -C\left[\frac{e^{-\lambda x}}{\lambda}\right]_0^{\infty} = \frac{C}{\lambda}$$

dove l'interpretazione del calcolo di $e^{-\lambda x}$ per $x = +\infty$ è quella di limite

$$\lim_{x \to +\infty} e^{-\lambda x} = 0.$$

Quindi l'integrale è finito per ogni $\lambda > 0$ e la funzione è una densità se $C = \lambda$. La densità così trovata si chiama densità esponenziale di parametro λ.



Densità esponenziale, $x \geq 0$, $\lambda = 1$

1.1.9 Probabilità associata ad una densità discreta

Definizione 8 *Si chiama densità di probabilità discreta ogni successione* $\{p_n\}_{n\in\mathbb{N}}$ *avente le seguenti due proprietà:*

$$p_n \geq 0 \quad \textit{per ogni } n$$
$$\sum_{n=0}^{\infty} p_n = 1.$$

Al posto degli integrali impropri qui serve la teoria delle serie a termini positivi. Si ricordi che una serie a termini positivi può solo convergere o divergere a $+\infty$.

Le due proprietà precedenti implicano

$$p_n \leq 1 \quad \text{per ogni } n.$$

Questo non era vero per le densità f (ad esempio, per l'esponenziale, $f(0) = \lambda$, che può assumere qualsiasi valore positivo).

Definizione 9 *Data una densità di probabilità discreta* $\{p_n\}_{n\in\mathbb{N}}$ *, per ogni insieme* $A \subset \mathbb{N}$ *poniamo*

$$P(A) = \sum_{n\in A} p_n.$$

Questo definisce una probabilità sullo spazio probabilizzabile $(\Omega, \mathcal{F})$ *dove* $\Omega = \mathbb{N}$ *ed* $\mathcal{F}$ *è la* σ*-algebra* $\mathcal{P}(\mathbb{N})$ *di tutte le parti.*

Anche $\sum_{n\in A}$ può essere una somma infinita; in ogni caso rientra nella teoria delle serie a termini positivi. La verifica che $P(A) \in [0,1]$ è identica al caso di una pdf f. La finita additività è elementare ma noiosa. La numerabile additività richiede un po' più di lavoro sulle serie numeriche a termini positivi, comunque non difficile, che però omettiamo.

Esempio 4 *Dati due numeri reali* C, p*, chiediamoci quando la successione* $\{p_n\}_{n\in\mathbb{N}}$ *definita da*

$$p_n = C(1-p)^n, \qquad n = 0, 1, 2, ...$$

è una densità di probabilità discreta. Per convenzione, se $p = 1$*, si intende che* $(1-p)^0 = 1$*, mentre ovviamente* $(1-p)^n = 0$ *per* $n \geq 1$*. Siccome per* p *pari il termine* $(1-p)^n$ *è positivo, dev'essere* $C \geq 0$*. A quel punto,* $(1-p)^n$ *dev'essere positivo per ogni* n*, quindi* $1-p \geq 0$*, cioè* $p \leq 1$*. Dobbiamo ora capire quando converge la serie* $\sum_{n=0}^{\infty}(1-p)^n$*. Essendo una serie geometrica, converge se e solo se* $|1-p| < 1$*, cioè se* $1-p < 1$ *(già sappiamo che* $1-p \geq 0$*), ovvero* $p > 0$*. In tale caso vale*

$$\sum_{n=0}^{\infty}(1-p)^n = \frac{1}{1-(1-p)} = \frac{1}{p}$$

e quindi dev'essere $C = p$*. Riassumendo, per* $p \in (0,1]$*,* $C = p$*, la successione data è una densità di probabilità discreta. Viene chiamata densità geometrica di parametro* p*, limitatamente al caso* $p \in (0,1)$*.*

Esempio 5 *Spazi di esiti equiprobabili. L'esempio più semplice ma anche assai utile è quello di un insieme finito Ω composto di N elementi, $\mathcal{F} = \mathcal{P}(\Omega)$ e P definita così:*

$$P(A) = \frac{|A|}{N} = \frac{|A|}{|\Omega|}$$

dove abbiamo indicato con $|A|$ la cardinalità di A, ovvero il numero di elementi di A. A parole, la probabilità di A è il rapporto tra il numero dei casi favorevoli e quello dei casi possibili.

Si può riconoscere che vale l'additività di P (e $P(\Omega) = 1$) quindi P è una probabilità. Sottolineiamo che se $\omega \in \Omega$ è un evento elementare, allora vale

$$P(\omega) = \frac{1}{N}.$$

Da qui deriva il nome di spazio di esiti equiprobabili. Per quanto semplice possa sembrare questo esempio, è abbastanza vero che ogni costruzione più elaborata del calcolo delle probabilità affonda le sue radici in qualche modo negli spazi equiprobabili.

Osservazione 2 *Spazi di probabilità finiti. Un po' più generale del precedente è il caso di un insieme finito Ω composto di N elementi, $\mathcal{F} = \mathcal{P}(\Omega)$, ma con P non necessariamente uniforme. Vedremo tra un attimo un esempio. Qui osserviamo solo una proprietà importante: la conoscenza di P (ovvero il valore di $P(A)$ per qualsiasi $A \subset \Omega$) equivale alla conoscenza del valore di P sugli eventi elementari. In altre parole, se conosciamo $P(\omega)$ per qualsiasi $\omega \in \Omega$, allora possiamo calcolare $P(A)$ per qualsiasi $A \subset \Omega$ tramite addizione:*

$$P(A) = \sum_{\omega \in A} P(\omega).$$

La formula vale per l'additività di P. La somma è finita, per ipotesi sullo spazio. Tuttavia quanto detto in questa osservazione vale esattamente anche nel caso di Ω infinito numerabile, nel qual caso la somma che calcola $P(A)$ può essere una serie numerica (comunque è una serie a termini positivi convergente).

Osservazione 3 *Insistendo sull'osservazione precedente, notiamo che per costruire un esempio di spazio probabilizzato finito, dopo aver specificato Ω e scelto $\mathcal{F} = \mathcal{P}(\Omega)$, basta introdurre una sequenza di numeri $\{p(\omega)\}_{\omega \in \Omega}$ tali che*

$$p(\omega) \in [0,1] \quad \text{per ogni } \omega \in \Omega$$

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

A partire da essi si definisce poi

$$P(A) = \sum_{\omega \in A} p(\omega)$$

per ogni $A \subset \Omega$ e si verifica facilmente che P è una probabilità.

Esempio 6 *Fissato un intero positivo n, consideriamo l'insieme Ω di tutte le sequenze $(x_1, ..., x_n)$ composte di zeri ed uni. A volte si usa scrivere*

$$\Omega = \{0, 1\}^n$$

ovvero l'insieme di tutte le applicazioni da un insieme di n elementi in $\{0, 1\}$. Ω è un insieme finito, con 2^n elementi. Definiamo un'interessante probabilità P su $\mathcal{F} = \mathcal{P}(\Omega)$. Per quanto detto nella precedente osservazione, basta che assegniamo la probabilità ad ogni sequenza $(x_1, ..., x_n)$ in modo da avere somma uno. Fissato un numero $p \in [0, 1]$, posto $q = 1 - p$, detto k il numero di uni nella sequenza $(x_1, ..., x_n)$, poniamo

$$p(x_1, ..., x_n) = p^k q^{n-k}.$$

Sono numeri in $[0, 1]$. La loro somma è pari a

$$\sum_{k=0}^{n} n_k p^k q^{n-k}$$

dove n_k è il numero di sequenze con k uni. Chiariremo in un paragrafo a parte che questo numero è il coefficiente binomiale $\binom{n}{k}$. Dobbiamo allora calcolare

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

Questa somma vale uno ricordando la formula del binomio di Newton:

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

Lo spazio probabilizzato appena introdotto è molto ricco e costituisce la base per un enorme numero di considerazioni teoriche e di applicazioni pratiche.

Osservazione 4 *Una probabilità P è definita su una σ-algebra $\mathcal{F}$, non su uno spazio Ω come in genere si è portati a dire intuitivamente. In genere non è pericoloso fare questa piccola confusione di linguaggio; piuttosto, in alcuni casi è utile rammentare questa specifica, quando si studiano problemi avanzati con diverse σ-algebra in azione contemporaneamente.*

1.1.10 Probabilità condizionale

Supponiamo di aver scelto una terna $(\Omega, \mathcal{F}, P)$ per descrivere un problema concreto. Supponiamo poi di venire a conoscenza di un'informazione aggiuntiva che prima ci era ignota, esprimibile nel fatto che un certo evento $B \in \mathcal{F}$ si è verificato.

Ad esempio, consideriamo nuovamente il problema della spedizione e ricezione di un simbolo 0,1 attraverso un canale di comunicazione, in cui inizialmente introduciamo lo schema $(\Omega, \mathcal{F}, P)$ quando non sappiamo né che simbolo è stato trasmesso né quale è stato ricevuto. Ricordiamo che Ω è l'insieme composto dai quattro elementi $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$. In

questo esempio $\mathcal{F} = \mathcal{P}(\Omega)$, mentre di P ancora non abbiamo parlato, ma supponiamo di averla fissata.

Come dicevamo, supponiamo che un evento B si sia verificato. Nell'esempio, potrebbe essere l'evento: "il simbolo ricevuto è 1". Questa è solo un'informazione parziale, non esaurisce ciò che vorremmo sapere del problema aleatorio, ma certamente è un'importante informazione in più.

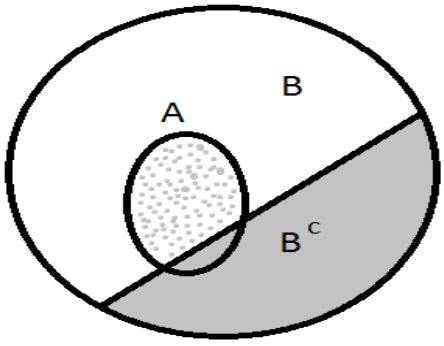
Matematicamente, accade questo: la nuova informazione contenuta nel fatto che B si è verificato, modifica la probabilità di tutti gli altri eventi. Ogni evento A aveva inizialmente probabilità $P(A)$; ora ha una nuova probabilità che indicheremo con

$$P(A|B)$$

(e leggeremo "probabilità di A sapendo B", o "condizionata a B"). La formula che è stata scelta per calcolarla, o se si vuole come sua definizione, è la seguente:

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

Come ogni definizione contiene un certo grado di arbitrarietà, ma è comunque ben motivata sia dalla sensatezza negli esempi, sia dal seguente ragionamento generale. Si pensi ad Ω. Nel momento in cui sappiamo che B si è verificato, l'insieme B^c non può più verificarsi, quindi il nostro universo si restringe a B stesso, diventa $\Omega' = B$. Preso un qualsiasi evento A, la parte di A in B^c non può più verificarsi, mentre sopravvive la parte di A in B, pari a $A \cap B$. In altre parole, nel passaggio (restrizione) da Ω ad Ω', l'insieme A si trasforma in $A \cap B$. Sarebbe naturale poi dire che la probabilità $P(A)$ si trasforma in $P(A \cap B)$. Però la nuova probabilità così trovata avrebbe il difetto di non valere 1 sul nuovo universo: $P(\Omega') = P(B)$, diverso da 1 in generale. Va allora normalizzata ad 1, dividendo per $P(B)$. Ecco come si arriva all'espressione $\frac{P(A \cap B)}{P(B)}$ partendo da $P(A)$.



Solo la parte a puntini sopravvive come eventualità quando sappiamo che vale B

Osserviamo che nella definizione di $P(A|B)$ bisogna supporre che sia $P(B) > 0$ per dare senso alla frazione. Tuttavia, quando $P(B) = 0$, anche $P(A \cap B) = 0$ (in quanto $A \cap B \subset B$), quindi l'espressione è del tipo $\frac{0}{0}$, che non ha un senso elementare, algebrico, ma potrebbe avere un senso limite, magari prendendo una successione di insiemi $B_n \to B$ con opportune

proprietà. In molti casi questo tipo di ragionamento funziona e produce nozioni utilissime di probabilità condizionata in un senso generalizzato. Per ora non approfondiamo questo argomento.

Ricordiamo che P era, rigorosamente parlando, una funzione. Analogamente è molto utile pensare a $P(.|B)$ come ad una funzione, per B fissato: funzione dell'evento A che mettiamo nell'espressione $P(A|B)$. Si dimostra che la funzione $P(.|B)$ (con B fissato) è una probabilità, σ-additiva.

1.1.11 Indipendenza

Prima di conoscere un certo B, un evento A ha probabilità $P(A)$. Dopo, ha probabilità $P(A|B)$.

Quando questi due valori sono uguali, ovvero

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

siamo portati a dire che B non influenza A. Un esempio semplice da capire è quello del lancio di due dadi: se B è l'evento "nel primo lancio esce 6" e A è l'evento "nel secondo lancio esce 6", è chiaro intuitivamente che B non può influenzare A in alcun modo.

Osservazione 5 *Un'osservazione semi-seria. Una credenza ingenua è che se in un lancio esce 6, nel successivo sia più difficile che esca di nuovo 6. Più formalmente, concordando che a priori la probabilità che al secondo lancio esca 6 è 1/6, alcuni pensano che, una volta noto che al primo lancio è uscito 6, la probabilità che esca 6 al secondo lancio è minore di 1/6. Questo è completamente assurdo se si pensa alla fisica del lancio del dado. Casomai, si potrebbe dubitare che valga proprio il contrario: se il dado non è perfetto, il fatto che sia uscito 6 al primo lancio potrebbe essere un indizio che il dado è sbilanciato a favore di certe facce, inclusa la faccia 6; ma allora al secondo lancio la probabilità che esca 6 è un po' maggiore di 1/6!*

La condizione $P(A|B) = P(A)$ sembra asimmetrica, mentre non lo è. Siccome (usando la simmetria di $A \cap B$)

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

da $P(A|B) = P(A)$ si ricava $P(B|A) = P(B)$, ovvero che A non influisce su B. Quindi si può parlare di *indipendenza* tra A e B, simmetricamente. Per dare una veste simmetrica anche alla formulazione matematica, basta osservare che l'uguaglianza

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

è equivalente alle precedenti (per esercizio). Oltre ad essere simmetrica ha il pregio di non obbligare alle specifiche del tipo $P(A) > 0$ o $P(B) > 0$ insite nella definizione di probabilità condizionale. Arriviamo quindi alla seguente definizione, nel caso di due eventi, che generalizziamo al caso di n eventi.

Definizione 10 *Due eventi A e B si dicono indipendenti se*

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

Gli eventi $A_1, ..., A_n$ si dicono indipendenti se

$$P(A_{i_1} \cap ... \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

per ogni scelta di $i_1 < ... < i_k \in \{1, ..., n\}$.

Osservazione 6 *Se A e B sono indipendenti allora anche A e B^c sono indipendenti. Quindi (cambiando nome agli insiemi) anche A^c e B, ed anche A^c e B^c. Queste affermazioni sono tutte equivalenti. Basta dimostrare la prima. Vale*

$$P(A \cap B^c) + P(A \cap B) = P(A)$$

da cui, supponendo A e B indipendenti,

$$P(A \cap B^c) + P(A)\,P(B) = P(A)$$

da cui

$$P(A \cap B^c) = P(A)\,(1 - P(B)) = P(A)\,P(B^c).$$

Quindi A e B^c sono indipendenti.

Osservazione 7 *Grazie al fatto che nella definizione di indipendenza di n eventi abbiamo preso gli indici $i_1 < ... < i_k$ in modo arbitrario (e non semplicemente $i_1 = 1$, ... , $i_n = n$), si può dimostrare che vale l'analogo dell'osservazione appena fatta anche nel caso di n eventi (cioè si possono sostituire alcuni degli eventi con i loro complementari). Vediamolo limitamente ad un esempio (il caso generale è solo simbolicamente più pesante):*

$$\begin{aligned} P(A \cap B \cap C^c) + P(A \cap B \cap C) &= P(A \cap B) \\ P(A \cap B \cap C^c) + P(A)\,P(B)\,P(C) &= P(A)\,P(B) \\ P(A \cap B \cap C^c) &= P(A)\,P(B)\,(1 - P(C)) \\ &= P(A)\,P(B)\,P(C^c). \end{aligned}$$

Questa invarianza per complementare è un requisito abbastanza irrinunciabile se si pensa agli esempi (come il seguente).

Esempio 7 *Supponiamo che un sistema S sia composto da tre sottosistemi S_1, S_2, S_3. La probabilità che S_i si rompa è p_i. Supponendo che il funzionare o meno dei sottosistemi sia indipendente, calcolare la probabilità che si rompa S. Soluzione: introduciamo gli eventi*

$$\begin{aligned} R &= \text{“il sistema } S \text{ si rompe”} \\ R_i &= \text{“il sistema } S_i \text{ si rompe”}, \qquad i = 1, 2, 3 \end{aligned}$$

ed immaginiamo, anche senza formalizzarlo, che questi siano eventi (sottoinsiemi) di un certo universo Ω su cui sia definita una certa probabilità P. Vale, per ipotesi,

$$P(R_i) = p_i, \qquad i = 1, 2, 3$$

ed inoltre gli eventi (eventualmente complementati) R_i sono indipendenti. Dobbiamo calcolare $P(R)$. Vale

$$P(R) = 1 - P(R^c) = 1 - P(R_1^c \cap R_2^c \cap R_3^c)$$

(l'evento R^c è la non rottura del sistema S, che avviene quando c'è la non rottura simultanea di tutti i sottosistemi)

$$= 1 - P(R_1^c) P(R_2^c) P(R_3^c) = 1 - (1-p_1)(1-p_2)(1-p_3).$$

Osservazione 8 *Nell'esempio precedente si osservi che R non è uguale a $R_1 \cap R_2 \cap R_3$, quindi non basta applicare l'indipendenza a questi eventi. Vale invece*

$$R = R_1 \cup R_2 \cup R_3$$

ma questi non sono eventi disgiunti, quindi non si può applicare la regola della somma. Si potrebbe scrivere R come unione disgiunta, elencando ad esempio tutti i casi che compongono R ($R_1 \cap R_2^c \cap R_3^c$ ecc.), usando poi l'additività. Questo è l'approccio più elementare (e si consiglia di tenerlo presente per i casi di emergenza) ma è lungo e faticosamente estendibile a numerosità maggiori.

1.1.12 Formula di fattorizzazione

Definizione 11 *Chiamiamo partizione (finita, misurabile) di Ω una famiglia di insiemi $B_1, ..., B_n \in \mathcal{F}$ tali che $B_i \cap B_j = \varnothing$ per ogni $i \neq j$ e $\bigcup_{i=1}^{n} B_i = \Omega$.*

Una partizione, in parole povere, è una suddivisione di Ω in insiemi disgiunti. Se $B \in \mathcal{F}$, gli insiemi B, B^c formano una partizione.

Teorema 1 (Formula di fattorizzazione) *Se $B_1, ..., B_n$ è una partizione di Ω e $P(B_i) > 0$ per ogni $i = 1, ..., n$, allora*

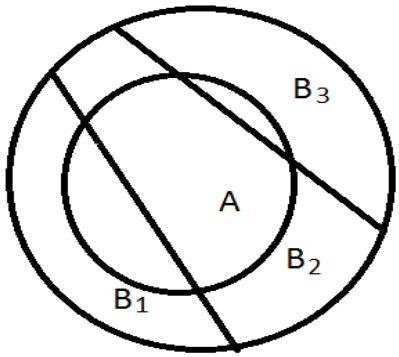
$$P(A) = \sum_{i=1}^{n} P(A|B_i) P(B_i)$$

per ogni evento A.

Proof. Per l'additività di P vale

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

in quanto gli insiemi $A \cap B_i$ sono disgiunti (lo sono i B_i) e vale $A = \bigcup_{i=1}^{n} (A \cap B_i)$. Inoltre, essendo per definizione $P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)}$, vale $P(A \cap B_i) = P(A|B_i) P(B_i)$. Sostituendo nella formula precedente si trova il risultato desiderato. ■



Raffigurazione della fattorizzazione di un evento A

Le partizioni, negli esempi, descrivono le diverse alternative. Vediamo un esempio.

Esempio 8 *Una ditta commercia vino bianco (B) e rosso (R), richiesti da clienti in Francia (F) e in Germania (G). 1/3 delle richieste arriva dalla Francia, 2/3 delle richieste dalla Germania. I 3/4 delle richieste provenienti dalla Francia sono di vino bianco, 1/4 delle richieste sono di vino rosso. Viceversa, 1/4 delle richieste provenienti dalla Germania sono di vino bianco, 3/4 delle richieste sono di vino rosso.*

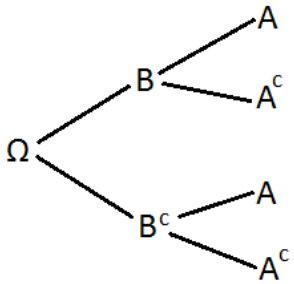
Calcolare la probabilità che un generico ordine riguardi il vino bianco.

Soluzione. Dati: $P(F) = 1/3$, $P(G) = 2/3$, $P(B|F) = 3/4$, $P(R|F) = 1/4$, $P(B|G) = 1/4$, $P(R|G) = 3/4$. Quindi

$$P(B) = P(B|F)P(F) + P(B|G)P(G) = \frac{3}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}.$$

In molti esempi, come il precedente, il problema concreto fa sì che sano note certe probabilità condizionali, piuttosto che certe probabilità. Si deve immaginare di avere Ω e la probabilità P, che però è incognita; P definisce le probabilità condizionali ed alcune di queste sonoo note dai dati dell'esempio; permettendo di risalire alle probabilità P.

Un'efficace raffigurazione della formula di fattorizzazione si ha disegnando l'*albero degli eventi*:



Nel disegno, per non creare confusione, non abbiamo indicato i valori lungo i rami. Su ogni ramo va scritta la probabilità condizionale ad esso relativa. Ad esempio, sul ramo che porta da Ω a B, va scritto $P(B)$ (che è $P(B|\Omega)$); sul ramo che porta da B ad A, va scritto

$P(A|B)$; e così via. Se si considera un percorso, ad esempio quello che porta da Ω a B e poi ad A, la sua probabilità è il prodotto delle probabilità condizionali lungo i rami, $P(A|B)\,P(B)$ in questo esempio. La probabilità (totale) di A si ottiene sommando lungo tutti i percorsi che portano ad A, cioè lungo i due percorsi

$$\begin{aligned}\Omega &\rightarrow B \rightarrow A\\ \Omega &\rightarrow B^c \rightarrow A.\end{aligned}$$

La formula di fattorizzazione viene anche detta *delle probabilità totali*, in quanto permette di calcolare la probabilità "totale" di un evento A a partire da quelle condizionali.

1.1.13 Formula di Bayes e formula di fattorizzazione

Teorema 2 (Formula di Bayes) *Se $P(A) > 0$ e $P(B) > 0$, allora*

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

Proof.

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

in quanto $P(A|B)\,P(B) = P(A \cap B)$ per definizione di $P(A|B)$. ■

In quasi tutti gli esempi, il valore del denominatore $P(A)$ non è noto in partenza, per cui va calcolato con la formula di fattorizzazione rispetto ad una partizione $B_1, ..., B_n$ (di cui B di solito è uno degli elementi). La formula di Bayes diventa:

$$P(B|A) = \frac{P(A|B)\,P(B)}{\sum_i P(A|B_i)\,P(B_i)}.$$

Esempio 9 *La preparazione di uno studente può essere scarsa, buona, ottima. Ciò non è direttamente misurabile, quindi lo studente viene sottoposto ad un test a crocette, che si intende superato se il punteggio è $\geq$10. Se la sua preparazione è scarsa, la probabilità che totalizzi almeno 10 negli esercizi è pari a 0.3. Se è buona, è pari a 0.8, se è ottima è pari a 0.995. Prima dello scritto il docente non ha informazioni sullo studente e decide di considerare equiprobabili le tre possibilità circa la sua preparazione. Supponiamo poi che lo studente esegua gli esercizi e prenda meno di 10; il docente, ora, che probabilità gli attribuisce di avere preparazione scarsa? E di avere una preparazione almeno buona (cioè buona o ottima)?*

Soluzione. Con ovvie notazioni per gli eventi in gioco,

$$\begin{aligned}P(<10) &= P(<10|S)P(S) + P(<10|B)P(B) + P(<10|O)P(O)\\ &= \frac{1}{3}\,(0.7 + 0.2 + 0.005) = 0.\,301\,67\end{aligned}$$

quindi

$$P(S|<10) = \frac{P(<10|S)P(S)}{P(<10)} = \frac{0.\,233\,33}{0.\,301\,67} = 0.\,773\,46.$$

Infine, usando la regola dell'evento complementare, la probabilità richiesta è $1 - 0.773\,46 = 0.\,226\,54$.

Osservazione 9 *Nei problemi "bayesiani", ci sono delle "probabilità a priori" e delle "probabilità a posteriori". Nell'esempio, la probabilità a priori che lo studente avesse preparazione scarsa era 1/3, a poteriori - cioè dopo aver osservato l'esito del test - è 0.773.*

Osservazione 10 *Accade spesso che le probabilità a priori vengano scelte uguali - 1/3 nell'esempio - come riflesso della mancanza di informazioni.*

La formula di Bayes permette di calcolare $P(B|A)$ a partire da $P(A|B)$ (ed altri due termini). E' interessante la sua struttura logica: se conosciamo come B influenza A, ovvero conosciamo $P(A|B)$, allora possiamo calcolare come A influenza B. C'è una sorta di inversione causale. Se immaginiamo che B sia una possibile causa ed A un possibile effetto, la logica normale è quella di conoscere come la causa B influenza l'effetto A, quindi conoscere $P(A|B)$. La formula di Bayes permette di risalire alle cause a partire da osservazioni sui loro effetti. Precisamente, osservato l'effetto A, permette di calcolare la probabilità che esso derivi dalla causa B.

In genere, in questo schema causa-effetto, si ha a che fare con diverse possibili cause, diverse alternative, che formano una partizione $B_1, ..., B_n$. Osservato A, vorremmo risalire alla causa che lo ha provocato. Tutte le cause B_i possono aver provocato A, quindi ciò che possiamo fare è calcolare le probabilità delle diverse cause B_i condizionate ad A e decidere che la causa è quella più probabile (è uno schema di teoria delle decisioni; si veda anche il Capitolo 6).

In quest'ottica, serve solo confrontare i valori di $P(B_i|A)$ al variare di $i = 1, ..., n$. Per questo scopo non serve calcolare il denominatore $P(A)$ della formula di Bayes, che è uguale per tutti. Basta quindi confrontare i numeri $P(A|B_i)\,P(B_i)$.

Poi, nel caso molto comune in cui gli eventi B_i siano equiprobabili, basta confrontare $P(A|B_i)$. Si pensi a tutte queste semplificazioni sull'alvero degli eventi.

1.1.14 Calcolo combinatorico

Il calcolo combinatorico fornisce idee e regole per calcolare la cardinalità $|A|$ di un insieme finito A, cosa essenziale quando di calcola la probabilità di un evento in uno spazio di esiti equiprobabili.

Ricordiamo il significato di n fattoriale:

$$n! := n\,(n-1)\cdots 2\cdot 1$$

e convenzionalmente $0! = 1$. Più formalmente $n!$ è definito per ricorrenza da

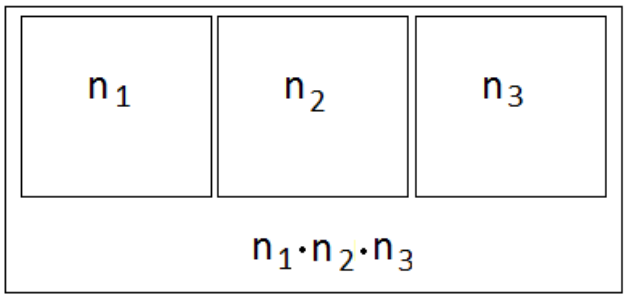
$$\begin{aligned} 1! &= 1 \\ n! &= n\cdot(n-1)! \qquad \text{per } n \geq 2. \end{aligned}$$

Ricordiamo inoltre il significato dei coefficienti binomiali:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n\,(n-1)\cdots(n-k+1)}{k!}$$

definiti per $n \geq 1, k = 0, ...n$. Vale $\binom{n}{k} = \binom{n}{n-k}$, $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{1} = n$ e così via. Non è ovvio che $\binom{n}{k}$ sia un intero; un modo di scoprirlo è attraverso un teorema che vedremo proa poco.

Alla base di molti fatti c'è il cosidetto *principio di enumerazione*. Esso asserisce che se si svolgono due esperimenti successivi, il primo con n possibili risultati diversi ed il secondo con m possibili risultati diversi, allora le coppie di risultati possibili sono $m \cdot n$. E' davvero un principio ovvio, ma permette di risolvere un grandissimo numero di problemi. Naturalmente si sottointende che vale anche per una sequenza formata da più di due esperimenti; ad esempio per tre esperimenti, se nel primo ci sono n_1 risultati possibili, nel secondo n_2 e nel terzo n_3, il numero totale di risultati possibili della terna di esperimenti è $n_1 n_2 n_3$. Vediamolo all'opera.



Principio di enumerazione

Esempio 10 *Quante sono le stringhe di n simboli, $(x_1, ..., x_n)$, in cui ciascun simbolo x_i può assumere k possibili valori diversi? Il risultato è*

$$k^n.$$

Infatti, usiamo il principio di enumerazione immaginando che la scelta del primo simbolo sia il primo esperimento, la scelta del secondo simbolo il secondo esperimento e così via. Nel primo esperimento ci sono k risultati possibili, nel secondo pure, e così via, per cui il numero di risultati possibili della sequenza di esperimenti è il prodotto $k \cdot k \cdots k = k^n$.

Esempio 11 *Sia Ω l'insieme di tutte le applicazioni $f : \{1, ..., n\} \to \{1, ..., k\}$. Allora*

$$|\Omega| = k^n.$$

Basta riconoscere che Ω è in corrispondenza biunivoca con l'insieme delle stringhe descritto nell'esempio precedente. Infatti, assegnare una funzione $f : \{1, ..., n\} \to \{1, ..., k\}$ equivale a dire, per ciascun eleemento del dominio (vedi ciascun simbolo x_i dell'esempio precedente), quale valore tra 1 e k esso assume.

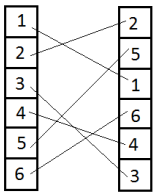
Esempio 12 *Dato un insieme finito Ω_0 con n elementi, detto $\Omega = \mathcal{P}(\Omega_0)$ l'insieme delle parti di Ω_0, vale*

$$|\mathcal{P}(\Omega_0)| = 2^n.$$

Infatti, numeriamo gli elementi di Ω_0 *come* $\omega_1, ..., \omega_n$*. Ogni parte* $A \subset \Omega_0$ *si può mettere in corrispondenza con la stringa di zeri ed uni* $(x_1, ..., x_n)$ *in cui* $x_i = 1$ *se* $\omega_i \in A$*; oppure in corrispondenza con la funzione* $f : \Omega_0 = \{\omega_1, ..., \omega_n\} \to \{0, 1\}$ *che vale* 1 *nei punti di* A *(detta anche indicatrice di* A*,* $f = 1_A$*). Queste due corrispondenze sono biunivoche. Quindi* $\mathcal{P}(\Omega_0)$ *ha tanti elementi quante sono le stringhe* $(x_1, ..., x_n)$ *di zeri ed uni, ovvero* 2^n*.*

Definizione 12 *Chiamiamo permutazione di* n *elementi una qualsiasi applicazione biunivoca* $f : \{1, ..., n\} \to \{1, ..., n\}$*.*

Colloquialmente, una permutazione di n elementi è un possibile scambio del loro ordine.



Sia Ω l'insieme di tutte le permutazioni. Vale

$$|\Omega| = n!$$

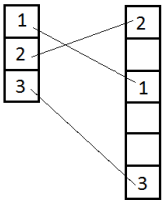
Per verificarlo basta pensare ai seguenti esperimenti: nel primo si sceglie dove mandare 1 e per questo ci sono n possibilità; nel secondo si sceglie dove mandare 2 e per questo ci sono $n - 1$ possibilità (la casella occupata da 1 non può più essere scelta); e così via.

Osservazione 11 *Nel principio di enumerazione, il numero* n_2 *può dipendere dal fatto che è stato svolto un primo esperimento - come in questo esempio delle permutazioni - ma non deve dipendere dall'esito del primo esperimento. Si pensi ad un gioco in cui prima si lancia un dado, poi, se è uscito un pari si lancia un secondo dado altrimenti una moneta. Il numero dei risultati possibili del secondo esperimento dipende dall'esito del primo. Non siamo quindi nell'ambito del principio di enumerazione.*

Esempio 13 *Dato un insieme di* n *oggetti diversi, in quanti modi diversi li possiamo ordinare? In altre parole, vogliamo costruire stringhe ordinate* $(x_1, ..., x_n)$ *in cui gli oggetti* $x_1, ..., x_n$ *sono diversi tra loro, presi da un insieme prefissato di* n *oggetti. Lo si può fare in* $n!$ *modi.*

Definizione 13 *Chiamiamo disposizione di* k *elementi in* n *posti una qualsiasi applicazione iniettiva* $f : \{1, ..., k\} \to \{1, ..., n\}$*. Dev'essere* $k \leq n$*.*

Colloquialmente, una disposizione di k elementi in n posti è un modo di disporre k oggetti diversi in n caselle diverse.



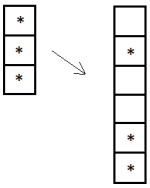
Sia Ω l'insieme di tutte le disposizioni di k elementi in n posti. Vale

$$|\Omega| = n\,(n-1)\cdots(n-k+1)$$

la verifica è identica al caso delle permutazioni. Si osservi che questo numero è il numeratore della riscrittura del coefficiente binomiale usata sopra.

Definizione 14 *Si chiama combinazione di k elementi in n posti ogni sottoinsieme A di $\{1,...,n\}$ avente cardinalità $|A| = k$. Dev'essere $k \leq n$.*

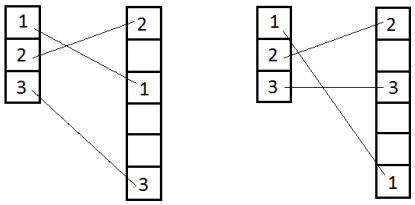
Colloquialmente, una combinazione di k elementi in n posti è una scelta di k posti tra gli n. Oppure un modo di disporre k oggetti uguali, indistinguibili, in n caselle diverse.



Teorema 3 *Sia Ω l'insieme di tutte le combinazioni di k elementi in n posti. Allora*

$$|\Omega| = \binom{n}{k}.$$

Proof. Sia Ω_0 l'insieme di tutte le disposizioni di k elementi in n posti. Introduciamo su Ω_0 una relazione di equivalenza: due disposizioni (cioè due applicazioni iniettive $f : \{1,...,k\} \to \{1,...,n\}$) sono equivalenti se hanno la stessa immagine; che individuano cioè lo stesso sottoinsieme del codominio $\{1,...,n\}$. Ad esempio, sono equivalenti quelle qui raffigurate:



Sia C una classe di equivalenza per questa relazione. Tutte le disposizioni della classe hanno la stessa immagine, e solo loro, per cui la classe si può mettere in corrispondenza biunivoca con l'immagine, cioè con un certo sottoinsieme A di $\{1,...,n\}$. Questi sottoinsiemi sono le combinazioni. Quindi le combinazioni sono tante quante le classi di equivalenza C. Indichiamo con x la loro cardinalità, che vogliamo calcolare.

Ogni classe C ha esattamente $k!$ elementi. Infatti, due disposizioni della stessa classe differiscono per una permutazione dell'immagine, cioè una permutazione di k elementi. Ci sono $k!$ modi di fare una tale permutazione.

Si noti che la cardinalità di C è la stessa per tutte le classi C. Allora, $k!$ (cardinalità di ogni classe) per x (numero di classi) è uguale al numero di elementi complessivi di Ω_0,

che è $n(n-1)\cdots(n-k+1)$. Ribadiamo questa idea, che consigliamo di raffigurare con un disegno: Ω_0 è suddiviso in x sottoinsiemi, le classi C, ciascuna fatta di $k!$ elementi. Quindi

$$n(n-1)\cdots(n-k+1) = x \cdot k!$$

ed allora

$$x = \frac{n(n-1)\cdots(n-k+1)}{k!} = \binom{n}{k}.$$

La dimostrazione è completa. ■

Esempio 14 *Consideriamo le 2^n stringhe $(x_1, ..., x_n)$ in cui ciascun simbolo x_i può assumere solo i valori 0 ed 1. Chiediamoci: dato $k \leq n$, quante di queste stringhe hanno k uni? La risposta è $\binom{n}{k}$ (quanti i sottoinsiemi di $\{1, ..., n\}$ aventi cardinalità k).*

Esempio 15 *Quante sono le commissioni di 5 membri che si possono formare partendo da 15 persone? Quante i sottoinsiemi di $\{1, ..., 15\}$ aventi cardinalità 5, quindi $\binom{15}{5}$. E se 7 delle 15 sono uomini, e la commissione si estrae a caso, che probabilità c'è che ci siano 3 uomini in commissione? La probabilità richiesta è $\frac{|A|}{|\Omega|}$ dove Ω è l'insieme delle commissioni possibili mentre A quello delle commissioni con 3 uomini. Allora $|\Omega| = \binom{15}{5}$. Per calcolare $|A|$ si pensi ad un primo esperimento in cui si scelgono tre uomini tra i 7 seguito da un secondo in cui si scelgono 2 donne tra le 8. Nel primo esperimento ci sono $\binom{7}{3}$ risultati possibili; nel secondo $\binom{8}{2}$. Per il principio di enumerazione, $|A| = \binom{7}{3}\binom{8}{2}$. In conclusione,*

$$\frac{|A|}{|\Omega|} = \frac{\binom{7}{3}\binom{8}{2}}{\binom{15}{5}}.$$

1.2 Variabili aleatorie e valori medi

1.2.1 Introduzione

Cosa sono le variabili aleatorie (abbreviato v.a. nel seguito)? La risposta a questa domanda è di gran lunga più sofisticata di molti altri elementi di teoria delle v.a. Quindi, per non partire subito con le cose più difficili, adottiamo una tattica pragmatica: ci accontentiamo di sviluppare un'intuizione pratica di cosa sia una v.a., introduciamo alcuni oggetti matematici che la descrivono (densità, ecc.) e cominciamo così a fare calcoli e vedere l'utilità pratica del concetto. In un secondo momento torneremo sull'aspetto fondazionale e daremo la definizione rigorosa di v.a., che costituirà anche il collegamento tra questo nuovo concetto e quello di spazio probabilizzato introdotto nella prima lezione.

L'idea intuitiva è semplice: chiamiamo v.a. ogni grandezza su cui non possiamo fare previsioni certe, ma di cui abbiamo informazioni probabilistiche nel senso specificato sotto col concetto di densità. Ad esempio, una v.a. è la durata della batteria di un portatile, il numero di esemplari di un certo prodotto che verranno richesti ad un negozio durante la prossima settimana, la quantità di traffico su un ramo della rete internet nelle prossime ore, e così via.

Indichiamo in genere con le lettere X, Y ecc. le v.a. Ad esse sono associati degli eventi che ci interessano in pratica, oppure in teoria. Ad esempio, può interessarci l'evento: $\{T > 10 \text{ ore}\}$ dove T è la durata della batteria di un portatile, oppure l'evento $\{N = 2\}$ dove N è il numero di lavastoviglie che verranno richeste ad un certo negozio. In generale possiamo pensare che gli eventi di interesse avranno la forma

$$\{X \in A\}$$

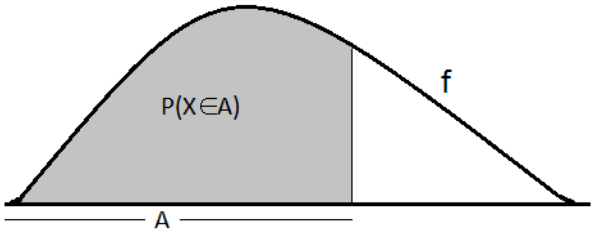
dove X è la v.a. che stiamo considerando ed A è un sottoinsieme dei numeri reali (o in certi casi dei numeri naturali, ad esempio).

1.2.2 V.a. continue e loro densità di probabilità

Abbiamo detto che ci interessano eventi del tipo $\{X \in A\}$ e quindi vorremo calcolarne la probabilità. Si chiamano *continue* quelle v.a. X a cui è associata una densità di probabilità f. La probabilità dell'evento $\{X \in A\}$ si calcola mediante un integrale di f:

$$P(X \in A) = \int_A f(x)\, dx$$

dove l'integrale è esteso all'insieme A.



Per una v.a. continua X, tutte le probabilità del tipo $P(X \in A)$ si calcolano mediante la densità f, quindi in un certo senso non serve avere una definizione rigorosa di v.a., è sufficiente il concetto di densità e la convenzione di interpretare l'integrale $\int_A f(x)\, dx$ come probabilità di un determinato evento. Per questo, entro certi limiti, si puòl fare a meno della definizione rigorosa di v.a. In quest'ottica, il simbolo X non descrive un oggetto matematico rigoroso, ma è solo un ausilio simbolico per abbreviare la scrittura di certi eventi e di certe probabilità. Ad esempio, invece di scrivere "probabilità che la batteria duri più di 10 ore", scriviamo sinteticamente $P(T > 10 \text{ ore})$. E' solo una scrittura convenzionale. Poi, per calcolare matematicamente questa probabilità, basta avere la densità f e calcolare $\int_{10}^{+\infty} f(x)\, dx$.

Nella definizione di densità di probabilità abbiamo omesso alcune precisazioni matematiche, che non approfondiamo in tutta la loro possibile generalità; accenniamo solo al fatto che bisogna richiedere che abbia senso calcolare l'integrale, quindi bisogna far riferimento ad una nozione di funzione integrabile. La versione facile di questa nozione è quella di funzione integrabile secondo Riemann, che abbraccia ad esempio le funzioni continue e qualcosa in

più; la verisone più matura richiederebbe invece la nozione di funzione integrabile secondo Lebesgue, che comprende più funzioni e si adatta meglio alle questioni coinvolgenti operazioni limite.

Esempio 16 *Una v.a. uniforme su* $[a,b]$ *è una v.a.* X *con densità* f *data da*

$$f(x) = \begin{cases} \frac{1}{b-a} & per \quad x \in [a,b] \\ 0 & per \quad x \notin [a,b] \end{cases} .$$

L'area sottesa da f *è uno per ragioni elementari.*

Esempio 17 *Una v.a. esponenziale di parametro* λ*, con* $\lambda > 0$*, è una v.a.* X *con densità* f *data da*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & per \quad x \geq 0 \\ 0 & per \quad x < 0 \end{cases} .$$

Scriveremo per brevità $X \sim Exp(\lambda)$*. Abbiamo già verificato che questa è una pdf. Tra le cose più significative delle v.a. esponenziali c'è la formula (valida per* $t \geq 0$*)*

$$P(X \geq t) = e^{-\lambda t}$$

che si dimostra calcolando l'integrale

$$P(X \geq t) = \int_t^\infty \lambda e^{-\lambda x} dx = -\left[e^{-\lambda x}\right]_t^\infty = e^{-\lambda t}.$$

La funzione $t \mapsto P(X \geq t)$ *viene a volte chiamata "affidabilità" (reliability), nell'omonima teoria.*

Esempio 18 *Una v.a. gaussiana, o normale, canonica è una v.a.* X *con densità* f *data da*

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

La verifica della proprietà di area uno è più complessa. Tralasciando i dettagli, si fonda sui seguenti calcoli:

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx\right)^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dxdy$$
$$= \int_0^{2\pi}\int_0^{+\infty} re^{-\frac{r^2}{2}} drd\theta = 2\pi \int_0^{+\infty} re^{-\frac{r^2}{2}} dr$$

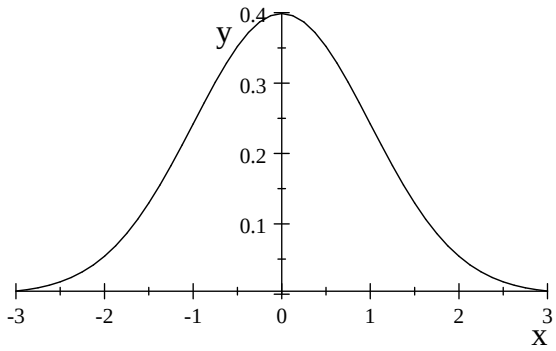
dove abbiamo usato il cambio di variabili in coordinate polari (il determinante jacobiano è r*). Essendo*

$$\int_0^{+\infty} re^{-\frac{r^2}{2}} dr = -\int_0^{+\infty} \frac{d}{dr} e^{-\frac{r^2}{2}} dr = -\left[e^{-\frac{r^2}{2}}\right]_0^\infty = 1$$

troviamo infine

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$$

che spiega il fattore $\frac{1}{\sqrt{2\pi}}$ *nella definizione di* f*.*



Densità gaussiana canonica

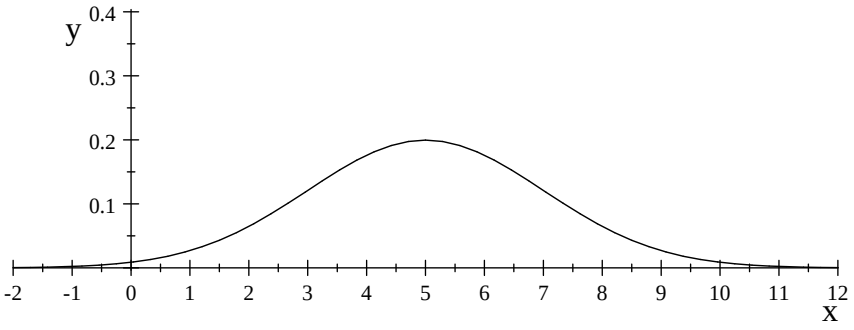
Osservazione 12 *Osserviamo che purtroppo non è possibile calcolare una primitiva di f tramite funzioni elementari, quindi il calcolo di probabilità gaussiane non è di tipo ananlitico, ma solo numerico (con l'uso di tavole o computer).*

Esempio 19 *Una v.a. gaussiana, o normale, di parametri μ e σ^2 (con $\sigma > 0$) è una v.a. X con densità f data da*

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

Scriveremo per brevità $X \sim N\left(\mu, \sigma^2\right)$. La verifica della proprietà di area uno si fa riconducendosi al caso canonico con il cambio di variabile $y = \frac{x-\mu}{\sigma}$:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \overset{y=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2}} \sigma dy$$
$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 1.$$

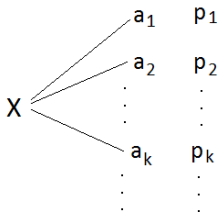


Densità gaussiana con $\mu = 5$ e $\sigma^2 = 4$

1.2.3 V.a. discrete

In un certo senso più elementari delle precedenti (ma in un'altro più singolari) sono le v.a. discrete, quelle ovvero che assumono solo un numero finito di valori $\{a_1, ..., a_N\}$ o al più una

quantità numerabile di valori $\{a_k\}_{k\in\mathbb{N}}$. Per specificare una v.a. X discreta bisogna indicare quali sono i suoi valori a_k e quali siano le corrispondenti probabilità, $p_k = P(X = a_k)$. I numeri $\{p_k\}_{k\in\mathbb{N}}$ devono essere una densità di probabilità discreta, cioè essere compresi in $[0,1]$ ed avere somma 1. Un semplice diagramma del tipo



riassume tutte le caratteristiche di una v.a. discreta.

In tutti gli esempi fondamentali i valori possibili sono un sottoinsieme dei numeri naturali $\mathbb{N} = \{0, 1, ...\}$, o del tipo $\{0, ..., N\}$ oppure $\mathbb{N}$ stesso. Scriveremo

$$p_k := P(X = k)$$

per ogni $k \in \mathbb{N}$. A partire dalla densità di probabilità discreta $\{p_k\}_{k\in\mathbb{N}}$ si calcolano probabilità più complesse semplicemente per somma (finita o infinita a seconda dei casi):

$$P(X \in A) = \sum_{k\in A} P(X = k) = \sum_{k\in A} p_k.$$

Esempio 20 *Una v.a. di Bernoulli di parametro p, con $p \in [0,1]$, è una v.a. X che assume solo i valori 0 ed 1, con densità discreta di probabilità data da $p_0 = 1 - p$, $p_1 = p$, o in altre parole*

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

Può essere utile una scrittura schematica del tipo

$$X = \begin{cases} 1 & \text{con probabilità} \quad p \\ 0 & \text{con probabilità} \quad 1-p \end{cases}.$$

La proprietà di somma uno è ovvia.

Esempio 21 *Una v.a. binomiale di parametri n e p, con n intero positivo e $p \in [0,1]$, è una v.a. X che assume i valori $k = 0, 1, ..., n$, con probabilità*

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

Scriveremo per brevità $X \sim B(n,p)$. La proprietà di somma uno deriva dalla formula del binomio di Newton:

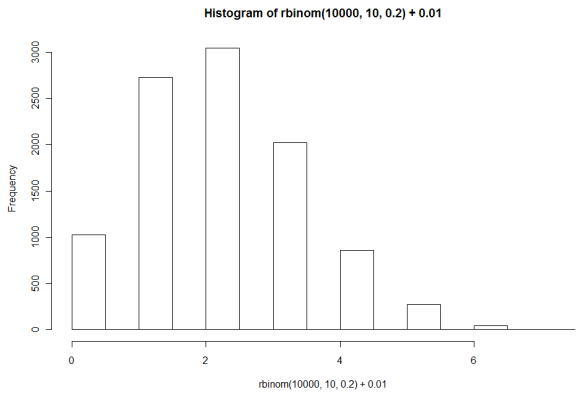
$$(a+b)^n = \binom{n}{k} a^k b^{n-k}.$$

Per questa formula,

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

quindi i numeri della definizione di v.a. binomiale sono effettivamente una densità discreta di probabilità. Nella figura si vede una $B(10, 0.2)$*; i valori numerici, per* $k = 0, 1, ..., 10$*, sono* 0.107*,* 0.268*,* 0.301*,* 0.201*,* 0.088*,* 0.026*,* 0.005*,* 7.8×10^{-4}*,* 7.3×10^{-5}*,* 4.0×10^{-6}*,* 1.0×10^{-7} *(si noti la piccolezza degli ultimi). Non riportiamo il grafico di una* $B(10, 0.5)$*, che, come si può immaginare, è simmetrico. Infine, il grafico di una* $B(10, 0.8)$ *è come quello della figura ma riflesso rispetto al punto centrale.*

```
hist(rbinom(10000,10,0.2)+0.01,11)
```



Densità di massa di una $B(10, 0.2)$

Osservazione 13 *Osserviamo che per* $n = 1$ *le v.a. binomiali sono v.a. di Bernoulli. Quindi possiamo indicare le Bernoulli scrivendo* $X \sim B(1, p)$*. Vedremo più avanti, nel Teorema 5, che la somma di* n *v.a. di Bernoulli* $B(1, p)$ *indipendenti è una* $B(n, p)$*.*

Esempio 22 *Una v.a. di Poisson di parametro* λ*, con* $\lambda > 0$*, è una v.a.* X *che assume tutti i valori interi non negativi con probabilità data dalla formula*

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

per ogni $k \in \mathbb{N}$*. Scriveremo* $X \sim \mathcal{P}(\lambda)$*. La proprietà di somma uno deriva dallo sviluppo in serie dell'esponenziale:*

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

Il seguente teorema stabilisce un legame fondamentale tra v.a. binomiali e di Poisson. Rimandiamo un po' più avanti la sua interpretazione, che svolgeremo congiutamente a vari discorsi interpretativi.

Teorema 4 (degli eventi rari) *Dato* $\lambda > 0$*, posto* $p_n = \frac{\lambda}{n}$ *(che di solito si scrive* $pn = \lambda$*), per ogni* $k \in \mathbb{N}$ *vale*

$$\lim_{n \to \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Proof. Fissato $k \in \mathbb{N}$, vale

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \frac{(1-p_n)^n}{(1-p_n)^k}$$
$$= \frac{\lambda^k}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \frac{(1-p_n)^n}{(1-p_n)^k}$$

ed ora basta osservare che per $n \to \infty$

$$\frac{n}{n} = 1, \quad \frac{n-1}{n} \to 1, \quad ..., \frac{n-k+1}{n} \to 1$$

(e sono un numero finito e *fissato* k di termini),

$$(1-p_n)^k = \left(1 - \frac{\lambda}{n}\right)^k \to 1^k = 1$$

mentre per un noto limite notevole

$$(1-p_n)^n = \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

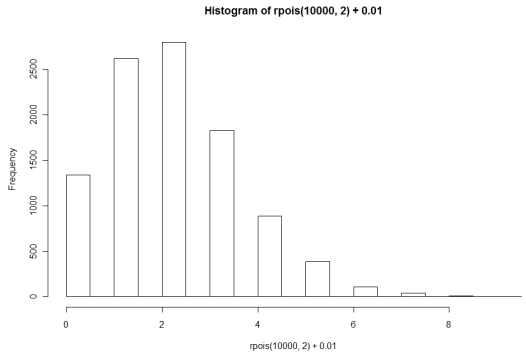
Mettendo insieme tutti questi limiti ed usando i teoremi sul limite di prodotto e rapporto di successioni, si ottiene il risultato desiderato. ■

A titolo di esempio, consideriamo una v.a. $\mathcal{P}(2)$. Essa è limite di $B(n,p)$ con $np = 2$. I valori

$$n = 10, \quad p = 0.2$$

sono ancora ben lontani intuitivamente da ciò che pensiamo essere il limite per n grande. Eppure i primi valori, per $k = 0, 1, ..., 10$ della $\mathcal{P}(2)$ sono 0.135, 0.270, 0.270, 0.180, 0.090, 0.036, 0.012, 0.003, 8.5×10^{-4}, 1.9×10^{-4}, 3.8×10^{-5}, che non si scostano molto da quelli riportati sopra per una $B(10, 0.2)$. Il grafico è riportato in figura. Qualche lieve differenza è ancora apprezzabile e fa capire intuitivamente alcune differenze di forma tra le due densità di massa.

```
hist(rpois(10000,2)+0.01)
```



Densità di massa di una $\mathcal{P}(2)$

Osservazione 14 *Il legame simbolico tra il parametro delle v.a. esponenziali e quello delle Poisson non è casuale. Vedremo più avanti un legame anche tra queste due classi, particolarmente interessante in quanto lega v.a. continue a v.a. discrete, e non attraverso operazioni limite, bensì operazioni logiche finite.*

Esempio 23 *Una v.a. geometrica di parametro p, con $p \in (0,1)$, è una v.a. X che assume tutti i valori interi non negativi con probabilità data dalla formula*

$$P(X = k) = (1-p)^k p$$

per ogni $k \in \mathbb{N}$ (questa è la densità geometrica, introdotta precedentemente). Queste v.a. sono un po' l'analogo nel discreto delle v.a. esponenziali. Non tracciamo la loro densità di massa, che si può facilmente immaginare per analogia con le v.a. esponenziali.

Esempio 24 *Per certe applicazioni è utile introdurre la cosidetta v.a. geometrica modificata (spesso chiamata anch'essa semplicemente v.a. geometrica). Una v.a. geometrica modificata di parametro p è una v.a. che assume i valori interi positivi $k = 1, 2, ...$ con probabilità*

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

1.2.4 Definizione di variabile aleatoria

Fino ad ora, per v.a. abbiamo inteso intuitivamente ogni grandezza casuale che incontriamo in qualche applicazione pratica. Se però ci sforziamo, di fronte ad un problema concreto, di costruire esplicitamente Ω, vediamo che le grandezze aleatorie si possono vedere come funzioni definite sul dominio Ω a valori reali.

Esempio 25 *Consideriamo n v.a. di Bernoulli di parametro p. Ad esempio, potremmo essere interessati allo studio di una banca avente n correntisti (es. 100), ciascuno dei quali, in una giornata generica, si presenta con probabilità p (es. $\frac{1}{5}$) per ritirare del denaro. Associamo ad ogni correntista una v.a. di Bernoulli che vale 1 se il correntista si presenta per ritirare denaro, 0 altrimenti. Abbiamo quindi n v.a. di Bernoulli, X_1 per il correntista numero 1, ecc. fino a X_n per il correntista numero 100. Il numero di richieste (in un dato giorno) è dato allora da*

$$S_n = X_1 + ... + X_n$$

in quanto ogni richiesta contribuisce con un 1 in questa somma, mentre le mancate richeste contribuiscono con 0.

Introduciamo lo spazio Ω dei possibili esiti. Un esito ω in questo problema corrisponde a sapere, per ogni correntista, se si è presentato o meno. Quindi, un esito è una stringa $\omega = (\omega_1, ..., \omega_n)$ in cui ω_1 vale 1 se il primo correntista si è presentato, zero altrimenti, e così via per gli altri ω_i. Ω è l'insieme di tutte queste sequenze.

Definito Ω, ad ogni esito ω possiamo associare diverse grandezze: ad esempio la grandezza

$$X_1(\omega) = \omega_1$$

che legge, di tutta l'informazione contenuta in ω, solo se il primo correntista si è presentato o meno. Oppure, ad esempio, la grandezza

$$S(\omega) = X_1(\omega) + ... + X_n(\omega) = \omega_1 + ... + \omega_n$$

che legge il numero di correntisti che si sono presentati, relativamente a quella sequenza ω. Vediamo che in questo modo abbiamo definito delle funzioni X_1, S, con dominio Ω, a valori reali. Esse corrispono esattamente, come significato pratico, alle omonime grandezze aleatorie introdotte prima a livello non rigoroso, mentre ora, come funzioni da Ω in $\mathbb{R}$, sono oggetti matematici ben precisi.

L'esempio mostra che è ragionevole definire come variabili aleatorie le *funzioni* definite su uno spazio Ω, a valori in qualche insieme. Manca ancora una precisazione, per arrivare alla definizione completa, ma prima svolgiamo qualche osservazione.

Con riferimento all'esempio, il simbolo S, prima senza significato matematico ma usato per comodità di scrittura, diventa ora l'usuale simbolo di funzione avente un significato matematico preciso: S è abbreviazione di $S(\omega)$, come f lo è di $f(x)$. Prima scrivevamo $\{S = k\}$ come simbolo abbreviato per intendere l'evento "k correntisti si presentano". Ora possiamo interpretare rigorosamente $\{S = k\}$ come evento in Ω, ovvero come sottoinsieme di Ω: è l'insieme di tutti i punti ω tali che $S(\omega) = k$. Detto altrimenti, ora il simbolo $\{S = k\}$ è semplicemente l'abbreviazione dell'espressione perfettamente rigorosa e significativa

$$\{\omega \in \Omega : S(\omega) = k\}.$$

Le variabili aleatorie sono funzioni. Quando nominiamo una v.a. X, sottointendiamo che ci sia uno spazio probabilizzato $(\Omega, \mathcal{F}, P)$ su cui X sia definita come funzione $\omega \mapsto X(\omega)$. Quando scriviamo un evento $\{X \in A\}$ intendiamo l'evento

$$\{\omega \in \Omega : X(\omega) \in A\}.$$

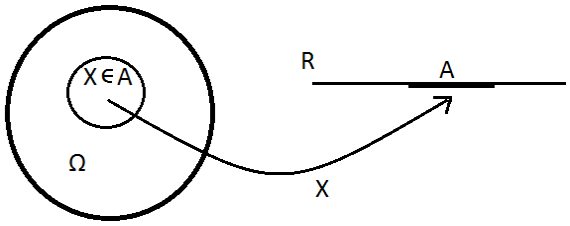
Quando scriviamo $P(X \in A)$ stiamo calcolando la probabilità P di questo evento. Come in vari esempi visti nella prima lezione, non sempre si esplicita lo spazio Ω quando si maneggiano delle variabili aleatorie; una cosa è l'impianto teorico, un'altra è la pratica con le sue scorciatoie e l'eliminazione della trattazione esplicita di tutti i dettagli a volte solo noiosi e non rilevanti per il risultato pratico. Notiamo però che nella nostra esperienza personale capita ogni tanto di doversi fermare e cercare di capire le cose con l'impianto rigoroso, di fronte a problemi non banali in cui una tratazione troppo intuitiva lascia qualche ansia circa la veridicità dei risultati (per motivi di sostanza, non puramente formali). In altre parole, *a volte pensare che la scrittura $\{X \in A\}$ sta per l'insieme degli $\omega \in \Omega$ tali che $X(\omega) \in A$, è molto utile per essere sicuri di ciò che si sta facendo.* Senza menzionare i casi in cui è invece indispensabile l'uso esplicito dello spazio Ω, come ad esempio nella legge forte dei grandi numeri.

Veniamo però alla definizione completa di v.a. Il problema è che, data una funzione $X : \Omega \to \mathbb{R}$, vogliamo calcolare $P(X \in A)$, quindi l'isieme $\{X \in A\} \subset \Omega$ deve appartenere alla famiglia $\mathcal{F}$. Quindi dovremo imporre la condizione $\{X \in A\} \in \mathcal{F}$. Come però abbiamo preso $\mathcal{F}$ invece che la famiglia di tutte le parti di Ω, per motivi analoghi non vogliamo necessariamente considerare tutti gli insiemi $A \subset \mathbb{R}$, nella richiesta precedente.

Fissiamo allora una σ-algebra $\mathcal{B}$ di sottoinsiemi di $\mathbb{R}$.

Definizione 15 *Chiamiamo v.a. su* $(\Omega, \mathcal{F}, P)$ *a valori in* $(\mathbb{R}, \mathcal{B})$ *ogni funzione* $X : \Omega \to \mathbb{R}$ *tale che* $\{X \in A\} \in \mathcal{F}$ *per ogni* $A \in \mathcal{B}$.

In genere, salvo avviso contrario, si prende come σ-algebra $\mathcal{B}$ quella dei boreliani. Questa è la *definizione di variabile aleatoria*, che illustriamo col seguente disegno:



Si noti che la somma di due v.a. X ed Y è ben definita se esse sono v.a. definite sullo stesso spazio $(\Omega, \mathcal{F}, P)$. Infatti si pone

$$(X + Y)(\omega) = X(\omega) + Y(\omega).$$

1.2.5 Legge di una v.a.

Data una v.a. X a valori reali definita su uno spazio probabilizzato $(\Omega, \mathcal{F}, P)$, questa induce una distribuzione di probabilità μ_X, detta *legge* (o *distribuzione*) di X, sui borelani di $\mathbb{R}$. Questa distribuzione di probabilità μ_X è definita semplicemente da

$$\mu_X(A) = P(X \in A).$$

In altre parole, le probabilità $P(X \in A)$ che abbiamo introdotto come i primi oggetti legati ad una v.a. X, si possono vedere come una *funzione di* A, definita per tutti i boreliani A di $\mathbb{R}$. Questa funzione la indichiamo con μ_X e la chiamamo legge di X. Si può verificare che soddisfa i requisiti di una probabilità (a valori in $[0, 1]$, $\mu_X(\mathbb{R}) = 1$, ed è σ-additiva).

E' utile farsi un'immagine mentale o grafica, anche se tracciare un disegno è piuttosto difficile. Si deve pensare che sull'insieme Ω sia distribuita una massa P, e che questa venga trasportata dalla funzione X in una massa μ_X distribuita su $\mathbb{R}$. Parlando intuitivamente, è come se la massa che sta in ogni punto ω venga trasportata da X nel punto $X(\omega)$ (l'immagine è matematicamente scorretta in quanto in molti casi i singoli punti ω hanno tutti massa nulla, quindi il ragionamento va sempre riferito a insiemi di punti). Se ad esempio due o più punti vengono trasformati da X nello stesso punto (X non iniettiva), le loro masse vanno a sommarsi nel punto di arrivo.

La probabilità μ_X è un po' astratta, quanto lo è P stessa, in relazione a problemi in cui tendamo ad interessarci solo delle densità delle v.a. in gioco e dei calcoli che si possono fare su di esse. Osserviamo allora che se X è una v.a. continua con densità $f(x)$, allora vale

$$\mu_X(A) = \int_A f(x)\, dx$$

mentre se X è una v.a. discreta sui numeri interi non negativi, con densità di massa $p(k)$, allora

$$\mu_X(A) = \sum_{k \in A} p(k).$$

Anzi, data una misura di probabilità μ sui boreliani di $\mathbb{R}$, anche a priori non associata ad una v.a. X, diremo che è continua se esiste una densità $f(x)$ per cui valga la prima formula precedente, discreta se vale la seconda. Ma esistono interessanti misure μ_X (associate ad altrettante v.a. X) che non sono né continue né discrete: miste nei casi più semplici, oppure del tutto inedite come le misure frattali.

Quando diremo che certe v.a. sono equidistribuite o identicamente distribuite (dette anche isonome), intenderemo che abbiano la stessa legge. Ad esempio, tutte esponenziali di parametro 3. Questo non significa che siano *uguali*, in quanto funzioni da Ω in $\mathbb{R}$. Pensiamo ai due risultati dei lanci di due dati. Descriviamo il primo con una v.a. X_1, il secondo con X_2. queste due v.a. hanno la stessa legge μ, che è una probabilità discreta sui numeri da 1 a 6, uniforme. Ma non sono la stessa v.a. Intuitivamente è chiaro, in quanto non corrispondono allo stesso esperimento. Matematicamente la differenza si apprezza se si introduce esplicitamente lo spazio Ω delle coppie (x, y) dei possibili risultati. Vale $X_1(x, y) = x$, $X_2(x, y) = y$, quindi sono due diverse funzioni.

1.2.6 Funzione di distribuzione (cdf) di una v.a.

Data una v.a. X, si chiama *funzione di distribuzione* (o *di ripartizione*) la funzione $x \mapsto F(x)$ definita da

$$F(x) = P(X \leq x).$$

Nel linguaggio ingegneristico si sottolinea che è la *cumulativa*: funzione di distribuzione cumulativa, abbreviata (seguendo il nome inglese) in *cdf*. Essa è una funzione da $\mathbb{R}$ in $[0, 1]$, è crescente (in senso debole), soddisfa

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to +\infty} F(x) = 1,$$

è continua a destra in ogni punto:

$$\lim_{x \to x_0^+} F(x) = F(x_0) \quad \forall x_0 \in \mathbb{R}.$$

La verifica di queste proprietà è facile ma richiede un po' di lavoro a partire dalla numerabile additività di μ. La probabilità degli intervalli è legata agli incrementi di F:

$$F(b) - F(a) = P(X \in (a, b]), \quad \forall a < b \in \mathbb{R}.$$

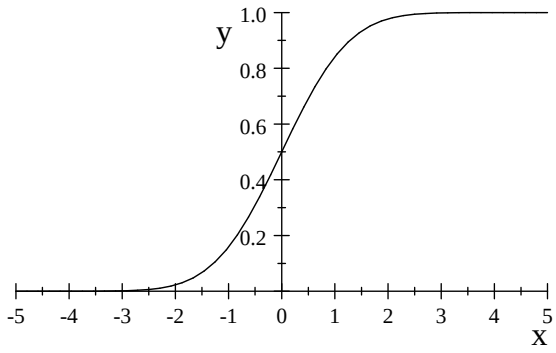


Grafico della cdf normale standard

Il limite sinistro di $f(x)$ esiste in ogni punto x_0, come per qualsiasi funzione crescente, ma può essere strettamente minore di $F(x_0)$, nel qual caso la funzione F è discontinua in x_0. In tale punto si verifica una concentrazione di massa per la μ, nel senso che $\mu(\{x_0\}) > 0$. Questa proprietà è tipica per le misure discrete, e si ritrova anche nelle cosidette distribuzioni miste, mentre per le misure definite da una densità di probabilità la massa dei singoli punti è nulla.

La funzione $F(x)$ porta il nome di funzione di distribuzione perché da un lato è una funzione e non una misura, dall'altro però dice tutto della distribuzione (legge) della v.a. a cui è associata. Spesso nella letteratura applicativa non viene mai introdotto il concetto di legge di una v.a., essendo un po' difficile, mentre si cerca di ricondurre tutto all'uso della funzione di distribuzione $F(x)$, oggetto più semplice, che in effetti è sufficiente per molti scopi.

Quando X ha densità $f(x)$, vale

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Graficamente, $F(x)$ misura l'area sottesa dal grafico di f, a sinistra del punto x. Nei punti in cui f è continua, per il teorema fondamentale del calcolo integrale abbiamo

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

Quindi, fa f si ricava F per integrazione, e da F si ricava f per derivazione.

Se X è una v.a. discreta sui numeri interi non negativi, con massa di probabilità p_k, vale

$$F(x) = \sum_{k \le x} p_k$$

e

$$p_k = F(k) - F(k-1).$$

1.2.7 V.A. indipendenti

Date due v.a. X, Y definite sullo stesso spazio probabilizzato $(\Omega, \mathcal{F}, P)$, diciamo che sono indipendenti se

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

per ogni coppia A, B di boreliani di $\mathbb{R}$. L'interpretazione è chiara: gli eventi $X \in A$ e $Y \in B$ che descrivono cosa può accadere in relazione ad X e Y, devono essere indipendenti.

Una famiglia $\{X_\alpha\}$ di v.a. è composta da v.a. indipendenti se per ogni sequenza $\{\alpha_k\}$ di indici e $\{A_k\}$ di boreliani, abbiamo

$$P\left(\bigcap_k (X_{\alpha_k} \in A_k)\right) = \prod_k P\left(X_{\alpha_k} \in A_k\right).$$

A livello quantitativo, c'è modo di descrivere l'indipendenza tramite oggetti come la densità o i valori medi che introdurremo? In parte sì, ma serve la densità congiunta, che descriveremo nel prossimo paragrafo.

Come applicazione del concetto rigoroso di v.a. e del concetto di indipendenza, dimostriamo il seguente teorema.

Teorema 5 *La somma di n Bernoulli indipendenti di parametro p è una $B(n,p)$.*

Proof. Il teorema vale per v.a. di Bernoulli definite su qualsiasi spazio probabilizzato ma per fare una dimostrazione più istruttiva mettiamoci in uno schema più preciso (si può dimostrare che questo non è restrittivo). Riprendiamo lo spazio

$$\Omega = \{0,1\}^n$$

dell'esempio 25 con la probabilità di una sequenza $\omega = (\omega_1, ..., \omega_n)$ data da $P(\omega) = p^{k(\omega)}(1-p)^{n-k(\omega)}$, dove $k(\omega)$ è il numero di uni nella sequenza, ovvero

$$k(\omega) = \sum_{i=1}^{n} \omega_i.$$

Si ricorderà che avevamo già introdotto questo spazio in passato, come esempio di spazio probabilizzato finito, diverso da quello equiprobabile. Su Ω introduciamo le v.a. X_i definite da

$$X_i(\omega) = \omega_i$$

dove continuiamo ad usare la convenzione $\omega = (\omega_1, ..., \omega_n)$.

Passo 1. Verifichiamo che le X_i sono v.a. di Bernoulli di parametro p indipendenti. La verifica è noiosa ed il lettore può ometterla. Data una stringa $x = (x_1, ..., x_n)$, vale

$$\begin{aligned} P(X_1 = x_1, ..., X_n = x_n) &= P(\omega \in \Omega : \omega_1 = x_1, ..., \omega_n = x_n) \\ &= P((x_1, ..., x_n)) = p^{k(x)}(1-p)^{n-k(x)} \end{aligned}$$

e d'altra parte

$$\begin{aligned} P(X_1 = x_1) &= P(\omega \in \Omega : \omega_1 = x_1) = \sum_{\omega \in \Omega : \omega_1 = x_1} P(\omega) \\ &= \sum_{(\omega_2, ..., \omega_n)} P(x_1, \omega_2, ..., \omega_n) = \sum_{(\omega_2, ..., \omega_n)} p^{k((x_1, \omega_2, ..., \omega_n))}(1-p)^{n-k((x_1, \omega_2, ..., \omega_n))} \\ &= p^{x_1}(1-p)^{1-x_1} \sum_{(\omega_2, ..., \omega_n)} p^{k((\omega_2, ..., \omega_n))}(1-p)^{(n-1)-k((\omega_2, ..., \omega_n))} = p^{x_1}(1-p)^{1-x_1} \end{aligned}$$

ed analogamente

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$$

da cui discende sia che

$$P(X_1 = x_1, ..., X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$$

cioè l'indipendenza, sia il fatto che le X_i sono v.a. di Bernoulli di parametro p.

Passo 2. Fatta questa verifica, introduciamo la v.a. $S_n = X_1 + ... + X_n$ e dimostriamo che è una v.a. binomiale $B(n,p)$. Calcoliamo $P(S = j)$. Osserviamo che $S = j$ equivale a dire che ci sono j uni. Quindi

$$P(S = j) = \sum_{\omega \in \{S=j\}} P(\omega) = \sum_{\omega : k(\omega) = j} p^{k(\omega)}(1-p)^{n-k(\omega)}$$

$$\sum_{\omega : k(\omega) = j} p^j (1-p)^{n-j} = p^j (1-p)^{n-j} \cdot |\{S = j\}|$$

dove $|\{S = j\}|$ indica la cardinalità dell'insieme $\{S = j\}$. Ma per la proposizione **??** sul numero di sequenze con j uni, vale

$$|\{S = j\}| = \binom{n}{j}.$$

La dimostrazione è completa. ■

Vediamo un altro esempio tipico di calcolo su v.a. indipendenti.

Esempio 26 *Supponiamo che un sistema elettronico sia composto di tre sottosistemi. Indichiamo con T il tempo di vita del sistema e con T_1, T_2, T_3 i tempi di vita dei tre sottosistemi. Vale*

$$T = \min(T_1, T_2, T_3).$$

Pertanto, eventi det tipo $\{T > t_0\}$ ("il sistema si romperà dopo il tempo t_0", cioè durerà almeno un tempo t_0) si possono riscrivere nella forma

$$\{T > t_0\} = \{T_1 > t_0, T_2 > t_0, T_3 > t_0\} = \bigcap_{i=1}^{3} \{T_i > t_0\}.$$

Supponiamo che le tre variabili aleatorie T_1, T_2, T_3 siano indipendenti. Allora

$$P(T > t_0) = \prod_{i=1}^{3} P(T_i > t_0).$$

Pertanto, se conosciamo la densità di probabilità del tempo di vita dei tre sottosistemi, possiamo calcolare $P(T > t_0)$.

1.2.8 Vettori aleatori ed altri enti aleatori

Una grandezza aleatoria a valori vettoriali

$$X = (X_1, ..., X_n)$$

in cui le componenti X_i sono v.a. a valori reali definite su uno stesso spazio probabilizzato $(\Omega, \mathcal{F}, P)$, può essere chiamata un *vettore aleatorio*. Un vettore aleatorio è quindi un'applicazione

$$\Omega \xrightarrow{X} \mathbb{R}^n$$

le cui componenti sono variabili aleatorie. Può essere la coppia posizione-velocità di una particella che si muove soggetta a variazioni casuali. Oppure semplicemente possono essere i valori uscenti da una sequenza di n esperimenti.

Analogamente, una grandezza aleatoria a valori in uno spazio di funzioni, ad esempio lo spazio delle funzioni continue su un intervallo $[0, T]$,

$$\Omega \xrightarrow{X} C\left([0, T]; \mathbb{R}\right)$$

può essere chiamata una *funzione aleatoria* (bisogna specificare una proprietà del tipo $\{X \in A\} \in \mathcal{F}$, ma tralasciamo questo particolare). Si pensi ad esempio al campo di velocità di un fluido turbolento, se decidiamo di descriverlo come campo aleatorio. Per chi conosce le distribuzioni, si possono introdurre le distribuzioni aleatorie. Similmente si possono introdurre le misure aleatorie, gli insiemi aleatori, ecc. In sintesi, anche se dedichiamo la maggior parte dei nostri sforzi allo studio di v.a. a valori reali, esistono generalizzazioni ad enti aleatori a valori in insiemi di oggetti diversi dai numeri reali ($\mathbb{R}^n$, spazi di funzioni, distribuzioni, misure, spazi di insiemi, ecc.). In genere queste generalizzazioni si appoggiano su concetti topologici, quindi è utile che ci sia un concetto di vicinanza in tali famiglie di oggetti. Dal punto di vista matematico, in genere si riesce a vincere la sfida di definire oggetti aleatori del tipo più disparato. Nelle scienze applicate questo può essere di grande interesse (descrivere forme o profili aleatori, concentrazioni di massa aleatorie, campi aleatori di interesse fisico, ecc.). Naturalmente poi c'è il problema di ridurre i gradi di libertà per tornare a descrizioni quantitativamente efficaci.

Esempio 27 *Dato uno spazio probabilizzato $(\Omega, \mathcal{F}, P)$, consideriamo un insieme $C(\omega) \subset \mathbb{R}^n$, indicizzato da $\omega \in \Omega$. Lo chiamamo insieme aleatorio se, preso un qualsiasi punto $x \in \mathbb{R}^n$, la funzione a valori reali*

$$\omega \mapsto d\left(x, C(\omega)\right)$$

è una varabile aleatoria. La notazione $d\left(x, C(\omega)\right)$ indica la distanza euclidea di x da $C(\omega)$, definita n generale da

$$d(x, A) = \inf_y d(x, y)$$

dove $d(x, y) = |x - y|$ è l'usuale distanza euclidea tra due punti. A titolo di esempio, $C(\omega)$ potrebbe descrivere come si presenta una struttura, inizialmente di forma C_0, dopo essere stata sollecitata da una trasformazione aleatoria. Detto così è astratto e probabilmente privo

di interesse pratico. Però, se è possibile parametrizzare le trasformazioni aleatorie che interessano in un esempio specifico, in modo da avere solo pochi parametri aleatori, $C(\omega)$ verrebbe a dipendere da pochi parametri aleatori, ad es. una coppia di v.a. gaussiane che descrivano torsione e dilatazione. Vediamo quindi che è possibile formalizzare matematicamente concetti anche piuttosto arditi, come quello di "forma aleatoria".

Esempio 28 *Indichiamo con $M_1^+(\mathbb{R}^n)$ l'insieme delle misure di probabilità sui boreliani di $\mathbb{R}^n$. Chiamiamo delta di Dirac in $x_0 \in \mathbb{R}^n$ la misura di probabilità δ_{x_0} definita da*

$$\delta_{x_0}(A) = \begin{cases} 1 & se \quad x_0 \in A \\ 0 & se \quad x_0 \notin A \end{cases}.$$

Intuitivamente, è una massa unitaria concentrata nel punto x_0. Supponiamo di studiare una dinamica aleatoria, a tempo discreto, che si svolge in $\mathbb{R}^n$. Indichiamo con X_1 la posizione al tempo $t = 1$, aleatoria, poi con X_2 la posizione al tempo $t = 2$, sempre aleatoria, e così via. Poi consideriamo, al tempo n, la media temporale

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Con questo simbolo abbiamo indicato una massa equidistribuita tra i punti X_i, per $i = 1, ..., n$. μ_n è una misura di probabilità, quindi un elemento di $M_1^+(\mathbb{R}^n)$, ed è aleatoria, in quanto lo sono i punti X_i. Abbiamo quindi una "misura aleatoria":

$$\Omega \overset{\mu_n}{\to} M_1^+(\mathbb{R}^n).$$

L'esempio non è artificioso: questa misura descrive il tempo trascorso dalla dinamica nelle diverse regioni dello spazio $\mathbb{R}^n$. Per $n \to \infty$ la musura aleatoria μ_n è legata al concetto di misura invariante (che descrive il regime stazionario) della dinamica.

Torniamo ai semplici vettori aleatori. Un vettore aleatorio $X = (X_1, ..., X_n)$ definisce una legge μ_X sui boreliani di $\mathbb{R}^n$, detta *legge congiunta* del vettore X. Per i boreliani prodotto essa è definita da

$$\mu_X(A_1 \times ... \times A_n) = P(X_1 \in A_1, ..., X_n \in A_n)$$

e per gli altri si riesce a definire con procedimenti di estensione che non stiamo a descrivere.

Questa legge congiunta può essere *continua*, ovvero avere una densità $f(x_1, ..., x_n)$ tale che

$$\mu_X(A_1 \times ... \times A_n) = \int_{A_1 \times ... \times A_n} f(x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

Vel cioè

$$P(X_1 \in A_1, ..., X_n \in A_n) = \int_{A_1 \times ... \times A_n} f(x_1, ..., x_n)\, dx_1 \cdots dx_n$$

e più in generale

$$P((X_1, ..., X_n) \in A) = \int_A f(x_1, ..., x_n)\, dx_1 \cdots dx_n$$

se A è un sottoinsieme (opportuno, ad es. boreliano) di $\mathbb{R}^n$. Quando esiste, $f(x_1, ..., x_n)$ si chiama *densità congiunta* del vettore aleatorio X.

Esempio 29 *Ad esempio,*

$$P\left(X_1 > X_2\right) = \int_{\{(x_1,x_2)\in\mathbb{R}^2 : x_1 > x_2\}} f\left(x_1, x_2\right) dx_1 dx_2.$$

In altri casi, la legge di $X = (X_1, ..., X_n)$ può essere *discreta*. Supponiamo per semplicità di notazione (in realtà non è restrittivo) che le singole variabili $X_1, ..., X_n$ assumano come valori possibili solo i valori $\{a_k\}_{k\in\mathbb{N}}$. Allora il vettore $X = (X_1, ..., X_n)$ può assumere ciascuno dei valori

$$(a_{k_1}, ..., a_{k_n})$$

con gli $a_{k_i} \in \{a_k\}_{k\in\mathbb{N}}$. Allora interesserà calcolare innanzi tutto le probabilità del tipo

$$p\left(a_{k_1}, ..., a_{k_n}\right) = P\left(X_1 = a_{k_1}, ..., X_n = a_{k_n}\right).$$

La famiglia di numeri $\left\{p\left(a_{k_1}, ..., a_{k_n}\right); a_{k_1}, ..., a_{k_n} \in \{a_k\}_{k\in\mathbb{N}}\right\}$ può essere chiamata *densità discreta* del vettore X.

Parallelamente sopravvivono i vecchi concetti per ciascuna delle v.a. X_i. La legge di X_1 si chiama ora *legge marginale* di X_1, e se ha densità $f_{X_1}(x_1)$ questa si dirà *densità marginale* di X_1, e così via per le altre.

Nasce allora la domanda circa il legame tra congiunta e marginali. Limitiamoci a discutere le densità.

Teorema 6 *In generale (quando le densità esistono), vale*

$$f_{X_1}(x_1) = \int_{\mathbb{R}^{n-1}} f\left(x_1, ..., x_n\right) dx_2 \cdots dx_n$$

e così per le altre. Quando $X_1, ..., X_n$ sono v.a. indipendenti, vale inoltre

$$f\left(x_1, ..., x_n\right) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

e vale anche il viceversa (se la densità congiunta è il prodotto delle marginali, allora le v.a. sono indipendenti).

Omettiamo la dimostrazione, non troppo difficile peraltro. Osserviamo come interpretazione che, mentre dalla congiunta è sempre possibile calcolare le marginali, viceversa dalle marginali è in genere molto difficile risalire alla congiunta, salvo nel caso di indipendenza. Questo non deve stupire: è come il problema di calcolare la probabilità di una intersezione $P\left(A \cap B\right)$. In generale, abiamo bisogno di conoscere ad esempio $P\left(A|B\right)$, che è un'informazione ben più complessa delle probabilità "marginali" $P(A)$ e $P(B)$.

Esempio 30 *Gaussiana multidimensionale canonica. Supponiamo che $X_1, ..., X_n$ siano v.a. indipendenti gaussiane canoniche, quindi tutte con densità (marginale) $\frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$. Allora il vettore aleatorio $X = (X_1, ..., X_n)$ ha densità congiunta data dal prodotto delle marginali (Teorema 6)*

$$f\left(x_1, ..., x_n\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{\left(x_1^2 + \cdots + x_n^2\right)}{2}\right)$$

che, usando la norma euclidea $|.|$ ed il prodotto scalare euclideo $\langle .,. \rangle$ e la notazione $x = (x_1, ..., x_n)$, possiamo scrivere anche nella forma più compatta

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{|x|^2}{2}\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{\langle x, x \rangle}{2}\right).$$

Questa è la gaussiana canonica in n dimensioni. Il suo grafico in dimensione 2 è una superficie a campana, simmetrica per rotazione.

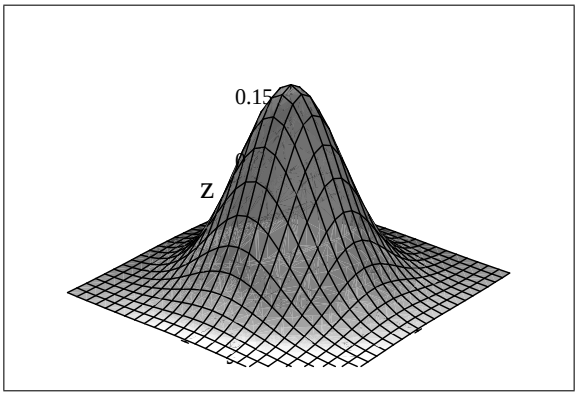


Grafico della normale standard in due dimensioni

1.2.9 Valori medi o attesi

Valori medi sperimentali

Dato un campione sperimentale $x_1, ..., x_n$, chiamiamo sua *media aritmetica* il numero

$$\overline{x} = \frac{x_1 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

A volte viene chiamata anche media sperimentale, o empirica, o anche in altri modi.

Data poi una una funzione $\varphi(x)$, possiamo considerare il campione $\varphi(x_1)$, ... , $\varphi(x_n)$ e calcolarne la media aritmetica

$$\overline{\varphi} = \frac{\varphi(x_1) + ... + \varphi(x_n)}{n}.$$

Ad esempio, presa come φ la funzione scarto quadratico (rispetto alla media $\overline{x}$)

$$\varphi(x) = (x - \overline{x})^2$$

si ottiene il numero

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

che potremmo chiamare scarto quadratico *medio*. In realtà, per un motivo che ora non è possibile anticipare, si preferisce il fattore $\frac{1}{n-1}$ di fronte alla precedente espressione, per cui si arriva ad introdurre il numero

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

detto appunto *scarto quadratico medio sperimentale*.

Così di seguito si potrebbero introdurre altri valori medi sperimentali. Citiamo solamente la frequenza empirica: supponiamo che i valori $x_1, ..., x_n$ (oppure i valori $\varphi(x_1)$, ... , $\varphi(x_n)$) siano tutti pari ad 1 o 0, col significato che si sta esaminando un evento, diciamo A, e vale $x_1 = 1$ se al primo esperimento si è avverato A, $x_1 = 0$ altrimenti, e così via per gli altri x_i. Allora la somma $x_1 + ... + x_n$ conta il numero di volte in cui si è avverato A (come in uno schema già visto con le v.a. di Bernoulli e binomiali), e quindi $\overline{x}$ rappresenta la frequenza relativa con cui si è avverato A. In questo contesto si preferisce allora una notazione del tipo $\overline{p}$ al posto di $\overline{x}$, che allude all'approssimazione di una probabilità, arrivando quindi a scrivere

$$\overline{p} = \frac{x_1 + ... + x_n}{n}$$

come definizione di *frequenza empirica* con cui si avvera l'evento A.

1.2.10 Valor atteso: suo calcolo con le densità

Data una v.a. $X : \Omega \to \mathbb{R}$, in potesi estremamente generali è possibile definire il concetto di valore atteso di X, che indicheremo con $E[X]$. A volte il valore atteso viene anche chiamato speranza o attesa, o semplicemente media, valor medio. Useremo molto spesso il termine media o valor medio, il più usato ad esempio nella letteratura fisica, anche se bisogna ammettere che può creare qualche frainteso con la media aritmetica di un campione sperimentale. Non diamo subito la definizione, piuttosto impegnativa, ma enunciamo un teorema di calcolo, valido quando X è una v.a. continua o discreta.

Teorema 7 *Se X è una v.a. continua con densità $f(x)$, e $\int_{-\infty}^{+\infty}|x|f(x)\,dx < \infty$, allora*

$$E[X] = \int_{-\infty}^{+\infty} xf(x)\,dx.$$

Se X è una v.a. discreta sui numeri interi non negativi, con densità di massa $p(k)$, e $\sum_{k=0}^{\infty} kp(k) < \infty$, allora

$$E[X] = \sum_{k=0}^{\infty} kp(k).$$

Se la v.a. discreta X assume i valori $a_1, a_2...$ invece che i numeri naturali, la formula diventa semplicemente

$$E[X] = \sum_{k=0}^{\infty} a_k p(k)$$

e vale se $\sum_{k=0}^{\infty} |a_k| \, p(k) < \infty$. A parole, il valore atteso è la *somma dei valori per le loro probabilità*, o la media dei valori pesati con le loro probabilità.

Non avendo dato la definizione, non possiamo ovviamente dimostrare il teorema. Osserviamo solo che a volte esso viene scelto come definizione, anche se questa impostazione è sia restrittiva (il valor medio si può definire anche per molte v.a. che non sono né continue né discrete), sia limitativa per quanto riguarda la possibilità di svolgere poi dimostrazioni rigorose di importanti teoremi.

Vediamo però una interpretazione intuitiva della seconda formula del teorema, per semplicità nel caso di una v.a. X discreta che assume solo un numero finito di valori $a_1, ..., a_M$. Vale $E[X] = \sum_{k=0}^{M} a_k p(k)$. Supponiamo di avere un campione sperimentale $x_1, ..., x_n$ estratto da questa v.a.; indichiamo con $\widehat{n}(k)$ è il numero di elementi di questo campione uguali ad a_k e con $\widehat{p}(k)$ il rapporto $\frac{\widehat{n}(k)}{n}$ cioè la percentuale degli elementi del campione che valgono a_k. Ci aspettiamo che $\widehat{p}(k)$ sia circa uguale a $p(k)$:

$$\widehat{p}(k) \sim p(k).$$

Ma allora, raggruppando la somma $x_1 + ... + x_n$ secondo i valori assunti dai vari elementi (scambiando ovviamente i termini)

$$\begin{aligned} x_1 + ... + x_n &= (a_1 + ... + a_1) + ... + (a_M + ... + a_M) \\ &= \widehat{n}(1) a_1 + ... + \widehat{n}(M) a_M \end{aligned}$$

otteniamo

$$\begin{aligned} \overline{x} = \frac{x_1 + ... + x_n}{n} = \frac{\widehat{n}(1) a_1 + ... + \widehat{n}(M) a_M}{n} = \frac{\widehat{n}(1)}{n} a_1 + ... + \frac{\widehat{n}(M)}{n} a_M \\ = \widehat{p}(1) a_1 + ... + \widehat{p}(M) a_M \sim p(1) a_1 + ... + p(M) a_M = E[X]. \end{aligned}$$

Abbiamo cioè verificato la seguente affermazione: se le percentuali sperimentali $\widehat{p}(k)$ sono circa uguali alle probabilità teoriche $p(k)$, allora la media aritmetica $\overline{x}$ è circa uguale alla media teorica $E[X]$.

Infine, si riconosce che l'espressione della media nel caso di v.a. continue è l'estensione naturale al continuo della formula per le v.a. discrete. Per tutte queste ragioni il risultato del teorema è molto naturale (e viene a volte preso come definizione di valo medio).

Il teorema precedente si generalizza a *funzioni di variabili aleatorie*:

Teorema 8 *Se X è una v.a. continua con densità $f(x)$, allora*

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

per ogni funzione $\varphi(x)$ per cui abbia senso l'integrale. Analogamente, se X è una v.a. discreta con densità di massa $p(k)$, allora

$$E[\varphi(X)] = \sum_{k=0}^{\infty} \varphi(k) p(k).$$

Il teorema si estende poi al caso di funzioni di vettori aleatori. Scriviamo l'enunciato solo nel caso di vettori continui; nel caso discreto vale un teorema simile, un po' noioso da scrivere in generale (si veda un esempio poco sotto).

Teorema 9 *Se $X = (X_1, ..., X_n)$ è un vettore aleatorio continuo con densità congiunta $f(x_1, ..., x_n)$, allora*

$$E\left[\varphi\left(X_1, ..., X_n\right)\right] = \int_{-\infty}^{\infty} \varphi\left(x_1, ..., x_n\right) f(x_1, ..., x_n) dx_1 ... dx_n.$$

per ogni funzione $\varphi\left(x_1, ..., x_n\right)$ per cui abbia senso l'integrale.

Proprietà del valor medio

Iniziamo con l'enunciare le proprietà più semplici. Per ora non diamo (quasi) mai la dimostrazione, in attesa di avere a disposizione la definizione rigorosa di valor medio, con la quale molte delle sue proprietà diventano abbastanza semplici da dimostrare. Evitiamo di appesantire gli enunciati con tutte le ipotesi ma, in sostanza, bisogna assumere che tutti i valori medi di cui si parla esistano finiti.

Linearità

Se X e Y sono due v.a. *qualsiasi* definite sullo stesso spazio probabilizzato e α, β e γ sono numeri reali, allora si ha:

$$E[\alpha X + \beta Y + \gamma] = \alpha E[X] + \beta E[Y] + \gamma \ .$$

Osservazione 15 *Ribadiamo il fatto che non è necessaria alcuna ipotesi di indipendenza delle variabili aleatorie X e Y.*

Osservazione 16 *La proprietà di linearità fa pensare che il valor medio sia un'operazione simile all'integrale. Con la definizione rigorosa vedremo che questo è profondamente vero. Invece, la scrittura integrale $E[X] = \int x f(x) dx$ è solo un riflesso di questo: non è per via di questa scrittura che $E[X]$ ha le proprietà di un'integrale. Si provi infatti ad immaginare una dimostrazione della linearità basata su $\int x f(x) dx$: bisognerebbe conoscere la densità $f_{\alpha X + \beta Y}$ in relazione alle densità f_X e f_Y. E' possibile ma intricata.*

Osservazione 17 *Dimostriamo la linearità nel caso di v.a. discrete, caso in cui è abbastanza intuitivo scrivere la densità discreta $p_{\alpha X + \beta Y + \gamma}$ in relazione alle densità discrete p_X e p_Y. Chiariamo alcune notazioni. Supponiamo che X assuma i valori $\{x_i\}$ con probabilità $\{p_X(i)\}$ mentre Y assuma i valori $\{y_j\}$ con probabilità $\{p_Y(j)\}$. Allora $Z = \alpha X + \beta Y + \gamma$ assume i valori z della forma $\alpha x_i + \beta y_j + \gamma$ al variare di tutte le coppie (i, j). Pertanto*

$$E\left[\alpha X + \beta Y + \gamma\right] = \sum_{ij} \left(\alpha x_i + \beta y_j + \gamma\right) P\left(X = x_i, Y = y_j\right).$$

La validità di questa identità è abbastanza intuitiva. Se si vuole tracciare una dimostrazione completa si può argomentare così. Il vettore aleatorio (X, Y) ha come densità congiunta i

valori $P(X = x_i, Y = y_j)$ al variare di tutte le coppie (x_i, y_j). Introduciamo la trasformazione $\varphi(x, y) = \alpha x + \beta y + \gamma$. Vale $Z = \varphi(X, Y)$. Per il teorema sulle trasformazioni di v.a. enunciato sopra, vale

$$E[\varphi(X, Y)] = \sum_{ij} \varphi(x_i, y_j) P(X = x_i, Y = y_j).$$

Questa è l'identità enunciata sopra.

Tornando alla linea principale della dimostrazione, in base all'identità scritta, vale

$$\begin{aligned} E[\alpha X + \beta Y + \gamma] &= \alpha \sum_{ij} x_i P(X = x_i, Y = y_j) + \beta \sum_{ij} y_j P(X = x_i, Y = y_j) + \gamma \sum_{ij} P(X = x_i, Y = y_j) \\ &= \alpha \sum_i x_i \sum_j P(X = x_i, Y = y_j) + \beta \sum_j y_j \sum_i P(X = x_i, Y = y_j) + \gamma \sum_{ij} P(X = x_i, Y = \end{aligned}$$

Osservando che

$$\begin{aligned} \sum_j P(X = x_i, Y = y_j) &= P(X = x_i) \\ \sum_i P(X = x_i, Y = y_j) &= P(Y = y_j) \\ \sum_{ij} P(X = x_i, Y = y_j) &= 1 \end{aligned}$$

troviamo

$$\begin{aligned} &= \alpha \sum_i x_i P(X = x_i) + \beta \sum_j y_j P(Y = y_j) + \gamma \\ &= \alpha E[X] + \beta E[Y] + \gamma. \end{aligned}$$

La dimostrazione è completa.

Positività

Se $X \geq 0$ (cioè $X(\omega) \geq 0$ per ogni $\omega \in \Omega$), allora $E[X] \geq 0$.

Osservazione 18 *Questa proprietà può invece essere enunciata anche ricorrendo alla densità di X, in quanto la condizione $X \geq 0$ si può formulare con $f_X(x) = 0$ per ogni $x < 0$. Ovviamente questo si può dire solo se X è una v.a. che ammette densità (continua o discreta).*

Monotonia

Se $X \leq Y$ (cioè $X(\omega) \leq Y(\omega)$ per ogni $\omega \in \Omega$), allora $E[X] \leq E[Y]$. Si vede facilmente, ragionando sulla differenza $Y - X$, che questa proprietà è equivalente alla positività.

1.2.11 Alcuni esempi

Riportiamo alcuni esempi di calcolo del valor medio, alcuni dei quali sfruttano qualche proprietà sopra enunciata.

Esempio 31 *Se $X \equiv c$, allora $E[X] = c$.*

Esempio 32 *Se $X \sim B(1,p)$, si ha che*

$$E[X] = p.$$

Infatti, dalla definizione, $E[X] = 1 \cdot p + 0 \cdot (1-p)$.

Esempio 33 *Se $X \sim B(n,p)$, cioè $P(X=k) = \binom{n}{k} p^k q^{n-k}$, si ha*

$$E[X] = np.$$

I calcoli diretti con la definizione

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}$$

sono laboriosi (si possono fare semplificando $k\binom{n}{k}$, quindi riconducendosi ad espressioni di tipo binomiale con $n-1$ e $k-1$). Meglio sfruttare la linearità del valor medio. Ricordando che la somma di n v.a. di Bernoulli $X_i \sim B(1,p)$ indipendenti è una binomiale $X \sim B(n,p)$, Teorema 5, vale

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = \underbrace{p + p + \cdots + p}_{n \text{ volte}} = np \ .$$

Notiamo che le X_i sono v.a. indipendenti, ma questa ipotesi non è necessaria per ricavare il risultato.

Esempio 34 *Se $X \sim \mathcal{P}(\lambda)$, (v.a. di Poisson di parametro λ), si ha*

$$E[X] = \lambda.$$

Ci si può arrivare dalla definizione

$$E[X] = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!}$$

scrivendo

$$k\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\lambda\frac{\lambda^{k-1}}{(k-1)!}$$

con un po' di calcoli laboriosi ma fattibili. Per convincersi invece in modo rapido del risultato conviene sfruttare il teorema degli eventi rari che stabilisce la convergenza della binomiale $B(n,p_n)$ alla Poisson $\mathcal{P}(\lambda)$ per $n \to \infty$, con $p_n = \lambda/n$. Siccome il valor medio di una $B(n,p_n)$ è $n \cdot p_n$ che vale λ, tutte le approssimanti $B(n,p_n)$ hanno valor medio λ, quindi è intuitivamente chiaro che la Poisson limite $\mathcal{P}(\lambda)$ deve avere anch'essa media λ. L'argomento non del tutto rigoroso, non disponendo in questo momento di opportuni teoremi limite sui valori medi, ma è convincente.

Esempio 35 *Se X è geometrica, $P(X=k) = p(1-p)^k$, $k = 0, 1, ...$, non è immediatamente facile calcolare il valor medio $\sum_{k=0}^{\infty} kp(1-p)^k$. Bisogna conoscere la regola*

$$\sum_{k=0}^{\infty} ka^{k-1} = \frac{1}{(1-a)^2}$$

che vale per ogni a tale che $|a| < 1$. Usando questa formula vale

$$\sum_{k=0}^{\infty} kp(1-p)^k = p(1-p)\sum_{k=0}^{\infty} k(1-p)^{k-1} = \frac{p(1-p)}{(1-(1-p))^2} = \frac{1-p}{p}.$$

La furmula precedente si dimostra derivando la formula nota $\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}$. Queste due funzioni della variabile a coincidono, quindi hanno uguale derivata. Per un teorema di analisi, di può passare la derivata sotto il segno di serie (nelle ipotesi di questo esempio, cioè $|a| < 1$):

$$\frac{d}{da}\sum_{k=0}^{\infty} a^k = \sum_{k=0}^{\infty} \frac{d}{da} a^k = \sum_{k=0}^{\infty} ka^{k-1}.$$

Siccome $\frac{d}{da}\frac{1}{1-a} = \frac{1}{(1-a)^2}$, si ottiene il risultato desiderato.

Esempio 36 *Se X è una v.a. uniforme nell'intervallo $[a, b]$ allora*

$$E[X] = \frac{a+b}{2}.$$

La dimostrazione di questo fatto, intuitivamente abbastanza evidente, è lasciata per esercizio.

Esempio 37 *Se X è una v.a. esponenziale di parametro λ, vale*

$$E[X] = \frac{1}{\lambda}.$$

Infatti

$$\begin{aligned} E[X] &= \int_{-\infty}^{+\infty} xf(x)dx = \int_0^{+\infty} x\lambda e^{-\lambda x}dx \\ &= \left[-xe^{-\lambda x}\right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x}dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^{+\infty} = \frac{1}{\lambda}. \end{aligned}$$

Esempio 38 *Se $X \sim N(\mu, \sigma^2)$, allora $E[X] = \mu$. Per ricavare il risultato, si può calcolare per esercizio l'integrale usando la densità della gaussiana, sfruttando la simmetria della gaussiana rispetto al punto $x = \mu$.*

1.2.12 Proprietà meno elementari del valor medio

In questa sezione enunciamo alcune proprietà che richiedono un po' di lavoro per essere dimostrate ed anche capite. Alcune di esse, pur essendo molto potenti e dal profondo significato, sono di uso corrente solo per chi vuole investigare gli aspetti teorici dell probabilità. Come sopra, negli enunciati che seguono bisogna assumere che tutti i valori medi di cui si parla esistano finiti.

1.2.13 Media di v.a. indipendenti

Teorema 10 *Se due v.a. X e Y sono indipendenti, il valor medio del prodotto è uguale al prodotto dei valori medi, cioè*

$$E[XY] = E[X] \cdot E[Y].$$

Per essere precisi a livello rigorso, assumendo semplicemente che X e Y abbiano valor medio finito, si trova dalla dimostrazione stessa che la v.a. XY ha anch'essa valor medio finito.

Osservazione 19 *Questa proprietà non ha simili tra i fatti elementari sugli integrali di funzioni di una variabile. Esiste invece una proprità che la ricorda nell'ambito degli integrali doppi: per la formula di riduzione, se $f(x,y) = g(x) \cdot h(y)$, vale*

$$\iint_{AB} f(x,y)dxdy = \int_A g(x)dx \int_B h(y)dy.$$

Una possibile dimostrazione rigorosa del teorema poggia proprio su questa proprietà, ma per completare la dimostrazione bisognerebbe capire come fare a passare da $E[XY]$ a un integrale doppio.

Osservazione 20 *Il teorema inverso è falso: $E[XY] = E[X] \cdot E[Y]$ non implica che X e Y sono indipendenti. Lo si può intuire dall'osservazione precedente: l'indipendenza equivale alla proprietà che la densità congiunta è il prodotto delle marginali, mentre l'uguaglianza integrale espressa da $E[XY] = E[X] \cdot E[Y]$ è solo una uguaglianza tra particolari integrali (riassunti) di tali densità.*

Osservazione 21 *Più avanti nel corso vedremo il concetto di vettore gaussiano. In quel momento potremo mostrare che se il vettore (X,Y) è gaussiano, allora la proprietà $E[XY] = E[X] \cdot E[Y]$ implica che X e Y sono indipendenti.*

Osservazione 22 *Dimostriamo la proprietà $E[XY] = E[X] \cdot E[Y]$ nel caso di v.a. discrete, usando le notazioni ed alcuni fatti che si trovano nell'osservazione 17. Vale*

$$E[XY] = \sum_{ij} x_i y_j P(X = x_i, Y = y_j).$$

Per l'indipendenza,

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j).$$

Quindi

$$\begin{aligned} E[XY] &= \sum_{ij} x_i y_j P(X = x_i) P(Y = y_j) = \sum_i x_i P(X = x_i) \sum_j y_j P(Y = y_j) \\ &= E[X] \cdot E[Y]. \end{aligned}$$

1.2.14 Disuguaglianza di Hölder

Date X e Y v.a. qualsiasi, se i valori medi della formula sono ben definiti, si ha

$$E[XY] \leq E\left[X^p\right]^{\frac{1}{p}} \cdot E\left[Y^q\right]^{\frac{1}{q}}$$

con $\frac{1}{p} + \frac{1}{q} = 1,\ p, q > 1$. Come esempio di applicazione, per $p = q = 1/2$ si ha

$$E[XY] \leq \sqrt{E[X^2]}\sqrt{E[Y^2]}.$$

Per capire l'utilità di questa disuguaglianza, si deve pensare al fatto che sappiamo scrivere un'uguaglianza per $E[XY]$ solo per v.a. indipendenti (più in generale scorrelate, si veda oltre). Quindi la disuguaglianza di Hölder ci permette almeno di scrivere una disuguaglianza, in generale.

Ha però il difetto di elevare a potenza le v.a., cosa che in certi ambiti è molto dannoso. Si pensi ad esempio ai problemi di chiusura in fluidodinamica. Quando si considera l'equazione di Navier-Stokes (che è non lineare) si tenta talvolta di ricavare da essa un'equazione per i valori medi della velocità, detta equazione di Reynolds, ma la presenza della nonlinearità fa sì che nell'operazione di media si ottengano valori medi di potenze che non sono riconducibili ai valori medi delle singole v.a. Detto $u(x) = (u_1(x), u_2(x), u_3(x))$ il campo di velocità, bisognerebbe saper esprimere il cosidetto tensore di Reynolds $E\left[u_i(x)u_j(x)\right]$ tramite prodotti del tipo $E\left[u_i(x)\right]E\left[u_j(x)\right]$, ma questo richiederebbe l'indipendenza, che è falsa in generale in questo problema. Purtroppo, anche se si usa la disuguaglianza di Hölder, questa, oltre ad essere una disuguaglianza (quindi servirebbe più che altro per trovare stime per l'equazione di Reynolds piuttosto che una chiusura della stessa), metterebbe in gioco momenti di ordine più elevato, come $E\left[u_i(x)^2\right]$.

1.2.15 Disuguaglianza di Jensen

Data una funzione ϕ convessa e una v.a. X, si ha (se esistono i valori medi considerati)

$$E[\phi(X)] \geq \phi\left(E[x]\right).$$

Ad esempio si ha che

$$E[X^2] \geq \left(E[x]\right)^2$$

che può anche essere dimostrata anche con la disuguaglianza di Hölder, e

$$E[e^X] \geq e^{E[x]}.$$

Questa disuguaglianza ammette una semplice interpretazione grafica.

1.2.16 Disuguaglianza di Chebyshev

Questa potente disuguaglianza che lega una probabilità a un valor medio è talvolta detta anche disuguaglianza di Markov. Se $X \geq 0$ e $a > 0$ si ha che

$$P(X > a) \leq \frac{E[X]}{a}.$$

Proof. Dobbiamo mostrare che $a \cdot P(X \geq a) \leq E[X]$. Mostriamolo nel caso particolare in cui la v.a. X ammetta densità $f(x)$. Poichè $X \geq 0$, si ha $f(x) = 0$ per $x < 0$, quindi

$$E[X] = \int_0^{+\infty} xf(x)dx = \int_0^a xf(x)dx + \int_a^{+\infty} xf(x)dx$$
$$\geq \int_a^{+\infty} xf(x)dx \geq a \int_a^{+\infty} f(x)dx = a \cdot P(X > a).$$

Abbiamo usato il fatto che $\int_0^a xf(x)dx \geq 0$ in quanto la funzione $xf(x)$ è ≥ 0 nell'intervallo d'integrazione $[0, a]$, ed il fatto che la funzione $xf(x)$ è $\geq af(x)$ nell'intervallo d'integrazione $[a, \infty)$. La dimostrazione è completa. ■

Prendendo al posto di X ed a vari esempi di v.a. e numeri positivi, si ottengono numerose conseguenze. Ecco alcuni esempi importanti.

Corollario 1 *Data una v.a. X avente media μ e un numero $a > 0$, si ha*

$$P(|X - \mu| > a) \leq \frac{E[|X - \mu|^2]}{a^2}.$$

Si ha infatti $P(|X - \mu| > a) = P(|X - \mu|^2 > a^2)$, a cui si può applicare la disuguaglianza di Chebyshev. Invece dell'elevamento al quadrato si può usare qualunque funzione monotona crescente ϕ sui positivi, che conservi la disuguaglianza:

$$P(\phi(|X - \mu|) > a) \leq \frac{E[\phi(|X - \mu|)]}{\phi(a)}.$$

Osservazione 23 *Questo corollario è utilissimo quando si usa $\phi = e^{\lambda x}$, $\lambda > 0$. In questo caso a volte si trovano stime dall'alto ottimali (nel senso che valgono analoghe stime dal basso). Vedremo la disuguaglianza di Chernoff.*

Osservazione 24 *Prendiamo ad esempio la semplice disuguaglianza*

$$P(|X - \mu| > a) \leq \frac{E[|X - \mu|]}{a}.$$

Questa (come le altre) ha un'interpretazione grafica: la somma delle due aree sotto le "code" della distribuzione f è abbastanza piccola e può essere controllata col valor medio di $|X - \mu|$. Queste disuguaglianze sono utili quando la f non è nota o non è semplice calcolare probabilità ad essa associate, ed è comunque necessario stimare l'area sotto le code della distribuzione.

1.2.17 Varianza e deviazione standard

Definizione 16 *Sia X una v.a. con valor medio μ finito. Chiamiamo varianza, o scarto quadratico medio, di X, il numero reale*

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

quando questo è finito (se è infinito, diremo che X ha varianza infinita).

La formula descrive appunto lo scarto dalla media μ, $X-\mu$, quadratico, $(X-\mu)^2$, medio, $E\left[(X-\mu)^2\right]$, cioè rimediato rispetto alla distribuzione di probabilità di X.

Si vede subito, sviluppando il quadrato ed usando la linearità del valor medio (ed il fatto che la media di una costante è la costante stessa) che

$$Var\left[X\right] = E\left[X^2\right] - \mu^2.$$

Osserviamo che, essendo la varianza pari alla media di una v.a. positiva, sicuramente

$$Var\left[X\right] \geq 0$$

e quindi

$$\mu^2 \leq E\left[X^2\right].$$

Questa disuguaglianza, vista di per sé, non sarebbe stata così elementare; si poteva però anche dimostrare ad esempio con la disuguaglianza di Hölder o di Jensen.

Osservazione 25 *Una delle disuguaglianze di Chebyshev si può ora riscrivere nella forma* $(a > 0)$

$$P(|X-\mu| > a) \leq \frac{Var\left[X\right]}{a^2}.$$

Osservazione 26 *Si osservi che* $Var\left[X\right] = 0$ *implica* $P(|X-\mu| > a) = 0$ *per ogni* $a > 0$. *Vale cioè* $P(|X-\mu| \leq a) = 1$ *per ogni* $a > 0$. *Intuitivamente, o con l'uso della* σ*-additività, si deduce* $P(X = \mu) = 1$, *cioè* X *è costante. Le uniche variabili per cui* $Var\left[X\right] = 0$, *o equivalentemente* $\mu^2 = E\left[X^2\right]$, *sono le costanti.*

La varianza fornisce un'indicazione media circa lo scarto rispetto a μ, e misura quindi il grado di aleatorietà, di dispersione, la deviazione rispetto al valor medio. E' quindi un indicatore importantissimo. In pratica sarebbe altrettanto importante un indicatore del tipo $E\left[|X-\mu|\right]$, ma questo offrirebbe ben poche possibilità di calcolo a causa del valore assoluto.

Dal punto di vista numerico, però, la varianza si comporta come un quadrato: se stiamo misurando grandezze in metri, con errori di misura dell'ordine dei 10 metri, la varizanza verrà un numero dell'ordine delle centinaia, misurato in metri quadri. Per questo è utile introdurre la *deviazione standard.*

Definizione 17 *Si chiama deviazione standard della v.a.* X *il numero*

$$\sigma\left[X\right] = \sqrt{Var\left[X\right]}$$

(ben definito quando la varianza esiste).

L'estrazione della radice quadrata ci riporta alla giusta unità di misura ed a valori comparabili con quelli in gioco. In un certo senso un po' vago, il quadrato che compare nella definizione di varianza e la radice quadrata introdotta ora si compensano; ciò avviene sicuramente alivello di unità di misura, e grosso m odo di ordine di grandezza delle quantità in gioco; ma non avviene in modo algebrico: essi non si semplificano, se non quando X è costante.

Come per μ, spesso useremo i simboli σ e σ^2 per indicare deviazione e varianza, se è chiaro dal contesto a quale v.a. ci si riferisca. A volte scriveremo anche μ_X e σ_X.

E' facile dimostrare, algebricamente, il seguente fatto.

Proposizione 1 *Dati due numeri reali α, β, vale*

$$\begin{aligned} Var[\alpha X + \beta] &= \alpha^2 Var[X] \\ \sigma[\alpha X] &= |\alpha| \sigma[X]. \end{aligned}$$

Proof. Anche se sipuò dare una dimostrazione più compatta, seguiamo una linea più lunga ma ovvia:

$$\begin{aligned} Var[\alpha X + \beta] &= E\left[(\alpha X + \beta)^2\right] - \mu^2_{\alpha X+\beta} = E\left[\alpha^2 X^2 + 2\alpha\beta X + \beta^2\right] - (\alpha\mu_X + \beta)^2 \\ &= \alpha^2 E\left[X^2\right] + 2\alpha\beta\mu_X + \beta^2 - \alpha^2\mu_X^2 - 2\alpha\beta\mu_X - \beta^2 \\ &= \alpha^2 E\left[X^2\right] - \alpha^2\mu_X^2 = \alpha^2 Var[X]. \end{aligned}$$

Estraendo poi la radice quadrata (e ricordando che $\sqrt{\alpha^2} = |\alpha|$) si ottiene la seconda formula. ■

L'interpretazione è semplice: le traslazioni β non modificano la varianza (come si intuisce pensando ad una densità f e ad una sua traslata); le moltiplicazioni per α hanno effetto quadratico sulla varianza, essendo la varianza un'espressione quadratica; su σ hanno l'effetto di moltiplicarla per $|\alpha|$ (la deviazione di $2X$ è due volte la deviazione di X, ad esempio). In questa interpretazione, se ci si vuole aiutare con un grafico, è essenziale ricordare che, in un piano cartesiano in cui raffiguriamo una densità, i valori della v.a. X sono i punti dell'asse delle ascisse (la v.a. X "vive" sull'asse delle ascisse), quindi operazioni del tipo $\alpha X + \beta$ vanno pensate come traslazioni e omotetie di tale asse (quindi del grafico di f, ma in orizzontale).

Circa la varianza della somma di v.a., vale il seguente fatto.

Proposizione 2 *Date due v.a. X ed Y, con varianza finita, vale in generale*

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov(X, Y).$$

Se inoltre X ed Y sono indipendenti (o almeno scorrelate, si veda più avanti), allora

$$Var[X + Y] = Var[X] + Var[Y].$$

Proof. Come sopra,

$$\begin{aligned} Var[X + Y] &= E\left[(X + Y)^2\right] - \mu^2_{X+Y} = E\left[X^2 + 2XY + Y^2\right] - (\mu_X + \mu_Y)^2 \\ &= E\left[X^2\right] + 2E[XY] + E\left[Y^2\right] - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2 \\ &= \left(E\left[X^2\right] - \mu_X^2\right) + \left(E\left[Y^2\right] - \mu_Y^2\right) + 2\left(E[XY] - 2\mu_X\mu_Y\right) \end{aligned}$$

da cui la tesi. ■

La definizione di $Cov(X, Y)$ e la spiegazione di questo risultato verranno date nel prossimo paragrafo. Algebricamente, la prima uguaglianza è semplicemente il fatto che il quadrato della

somma è pari alla somma dei quadrati più il doppio prodotto. La seconda deriva dal fatto che per variabili indipedenti vale $Cov(X,Y)=0$.

Dall'ultima affermazione della proposizione si trova, se X ed Y sono indipendenti,

$$\sigma[X+Y]=\sqrt{\sigma^2[X]+\sigma^2[Y]}.$$

Se invece prendessimo $Y=X$, troveremmo

$$\sigma[2X]=2\sigma[X].$$

Consideriamo due v.a. X ed Y aventi uguale σ. Se sono indipendenti, allora

$$\sigma[X+Y]=\sqrt{2}\sigma$$

mentre se sono uguali

$$\sigma[2X]=2\sigma.$$

La variabilità della somma di grandezze indipendenti è inferiore a quella di grandezze uguali. Per questo, se si possiede una ricchezza V e la si vuole investire in attività che contengono un rischio (cioè tali per cui il valore della ricchezza può variare aleatoriamente), conviene suddividerla in parti ed investire le parti in attività indipendenti. In questo modo il rischio diminuisce rispetto ad un singolo investimento globale.

1.2.18 Covarianza e coefficiente di correlazione

Definizione 18 *Date due v.a. X,Y, si chiama covarianza tra X ed Y il numero*

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

dove μ_X e μ_Y sono le medie di X ed Y. La definizione ha senso se μ_X e μ_Y sono finiti ed il valor medio complessivo è finito, cosa che accade ad esempio se si suppone che sia $E\left[X^2\right]<\infty$ e $E\left[Y^2\right]<\infty$.

La definizione è quindi analoga, algebricamente, a quella di varianza, e risulta infatti

$$Var[X]=Cov(X,X)$$

e

$$Cov(X,Y)=E[XY]-\mu_X\mu_Y$$

come per la varianza. Però il numero $Cov(X,Y)$ può avere segno qualsiasi. Ad esempio, se $\mu_X=0$ e prendiamo $Y=-X$, vale $Cov(X,Y)=-E\left[X^2\right]$.

Anche la covarianza soffre dei problemi di scala illustrati per la varianza. Qui, non potendo prendere la radice quadrata ($Cov(X,Y)$ non è sempre positiva), si normalizza in quest'altro modo, dividendo per le deviazioni standard.

Definizione 19 *Chiamiamo coefficiente di correlazione tra X ed Y il numero definito da*

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

Si noti che, per la disuguaglianza di Hölder,

$$|Cov(X,Y)| \leq \sqrt{E\left[(X-\mu_X)^2\right] E\left[(Y-\mu_Y)^2\right]}$$

e quindi $|\rho(X,Y)| \leq 1$. Questo dimostra la prima delle seguenti proprietà, che tutte insieme chiariscono l'aspetto di universalità, o invarianza per cambio di unità di misura, di ρ, a differenza della covarianza.

Proposizione 3 *Vale*

$$-1 \leq \rho(X,Y) \leq 1.$$

Vale inoltre

$$Cov(aX,bY) = abCov(X,Y)$$

per ogni $a,b \in \mathbb{R}$, *e*

$$\rho(aX,bY) = \rho(X,Y)$$

per ogni $a,b>0$.

Proof. Abbiamo già visto come mai $-1 \leq \rho(X,Y) \leq 1$. Dimostriamo la seconda proprietà. Vale

$$\begin{aligned} Cov(aX,bY) &= E[(aX-\mu_{aX})(bY-\mu_{bY})] = E[(aX-a\mu_X)(bY-b\mu_Y)] \\ &= abE[(X-\mu_X)(Y-\mu_Y)] = abCov(X,Y). \end{aligned}$$

Vale poi

$$\rho(aX,bY) = \frac{Cov(aX,bY)}{\sqrt{Var[aX]\,Var[bY]}} = \frac{abCov(X,Y)}{\sqrt{a^2b^2}\sqrt{Var[X]\,Var[Y]}} = \frac{ab}{|ab|}\rho(X,Y)$$

e quindi la formula desiderata, se $a,b>0$. ■

Nello stesso modo si dimostra la seguente proprietà, che in un certo senso è la linearità della convarianza nei suoi argomenti. Si noti che le costanti additive spariscono, come per la varianza.

Proposizione 4

$$Cov(aX+bY+c,Z) = aCov(X,Z) + bCov(Y,Z)$$

$$Cov(X,\alpha Y+\beta Z+\gamma) = \alpha Cov(X,Y) + \beta Cov(X,Z).$$

Proof. Basta dimostrare la prima in quanto la covarianza è simmetrica. Vale

$$\begin{aligned} Cov(aX+bY+c,Z) &= E\left[\left(aX+bY+c-\mu_{aX+bY+c}\right)(Z-\mu_Z)\right] \\ &= E[(a(X-\mu_X)+b(Y-\mu_Y))(Z-\mu_Z)] \\ &= aCov(X,Z) + bCov(Y,Z). \end{aligned}$$

■

Ricordiamo che se X ed Y sono v.a. indipendenti, allora $E[XY] = \mu_X\mu_Y$ (mentre il viceversa non è vero in generale). Ne discende subito il seguente risultato.

Teorema 11 *Se X ed Y sono v.a. indipendenti, allora*

$$Cov(X,Y)=0, \quad \rho(X,Y)=0.$$

Viceversa, se $Cov(X,Y)=0$, non è detto che X ed Y siano indipendenti. Se però (X,Y) è gaussiano (definizione che daremo nel seguito) e $Cov(X,Y)=0$, allora X e Y sono indipendenti.

Definizione 20 *Diciamo che X e Y sono scorrelate se hanno correlazione nulla, $\rho(X,Y)=0$, o equivalentemente se $Cov(X,Y)=0$.*

Quindi l'indipendenza implica la scorrelazione.

A livello numerico su dati sperimentali, se la correlazione è molto vicino a zero, questo è un buon indicatore di indipendenza, o più precisamente di scorrelazione (invece, dipendendo il numero $Cov(X,Y)$ dalla scala scelta, la sua vicinanza a zero è meno assoluta, quindi può trarre in inganno). Precisiamo cosa intendiamo con correlazione di dati sperimentali. Stiamo pensando di avere n coppie di valori sperimentali (x_1,y_1), ... , (x_n,y_n), o più espressivamente una tabella

	X	Y
1	x_1	y_1
...	...	...
...	...	...
n	x_n	y_n

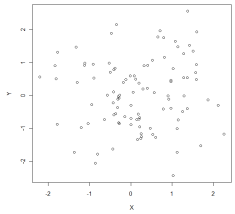
in cui le colonne corrispondono alle variabili e le righe agli "individui" (unità sperimentali, unità osservate). Di questi dati sperimentali possiamo calcolare la *varianza empirica* ed il *coefficiente di correlazione empirico* definiti da

$$\widehat{Cov}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}), \quad \widehat{\rho}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Questi indicatori sono buone stime di quelli teorici, ad esempio per via della legge dei grandi numeri, che descriveremo nella prossima lezione. Fatte queste premesse, la vicinanza a zero di $\widehat{\rho}$ si interpreta come sintomo di indipendenza o comunque bassa dipendenza, la vicinanza ad 1 come elevato legame positivo, a -1 come elevato legame negativo.

Esaminiamo questi fatti per mezzo del software R. Innanzi tutto generiamo due campioni di cardinalità 100, con distribuzione gaussiana standard, mostriamo le coppie (x_i,y_i) nel piano cartesiano e calcoliamo la correlazione empirica:

```
X=rnorm(100); Y=rnorm(100)
cor(X,Y)
[1] 0.06068838
plot(X,Y)
```

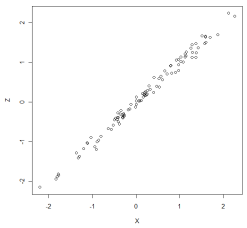


Questa è una situazione a correlazione sostanzialmente nulla. Costruiamo invece un campione Z simile a X ma un po' perturbato in modo aleatorio:

```
Z=X+0.1*rnorm(100)
cor(X,Z)
```

[1] 0.9949628

```
plot(X,Z)
```

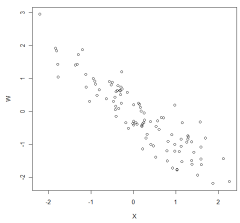


Questa è una situazione ad elevatissima correlazione positiva. Proviamo poi

```
W=-X+0.5*rnorm(100)
cor(X,W)
```

[1] -0.8987381

```
plot(X,W)
```



Quest'ultima è una situazione a moderata/elevata correlazione negativa. Si noti che il coefficiente di correlazione non è esattamente uguale al coefficiente angolare, come si potrebbe pensare dai nomi. Si veda sotto il legame.

Menzioniamo infine il fatto che il numero $Cov(X, Y)$ descrive bene l'eventuale legame *lineare* tra X ed Y (mentre è meno preciso per legami non lineari). Si può ad esempio dimostrare facilmente che, se X ed Y sono legate dalla relazione lineare

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

dove ε (chiamato "errore") è una v.a. indipendente da X, allora il coefficiente α che descrive

il legame di proporzionalità lineare tra le variabili è dato da

$$\alpha = \frac{Cov(X,Y)}{Var[X]}.$$

Lo si può verificare calcolando

$$Cov(X,Y) = Cov(X, \alpha X + \beta + \varepsilon)$$

ed applicando la linearità della covarianza nella seconda variabile.

1.2.19 Esempi

Esempio 39 *Se $X \sim B(1,p)$, si vede subito che anche $X^2 \sim B(1,p)$, quindi $E\left[X^2\right] = p$. Pertanto*

$$E\left[X^2\right] - \mu^2 = p - p^2 = pq.$$

Per una Bernoulli di parametro p vale allora

$$Var[X] = pq$$

e $\sigma = \sqrt{pq}$.

Esempio 40 *Se $X \sim B(n,p)$, usando il fatto che la somma di n v.a. $B(1,p)$ indipendenti è una $B(n,p)$, e ricordando che la varianza della somma di v.a. indipendenti è uguale alla somma delle varianze, troviamo*

$$Var[X] = npq$$

e

$$\sigma = \sqrt{n}\sqrt{pq}.$$

Quest'ultimo fatto era già stato anticipato in un esempio della lezione 2, riguardo al fatto che per n grande la binomiale si "concentra" intorno alla propria media.

Esempio 41 *Se $X \sim \mathcal{P}(\lambda)$, vale*

$$Var[X] = \lambda.$$

Questo fatto si può dimostrare rigorosamente usando la densità di massa, ma richiede un certo numero di calcoli un po' noiosi. Accontentiamoci di accettare il risultato sulla base del seguente ragionamento sensato (ma non rigoroso): prendiamo una v.a. $X_n \sim B(n,p_n)$, con $\lambda = np_n$. Se n è grande, sappiamo che la legge di X_n approssima la legge di X; allora anche la varianza di X_n, che è np_nq_n dovrebbe approssimare la varianza di X; ma $np_nq_n = \lambda q_n$ e $q_n \to 1$ per $n \to \infty$ (in quanto $q_n = 1 - p_n = 1 - \frac{\lambda}{n}$).

Esempio 42 *Se $X \sim N\left(\mu, \sigma^2\right)$, vale*

$$Var[X] = \sigma^2.$$

Nel prossimo paragrafo svolgeremo un conto di questo tipo ma più complesso, per cui ora omettiamo la verifica. Quindi i due parametri μ e σ^2 della normale $N\left(\mu, \sigma^2\right)$ sono la sua media e la sua varianza (come le notazioni lasciavano pensare).

1.2.20 Momenti

Chiamiamo *momento di ordine* n di X il numero

$$M_n := E[X^n].$$

A volte, a seconda delle utilità specifiche, si trova in letteratura il nome di momento di ordine n attribuito a quantità lievemente diverse, come

$$E[|X|^n]$$

oppure

$$E[(X-\mu)^n] \text{ o infine } E[|X-\mu|^n].$$

La ragione del valore assoluto è che se X ha distribuzione simmetrica, per n dispari vale $E[X^n] = 0$, fatto che quantitativamente può non essere molto istruttivo (dipende da cosa si vuol evidenziare con quella grandezza). La ragione della centratura con μ è simile a quella per cui si centra la definizione di varianza (si vuole capire lo scostamento dalla media e non l'ampiezza assoluta dei valori).

Mentre è evidente l'interesse per media e varianza (ad esempio sono i parametri delle gaussiane), meno chiaro è come utilizzare i momenti di ordine superiore a due. Ad esempio, vengono a volte utilizzati in statistica per confrontare due distribuzioni sperimentali, o una distribuzione sperimentale con un modello ipotizzato, ad esempio quello gaussiano. Il confronto dei momenti di ordine elevato mette in evidenza possibili differenze significative tra le code, tra le probabilità di valori un po' alti. Cerchiamo di apprezzare questo fatto con un esempio.

Supponiamo di avere un istogramma sperimentale e di cercare una densità $f(x)$ che lo descriva. Supponiamo di aver individuato due densità $f_1(x)$ ed $f_2(x)$ che descrivono bene l'istogramma nella parte centrale, ma abbiamo dei dubbi sulle code. Per semplicità, supponiamo di studiare una v.a. positiva, quindi solo con la coda a destra. Per schematizzare ulteriormente a titolo di esempio, abbandoniamo le densità e supponiamo di aver scelto come possibili modelli due v.a. discrete, entrambe con solo i valori 2 e 10. La prima, X_1, assume il valore 2 con probabilità 0.99 e 10 con probabilità 0.01. La seconda, X_2, assume 2 con probabilità 0.999 e 10 con probabilità 0.001. I loro momenti di ordine n sono

$$E[X_1^n] = 2^n \cdot 0.99 + 10^n \cdot 0.01$$

$$E[X_2^n] = 2^n \cdot 0.999 + 10^n \cdot 0.001.$$

Vediamo allora che per valori bassi di n i momenti sono abbastanza simili;

$$E[X_1] = 2.08, \quad E\left[X_1^2\right] = 4.96$$

$$E[X_2] = 2.008, \quad E\left[X_2^2\right] = 4.096$$

e quindi è possibile che, sulla base di stime empiriche di media e varianza, non siamo in grado di decidere quale delle due distribuzioni sia la migliore. Invece i momenti di ordine più elevato divergono tra loro: ad esempio

$$E\left[X_1^4\right] = 115.84, \quad E\left[X_2^4\right] = 25.984.$$

Essi quindi diventano indicatori numerici discriminanti. Va però osservato che, per le stesse ragioni, sono molto più sensibili dei momenti di ordine basso rispetto a piccole variazioni casuali, come l'errore statistico dovuto a pochi dati sperimentali, oppure vari errori numerici di approssimazione nella raccolta dati ecc. Quindi la diversità, magari estremamente marcata, tra i momenti di ordine elevato di due campioni sperimentali va usata con cautela. In statistica cercheremo di capire gli intervalli di confidenza per tali indicatori.

A titolo di esempio, calcoliamo i momenti di una gaussiana, per capirne il comportamento al crescere di n.

Osservazione 27 *Se $X \sim N\left(\mu, \sigma^2\right)$, allora $E\left[(X-\mu)^{2n+1}\right] = 0$, mentre*

$$E\left[(X-\mu)^{2n}\right] = C_n\left(\sigma^2\right)^n$$

dove

$$C_n = (2n-1)(2n-3)\cdots 3\cdot 1.$$

Infatti, limitando (senza restrizione) la verifica al caso $\mu = 0$, vale

$$E\left[X^{2n}\right] = \int_{-\infty}^{\infty} x^{2n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}dx \overset{x=\sigma y}{=} \int_{-\infty}^{\infty}\sigma^{2n}y^{2n}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy = \sigma^{2n}E\left[Z^{2n}\right]$$

dove $Z \sim N(0,1)$, e per questa vale

$$\begin{aligned} E\left[Z^{2n}\right] &= \frac{-1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{2n-1}(-x)\,e^{-\frac{x^2}{2}}dx \\ &= \left[\frac{-1}{\sqrt{2\pi}}x^{2n-1}e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(2n-1)\,x^{2n-2}e^{-\frac{x^2}{2}}dx \\ &= (2n-1)\,E\left[Z^{2n-2}\right]. \end{aligned}$$

Iterando,

$$E\left[Z^{2n}\right] = (2n-1)(2n-3)\,E\left[Z^{2n-4}\right] = ... = C_n.$$

Osservazione 28 *Volendo sintetizzare graficamente questo risultato, si può osservare che la grandezza*

$$y_n := \log\frac{E\left[(X-\mu)^{2n}\right]}{C_n}$$

cresce linearmente in n:

$$y_n = n\log\sigma^2$$

quindi se riportiamo in un grafico in ascissa gli interi n ed in ordinata i numeri y_n per una gaussiana troviamo punti su una retta passante per l'origine, di coefficiente angolare $\log\sigma^2$. In questo modo, la visualizzazione dei numeri y_n per un'altra distribuzione oppure per un campione sperimentale, mette subito in evidenza l'eventuale scostamento dalla gaussianità.

1.2.21 La funzione generatrice dei momenti

Definizione 21 *Data una v.a. X, si chiama sua funzione generatrice dei momenti la funzione $\varphi_X(t)$ definita da*

$$\varphi_X(t) = E\left[e^{tX}\right]$$

per tutti i valori t per cui tale valore atteso è finito.

La funzione generatrice non è sempre definita per ogni $t \in \mathbb{R}$, come vedremo ad esempio per le v.a. esponenziali. Osserviamo che $\varphi_X(0) = 0$, semplice fatto che a volte si usa per stabilire che certe funzioni non sono funzioni generatrici.

Nel caso di una v.a. X con densità $f(x)$ vale

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

(forse alcuni riconosceranno la trasformata di Laplace di f, in questa espressione) mentre nel caso discreto a valori interi positivi vale

$$\varphi_X(t) = \sum_{n=0}^{\infty} e^{tn} p(n).$$

Vale il seguente fatto:

Teorema 12 *Se due v.a. hanno la stessa funzione generatrice, per t in un intervallo aperto non vuoto, allora hanno la stessa legge.*

La dimostrazione non è semplice ed è legata ai problemi di inversione della trasformata di Fourier, che non esponiamo. Dimostriamo invece un fatto semplice ma importante:

Proposizione 5 *Se X ed Y sono indipendenti allora*

$$\varphi_{X+Y}(t) = \varphi_X(t)\,\varphi_Y(t).$$

Proof.

$$\varphi_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E\left[e^{tX}e^{tY}\right]$$

ed ora, per l'indipendenza (di X ed Y, che implica quella di e^{tX} ed e^{tY})

$$= E\left[e^{tX}\right] E\left[e^{tY}\right] = \varphi_X(t)\,\varphi_Y(t).$$

■

Esercizio 1 *Mostrare che, se α, β sono due numeri reali, allora*

$$\varphi_{\alpha X+\beta}(t) = \varphi_X(\alpha t)\, e^{\beta t}.$$

Esercizio 2 *Mostrare che, se X ed Y sono v.a. indipendenti ed a, b, c sono numeri reali, allora*

$$\varphi_{aX+bY+c}(t) = \varphi_X(at)\,\varphi_Y(bt)\, e^{ct}.$$

Esempio 43 *La funzione generatrice una Bernoulli $X \sim B(1,p)$ è*

$$\varphi_X(t) = pe^t + q.$$

Esempio 44 *Sia $X \sim B(n,p)$ una binomiale, della forma $X = X_1 + ... + X_n$ con $X_1, ..., X_n$ Bernoulli $B(1,p)$ indipendenti. Allora, per la proposizione applicata iterativamente,*

$$\varphi_X(t) = \left(pe^t + q\right)^n.$$

Siccome la generatrice dipende solo dalla legge, il risultato vale anche se la binomiale, a priori, non è espressa in tale forma.

Esempio 45 *La funzione generatrice di una v.a. X di Poisson, $X \sim \mathcal{P}(\lambda)$, è*

$$\varphi_X(t) = e^{\lambda\left(e^t-1\right)}.$$

Si può calcolare dalla definizione:

$$\varphi_X(t) = \sum_{n=0}^{\infty} e^{tn} e^{-\lambda} \frac{\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\left(\lambda e^t\right)^n}{n!} = e^{-\lambda} e^{\lambda e^t}.$$

Esercizio 3 *Verificare che il limite delle generatrici di binomiali $X_n \sim B\left(n, \frac{\lambda}{n}\right)$ è uguale alla generatrice di una $X \sim \mathcal{P}(\lambda)$.*

Esempio 46 *Se X è una v.a. esponenziale, $X \sim Exp(\lambda)$, allora*

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} e^{(t-\lambda)x} dx$$

dove, osserviamo fin da ora, questo integrale converge (ha valore finito) se e solo se $t - \lambda < 0$, cioè se

$$t < \lambda.$$

Per questi valori di t troviamo

$$\varphi_X(t) = \lambda \left[\frac{e^{(t-\lambda)x}}{t-\lambda}\right]_0^{\infty} = \frac{\lambda}{\lambda - t}.$$

In questo esempio la funzione generatrice non è definita per tutti i valori di t.

Esempio 47 *Una v.a. $X \sim N(0,1)$ ha funzione generatrice*

$$\varphi_X(t) = e^{t^2/2}.$$

Infatti

$$\begin{aligned} E\left[e^{tX}\right] &= \frac{1}{\sqrt{2\pi}} \int e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int e^{\frac{t^2}{2}} e^{-\frac{\left(x^2-2tx+t^2\right)}{2}} dx \\ &= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \int e^{-\frac{(x-t)^2}{2}} dx = e^{\frac{t^2}{2}}. \end{aligned}$$

Esempio 48 *Più in generale, una v.a.* $X \sim N\left(\mu, \sigma^2\right)$ *ha funzione generatrice*

$$\varphi_X(t) = e^{t\mu + \frac{\sigma^2 t^2}{2}}.$$

Infatti, con gli stessi calcoli fatti nel caso standard ma più laboriosi,

$$\begin{aligned} E\left[e^{tX}\right] &= \frac{1}{\sigma\sqrt{2\pi}} \int e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int e^{\frac{\left(t\sigma^2+\mu\right)^2-\mu^2}{2\sigma^2}} e^{-\frac{\left(x^2-2\left(\sigma^2 t+\mu\right)x+\left(t\sigma^2+\mu\right)^2\right)}{2\sigma^2}} dx \\ &= \frac{e^{\frac{\sigma^2 t^2}{2}+t\mu}}{\sigma\sqrt{2\pi}} \int e^{-\frac{\left(x-t\sigma^2-\mu\right)^2}{2\sigma^2}} dx = e^{t\mu+\frac{\sigma^2 t^2}{2}}. \end{aligned}$$

Il motivo del nome "generatrice dei momenti" sta nel fatto che derivando la funzione generatrice e calcolando le derivate in zero si ottengono i momenti.

Teorema 13 *Se la funzione generatrice* $\varphi_X(t)$ *è definita in un intervallo aperto non vuoto contenente l'origine allora è infinite volte derivabile in zero e vale*

$$\varphi_X^{(n)}(0) = E\left[X^n\right].$$

Non diamo la dimostrazione completa che fa uso di teoremi di scambio tra derivate e integrali che non trattiamo nel corso, ma riportiamo solo formalmente i seguenti passaggi (naturali ed in realtà anche rigorosi)

$$\begin{aligned} \frac{d}{dt} E\left[e^{tX}\right] &= E\left[\frac{d}{dt} e^{tX}\right] = E\left[Xe^{tX}\right] \\ \frac{d^2}{dt^2} E\left[e^{tX}\right] &= E\left[\frac{d^2}{dt^2} e^{tX}\right] = E\left[X^2 e^{tX}\right] \end{aligned}$$

e così via, per cui

$$\begin{aligned} \varphi_X'(0) &= E[X] \\ \varphi_X''(0) &= E\left[X^2\right] \end{aligned}$$

e così via. Con queste regole si possono ritrovare numerosi valori medi calcolati fino ad ora. Vediamo a titolo di esempio il caso delle geometriche, che era risultato piuttosto difficile.

Esempio 49 *Ricordiamo che chiamiamo geometrica di parametro* p *una v.a. tale che*

$$P(X = n) = (1-p)^n p \qquad \text{per } n = 0, 1, ...$$

Allora

$$\varphi_X(t) = \sum_{n=0}^{\infty} e^{tn} (1-p)^n p = p \sum_{n=0}^{\infty} \left[(1-p)\, e^t\right]^n = \frac{p}{1-(1-p)\, e^t}$$

dove l'ultimo passaggio vale se $(1-p)\,e^t < 1$, *quindi se* $e^t < \frac{1}{1-p}$, *ovvero* $t < \log\left(\frac{1}{1-p}\right)$. *Allora*

$$\varphi'_X(t) = \frac{p\,(1-p)\,e^t}{(1-(1-p)\,e^t)^2}$$

da cui

$$\varphi'_X(0) = \frac{1-p}{p}.$$

Esempio 50 *Se* X' *è geometrica modificata, allora* $X = X' - 1$ *è geometrica, quindi*

$$E\left[X'\right] = E\left[X+1\right] = \frac{1-p}{p} + 1 = \frac{1}{p}.$$

La media di una geometrica modificata è $\frac{1}{p}$.

Osservazione 29 *Con calcoli un po' più laboriosi si verifica che la varianza di una geometrica è* $\frac{1-p}{p^2}$. *Allora anche la varianza di una geometrica modificata è* $\frac{1-p}{p^2}$, *in quanto le due differiscono per una costante.*

1.2.22 Definizione generale di valor medio

Sia $X : \Omega \to [0,\infty)$ una v.a. non negativa. Per ogni numero della forma $\frac{k}{2^n}$, con $n, k \in \mathbb{N}$, consideriamo l'evento

$$A_{n,k} = \left\{ X \in \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right) \right\}$$

e introduciamo la v.a.

$$X_n = \sum_{k=0}^{\infty} \frac{k}{2^n} 1_{A_{n,k}}$$

dove 1_A è la funzione indicatrice di A (che vale uno in A e zero fuori). La funzione $X_n : \Omega \to [0,\infty)$ è "costante a tratti", per così dire; prende il valore $\frac{k}{2^n}$ sull'insieme $A_{n,k}$, dove $X \geq \frac{k}{2^n}$, quindi in ogni punto di Ω vale

$$X_n \leq X.$$

Definiamo

$$E\left[X_n\right] := \sum_{k=0}^{\infty} \frac{k}{2^n} P\left(A_{n,k}\right)$$

dove la serie, essendo a termini positivi, può o convergere o divergere a più infinito. Poi definiamo

$$E\left[X\right] = \lim_{n\to\infty} E\left[X_n\right].$$

Questo limite esiste sempre in quanto la successione numerica $E\left[X_n\right]$ è monotona non decrescente: invece di dimostrarlo algebricamente, suggeriamo di rendersi conto graficamente del fatto che $X_{n+1} \geq X_n$, da cui segue la monotonia delle serie. Il limite che definisce $E\left[X\right]$ può nuovamente essere finito oppure uguale a $+\infty$.

In questo modo abbiamo definito la media di una v.a. positiva *qualsiasi* (accettando anche il valore $+\infty$). Si dice poi che una tale v.a. X *ha media finita*, o *è integrabile*, se risulta $E[X] < \infty$.

Data poi una v.a. $X : \Omega \to \mathbb{R}$ (non necessariamente positiva), la si può scrivere come differenza di v.a. positive:

$$X = X^+ - X^-$$

dove

$$X^+ = \max\{X, 0\}, \quad X^- = X^+ - X.$$

Entrambi i valori medi $E[X^+]$ e $E[X^-]$ sono ben definiti e sarebbe naturale definire $E[X] = E[X^+] - E[X^-]$, ma se entrambi fossero pari a $+\infty$ troveremmo una forma indeterminata. Si stabilisce allora che, se almeno uno dei due, tra $E[X^+]$ e $E[X^-]$, è finito, si pone

$$E[X] = E\left[X^+\right] - E\left[X^-\right]$$

con le usuali convenzioni tra somma di numeri finiti ed infiniti.

Con questa definizione abbiamo introdotto il valor medio per una grandissima classe di v.a. e come risultato possiamo trovare un numero reale oppure $+\infty$ oppure $-\infty$. Diremo poi che X *ha media finita*, o *è integrabile*, se risulta $E[X^+] < \infty$ e $E[X^-] < \infty$, nel qual caso $E[X]$ è un numero reale. Questa condizione equivale a $E[|X|] < \infty$.

1.2.23 Proprietà generali

Sviluppare rigorosamente tutta la teoria dell'integrazione occuperebbe l'intero corso, quindi ci limitiamo ad indicare qualche traccia.

La definizione di $E[X]$ nel caso $X \geq 0$ è vagamente simile alle definizioni di integrale ben note nei corsi di analisi di base, per funzioni reali di una variabile reale. E' quindi intuitivamente chiaro che varranno le proprietà generali note in quell'ambito, che sono la linearità, la positività (o equivalentemente la monotonia), e ad esempio l'additività rispetto a decomposizioni del dominio:

$$E[X \cdot 1_{A \cup B}] = E[X \cdot 1_A] + E[X \cdot 1_B]$$

se $A \cap B = \emptyset$. Così di seguito, tutte le proprietà anche meno banali si possono dimostrare usando la definizione. A titolo di esempio, discutiamo la disuguaglianza di Chebishev. Data $X \geq 0$ e le sue approssimanti X_n, preso un qualsiasi $k_0 > 0$, dalla definizione di $E[X_n]$ abbiamo

$$\begin{aligned} E[X_n] &\geq \sum_{k=k_0}^{\infty} \frac{k}{2^n} P(A_{n,k}) \geq \frac{k_0}{2^n} \sum_{k=k_0}^{\infty} P(A_{n,k}) \\ &= \frac{k_0}{2^n} P\left(\bigcup_{k \geq k_0} A_{n,k}\right) = \frac{k_0}{2^n} P\left(X \geq \frac{k_0}{2^n}\right). \end{aligned}$$

Inoltre, essendo $E[X_n]$ non decrescente, $E[X] \geq E[X_n]$, quindi

$$E[X] \geq \frac{k_0}{2^n} P\left(X \geq \frac{k_0}{2^n}\right).$$

Questa disuguaglianza vale per ogni $n, k_0 > 0$, quindi vale per ogni numero reale positivo a della forma $\frac{k_0}{2^n}$:

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

A questo punto, con un ragionamento limite che non discutiamo in dettaglio, è facile passare ad ogni numero reale positivo a, completando la dimostrazione.

Infine, supponiamo che X abbia densità $f(x)$, nel senso che valga

$$P(X \in I) = \int_I f(x)dx$$

per ogni boreliano I, quindi in particolare per ogni intervallo I. Risulta allora, sempre nell'ipotesi $X \geq 0$,

$$E[X_n] = \sum_{k=0}^{\infty} \frac{k}{2^n} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} f(x)dx.$$

Trascurando il rigore, che qui non è il nostro scopo, osserviamo che, quando n è grande e quindi l'intervallo $\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)$ è piccolo, in tale intervallo di integrazione la funzione x è circa uguale a $\frac{k}{2^n}$, quindi approssimativamente

$$E[X_n] \sim \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} xf(x)dx = \int_{-\infty}^{\infty} xf(x)dx$$

e questa approssimazione diventa sempre più precisa se $n \to \infty$. Quindi ci aspettiamo che sia

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

che è appunto uno dei teoremi forndamentali per il calcolo dei valori medi. Con un ragionamento simile si trova la formula più elaborata

$$E[\varphi(X_1, ..., X_n)] = \int_{-\infty}^{\infty} \varphi(x_1, ..., x_n) f(x_1, ..., x_n)dx_1...dx_n.$$

1.3 Esempi

Abbiamo già visto, nelle sezioni precedenti, alcuni esempi di v.a. discrete e continue ed alcuni loro legami. Usando le solite notazioni, cioè $B(1,p)$ per le Bernoulli, $B(n,p)$ per le binomiali, $\mathcal{P}(\lambda)$ per le Poisson, $Exp(\lambda)$ per le esponenziali, $N\left(\mu, \sigma^2\right)$ per le gaussiane, riassumiamo alcuni fatti salienti con la seguente tabella:

	media	varianza	generatrice
Bernoulli	p	pq	$pe^t + q$
binomiale	np	npq	$\left(pe^t + q\right)^n$
Poisson	λ	λ	$e^{\lambda\left(e^t - 1\right)}$
esponenziale	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - t}$
gaussiana	μ	σ^2	$e^{t\mu + \frac{\sigma^2 t^2}{2}}$

Abbiamo inoltre visto che la somma di n Bernoulli $B(1,p)$ è una binomiale $B(n,p)$; e che il limite di binomiali $B(n,p_n)$ quando $n \to \infty$ e $np_n = \lambda$ è una $\mathcal{P}(\lambda)$. Cominciamo ad approfondire altri fatti riguardanti queste variabili ed il loro legami, poi vedremo anche altri esempi di variabili.

1.3.1 Una proprietà di concentrazione delle binomiali

Con p fissato (quindi diversamente dal regime del teorema degli eventi rari), cerchiamo di capire cosa accade ad una binomiale $B(n,p)$ per n elevato. La media np diventa grande, linearmente in n. La deviazione standard σ, che misura nella stessa scala della media le variazioni rispetto alla media stessa, vale $\sqrt{n}\sqrt{pq}$, quindi cresce anch'essa con n, ma solo come una radice quadrata, molto meno che la media. Ad esempio, se $n = 100^2 = 10000$, e per esemplificare prendiamo $p = \frac{1}{2}$, quindi $pq = \frac{1}{4}$, vale

$$\mu = 10000 \cdot \frac{1}{4}$$
$$\sigma = 100 \cdot \frac{1}{2}.$$

La variabile $B(n,p)$ è incredibilmente concentrata attorno alla sua media. Percepiamo con un esempio le conseguenze pratiche di questo fatto.

Esempio 51 *Una banca ha 1000 conti correnti aperti. Attribuisce i numeri da 1 a 1000 ai suoi correntisti. La direzione della banca vuole conoscere il numero medio di correntisti che si presenta nell'arco di una giornata, e la probabilità che si presentino più di k correntisti, al variare di k, per poter dimensionare le scorte e gli sportelli aperti.*

Bisogna operare delle idealizzazioni, tenendo quindi presente che il risultato sarà un'approssimazione della realtà. Come vedremo, ci servirà supporre che i 1000 correntisti si comportino in modo indipendente, che ciascuno si presenti al più una volta al giorno, e che la probabilità p che il singolo correntista si presenti sia uguale per tutti i correntisti. Supponiamo inoltre di conoscere questa probabilità p; per fare i conti, supponiamo valga

$$p = \frac{1}{5}$$

(che corrisponde intuitivamente a dire che ogni correntista si presenta mediamente una volta alla settimana).

La banca associa ad ogni correntista una v.a. di Bernoulli, X_1 per il primo, e così via fino ad X_{1000} per l'ultimo. La v.a. X_1 vale 1 se il correntista si presenta in banca durante il giorno in questione, 0 altrimenti. Vale $p = P(X_k = 1)$ per ogni correntista k. Finalmente, la nuova v.a. definita da $S = X_1 + ... + X_{1000}$ rappresenta il numero di correntisti che si presentano in banca (infatti i vari addendi valgono 1 per ogni correntista che si presenta, zero altrimenti). Pertanto S descrive ciò che interessa alla banca. Per il teorema sul legame tra Bernoulli e binomiale, $S \sim B(1000, \frac{1}{5})$. Il numero medio di correntisti al giorno vale quindi

$$E[S] = np = \frac{1000}{5} = 200$$

come ci si poteva aspettare intuitivamente dal fatto che ogni correntista visita la banca in media una volta alla settimana. Questo risultato medio quindi non sorprende, non è un grosso successo della teoria.

Invece, assai meno banale sarebbe calcolare la probabilità che S superi un generico valore k. Ad esempio, visto che il numero medio è 200, ci chiediamo: quante volte, in percentuale, il numero di clienti sarà maggiore di 300? Si provi a immaginare intuitivamente il risultato: si vedrà che il risultato rigoroso è davvero sorprendente.

Dobbiamo calcolare

$$P(S > 300).$$

Vale allora

$$\begin{aligned} P(S > 300) &= \sum_{k=301}^{1000} \binom{1000}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{1000-k} \\ &= 1 - \sum_{k=0}^{300} \binom{1000}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{1000-k} \\ &= 2.201\,7 \times 10^{-14}. \end{aligned}$$

E' una probabilità assolutamente irrisoria! E' sostanzialmente impossibile che si presentino più di 300 correntisti.

Esempio 52 *Il risultato precedente pone le basi per una gestione assai economica delle risorse, difficilmente immaginabile senza la matematica. Risolviamo il seguente problema di "soglia di sicurezza". Decidiamo di accettare il rischio di non poter accontentare tutti i clienti una volta su 1000 (un giorno su tre anni, che è poco se si tiene anche conto che non si tratta di scontentare tutti i clienti di quel giorno sfortunato, ma solo i pochi ultimi in sovrappiù, e che forse in tale situazione eccezionale saremo in grado di porre rimedio con l'aiuto di un'altra filiale). Ci chiediamo: qual'è il numero intero k_0 tale che*

$$P\left(S > k_0\right) \leq \frac{1}{1000}?$$

Il numero k_0 è la soglia di sicurezza al 99,9%. O per tentativi o con l'uso del software `R`*, si può trovare*

$$k_0 = 248.$$

Si noti che è un numero straordinariamente vicino alla media, rispetto al migliaio di potenziali correntisti.

La deviazione standard della binomiale S dell'esempio vale

$$\sigma_S = \sqrt{1000 \cdot \frac{1}{5}\frac{4}{5}} = 12.649.$$

E' un numero molto piccolo rispetto al migliaio. Il numero 48, l'eccedenza di k_0 rispetto alla media 200, è circa 4 volte σ_S. Intuitivamente, questa è una conferma del fatto che il risultato sorprendente dell'esempio è giusto, non è un errore di calcolo.

1.3.2 Sul teorema degli eventi rari per v.a. di Poisson

Ricordiamo che le probabilità delle binomiali $B(n, p_n)$ tendono a quelle della Poisson $\mathcal{P}(\lambda)$ se $n \to \infty$ e $n \cdot p_n = \lambda$ (o anche solo se $n \cdot p_n \to \lambda$, come si dimostra con piccole complicazioni in più).

Questo teorema porta il nome di *teorema degli eventi rari*, per il motivo seguente. Si deve immaginare una sequenza molto lunga di esperimenti, ciascuno avente due esiti possibili che denominiamo "successo" o "insuccesso" e codifichiamo coi numeri 1 e 0 rispettivamente. Chiamiamo p la probabilità di successo. Il numero di successi è una v.a. binomiale. Se p è molto piccolo, i successi sono *rari*. Questo però è compensato dal fatto che il numero di prove n tende all'infinito. Il teorema dice che si può usare la distribuzione di Poisson al posto della binomiale.

Naturalmente nelle applicazioni pratiche non c'è nessun limite $n \to 0$, $p \to 0$. Ci chiediamo allora quando, per n grande ma fissato e p piccolo ma fissato, l'approssimazione di una binomiale con una Poisson fornisca risultati soddisfacenti. Il criterio, se pur vago, è che $\lambda = np$ sia un numero moderato e simultaneamente n sia grande e p piccolo. Nell'esempio della banca, $n = 1000$ è sicuramente grande, $p = \frac{1}{5}$ non è molto piccolo ma potrebbe sembrarlo abbastanza, ma $\lambda = np = 200$ è sicuramente troppo grande. Ad esempio,

$$1 - \sum_{k=0}^{248} \binom{1000}{k} \left(\frac{1}{5}\right)^k \left(\frac{4}{5}\right)^{1000-k} = 9.296\,5 \times 10^{-5}$$

$$1 - \sum_{k=0}^{248} e^{-200} \frac{200^k}{k!} = 4.588\,8 \times 10^{-4}.$$

Si osservi però che l'errore, per quanto grosso, è solo alla quarta cifra decimale, quindi è comunque contenuto. Se però effettuiamo un esperimento numerico con un λ più moderato, es.

$$n = 100, \quad p = \frac{1}{50}, \quad \lambda = 2$$

troviamo ad esempio

$$1 - \sum_{k=0}^{8} \binom{100}{k} \left(\frac{1}{50}\right)^k \left(\frac{49}{50}\right)^{100-k} = 1.893\,4 \times 10^{-4}$$

$$1 - \sum_{k=0}^{8} e^{-2} \frac{2^k}{k!} = 2.374\,5 \times 10^{-4}$$

cioè i due numeri coincidono quasi anche alla quarta cifra decimale.

1.3.3 Identificazione di un modello di Poisson piuttosto che di uno binomiale

Visto che grandezze aleatorie quali "il numero di persone che chiedono un certo servizio" possono essere descritte abbastanza realisticamente sia da v.a. binomiali sia di Poisson, quali conviene usare? Il modello di Poisson risulta vincente. Oltre ad essere più semplice

sia come formula analitica sia per il calcolo numerico, è più conveniente dal punto di vista dell'*identificazione* del modello, o più propriamente della *stima dei parametri* del modello. Vediamo il motivo.

Supponiamo di essere i gestori di un certo servizio. In alcuni casi particolari conosciamo il numero $n_{\max}$ di *potenziali* clienti, in altri casi no: si pensi al numero di correntisti di una banca (il numero complessivo è noto) ed al numero di coloro che potrebbero recarsi ad un distributore per un rifornimento (ignoto). Come gestori, vorremmo creare un modello matematico del numero aleatorio X di persone che effettivamente chiedono il nostro servizio, in un certo lasso di tempo (es. un giorno): da tale modello potremo poi calcolare grandezze medie e probabilità come quelle degli esempi del paragrafo precedente. Come identifichiamo un buon modello?

Chiediamoci quali dati reali, sperimentali, possiamo raccogliere, per decidere tra binomiale e Poisson e stimare i parametri. Il dato più semplice è il numero di clienti, in n casi simili a quello in questione, quindi un campione $x_1, ..., x_n$ estratto dalla v.a. X. Si tratta di registrare per n giorni il numero realmente accaduto di clienti che si sono presentati. Con esso possiamo calcolare la media aritmetica $\overline{x} = \frac{x_1+...+x_n}{n}$ e considerarla come approssimazione sperimentale della media vera $E[X]$. Ma allora ecco la risposta: dai dati sperimentali stimiamo direttamente il parametro $\lambda = E[X]$ se stiamo ipotizzando un modello di Poisson, mentre non stimiamo direttamente né $n_{\max}$ né p ma solo il prodotto $n_{\max}p$ se stiamo ipotizzando un modello binomiale. Ovviamente, se $n_{\max}$ ci è noto, usando $\overline{x}$ possiamo stimare p tramite il numero $\frac{\overline{x}}{n_{\max}}$. Ma se $n_{\max}$ non è noto, non possiamo risalire a p, per lo meno non in questo modo. In conclusione, ci sono varie ragioni per affermare che dai dati sperimentali è più naturale stimare il parametro λ di un modello di Poisson, che quindi risulta preferibile.

1.3.4 Processo di Bernoulli, ricorrenze, v.a. geometriche

Definizione 22 *Chiamiamo processo di Bernoulli di parametro p una successione (X_n) di v.a. indipendenti $B(1,p)$. Le v.a. X_n vengono pensate come "prove", esperimenti. Quando vale $X_n = 1$, si parla di successo al prova n-esima. Il numero di successi nelle prime n prove è la v.a. $S_n = X_1 + ... + X_n$. L'istante del primo successo è la v.a. $T = \min\{n : X_n = 1\}$.*

Osservazione 30 *Applichiamo alcuni teoremi noti: i) il numero di successi nelle prime n prove, S_n, è una binomiale $B(n,p)$; ii) per n grande e p piccolo, è approssimativamente una Poisson $\mathcal{P}(\lambda)$ con $\lambda = np$.*

Vediamo di capire queste definizioni in un esempio. Studiamo una zona costiera in cui il tempo cambia rapidamente ed esaminiamo i giorni di pioggia rispetto a quelli in cui non c'è alcuna precipitazione. Se supponiamo che i giorni siano indipendenti dal punto di vista della pioggia e che ci sia la stessa probabilità di pioggia in tutti i giorni, il nostro esame dei giorni di pioggia definisce un processo di Bernoulli, in cui la v.a. X_n vale 1 se il giorno n-esimo piove (è necessario fissare un giorno di inizio).

La v.a. $S_n = X_1 + ... + X_n$ rappresenta, nell'esempio, il numero di giorni di pioggia tra i primi n giorni. E' binomiale. Se la pioggia è relativamente rara, possiamo descrivere tale numero di giorni di pioggia, approssimativamente, con una Poisson.

Introduciamo poi alcune variabili che descrivono gli intertempi tra un giorno di pioggia e l'altro. Iniziamo le osservazioni un certo giorno, chiamato giorno 1. Indichiamo con T_1 (intero ≥ 1) il numero d'ordine del primo giorno di pioggia ($T_1 = 1$ significa che il giorno 1 c'è già pioggia, $T_1 = 2$ significa che il primo giorno non c'è pioggia mentre il secondo sì, ecc.). Poi indichiamo con T_2 (intero ≥ 1) il numero di giorni, dopo il primo giorno di pioggia, da attendere prima del secondo giorno di pioggia ($T_2 = 1$ significa che c'è pioggia già il giorno successivo a quello del primo giorno di pioggia, e così via). Proseguiamo così ad introdurre gli *intertempi* T_k. Se li sommiamo, $T_1 + ... + T_k$ è il k-esimo giorno di pioggia.

Esempio 53 *Supponiamo che il processo di Bernoulli abbia dato i seguenti valori:*

$$0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, ...$$

Allora $T_1 = 6$, $T_2 = 3$, $T_3 = 8$.

Definizione 23 *Ricordiamo che abbiamo chiamato v.a. geometrica di parametro p una v.a. discreta X, a valori interi non negativi, tale che*

$$P(X = n) = p(1-p)^n \qquad per\ n = 0, 1, ...$$

Chiamiamo poi v.a. geometrica modificata di parametro p una v.a. discreta X', a valori positivi, tale che

$$P\left(X' = n\right) = p(1-p)^{n-1} \qquad per\ n = 1, ...$$

Osservazione 31 *Per le geometriche avevamo dimostrato che $E[X] = \frac{1-p}{p}$. Per le geometriche modificate vale*

$$E\left[X'\right] = \frac{1}{p}.$$

Infatti, se X' è geometrica modificata, allora $X = X' - 1$ è geometrica, quindi

$$E\left[X'\right] = E[X + 1] = \frac{1-p}{p} + 1 = \frac{1}{p}.$$

Vale il seguente fatto:

Teorema 14 *Le v.a. $T_1, T_2, ..., T_i, ...$ sono indipendenti, geometriche modificate di parametro p.*

Proof. Cominciamo dimostrando che T_1 è geometrica. Vale $T_1 = 1$ se e solo se esce subito uno, cosa che avviene con probabilità p; vale poi, per $k \geq 2$, $T_1 = k$ se e solo se per $k-1$ volte esce zero, ed alla k-esima esce uno. Questa sequenza ha probabilità $q^{k-1}p$.

Mostriamo ora che T_2 è geometrica ed è indipendente da T_1 (l'indipendenza è intuitivamente ovvia, ma siccome le v.a. non vengono introdotte da noi nel definire il modello ma sono

derivate da altre, almeno per questa volta verifichiamo rigorosamente che sono indipendenti). Vale

$$\begin{aligned}&P(T_1 = k, T_2 = h)\\&= P(X_1 = 0, ..., X_{k-1} = 0, X_k = 1, X_{k+1} = 0, ..., X_{k+h-1} = 0, X_{k+h} = 1)\\&= q^{k-1}pq^{h-1}p = P(T_1 = k) \cdot q^{h-1}p.\end{aligned}$$

Quindi

$$\begin{aligned}P(T_2 = h) &= \sum_{k=1}^{\infty} P(T_1 = k, T_2 = h)\\&= q^{h-1}p \sum_{k=1}^{\infty} P(T_1 = k) = q^{h-1}p.\end{aligned}$$

Questo dimostra che T_2 è geometrica modificata di parametro p; inoltre, messa nell'uguaglianza precedente fornisce

$$P(T_1 = k, T_2 = h) = P(T_1 = k)\, P(T_2 = h)$$

per ogni k, h, quindi T_1 e T_2 sono indipendenti. La dimostrazione per T_3 ecc. è solo più lunga e la omettiamo. ■

Tra le conseguenze c'è il fatto (intuitivamente plausibile) che il tempo medio tra un giorno di pioggia ed un altro è

$$E[T] = \frac{1}{p}.$$

1.3.5 Tempo del k-esimo evento: binomiale negativa

Infine, consideriamo il tempo del k-esimo giorno di pioggia:

$$\tau_k = T_1 + ... + T_k.$$

Essa è una v.a. che assume i valori $k, k+1, ...$ Calcoliamone la massa di probabilità $P(\tau_k = k + h)$. L'evento $\tau_k = k + h$ accade quando $X_{k+h} = 1$, e tra le precedenti v.a. $X_1, ..., X_{k+h-1}$ ce ne sono esattamente $k-1$ pari ad uno. Ci sono $\binom{k+h-1}{k-1}$ modi di scegliere i tempi in cui questo accade; per ciascuna scelta, la probabilità di avere esattamente $k-1$ uni in quelle posizioni scelte più $X_{k+h} = 1$ è $p^k q^h$. Quindi

$$P(\tau_k = k + h) = \binom{k+h-1}{k-1} p^k q^h.$$

Questa è chiamata distribuzione *binomiale negativa* di parametri k e p. La binomiale negativa di parametri k e p è la distribuzione della somma di k v.a. geometriche di parametro q.

La formula precedente si può anche scrivere nella forma

$$P(\tau_k = j) = \binom{j-1}{k-1} p^{j-h} q^h$$

per $j = k, k+1, ...$

A dispetto della complicazione della formula, facilissimo è calcolare media e varianza di una binomiale negativa di parametri k e p:

$$\mu = \frac{k}{p}, \quad \sigma^2 = k\frac{q}{p^2}.$$

Basta infatti usare il fatto che la binomiale negativa di parametri k e p è somma di k v.a. geometriche modificate di parametro p, indipendenti (serve solo per la varianza).

Anticipiamo che, sviluppando analoghe idee a tempo continuo, si possono usare le v.a. esponenziali al posto delle geometriche, e v.a. di Erlang al posto delle binomiali negative.

1.3.6 Teoremi sulle v.a. esponenziali

In questa sezione sia T una v.a. esponenziale di parametro λ, $T \sim Exp(\lambda)$, cioè con densità

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{per} \quad t \geq 0 \\ 0 & \text{per} \quad t < 0 \end{cases}.$$

Abbiamo usato la lettera T (ma ogni altra è lecita) in quanto l'ambito tipico di applicazione delle v.a. esponenziali è ai *tempi* di vita, di funzionamento, di attesa ecc., per sistemi di vario tipo. La v.a. T rappresenta cioè, in molte applicazioni, l'istante in cui un certo sistema termina il suo lavoro, o si rompe, o una persona arriva in un sistema, e così via. Attraverso le proprietà delle v.a. esponenziali (in particolare la proprietà di assenza di memoria) capiremo quando il loro uso per descrivere tempi aleatori è giustificato con buona approssimazione oppure no.

La formula

$$P(T > t) = e^{-\lambda t}$$

è particolarmente elegante. La funzione $t \mapsto P(T > t)$ è detta funzione di *sopravvivenza* o di *affidabilità*. Se T è il tempo di vita o funzionamento di un sistema, $P(T > t)$ rappresenta la probabilità che il sistema funzioni ancora all'istante t. Se, a parità di t, questa funzione assume valori più grandi per un sistema piuttosto che un altro, il primo ha un miglior grado di sopravvivenza, una maggiore affidabilità.

Proprietà di assenza di memoria della legge esponenziale

Una proprietà importante della legge esponenziale è rappresentata dalla cosiddetta *assenza di memoria*. Per illustrarla intuitivamente, facciamo riferimento al tempo di vita di un sistema. La proprietà di assenza di memoria si manifesta quando qualunque sia il tempo trascorso, il tempo residuo di vita non è affetto dal passato e ha la stessa distribuzione del tempo di vita originario. In altre parole, l'oggetto non subisce logoramento, per cui la sua propensione statistica a rompersi resta invariata. Ovviemente da questo si vede che l'ipotesi di esponenzialità è piuttosto ideale nella pratica, ma la sua comodità matematica fa sì che la si supponga in molti contesti.

Esempio 54 *Attraverso Internet richiediamo un servizio, che può essere espletato solo quando il servente è libero. Supponiamo che la nostra richiesta non venga messa in una coda, ma che venga reiterata ogni secondo. Quando il servente si libera, prende la prima richiesta che gli arriva; se la nostra reiterazione gli arriva un istante dopo, viene scartata, e si continua ad aspettare. In questa situazione, anche se abbiamo aspettato inutilmente per 10 minuti, le nostre chances di essere accettati dall'operatore non sono aumentate: non c'è traccia nella memoria dell'operatore della nostra attesa più o meno lunga. In questo caso il tempo di attesa di connessione al servizio è molto plausibilmente esponenziale.*

Teorema 15 *Se T è esponenziale di parametro λ allora, per ogni $t, s \geq 0$, vale*

$$P\left(T > t+s | T > t\right) = P\left(T > s\right).$$

In altre parole, arrivati al tempo t ed osservato che siamo ancora in attesa ($T > t$), la probabilità che l'evento accada dopo un tempo s è uguale alla probabilità che inizialmente l'evento accadesse dopo un tempo s.

Proof. Vale

$$P\left(T > t+s | T > t\right) = \frac{P\left(T > t+s, T > t\right)}{P\left(T > t\right)}$$

ma il sistema

$$\begin{cases} T > t+s \\ T > t \end{cases}$$

equivale alla singola condizione $T > t+s$, quindi

$$= \frac{P\left(T > t+s\right)}{P\left(T > t\right)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P\left(T > s\right).$$

La dimostrazione è completa. ■

Sul minimo di v.a. esponenziali

Date due v.a. esponenziali T_1 e T_2 indipendenti, di parametri λ_1 e λ_2 rispettivamente, consideriamo la v.a. $T = \min\left(T_1, T_2\right)$. Ad esempio, se siamo i primi in coda ad una banca ed abbiamo davanti a noi due possibili sportelli, entrambi occupati, ciascuno che si libererà dopo un tempo esponenziale, T indica l'istante in cui si libererà il primo dei due, cioè l'istante in cui inizierà il nostro servizio.

La v.a. T ha densità esponenziale di parametro $\lambda_1 + \lambda_2$.

Per dimostrarlo calcoliamo il complementare della funzione di distribuzione di T:

$$\begin{aligned} P(T > t) &= P(\min\left(T_1, T_2\right) > t) = P(T_1 > t, T_2 > t) \\ &= P(T_1 > t) P(T_2 > t) = e^{\lambda_1 t} e^{\lambda_2 t} = e^{(\lambda_1 + \lambda_2) t}. \end{aligned}$$

Questo dimostra quanto volevamo. In generale vale:

Proposizione 6 *Se T_1,...,T_n sono v.a. esponenziali indipendenti, di parametri λ_1,...,λ_n, allora la v.a. $T = \min\left(T_1, ..., T_n\right)$ è esponenziale di parametro $\lambda_1 + ... + \lambda_n$.*

1.3.7 Proprietà delle gaussiane

Si dice che una classe di v.a. ha la proprietà di riproducibilità, o che le v.a. sono autoriproducenti, se prese due v.a. X ed Y di quella classe, indipendenti, allora $X+Y$ sta ancora nella stessa classe.

Le v.a. gaussiane godono di questa proprietà. Anche altre classi hanno questa proprietà (ad esempio le Poisson) ma le gaussiane la soddisfano in una forma ancora più forte, in cui oltre che la somma si possono considerare anche le combinazioni lineari, anzi affini.

Teorema 16 *Se X ed Y sono gaussiane indipendenti ed a, b, c sono numeri reali, allora $aX+bY+c$ è gaussiana. La sua media e la sua varianza sono date da*

$$\mu_{aX+bY+c} = a\mu_X + b\mu_Y + c$$
$$\sigma^2_{aX+bY+c} = a^2\sigma^2_X + b^2\sigma^2_Y.$$

Proof. Le funzioni generatrici di X ed Y sono

$$\varphi_X(t) = e^{\mu_X t + \frac{t^2\sigma_X^2}{2}}, \qquad \varphi_Y(t) = e^{\mu_Y t + \frac{t^2\sigma_Y^2}{2}}$$

e quindi, per l'Esercizio 2,

$$\begin{aligned}\varphi_{aX+bY+c}(t) &= \varphi_X(at)\,\varphi_Y(bt)\,e^{ct} = e^{\mu_X t + \frac{t^2\sigma_X^2}{2}} e^{\mu_Y t + \frac{t^2\sigma_Y^2}{2}} e^{ct} \\ &= e^{(\mu_X+\mu_Y+c)t + \frac{t^2\left(\sigma_X^2+\sigma_Y^2\right)}{2}}\end{aligned}$$

che è la generatrice di una gaussiana, quindi $aX+bY+c$ è gaussiana. Le formule per la sua media e varianza si leggono anche da qui, oppure si ottengono con facili calcoli sui valori medi. ■

Osservazione 32 *Le formule per media e varianza di $aX+bY+c$ valgono anche senza l'ipotesi di gaussinità e si dimostrano facilmente usando le proprietà dei valori medi. Quindi l'affermazione non ovvia del teorema è la gaussianità di $aX+bY+c$.*

Esercizio 4 *Dimostrare che le binomiali di parametro p fissato (mentre la numerosità n è libera) sono autoriproducenti. Si osservi che lo si può anche capire ad esempio facendo riferimento al teorema che le lega alle Bernoulli.*

Esercizio 5 *Dimostrare che le Poisson sono autoriproducenti, e precisamente la somma di una $\mathcal{P}(\lambda)$ ed una $\mathcal{P}(\lambda')$ indipendenti è una $\mathcal{P}(\lambda+\lambda')$.*

Tra le conseguenze semplici del teorema c'è che se X è gaussiana ed a, b sono numeri reali, allora $aX+b$ è gaussiana.

Definizione 24 *Data una v.a. X che ha media μ e varianza σ^2 finite, chiamiamo standardizzata di X la v.a.*

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

Essa ha media nulla e varianza unitaria.

Corollario 2 *Se X è gaussiana $N\left(\mu, \sigma^2\right)$, allora la sua standardizzata Z è una normale standard. Inoltre, vale la rappresentazione*

$$X = \mu + \sigma Z.$$

La dimostrazione è ovvia, ma il contenuto è della massima importanza. Si noti inoltre che l'espressione $\frac{x-\mu}{\sigma}$ (la standardizzazione) compare continuamente nei calcoli sulle gaussiane. Appare anche nel risultato che esporremo tra un momento, che si usa continuamente in statistica.

Definizione 25 *Indichiamo con $\Phi(x)$ la cdf della normale standard $N(0,1)$.*

Il suo grafico è già stato disegnato al paragrafo 1.2.6. Si può anche ottenere con R coi comandi

```
x<-(-500:500)/100
plot(x,pnorm(x))
```

Si vede dal grafico (e di verifica senza difficoltà) che vale

$$\Phi(-x) = 1 - \Phi(x)$$

una sorta di disparità rispetto al punto di coordiante $(0, 1/2)$. Questa regola è essenziale nell'uso delle tavole. La funzione $\Phi(x)$ è spesso tabulata al termine dei libri di testo, ma vengono dati i valori solo per $x > 0$. I valori per $x < 0$ si calcolano con la formula $\Phi(-x) = 1 - \Phi(x)$.

Proposizione 7 *Sia $F_{\mu,\sigma^2}(x)$ la cdf di una $X \sim N\left(\mu, \sigma^2\right)$. Allora*

$$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Proof.

$$F_{\mu,\sigma^2}(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

■

Definizione 26 *Se X è una v.a. con cdf $F(x)$ strettamente crescente e continua, dato $\alpha \in (0,1)$ esiste uno ed un solo numero $q_\alpha \in \mathbb{R}$ tale che*

$$F(q_\alpha) = \alpha.$$

La funzione $\alpha \mapsto q_\alpha$ è la funzione inversa di $x \mapsto F(x)$. Il numero q_α si dice quantile di ordine α.

La definizione di può estendere facilmente ad alcune situazioni in cui F non è strettamente crescente. Ad esempio, come accade per le v.a. esponenziali, $F(x)$ è nulla per $x < 0$ e poi è strettamente crescente. Allora, dato $\alpha \in (0,1)$, esiste uno ed un solo numero $q_\alpha > 0$ tale che $F(q_\alpha) = \alpha$, e quello viene preso come quantile. Invece che dare una complicata definizione generale, si ragioni caso per caso in questo modo, per definire i quantili, nelle situazioni in cui è chiaro cosa si deve fare.

I quantili gaussiani intervengono continuamente in statistica, o nel calcolo di soglie (anche quelli non gaussiani, solo che sono meno frequenti). Nel caso gaussiano vale anche per i quantili una formula di riduzione dal caso generale a quello standard, simile a quello delle cdf. La formula "ricopia" la struttura $X = \mu + \sigma Z$ vista sopra.

Proposizione 8 *Sia q_α^{μ,σ^2} il quantile di ordine α di una $X \sim N(\mu, \sigma^2)$ e sia q_α il quantile di ordine α della normale standard. Allora*

$$q_\alpha^{\mu,\sigma^2} = \mu + \sigma q_\alpha.$$

Proof. Il numero q_α^{μ,σ^2} è definito dall'equazione $F_{\mu,\sigma^2}\left(q_\alpha^{\mu,\sigma^2}\right) = \alpha$, che si può riscrivere

$$\Phi\left(\frac{q_\alpha^{\mu,\sigma^2} - \mu}{\sigma}\right) = \alpha.$$

Ma allora $\frac{q_\alpha^{\mu,\sigma^2} - \mu}{\sigma}$ è il quantile di ordine α della normale standard, cioè

$$\frac{q_\alpha^{\mu,\sigma^2} - \mu}{\sigma} = q_\alpha$$

da cui si ricava subito il risultato desiderato. ■

Per i quantili della normale standard vale la formula

$$q_{1-\alpha} = -q_\alpha$$

a volte utile, di nuovo legata alla disparità di Φ. Ricorrono spesso i seguenti quantili:

$$\begin{array}{lcllcl} q_{0.90} & = & 1.2815 & q_{0.975} & = & 1.9599 \\ q_{0.95} & = & 1.6448 & q_{0.99} & = & 2.3263 \end{array}$$

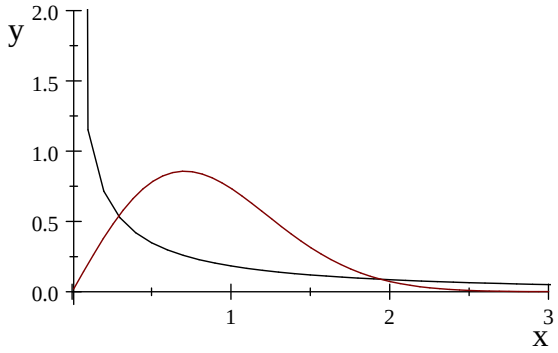
1.3.8 Variabili di Weibull

La densità Weibull di parametri $s > 0$ (detto *scala*) e $a > 0$ (detto *forma*) è data da

$$f(x) = \begin{cases} \frac{a}{s}\left(\frac{x}{s}\right)^{a-1} e^{-\left(\frac{x}{s}\right)^a} & \text{per} \quad x \geq 0 \\ 0 & \text{per} \quad x < 0 \end{cases}$$

Per $a = 1$ diventa $\frac{1}{s}e^{-\frac{x}{s}}$ ovvero è una esponenziale di parametro $\lambda = \frac{1}{s}$.

Ecco il grafico per $(s, a) = (1, 0.5)$ e $(s, a) = (1, 2)$:



Come nascono queste espressioni per la densità? Dalla funzione di ripartizione. Prendiamo (invece di $1 - e^{-\lambda x}$)

$$F(x) = 1 - e^{-\left(\frac{x}{s}\right)^a}, \quad x > 0$$

(è solo un altro modo di scrivere $F(x) = 1 - e^{-(\lambda x)^a}$). Vale

$$F'(x) = -e^{-\left(\frac{x}{s}\right)^a}\left(-a\left(\frac{x}{s}\right)^{a-1}\frac{1}{s}\right) = \frac{a}{s}\left(\frac{x}{s}\right)^{a-1} e^{-\left(\frac{x}{s}\right)^a}.$$

La media di una Weibull è

$$\mu = s \cdot \Gamma\left(1 + \frac{1}{a}\right).$$

da cui vediamo che la scala non è esattamente la media, ma è proporzionale. *Esempio.* La deviazione standard vale

$$\sigma = s \cdot \sqrt{\Gamma\left(1 + \frac{2}{a}\right) - \Gamma\left(1 + \frac{1}{a}\right)^2}.$$

Simile alla media $\mu = s \cdot \Gamma\left(1 + \frac{1}{a}\right)$, però con un legame meno facile da interpretare.

Le Weibull si incontrano ad esempio in ingegneria meccanica, nello studio dei fenomeni di fatica, dove descrivono il numero di cicli a cui si rompe una struttura; più in generale, vengono usate per descrivere tempi di vita, come generalizzazione delle esponenziali. Tra i tanti usi, le Weibull possonoanche servire per modellare una coda che descriva bene dei dati sperimentali. Le funzioni del tipo $1 - e^{-(\lambda x)^b}$ sono una delle classi più naturali e versatili.

Esempio 55 *Supponiamo di esaminare il tempo di vita di un componente meccanico o elettronico. L'insieme S degli esiti possibili è la semiretta $[0, \infty)$. Supponiamo ci serva che il componente duri per almeno 1 anno, ovvero 365 giorni. Usiamo i giorni come unità di misura. Indichiamo simbolicamente con T il tempi di vita e scriviamo*

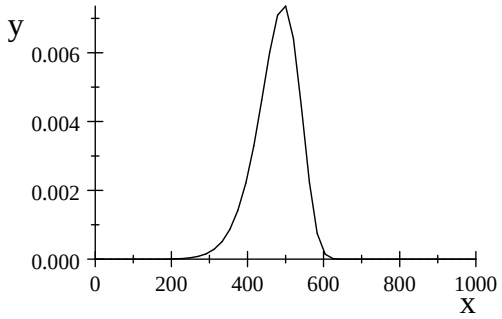
$$P(T > 365)$$

per indicare la probabilità che il componente duri più di 365 giorni. Nella pratica, il problema principale è conoscere la densità di probabilità giusta (o meglio, una ragionevolmente aderente

alla realtà). Ora, a titolo di esempio, supponiamo di conoscerla: una Weibull di parametri $(a,b) = (500, 10)$, $f(x) = \frac{10}{500}\left(\frac{x}{500}\right)^{10-1} e^{-\left(\frac{x}{500}\right)^{10}}$ *. Intuitivamente, significa che sappiamo che la vita media si aggira intorno a 500 giorni, con una certa aleatorietà. Vale* $F(x) = 1 - e^{-\left(\frac{x}{500}\right)^{10}}$*, quindi*

$$P(T > 365) = e^{-\left(\frac{365}{500}\right)^{10}} = 0.957\,93$$

in quanto $P(T > t) = 1 - F(t)$*. Questo è un esempio di calcolo della survival function (abbiamo calcolato la probabilità che il componente sopravviva almeno 365 giorni).*



1.3.9 Densità Gamma

La densità Gamma di parametri $s > 0$ (detto *scala*) e $a > 0$ (detto *forma*) è definita da

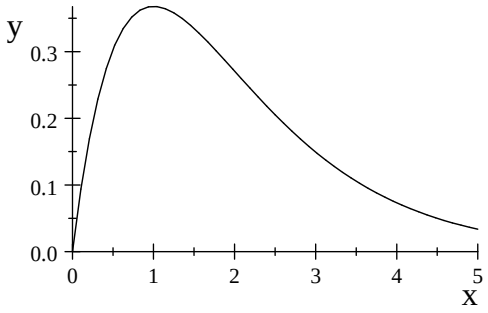
$$f(x) = \begin{cases} \frac{s}{\Gamma(a)}\left(\frac{x}{s}\right)^{a-1} e^{-\frac{x}{s}} & \text{per} \quad x \geq 0 \\ 0 & \text{per} \quad x < 0 \end{cases}$$

(dove $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$, la funzione gamma).

Si confronti con $\frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^b}$: molto simile ma il decadimento della Gamma è sempre esponenziale (potenza uno) e la potenza $a-1$ della x è sganciata dall'esponenziale. La sua provenienza non è da una F, ma da ragionamenti diretti: perturbare con un polinomio un esponenziale (oppure nasce sommando i quadrati di gaussiane indipendenti).

$a = 1$: è la densità esponenziale di parametro $\lambda = \frac{1}{s}$ (unica intersezione con la classe Weibull).

Per $a = 2$, $s = 1$:



Si suggerisce di esplorare l'help di R relativamente alle distribuzioni Weibull e Gamma, percependo il significato intuitivo dei due parametri (raffigurare alcune densità e magari le corrispondenti cumulative - con `pweibull`, `pgamma`).

Si dimostra che la media vale

$$\mu = a \cdot s.$$

Notare che il fattore di scala non è esattamente la media, come si potrebbe pensare; è però proporzionale. La deviazione standard vale

$$\sigma = \sqrt{a} \cdot s = \frac{\mu}{a}$$

(in genere quindi c'è una notevole variabilità, eliminabile solo con una forma speciale)

1.3.10 Densità Beta

Si chiama densità Beta di parametri $\alpha_1 \alpha_2 > 0$ la funzione

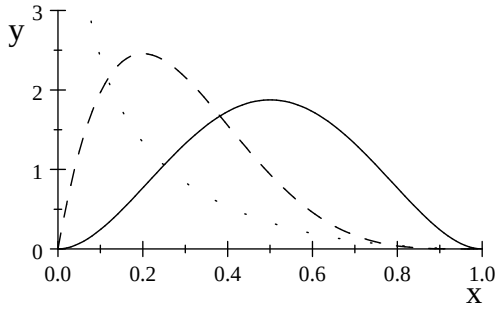
$$f(x) = \begin{cases} Cx^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1} & \text{per} \quad x \in (0,1) \\ 0 & \text{altrimenti} \end{cases}$$

dove C è la costate di normalizzazione, che si dimostra essere pari a

$$C = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}.$$

Ecco il grafico per:

- $\alpha_1 = \alpha_2 = 3$ (linea continua)
- $\alpha_1 = 2$, $\alpha_2 = 5$ (tratteggiata)
- $\alpha_1 = 0.5$, $\alpha_2 = 5$ (a punti).



Queste densità possono essere usate per quantificare la nostra fiducia nel valore di una grandezza aleatoria p che sia compresa tra 0 ed 1, ad esempio una frequenza relativa o una probabilità.

1.3.11 Code pesanti; distribuzione log-normale

Si dice che una v.a. X ha coda pesante (heavy tail) se la sua densità decade meno che esponenzialmente. Un caso limite è

$$f(x) = \frac{C}{1+x^{\alpha}}$$

con α positivo ma piccolo, $\alpha \in (1,2)$. Serve $\alpha > 1$ per avere una densità (altrimenti l'integrale diverge). Vale

$$\mu = \int_0^{\infty} x \frac{C}{1+x^{\alpha}} dx = +\infty \quad \text{se } \alpha \in (1,2)$$

(infatti $x\frac{C}{1+x^{\alpha}} \sim \frac{C}{x^{\alpha-1}}$ all'infinito, ed $\alpha - 1 \in (0,1)$ non è una potenza integrabile).

Quindi esistono v.a. a media infinita, pur assumendo valori finiti.

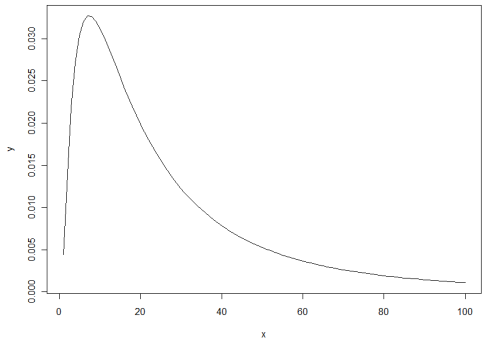
Tra gli esempi che si incontrano spesso nelle applicazioni ci sono le log-normali. Se X è una v.a. gaussiana o normale, la v.a.

$$Y = e^X$$

è detta *log-normale* (una log-normale è una variabile il cui logaritmo è normale). Essere ad esponente provoca l'occorrenza di valori enormi, ogni tanto. Nel senso: se tipicamente X vale circa 2-4, ma ogni tanto assume un valore dell'ordine di 5, i valori di Y saranno tipicamente del tipo 7-55, ma ogni tanto anche dell'ordine di 150.

I parametri di una log-normale sono media e deviazione della normale corrispondente. Per mimare i numeri appena dati, prendiamo una gaussiana di media 3 e deviazione 1. Ecco il grafico della relativa log-normale:

```
x<-1:100
y<- dlnorm(x,3,1)
plot(x,y)
```



Che queste densità abbiano coda pesante si intuisce dalla definizione, e dal grafico. Comunque, si dimostra che la densità è data da

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right)$$

con $x > 0$. Quindi esponenziale e logaritmo in qualche modo si compensano ed il decadimento diventa polinomiale.

1.3.12 Skewness e kurtosis

Esse sono i momenti standardizzati di ordine 3 e 4:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}, \quad \frac{\mu_4}{\sigma^4}$$

oppure, più spesso, per kurtosis, si intende la *kurtosi in eccesso*

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

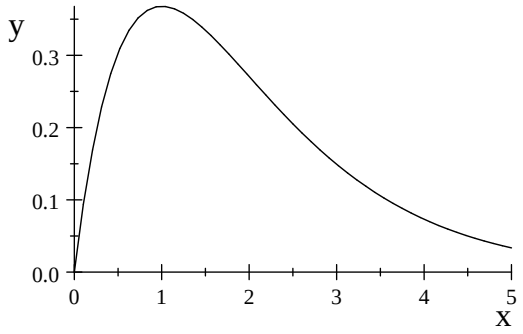
La skewness misura l'asimmetria. Infatti, se f è simmetrica, $\mu_3 = 0$.

Esempio 56 *X gaussiana: $\gamma_1 = \gamma_2 = 0$. La kurtosis (in eccesso) è una misura della deviazione dalla normalità.*

Esempio 57 *X gamma (a = forma):*

$$\gamma_1 = \frac{2}{\sqrt{a}}, \quad \gamma_2 = \frac{6}{a}$$

cioè dipendono entrambe solo dalla forma. Ecco ad es. $a = 2$ ($s = 1$):



1.4 Teoremi limite

1.4.1 Convergenze di variabili aleatorie

Convergenze in probabilità ed in media quadratica

Definizione 27 *Diciamo che una successione* $\{Y_n\}$ *di v.a. converge in media quadratica ad una v.a.* Y *se*

$$\lim_{n\to\infty} E\left[(Y_n - Y)^2\right] = 0.$$

Diciamo invece che converge in probabilità se

$$\lim_{n\to\infty} P\left(|Y_n - Y| > \varepsilon\right) = 0$$

per ogni $\varepsilon > 0$.

Lemma 1 *La convergenza in media quadratica implica la convergenza in probabiltà. Inoltre, quantitativamente, se per una certa successione numerica* $\{\alpha_n\}$ *vale*

$$E\left[(Y_n - Y)^2\right] \leq \alpha_n$$

allora vale

$$P\left(|Y_n - Y| > \varepsilon\right) \leq \frac{\alpha_n}{\varepsilon^2}.$$

Proof. Siccome

$$P\left(|Y_n - Y| > \varepsilon\right) = P\left(|Y_n - Y|^2 > \varepsilon^2\right),$$

per la disuguaglianza di Chebishev vale

$$P\left(|Y_n - Y| > \varepsilon\right) \leq \frac{E\left[(Y_n - Y)^2\right]}{\varepsilon^2}$$

da cui discendono tutte le affermazioni. ■

Convergenza quasi certa

Infine, esiste un altro concetto di convergenza, più delicato: la *convergenza quasi certa*.

Definizione 28 *Data una successione Y_n di v.a. ed una v.a. Y, tutte definite sullo stesso spazio probabilizzato $(\Omega, \mathcal{F}, P)$, diciamo che Y_n converge quasi certamente ad Y se*

$$P\left(\lim_{n\to\infty} Y_n = Y\right) = 1.$$

E' un concetto più delicato dei precedenti. Nella definizione si considera un evento che coinvolge simultaneamente infinite v.a.:

$$\left\{\lim_{n\to\infty} Y_n = Y\right\} := \left\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\}.$$

Grazie al fatto che $\mathcal{F}$ è chiusa per operazioni numerabili, si può mostrare che questo insieme è un evento, cioè appartiene ad $\mathcal{F}$, per cui se ne può calcolare la probabilità. La definizione richiede allora che tale probabilità sia pari ad uno.

Si dimostra che la convergenza quasi certa implica quella in probabilità, mentre il viceversa non è vero.

La convergenza quasi certa implica un salto concettuale e la necessità di strutture matematiche più complesse. Negli altri due tipi di convergenza, fissato n finito, si calcolano i numeri $E\left[|Y_n - Y|^2\right]$ e $P(|Y_n - Y| > \varepsilon)$. Solo di questi numeri si fa poi, eventualmente, il limite per $n \to \infty$. Per definire i numeri $E\left[|Y_n - Y|^2\right]$ e $P(|Y_n - Y| > \varepsilon)$ sono sufficienti spazi $(\Omega, \mathcal{F}, P)$ elementari.

Ben diversa è la convergenza quasi certa. Nella sua formulazione compaiono simultaneamente infinite v.a., dovendosi considerare l'evento $\{\lim_{n\to\infty} Y_n = Y\}$. Quindi lo spazio Ω deve essere più complesso. Esso deve contenere i possibili esiti che riguardano simultaneamente infinite variabili aleatorie. La trattazione rigorosa di questo argomento esula da questo corso.

Convergenza in legge

Introduciamo il concetto di convergenza in legge, detto anche convergenza debole o convergenza in distribuzione.

Definizione 29 *Una successione di v.a. (X_n) aventi funzione di distribuzione $(F_n(t))$ converge in legge ad una v.a. X con funzione di distribuzione $F(t)$ se*

$$F_n(t) \to F(t)$$

per ogni t in cui $F(t)$ è continua.

Si vede bene che, a differenza delle altre nozioni di convergenza stocastica viste fino ad ora (quasi certa, in probabilità, in media quadratica), la convergenza in legge dipende solo dalla legge delle v.a. e non dalle variabili in quanto tali. Si può quindi formulare una definizione

di convergenza in legge per una successione (μ_n) di misure di probabilità ed una misura di probabilità limite μ, richiedendo che

$$\mu_n\left((-\infty,t]\right)\to\mu\left((-\infty,t]\right)$$

per ogni t che sia punto di continuità di $\mu\left((-\infty,t]\right)$. Si dimostra che questa nozione equivale a richiedere

$$\int_X f\left(x\right)\mu_n\left(dx\right)\to\int_X f\left(x\right)\mu\left(dx\right)$$

per ogni funzione continua e limitata $f:X\to\mathbb{R}$, dove qui $X=\mathbb{R}$. Questo modo di vedere le cose è utile anche per le generalizzazioni a spazi metrici X diversi da $\mathbb{R}$.

Vale infine un teorema di convergenza legato alle funzioni generatrici. Se $\varphi_{X_n}\left(t\right)$ converge a $\varphi_X\left(t\right)$ in ogni punto t di un intervallo aperto non vuoto, allora X_n converge a X in legge.

1.4.2 Legge debole dei grandi numeri

Data una successione $X_1,X_2,...,X_n,...$ di v.a. (definite tutte sullo stesso spazio probabilizzato $(\Omega,\mathcal{F},P)$), scriveremo

$$\overline{X}_n=\frac{X_1+...+X_n}{n}.$$

Esercizio 6 *Se le v.a. $X_1,X_2,...,X_n,...$ hanno media μ (la stessa per tutte), allora*

$$E\left[\overline{X}_n\right]=\mu.$$

Esercizio 7 *Se inoltre $X_1,X_2,...,X_n,...$ sono indipendenti ed hanno varianze $\sigma_1^2,\sigma_2^2,...,\sigma_n^2,...$ finite, allora*

$$Var\left[\overline{X}_n\right]=\frac{\sigma_1^2+...+\sigma_n^2}{n^2}.$$

In particolare, se le varianze sono equilimitate da una costante $C>0$, ovvero

$$\sigma_n^2\leq C$$

per ogni n, allora

$$Var\left[\overline{X}_n\right]\leq\frac{C}{n}.$$

Soluzione. Le costanti escono al quadrato, per cui $Var\left[\overline{X}_n\right]=\frac{1}{n^2}Var\left[X_1+...+X_n\right]$. Essendo indipendenti, vale poi

$$Var\left[\overline{X}_n\right]=\frac{1}{n^2}\left(Var\left[X_1\right]+...Var\left[X_n\right]\right)=\frac{\sigma_1^2+...+\sigma_n^2}{n^2}.$$

L'ultima affermazione è di conseguenza ovvia.

Da tutti questi fatti è immediato dedurre la seguente versione della *Legge Debole dei Grandi Numeri* (LGN debole).

Teorema 17 *Sia $X_1, X_2, ..., X_n, ...$ una successione di v.a. indipendenti, con media μ e varianze $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2, ...$ equilimitate da una costante $C > 0$. Allora vale la convergenza in media quadratica*

$$\lim_{n\to\infty} E\left[\left(\overline{X}_n - \mu\right)^2\right] = 0 \tag{1.2}$$

e la convergenza in probabilità

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0 \tag{1.3}$$

per ogni $\varepsilon > 0$. Più precisamente, per ogni n vale

$$E\left[\left(\overline{X}_n - \mu\right)^2\right] \leq \frac{C}{n} \tag{1.4}$$

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \leq \frac{C}{\varepsilon^2 n}. \tag{1.5}$$

Proof. Vale

$$Var\left[\overline{X}_n\right] = E\left[\left(\overline{X}_n - \mu_{\overline{X}_n}\right)^2\right]$$

ma $\mu_{\overline{X}_n} = \mu$ (per il primo esercizio) quindi

$$E\left[\left(\overline{X}_n - \mu\right)^2\right] = Var\left[\overline{X}_n\right] \leq \frac{C}{n}$$

(per il secondo esercizio), da cui segue la convergenza in media quadratica. Quella in probabilità è poi conseguenza del fatto generale riportato nel paragrafo precedente. ■

Questo argomento verrà ripreso nel capitolo sui processi stocastici, a proposito dei processi stazionari ed ergodici.

Corollario 3 *Sia $X_1, X_2, ..., X_n, ...$ una successione di v.a. indipendenti ed identicamente distribuite (i.i.id.), con media μ e varianza σ^2. Allora valgono le affermazioni del teorema.*

Si può anche dimostrare che, senza l'ipotesi $\sigma^2 < \infty$, si ottiene ancora la convergenza in probabilità.

La LGN si applica in mille ambiti a problemi particolari. A livello più generale, in statistica, si può osservare che essa è alla base del legame tra molti stimatori ed i corrispondenti parametri. Oltre al caso ovvio, enunciato direttamente dalla LGN stessa, del legame tra lo stimatore $\overline{X}_n$ e la media μ, citiamo il legame tra

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

e la varianza σ^2, che si riconduce alla LGN introducendo le v.a.

$$Y_i = (X_i - \mu)^2$$

ed applicando ad esse la LGN (per cui abbiamo che $\frac{1}{n}\sum_{i=1}^{n} Y_i$ converge a $E[Y_1] = \sigma^2$). Con manipolazioni algebriche e la convergenza di $\overline{X}_n$ a μ, si vede poi che anche lo stimatore S^2, più usato, converge a σ^2. Infine, introducendo le v.a. $Z_i = (X_i - \mu_X)(Y_i - \mu_Y)$, si vede che lo stimatore $\widehat{Cov}_{XY}$ introdotto nella lezione 3 converge a $Cov(X,Y)$. E così per tanti altri esempi.

Vedremo tra poco la cosidetta legge forte dei grandi numeri. Non si dovrà però pensare necessariamente che essa rimpiazzi la legge debole. Infatti, nella legge debole è contenuta anche una *stima quantitativa dell'errore* che si commette approssimando μ con $\overline{X}_n$, cosa che si perderà nella legge forte. Il seguente esempio mostra un uso di tale stima quantitativa.

Esempio 58 *Sia T la v.a. "durata della batteria del PC". Supponiamo di non conoscere la legge di T e di voler stimare la media $E[T]$ tramite esperiementi. In 20 sessioni di lavoro al PC misuriamo la durata, ottenendo il campione sperimentale $t_1, ..., t_{20}$. Supponiamo che la media e deviazione empiriche di tale campione siano risp. $\bar{t} = 3$ h e $S = 1$ h. In prima approssimazione riteniamo quindi che $\bar{t} = 3$ h sia una discreta approssimazione di $E[T]$, per la LGN. In più di questo, però, possiamo dire che*

$$P\left(\left|\overline{T}_{20} - E[T]\right| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2 20}.$$

Se approssimiamo σ^2 con S^2, troviamo

$$P\left(\left|\overline{T}_{20} - E[T]\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2 20}.$$

Ad esempio, per $\varepsilon = 30$ min, risulta $\frac{1}{\varepsilon^2 20} = \frac{1}{5} = 0.2$. Quindi possiamo affermare che con probabilità 0.8 gli esperimenti dovevano fornire un valore $\bar{t}$ tale che

$$E[T] = \bar{t} \pm 30 \ min.$$

A causa di questa affermazione e dei nostri risultati sperimentali, confidiamo all'80% che valga

$$E[T] = 180 \pm 30 \ min.$$

Questo è un esempio di intervallo di confidenza.

L'esempio ora descritto mette in luce il fatto che, quantitativamente, la stima con $\frac{1}{n}$ è piuttosto povera. Essa però si può migliorare, a patto di conoscere altre grandezze medie legate alle variabili in gioco.

Esercizio 8 *Date X_n i.i.d., supponiamo che siano simmetriche rispetto alla media μ, e che sia*

$$\theta^4 := E\left[(X - \mu)^4\right] < \infty.$$

Allora

$$E\left[\left(\overline{X}_n - \mu\right)^4\right] = \frac{n\theta^4 + \binom{4}{2} n(n-1)\sigma^4}{n^4}$$

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \leq \frac{n\theta^4 + \binom{4}{2} n(n-1)\sigma^4}{\varepsilon^4 n^4}.$$

Quindi queste grandezze tendono a zero come $\frac{1}{n^2}$.

Esempio 59 *Riprendendo l'esempio precedente, supponiamo per semplicità la simmetria, e supponiamo che dai dati si possa stimare* $\widehat{\theta}^4 \sim 5$. *Allora*

$$P\left(\left|\overline{X}_{20} - \mu\right| > 30min\right) \leq \frac{20 \cdot 5 + \binom{4}{2} 20 \cdot 19}{5^4} = 3.808.$$

Questa stima non serve a nulla. Abbiamo mostrato questo risultato negativo per chiarire che le costanti davanti agli infinitesimi possono vanificarne l'uso pratico.

Esempio 60 *Valutiamo però l'intervallo di confidenza con* $\varepsilon = 1\ h$. *Col primo metodo avremmo scoperto*

$$E[T] = 180 \pm 60\ min$$

con confidenza $1 - \frac{1}{20} = 0.95$. *Ora invece vale*

$$P\left(\left|\overline{X}_{20} - \mu\right| > 1h\right) \leq \frac{20 \cdot 5 + \binom{4}{2} 20 \cdot 19}{20^4} = 0.015.$$

Quindi l'affermazione $E[T] = 180 \pm 60$ *min vale in realtà con confidenza* $1 - 0.015 = .985$.

Questo esercizio fa capire che le stime (1.4) e (1.5) non sono ottimali, in generale: sotto opportune ipotesi di maggior integrabilità di X il decadimento è più rapido. Nel seguito della lezione si dimostrerà un teorema di decadimento esponenziale.

1.4.3 Legge forte dei grandi numeri

Una LGN relativamente alla converganza quasi certa viene detta legge *forte* dei grandi numeri (LGN forte).

Teorema 18 *Sia* $X_1, X_2, ..., X_n, ...$ *una successione di v.a. indipendenti ed identicamente distribuite, con media* μ *finita. Allora vale la LGN forte.*

Vale anche il seguente teorema (di Ratchmann):

Teorema 19 *Sia* $X_1, X_2, ..., X_n, ...$ *una successione di v.a. scorrelate (*$Cov(X_i, X_j) = 0$ *per ogni* $i \neq j$*), con* $\lim_{n\to\infty} E[X_n] = \mu$ *e varianze equilimitate. Allora vale la LGN forte.*

Le dimostrazioni sono complesse e le omettiamo. Cerchiamo invece di apprezzare la differenza di informazione pratica che fornisce la LGN forte rispetto a quella debole. In genere tutti noi abbiamo la seguente convinzione: che se lanciamo una moneta per un gran numero di volte, per circa la metà di volte verrà testa; e che se continuassimo all'infinito i lanci, la frequenza relativa (numero di teste diviso numero di lanci) tenderebbe esattamente ad $\frac{1}{2}$. Il procedimento, pur ipotetico, di continuare i lanci all'infinito e studiare il limite delle frequenze relative corrisponde esattamente alla legge forte. Infatti, si sta considerando *una ben precisa storia* (sequenza) *infinita* ω, quella che accade continuando all'infinito i lanci, e relativamente a quella si stanno calcolando le medie parziali $\overline{X}_n(\omega)$ e se ne studia il limite per $n \to \infty$. Solo il concetto di convergenza quasi certa e la LGN forte esaminano questo tipo di procedimento.

Invece le leggi deboli ci dicono che se facciamo ad es. 1000 lanci, la probabilità che $\overline{X}_n$ disti da $\frac{1}{2}$ più di ε è minore di $\frac{pq}{\varepsilon^2 100} = \frac{1}{\varepsilon^2 400}$. Quindi abbiamo tale confidenza $(.00\,25 \cdot \varepsilon^{-2})$ che $\overline{X}_n(\omega)$, relativo alla storia ω che si sta avverando, disti da $\frac{1}{2}$ più di ε. Se aumentiamo n, aumenta la nostra confidenza, ma non possiamo dire che $\overline{X}_n(\omega)$, relativo alla nostra storia ω, si stia effettivamente avvicinando a $\frac{1}{2}$.

1.4.4 Stima di Chernoff (grandi deviazioni)

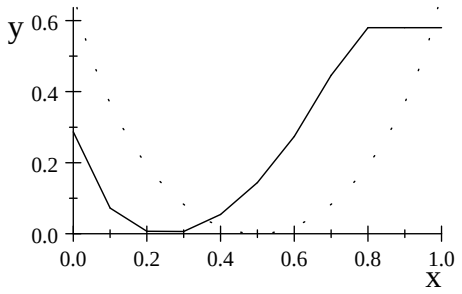
In questo paragrafo mostriamo stime esponenziali per le probabilità di errore tra media aritmetica e media teorica.

Per ogni coppia di numeri $\alpha, p \in (0,1)$, introduciamo l'*entropia relativa*

$$h(\alpha||p) = \alpha \log \frac{\alpha}{p} + (1-\alpha) \log \frac{(1-\alpha)}{(1-p)}$$

detta anche distanza (o divergenza) di Kullback-Leibler.

Per $\alpha > p$ vale $h(\alpha||p) > 0$, come si deduce ad esempio dalla dimostrazione del seguente teorema. Ecco il grafico di $\alpha \mapsto h(\alpha||\frac{1}{2})$ e di $\alpha \mapsto h(\alpha||\frac{1}{4})$, a titolo di esempio.



$\alpha \mapsto h(\alpha||\frac{1}{4})$ (linea intera) e $\alpha \mapsto h(\alpha||\frac{1}{2})$ (tratteggiata)

Altre proprietà generali, ben visibili negli esempi, sono che $h(p||p) = 0$ e che $h(\alpha||p)$ è convessa in α.

Data $S_n \sim B(n,p)$, ricordiamo che la sua media è np. Inoltre sappiamo che la sua deviazione standard σ è molto più piccola delle grandezze che crescono con n (come il range n e la media np): essa vale $\sqrt{n}\sqrt{pq}$. Quindi S_n si concentra, per così dire, attorno alla sua media. Preso allora un numero $\alpha > p$, la probabilità della coda $P(S_n \geq n\alpha)$ dovrebbe essere molto piccola. Dimostriamo che è esponenzialmente piccola.

Teorema 20 *Se $S_n \sim B(n,p)$, allora, per ogni $\alpha > p$, vale*

$$P(S_n \geq n\alpha) \leq e^{-nh(\alpha||p)}.$$

Inoltre, per ogni $\beta < p$, vale

$$P(S_n \leq n\beta) \leq e^{-nh(\beta||p)}.$$

Quindi vale anche

$$P\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leq e^{-nh(p+\varepsilon||p)} + e^{-nh(p-\varepsilon||p)}.$$

Proof. Dimostriamo solo la prima disuguaglianza; la seconda è analoga, considerando le v.a. $Y_i = -X_i$.

Per ogni $\lambda > 0$ vale

$$P\left(S_n \geq n\alpha\right) = P\left(\exp \lambda S_n \geq \exp \lambda n\alpha\right).$$

Per la disuguaglianza di Chebishev, allora,

$$P\left(S_n \geq n\alpha\right) \leq \exp\left(-\lambda n\alpha\right) E\left[\exp \lambda S_n\right].$$

Per l'indipendenza delle v.a. vale

$$\begin{aligned} E\left[\exp \lambda S_n\right] &= E\left[\exp \lambda X_1 \cdots \exp \lambda X_n\right] = E\left[\exp \lambda X_1\right] \cdots E\left[\exp \lambda X_n\right] \\ &= E\left[\exp \lambda X_1\right]^n = \left(pe^{\lambda} + q\right)^n. \end{aligned}$$

Quindi

$$\begin{aligned} P\left(S_n \geq n\alpha\right) &\leq \exp\left(-\lambda n\alpha\right) \cdot \left(pe^{\lambda} + q\right)^n \\ &= \exp\left[(-n)\left(\lambda\alpha - \log\left(pe^{\lambda} + q\right)\right)\right]. \end{aligned}$$

Questa disuguaglianza vale per ogni $\lambda > 0$. Quindi vale anche

$$\begin{aligned} P\left(S_n \geq n\alpha\right) &\leq \inf_{\lambda > 0} \exp\left[(-n)\left(\lambda\alpha - \log\left(pe^{\lambda} + q\right)\right)\right] \\ &= \exp\left[(-n) \sup_{\lambda > 0}\left(\lambda\alpha - \log\left(pe^{\lambda} + q\right)\right)\right]. \end{aligned}$$

Calcoliamo questo estremo superiore.

Consideriamo la funzione

$$f\left(\lambda\right) = \lambda\alpha - \log\left(pe^{\lambda} + q\right)$$

definita per $\lambda \geq 0$. Vale $f\left(0\right) = 0$,

$$f'\left(\lambda\right) = \alpha - \frac{pe^{\lambda}}{pe^{\lambda} + q} = \frac{\alpha q - \left(1 - \alpha\right) pe^{\lambda}}{pe^{\lambda} + q}$$

quindi $f'\left(0\right) = \alpha q - \left(1 - \alpha\right) p$. Avendo supposto $\alpha > p$, $q > \left(1 - \alpha\right)$, quindi $f'\left(0\right) > 0$. Vale inoltre $\lim_{\lambda \to \infty} f\left(\lambda\right) = -\infty$, quindi ci aspettiamo un massimo assoluto per $\lambda > 0$. Vale $f'\left(\lambda\right) = 0$ se

$$e^{\lambda} = \frac{\alpha q}{\left(1 - \alpha\right) p}$$

quindi per il solo valore

$$\lambda = \log \frac{\alpha q}{\left(1 - \alpha\right) p}$$

(che è positivo in quanto $\frac{\alpha q}{(1-\alpha)p} > 1$, come già osservato sopra). Quindi

$$\sup_{\lambda>0}\left(\lambda\alpha - \log\left(pe^{\lambda}+q\right)\right) = \alpha\log\frac{\alpha q}{(1-\alpha)\,p} - \log\left(\frac{\alpha q}{(1-\alpha)}+q\right)$$
$$= h(\alpha||p).$$

La dimostrazione è completa. ■

Nel teorema precedente, l'ipotesi $S_n \sim B(n,p)$ gioca solo un ruolo marginale. Presa una successione $X_1, X_2, ..., X_n, ...$ di v.a. indipendenti ed identicamente distribuite come X, supponiamo che per ogni $\lambda > 0$ sia $E\left[e^{\lambda X}\right] < \infty$. Osserviamo che, posto

$$S_n = X_1 + ... + X_n$$

vale

$$E\left[e^{\lambda S_n}\right] = E\left[e^{\lambda X}\right]^n.$$

Ripetendo i passaggi della dimostrazione si trova

$$P(S_n \geq n\alpha) \leq \exp(-\lambda n\alpha)\cdot E\left[e^{\lambda X}\right]^n$$
$$= \exp\left[(-n)\left(\lambda\alpha - \log E\left[e^{\lambda X}\right]\right)\right].$$

Introduciamo la funzione

$$\Lambda(\lambda) = \log E\left[e^{\lambda X}\right]$$

e la funzione

$$\Lambda^*(a) = \sup_{\lambda>0}(\lambda\alpha - \Lambda(\lambda))$$

detta *trasformata di Legendre-Fenchel* di $\Lambda(\lambda)$. In realtà, nella definizione tradizionale si deve prendere l'estremo superiore su tutti i λ, ma qui si può dimostrare che è inessenziale. Si scopre allora:

Teorema 21 *Presa una successione $X_1, X_2, ..., X_n, ...$ di v.a. indipendenti ed identicamente distribuite come X, con $E\left[e^{\lambda X}\right] < \infty$ per ogni $\lambda > 0$, definita $\Lambda^*(a)$ come sopra, per ogni $\alpha > \mu = E[X]$, vale*

$$P(S_n \geq n\alpha) \leq e^{-n\Lambda^*(a)}.$$

Inoltre, per ogni $\beta < E[X]$, vale

$$P(S_n \leq n\beta) \leq e^{-n\Lambda^*(\beta)}.$$

Quindi vale anche

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \leq e^{-n\Lambda^*(\mu+\varepsilon)} + e^{-n\Lambda^*(\mu-\varepsilon)}.$$

Infine, con un argomento più complicato si può dimostrare una stima dal basso dello stesso tipo, di tipo però asintotico:

$$P(S_n \geq n\alpha) \geq c_n e^{-n\Lambda^*(a)}$$

con $\lim_{n\to\infty}\frac{1}{n}\log c_n = 0$. In verità questa stima dal basso non vale proprio in tutti i punti $\alpha > E[X]$. Per tutte queste piccole difficoltà tecniche, non approfondiamo ulteriormente la stima dal basso e rimandiamo per una trattazione più esauriente ai testi della *teoria delle grandi deviazioni*.

1.4.5 Teorema limite centrale

Esercizio 9 *Siano $X_1, ..., X_n$ v.a. indipendenti, identicamente distribuite, con varianza finita σ^2 e media μ. Allora*

$$Z_n := \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

ha media zero e varianza uno.

Il teorema limite centrale vale sotto diverse ipotesi. La seguente versione porta il nome di Teorema di P. Lévy (o Lindeberg-Lévy). Nel caso particolare in cui le v.a. X_n siano delle Bernoulli, esso porta il nome di Teorema di De Moivre-Laplace. e si può dimostrare per via combinatorica.

Teorema 22 *Sia (X_n) una successione di v.a. indipendenti, identicamente distribuite, con varianza finita σ^2 e media μ. Allora la v.a.*

$$Z_n := \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

converge in legge ad una gaussiana canonica $N(0,1)$. In altre parole, per ogni $a < b$ vale

$$\lim_{n\to\infty} P\left(a \leq \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a)$$

dove Φ indica la cdf normale standard.

Prima di procedere alla dimostrazione, osserviamo che, in base all'esercizio preposto al teorema, la v.a. Z_n ha media zero e varianza uno. Però non è in generale gaussiana e non è ovvio che lo diventi al limite per $n \to \infty$. Questa è la parte difficile del teorema.

Proof. Calcoliamo la funzione generatrice $\varphi_n(t)$ di Z_n e mostriamo che, per ogni t, essa converge a $e^{-t^2/2}$. Questo implica la convergenza in legge di Z_n alla $N(0,1)$.

Osserviamo che

$$Z_n = \frac{\frac{X_1-\mu}{\sigma} + ... + \frac{X_n-\mu}{\sigma}}{\sqrt{n}}$$

dove le v.a. $Y_n = \frac{X_n-\mu}{\sigma}$ sono indipendenti ed hanno media zero e varianza uno. Quindi basta dimostrare il teorema in questo caso.

Supponiamo allora $\mu = 0$, $\sigma = 1$. Abbiamo

$$\varphi_n(t) = \varphi_{X_1+...+X_n}\left(\frac{t}{\sqrt{n}}\right) = \varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right)^n.$$

Allora, usando lo sviluppo di Taylor di $\varphi_{X_1}(t)$ ed il fatto che $E[X_1] = 0$ e $E\left[X_1^2\right] = 1$, vale

$$\varphi_{X_1}(t) = 1 + \frac{t^2}{2} + o\left(t^2\right).$$

Quindi

$$\varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Pertanto vale

$$\varphi_n(t) = \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Passando ai logaritmi abbiamo

$$\log \varphi_n(t) = n \log\left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)$$

ed usando il limite notevole $\lim_{x\to 0} \frac{\log(1+x)}{x} = 1$ si ottiene (t è fissato)

$$\begin{aligned}\lim_{n\to\infty} \log \varphi_n(t) &= \lim_{n\to\infty} n\left(\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) \frac{\log\left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)}{\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)} \\ &= \lim_{n\to\infty} n\left(\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) = \frac{t^2}{2}.\end{aligned}$$

Quindi

$$\lim_{n\to\infty} \varphi_n(t) = e^{\frac{t^2}{2}}.$$

La dimostrazione è completa. ■

Il caso particolare in cui le v.a. X_n sono Bernoulli $B(1,p)$ è particolarmente rilevante. In questo caso $S_n := X_1 + ... + X_n$ è una binomiale $B(n,p)$ ed il teorema dice che

$$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converge in legge ad una gaussiana canonica $N(0,1)$. E' un teorema di *convergenza della binomiale alla gaussiana*, che si affianca al teorema degli eventi rari. Qui, per vedere la convergenza, bisogna standardizzare la binomiale S_n (Z_n è la sua standardizzata).

Può sembrare assurdo che le binomiali approssimino contemporaneamente sia le Poisson sia le gaussiane. In effetti il regime limite è molto diverso: nel teorema degli eventi rari p non è fissato, tende a zero come $\frac{\lambda}{n}$, mentre nel teorema limite centrale è fissato. Se però non si considera il limite vero e proprio ma solo l'approssimazione per valori grandi o piccoli dei parametri in gioco, ci sono effettivamente delle situazioni in cui le due approssimazioni sono abbastanza buone entrambe e si sovrappongono un po'. A parte questo, il consiglio è di usare il teorema degli eventi rari quando il prodotto np è un numero dell'ordine dell'unità (es. 5), ed n è ovviamente non troppo piccolo (es. 20, 30). Se ad esempio $np = 3$ e $n = 30$, allora $p = \frac{3}{30} = 0.1$, è piuttosto piccolo. In queste situazioni l'uso del teorema limite centrale non prodice risultati molto precisi. Meglio che p sia più "interno" all'intervallo $(0,1)$, non così estremo, per una buona applicazione del TLC (ma se n è più grande, allora si possono accettare p più piccoli).

Infine, sempre nell'ambito dell'approssimazione gaussiana della binomiale, se si vuole un risultato più preciso conviene usare la *correzione di continuità*. Supponiamo di dover calcolare $P(S_n \leq 25)$. Siccome S_n assume valori solo negli interi, questo è uguale a $P(S_n < 26)$. Le

due approssimazioni darebbero

$$P(S_n \leq 25) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{25 - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{25 - n\mu}{\sigma\sqrt{n}}\right)$$
$$P(S_n < 26) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{26 - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{26 - n\mu}{\sigma\sqrt{n}}\right)$$

per cui in genere si ottiene un risultato più preciso prendendo

$$P(S_n \leq 25) \approx \Phi\left(\frac{25.5 - n\mu}{\sigma\sqrt{n}}\right).$$

1.4.6 Distribuzione del limite di massimi

Cominciamo da un caso particolare. Siano $X_1, ..., X_n, ...$ v.a. $Exp(\lambda)$ indipendenti, per cui

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0.$$

Indichiamo con M_n la v.a.

$$M_n = \max\{X_1, ..., X_n\}.$$

Che distribuzione ha M_n? Indichiamo con $F_n(x)$ la funzione di distribuzione di M_n. Vale (e questo è vero indipendentemente dalla legge di X)

$$F_n(x) = F(x)^n.$$

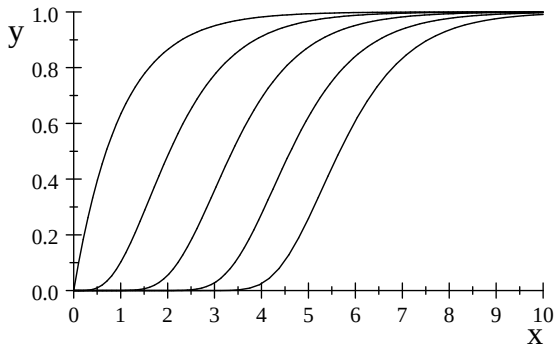
Infatti

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, ..., X_n \leq x) \\ &= P(X_1 \leq x) \cdots P(X_n \leq x) = P(X \leq x)^n. \end{aligned}$$

Usando poi il fatto che X è esponenziale, troviamo, per $x \geq 0$,

$$F_n(x) = \left(1 - e^{-\lambda x}\right)^n.$$

Nella figura si vedono i grafici, per $\lambda = 1$, per diversi valori di n.



$F_n(x)$ per n =1, 5, 20, 70, 200

Essendo $F(x) < 1$ per ogni x, $F(x)^n \to 0$ per $n \to \infty$, per cui il grafico di $F_n(x)$ si sposta verso destra, per così dire, al crescere di n. Questo è coerente con l'intuizione che i valori tipici di M_n sono maggiori di quelli di X, e diventano sempre più grandi al crescere di n. Si noti che lo spostamento (drift) non si arresta mai: diventa sempre più lento, ma non può arrestarsi, essendo $F(x) < 1$ per ogni $x > 0$.

Un fatto che si può intuire dai grafici è che la forma di $F_n(x)$ tende ad assestarsi, per quanto continui a slittare verso destra. Matematicamente, sembra che ci sia una successione $b_n \to \infty$ di traslazioni ed una funzione "limite" $G(x)$ tali che

$$F_n(x) \sim G(x - b_n)$$

ovvero rigorosamente

$$\lim_{n\to\infty} F_n(x + b_n) = G(x).$$

Dimostriamolo. Basta prendere

$$b_n = \frac{1}{\lambda} \log n.$$

Infatti

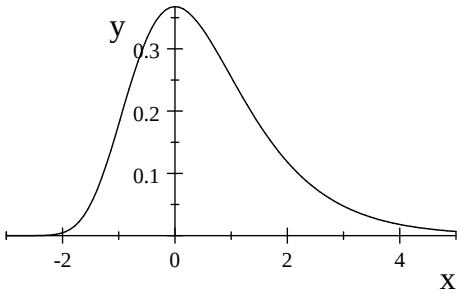
$$\begin{aligned} F_n(x + b_n) &= \left(1 - e^{-\lambda\left(x + \frac{1}{\lambda}\log n\right)}\right)^n \\ &= \left(1 - \frac{e^{-\lambda x}}{n}\right)^n \to G(x) \end{aligned}$$

con

$$G(x) = e^{-e^{-\lambda x}}.$$

Questa è detta *distribuzione di Gumbel*. E' una funzione di distribuzione, corrispondente alla densità

$$g(x) = \lambda e^{-e^{-\lambda x}} e^{-\lambda x} = \lambda e^{-\lambda x - e^{-\lambda x}}.$$



Densità di Gumbel per $\lambda = 1$

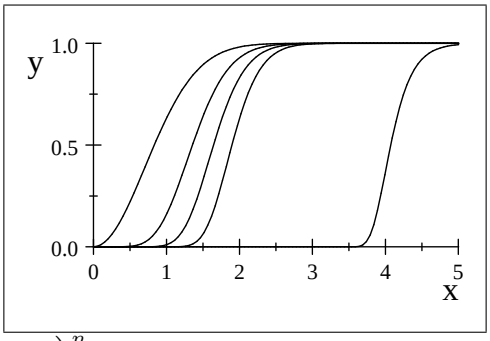
Si trova la distribuzione di Gumbel a partire da varie distribuzioni per X, non solo per l'esponenziale. A livello grafico, osserviamo ad esempio quanto accade partendo da una $F(x)$ che si avvicina ad 1 in modo esponenziale quadratico, come accade per le gaussiane. Per semplicità prendiamo

$$F(x) \sim 1 - e^{-x^2}.$$

Raffiguriamo $F(x)^n$ per n crescente e, traslata con $b = 4$ per esigenze visive, la Gumbel con con $\lambda = 5$:

$$G(x) = e^{-e^{-5(x-4)}}.$$

E' visivamente chiaro che $F(x)^n$ tende a $G(x)$.



$\left(1 - e^{-x^2}\right)^n$ per n =1, 4, 10, 25 e Gumbel (traslata)

Bisogna invece tener conto che, se si parte da distribuzioni $F(x)$ radicalmente diverse, si possono trovare al limite, per $F(x)^n$, due altri tipi di forme. Vediamolo attraverso due esempi. Se

$$F(x) = \begin{cases} 1 - x^{-\alpha} & \text{per} \quad x \geq 1 \\ 0 & \text{per} \quad x < 1 \end{cases}$$

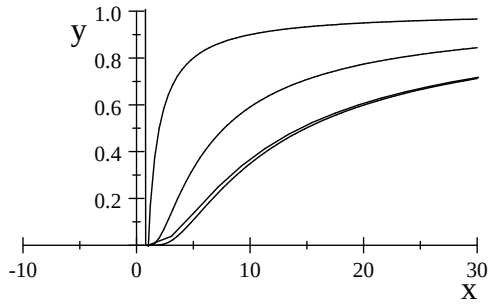
con $\alpha > 0$, si trova, per $n \to \infty$,

$$F(x)^n \sim G\left(\frac{x}{n^{1/\alpha}}\right)$$

dove $G(x)$ è la distribuzione di Frechet

$$G(x) = \begin{cases} e^{-x^{-\alpha}} & \text{per} \quad x \geq 1 \\ 0 & \text{per} \quad x < 0 \end{cases}.$$

A titolo di esempio, per $\alpha = 1$, tracciamo i grafici di $F(x)$, $F(x)^5$, $F(x)^{10}$, e della Frechet $e^{-\left(\frac{x}{10}\right)^{-1}}$, che praticamente coincide con $F(x)^{10}$.



Convergenza alla distribuzione di Frechet

L'ultimo tipo di distribuzione si trova prendendo ad esempio

$$F(x) = \begin{cases} 0 & \text{per} \quad x \leq 0 \\ 1-(1-x)^{\alpha} & \text{per} \quad 0 < x < 1 \\ 1 & \text{per} \quad x \geq 1 \end{cases}$$

con $\alpha > 0$. Si trova, per $n \to \infty$,

$$F(x)^n \sim G\left(n^{1/\alpha}(x-1)\right)$$

dove $G(x)$ è la distribuzione del massimo di terzo tipo

$$G(x) = \begin{cases} e^{-(-x)^{\alpha}} & \text{per} \quad x < 0 \\ 1 & \text{per} \quad x \geq 0 \end{cases}.$$

1.5 Approfondimenti sui vettori aleatori

1.5.1 Trasformazione di densità

Esercizio 10 *Se X ha cdf $F_X(x)$ e g è strettamente crescente e continua, allora $Y = g(X)$ ha cdf*

$$F_Y(y) = F_X\left(g^{-1}(y)\right)$$

per tutte le y nell'immagine di g. Se g è strettamente decrescente e continua, la formula è

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$$

Soluzione:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right).$$

La seconda è identica.

Esercizio 11 *Se X ha una pdf continua $f_X(x)$ e g è strettamente crescente e differenziabile, allora $Y = g(X)$ ha pdf*

$$f_Y(y) = \frac{f_X\left(g^{-1}(y)\right)}{g'\left(g^{-1}(y)\right)} = \left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}$$

per tutte le y nell'immagine di g. Se g è decrescente e differenziabile, la formula è

$$f_Y(y) = -\left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}.$$

Soluzione: Vale

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(g^{-1}(y)\right) = F_X'\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) \\ &= f_X\left(g^{-1}(y)\right)\frac{1}{g'\left(g^{-1}(y)\right)} = \left.\frac{f_X(x)}{g'(x)}\right|_{y=g(x)}. \end{aligned}$$

La seconda è identica.

Quindi, in generale, abbiamo:

Proposizione 9 *Se g è monotona e differenziabile, la trasformazione di densità è data da*

$$f_Y(y) = \left.\frac{f_X(x)}{|g'(x)|}\right|_{y=g(x)}$$

Osservazione 33 *Se g non è monotona, sotto ipotesi opportune la formula si generalizza a*

$$f_Y(y) = \sum_{x:y=g(x)} \frac{f_X(x)}{|g'(x)|}.$$

Esercizio 12 *Se X è una v.a. esponenziale di parametro λ, trovare la densità di $Y = X^2$ seguendo il metodo di risoluzione degli esercizi precedenti e confrontare il risultato con la formula generale.*

Osservazione 34 *Una seconda dimostrazione della formula precedente proviene dalla seguente caratterizzazione delle densità: f è la densità di X se e solo se*

$$E[h(X)] = \int_{\mathbb{R}} h(x) f(x)\, dx$$

per tutte le funzioni continue e limitate h. Usiamo questo fatto per dimostrare che $f_Y(y) = \left.\frac{f_X(x)}{|g'(x)|}\right|_{y=g(x)}$ è la densità di $Y = g(X)$. Calcoliamo $E[h(Y)]$ per una generica funzione continua e limitata h. Dalla definizione di Y e dalla caratterizzazione precedente applicata a X, abbiamo

$$E[h(Y)] = E[h(g(X))] = \int_{\mathbb{R}} h(g(x)) f(x)\, dx.$$

Usiamo il teorema di cambio di variabile negli integrali, con $y = g(x)$, se g è monotona, biunivoca e differenziabile. Abbiamo $x = g^{-1}(y)$, $dx = \frac{1}{|g'(g^{-1}(y))|} dy$ (abbiamo scritto il valore assoluto per non cambiare gli estremi di integrazione) così che

$$\int_{\mathbb{R}} h(g(x)) f(x)\, dx = \int_{\mathbb{R}} h(y) f\left(g^{-1}(y)\right) \frac{1}{|g'(g^{-1}(y))|} dy.$$

Se poniamo $f_Y(y) := \left.\frac{f_X(x)}{|g'(x)|}\right|_{y=g(x)}$ abbiamo dimostrato che

$$E[h(Y)] = \int_{\mathbb{R}} h(y) f_Y(y)\, dy$$

per ogni funzione continua e limitata h. Usando di nuovo la caratterizzazione, deduciamo che $f_Y(y)$ è la densità di Y. Questa dimostrazione è basata sul cambio di variabile negli integrali.

Osservazione 35 *La stessa dimostrazione funziona nel caso multidimensionale, in cui non riusciamo più a lavorare con le cdf. Bisogna usare il teorema di cambio di variabile negli integrali multipli. Ricordiamo che in esso al posto di $dy = g'(x)dx$ si deve usare $dy =$*

$|\det Dg(x)|\,dx$ *dove* Dg *è la matrice jacobiana (la matrice delle derivate prime) della trasformazione* $g:\mathbb{R}^n\to\mathbb{R}^n$. *In realtà abbiamo bisogno della trasformazione inversa, quindi usiamo la formula*

$$dx=\left|\det Dg^{-1}(y)\right|dy=\frac{1}{|\det Dg(g^{-1}(y))|}dy.$$

Con gli stessi passaggi visti sopra nel caso 1-dimensionale, otteniamo il seguente risultato.

Proposizione 10 *Se* g *è biunivoca e differenziabile con matrice jacobiana invertibile e* $Y=g(X)$, *allora*

$$f_Y(y)=\left.\frac{f_X(x)}{|\det Dg(x)|}\right|_{y=g(x)}.$$

Corollario 4 *Sia* $X=(X_1,...,X_n)$ *un vettore casuale,* A *una matrice* $n\times n$ *invertibile,* $b\in\mathbb{R}^n$, *ed* $Y=(Y_1,...,Y_n)$ *un vettore casuale definito da*

$$Y=AX+b.$$

Se X *ha densità congiunta* $f_X(x)$ *allora anche* Y *ha densità congiunta, data da*

$$f_Y(y)=\frac{f_X\left(A^{-1}(y-b)\right)}{|\det A|}.$$

Proof. La trasformazione $g(x)=Ax+b$ è invertibile, con inversa $g^{-1}(y)=A^{-1}(y-b)$. La matrice jacobiana di $g(x)$ è A, costante. Basta allora sostituire questi fatti nella formula precedente. ■

Esercizio 13 *Se* X *(in* $\mathbb{R}^n$*) ha densità* $f_X(x)$ *e* $Y=UX$, *dove* U *è una trasformazione ortogonale di* $\mathbb{R}^n$ *(ovvero* $U^{-1}=U^T$*), allora* Y *ha densità*

$$f_Y(y)=f_X\left(U^Ty\right).$$

Soluzione. Le trasformazioni ortogonali sono invertibili ed hanno determinante pari a ±1, in quanto

$$1=\det I_d=\det\left(UU^T\right)=\det(U)\det\left(U^T\right)=\det(U)^2.$$

Basta quindi sostituire nella formula precedente.

1.5.2 Trasformazione lineare dei momenti

La soluzione dei seguenti esercizi è basata sulla linearità del valore atteso (e quindi della covarianza, rispetto a ciascuno dei suoi argomenti)

Esercizio 14 *Sia* $X=(X_1,...,X_n)$ *un vettore casuale,* A *una matrice* $n\times d$, *cioè* $A:\mathbb{R}^n\to\mathbb{R}^d$, $b\in\mathbb{R}^d$, *ed* $Y=(Y_1,...,Y_d)$ *un vettore casuale definito da*

$$Y=AX+b.$$

Sia $\mu^X=\left(\mu_1^X,...,\mu_n^X\right)$ *il vettore dei valori medi di* X, *ovvero* $\mu_i^X=E[X_i]$ *e sia* $\mu^Y=\left(\mu_1^Y,...,\mu_d^Y\right)$ *il vettore dei valori medi di* Y. *Allora*

$$\mu^Y=A\mu^X+b.$$

Soluzione. L'identità $Y = AX + b$, per componenti significa

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j + b_i.$$

Pertanto, per la linearità del valor medio,

$$E[Y_i] = E\left[\sum_{j=1}^{n} A_{ij} X_j + b_i\right] = \sum_{j=1}^{n} A_{ij} E[X_j] + b_i$$

che è la versione per componenti dell'identità da dimostrare.

Esercizio 15 *Sotto le stesse ipotesi, se Q^X e Q^Y sono le matrici di covarianza di X ed Y, allora*

$$Q^Y = AQ^X A^T.$$

Soluzione. Sempre usando l'identità per componenti scritta sopra,

$$\begin{aligned} Q^Y_{ij} &= Cov(Y_i, Y_j) = Cov\left(\sum_{i'=1}^{n} A_{ii'} X_{i'} + b_i, \sum_{j'=1}^{n} A_{jj'} X_{j'} + b_j\right) \\ &= \sum_{i'=1}^{n} A_{ii'} Cov\left(X_{i'}, \sum_{j'=1}^{n} A_{jj'} X_{j'} + b_j\right) \\ &= \sum_{i'=1}^{n} A_{ii'} \sum_{j'=1}^{n} A_{jj'} Cov\left(X_{i'}, X_{j'}\right) = \sum_{i'=1}^{n} \sum_{j'=1}^{n} A_{ii'} Q^X_{i'j'} A_{jj'} \end{aligned}$$

avendo usato la linearità della covarianza nelle due componenti. Ricordiamo che, date due matrici A e B, vale $(AB)_{ij} = \sum_k A_{ik} B_{kj}$. Allora

$$\sum_{i'=1}^{n} \sum_{j'=1}^{n} A_{ii'} Q^X_{i'j'} A_{jj'} = \sum_{j'=1}^{n} \left(AQ^X\right)_{ij'} A_{jj'}.$$

Per interpretare anche quest'ultimo come prodotto tra matrici bisogna trasporre A:

$$= \sum_{j'=1}^{n} \left(AQ^X\right)_{ij'} A^T_{j'j} = \left(AQ^X A^T\right)_{ij}.$$

L'esercizio è risolto.

1.5.3 Sulle matrici di covarianza

La matrice di covarianza Q di un vettore $X = (X_1, ..., X_n)$, definita da $Q_{ij} = Cov(X_i, X_j)$, è simmetrica:

$$Q_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i) = Q_{ji}$$

e definita non-negativa:

$$x^T Q x = \sum_{i,j=1}^{n} Q_{ij} x_i x_j = \sum_{i,j=1}^{n} Cov\left(X_i, X_j\right) x_i x_j = \sum_{i,j=1}^{n} Cov\left(x_i X_i, x_j X_j\right)$$
$$= Cov\left(\sum_{i=1}^{n} x_i X_i, \sum_{j=1}^{n} x_j X_j\right) = Var\left[W\right]$$

dove $W = \sum_{i=1}^{n} x_i X_i$.

Il teorema spettrale afferma che ogni matrice simmetrica Q può essere diagonalizzata, nel senso che esiste una base ortonormale $e_1, ..., e_n$ di $\mathbb{R}^n$ in cui Q assume la forma

$$Q_e = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}.$$

Inoltre, i numeri λ_i sulla diagonale sono gli autovalori di Q ed i vettori e_i sono i corrispondenti autovettori. Dal punto di vista algebrico, quanto detto significa che esiste una matrice ortogonale U ed una matrice diagonale Q_e tali che

$$Q = U Q_e U^T.$$

Inoltre, U ha come colonne una base ortonormale di autovettori e_i corrispondenti ai λ_i:

$$U = \left(e_1, ..., e_n\right).$$

Esercizio 16 *Verificare che, se $U = (e_1, ..., e_n)$, Q_e è come sopra e $Q = U Q_e U^T$, allora $Qe_1 = \lambda_1 e_1$ (lo stesso vale per gli altri autovettori).*

Soluzione. Detto u_1 il vettore $\begin{pmatrix} 1 & 0 & ... & 0 \end{pmatrix}^T$, vale $Qe_1 = UQ_eU^Te_1 = UQ_eu_1$, in quanto nei prodotti riga per colonna di U^Te_1 si devono fare i prodotti scalari degli e_i con e_1, tutti zero tranne il primo. Quindi $Qe_1 = UQ_eu_1 = \lambda_1 U u_1 = \lambda_1 e_1$.

Siccome una matrice di covarianza Q è anche definita non-negativa, vale

$$\lambda_i \geq 0, \qquad i = 1, ..., n.$$

Usando entrambi questi fatti si può definire la radice quadrata di Q, cioè una matrice simmetrica che indicheremo con $\sqrt{Q}$, tale che $\left(\sqrt{Q}\right)^2 = Q$. Infatti, in primo luogo possiamo definire facilmente la radice quadrata di Q_e, ponendo

$$\sqrt{Q_e} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sqrt{\lambda_n} \end{pmatrix}.$$

Si vede subito che questa è simmetrica ed il suo quadrato è Q_e. Tramite questa matrice, "tornando indietro", possiamo definire

$$\sqrt{Q} := U \sqrt{Q_e} U^T.$$

Si verifica facilmente che la matrice $\sqrt{Q}$ è simmetrica ed il suo quadrato è uguale a Q. Infatti Abbiamo

$$\left(\sqrt{Q}\right)^T = U\left(\sqrt{Q_e}\right)^T U^T = U\sqrt{Q_e}U^T = \sqrt{Q}$$

e

$$\left(\sqrt{Q}\right)^2 = U\sqrt{Q_e}U^T U\sqrt{Q_e}U^T = U\sqrt{Q_e}\sqrt{Q_e}U^T = UQ_eU^T = Q$$

in quanto $U^TU = Id$.

Commenti di algebra lineare

Osservazione 36 *Per capire a fondo questo teorema, soprattutto dal punto di vista geometrico, ricordiamo alcuni fatti di algebra lineare. $\mathbb{R}^n$ è uno spazio vettoriale con prodotto scalare $\langle .,. \rangle$, cioè un insieme di elementi (vettori) con certe operazioni (somma di vettori, moltiplicazione per numeri reali, prodotto scalare tra vettori) e certe proprietà. Possiamo chiamare oggetti intrinseci gli oggetti definiti in questi termini, al contrario di quelli definiti tramite coordinate rispetto ad una base. Un vettore $x \in \mathbb{R}^n$ è un oggetto intrinseco; quando lo scriviamo nella forma $(x_1, ..., x_n)$ rispetto ad una base, questa scrittura non è intrinseca, dipende dalla base. Data una base ortonormale $u_1, ..., u_n$, le componenti di un vettore $x \in \mathbb{R}^n$ in tale base sono i numeri $\langle x, u_j \rangle$, $j = 1, ..., n$. Un'applicazione lineare L in $\mathbb{R}^n$ è un oggetto intrinseco: è una funzione $L : \mathbb{R}^n \to \mathbb{R}^n$ tale che $L(\alpha v + \beta w) = \alpha Lv + \beta Lw$ per ogni $v, w \in \mathbb{R}^n$ ed ogni $\alpha, \beta \in \mathbb{R}$. Data la base $u_1, ..., u_n$, L può essere rappresentata tramite la matrice di componenti $\langle Lu_i, u_j \rangle$; questa matrice non è intrinseca. Scriveremo a volte $y^T \cdot x$ al posto di $\langle x, y \rangle$ (o $\langle y, x \rangle$).*

Osservazione 37 *Dopo questi commenti di carattere generale, riconosciamo che una matrice rappresenta un'applicazione lineare relativamente ad una base specificata. Quindi, data la base canonica di $\mathbb{R}^n$, che indicheremo con $u_1, ..., u_n$, data la matrice Q, è definita una ben precisa applicazione lineare $L : \mathbb{R}^n \to \mathbb{R}^n$; e viceversa, data L e data una qualsiasi base $e_1, ..., e_n$ di $\mathbb{R}^n$, L si scrive in questa base tramite una matrice. Il teorema spettale afferma che se Q era simmetrica, allora esiste una base ortonormale $e_1, ..., e_n$ in cui la rappresentazione matriciale Q_e di L è diagonale.*

Osservazione 38 *Ricordiamo alcuni altri fatti di algebra lineare. Partiamo da una base ortonormale $u_1, ..., u_n$, che chiameremo canonica o base originaria. Sia $e_1, ..., e_n$ un'altra base ortonormale. Il vettore u_1, nella base canonica, ha componenti*

$$u_1 = \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix}$$

e così via per gli altri vettori. Ogni vettore e_j ha certe componenti nella base canonica. Indichiamo con U la matrice la cui prima colonna ha le componenti di e_1, la seconda quelle

di e_2 e così via. Potremmo scrivere $U = (e_1, ..., e_n)$. Vale anche $U_{ij} = e_j^T \cdot u_i$. Quindi

$$U \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix} = e_1$$

e così via, cioè U rappresenta l'applicazione lineare che trasforma la base canonica in $e_1, ..., e_n$

$$Uu_i = e_i, \qquad i = 1, ..., n.$$

Essa è una trasformazione ortogonale:

$$U^{-1} = U^T.$$

Infatti, U^{-1} trasforma $e_1, ..., e_n$ nella base canonica (invertendo quanto appena detto su U), e U^T fa lo stesso:

$$U^T e_1 = \begin{pmatrix} e_1^T \cdot e_1 \\ e_2^T \cdot e_1 \\ ... \\ e_n^T \cdot e_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix}$$

e così via. Ricordiamo che le trasformazioni ortogonali sono isometrie, come le rotazioni o le riflessioni.

Osservazione 39 *Torniamo alla matrice di covarianza Q ed alla matrice Q_e data dal teorema spettrale: sappiamo che Q_e è diagonale e rappresenta la stessa trasformazione lineare L, nella nuova base $e_1, ..., e_n$. Supponiamo di non sapere altro che questo, cioè che rappresentano la stessa trasformazione lineare L e che Q_e ha la forma*

$$Q_e = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}.$$

Da questo deduciamo alcuni fatti:

i)

$$Q_e = UQU^T$$

ii) gli elementi sulla diagonale λ_j sono autovalori di L, con autovettori e_j

iii) $\lambda_j \geq 0$, $j = 1, ..., n$.

Per dimostrare (i), ricordiamo che abbiamo appena visto che

$$(Q_e)_{ij} = e_j^T \cdot Le_i \text{ e } Q_{ij} = u_j^T \cdot Lu_i.$$

Inoltre, $U_{ij} = e_j^T \cdot u_i$, quindi $e_j = \sum_{k=1}^n U_{kj} u_k$, e di conseguenza

$$(Q_e)_{ij} = e_j^T \cdot Le_i = \sum_{k,k'=1}^n U_{ki} U_{k'j} u_{k'}^T \cdot Lu_k = \sum_{k,k'=1}^n U_{ki} Q_{ij} U_{k'j} = \left(UQU^T\right)_{ij}.$$

Per dimostrare (ii), scriviamo il vettore Le_1 *nella base* $e_1, ..., e_n$*:* e_i *è il vettore* $\begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix}$*, l'applicazione* L *è rappresentata da* Q_e*, quindi* Le_1 *è uguale a*

$$Q_e \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ 0 \\ ... \\ 0 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ ... \\ 0 \end{pmatrix}$$

che è $\lambda_1 e_1$ *nella base* $e_1, ..., e_n$*. Abbiamo verificato che* $Le_1 = \lambda_1 e_1$*, cioè che* λ_1 *è un autovalore e che* e_1 *è il corrispondente autovettore. La dimostrazione per* λ_2*, ecc. è la stessa.*

Per dimostrare (iii), basta osservare che, nella base $e_1, ..., e_n$*,*

$$e_j^T Q_e e_j = \lambda_j.$$

Ma

$$e_j^T Q_e e_j = e_j^T U Q U^T e_j = v^T Q v \geq 0$$

dove $v = U^T e_j$*, avendo usato la proprietà che* Q *è definita non-negativa. Quindi* $\lambda_j \geq 0$*.*

1.5.4 Vettori gaussiani

Ricordiamo che una v.a. gaussiana o normale $N\left(\mu, \sigma^2\right)$ è una v.a. con densità di probabilità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|x-\mu|^2}{2\sigma^2}\right).$$

Si dimostra che μ è la media e σ^2 la varianza. La normale standard è il caso $\mu = 0$, $\sigma^2 = 1$. Se Z è una normale standard allora $\mu + \sigma Z$ è $N\left(\mu, \sigma^2\right)$, ed ogni gaussiana $N\left(\mu, \sigma^2\right)$ si può scrivere nella forma $\mu + \sigma Z$ con $Z \sim N(0, 1)$.

Si può dare la definizione di *vettore* gaussiano, o gaussiana multidimensionale, in più modi, generalizzando o l'espressione per la densità oppure la proprietà che $\mu + \sigma Z$ è una $N\left(\mu, \sigma^2\right)$. Vediamoli entrambi e la loro equivalenza (valida sotto una certa ipotesi).

Definizione tramite trasformazione lineare di un vettore normale standard

Definizione 30 *i) Chiamiamo vettore normale standard in d dimensioni un vettore aleatorio* $Z = (Z_1, ..., Z_d)$ *con densità congiunta*

$$f(z_1, ..., z_d) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} = \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{z_1^2 + ... + z_d^2}{2}}.$$

ii) Tutti gli altri vettori gaussiani $X = (X_1, ..., X_n)$ *(in dimensione generica* n*) si ottengono da quelli standard tramite le trasformazioni affini:*

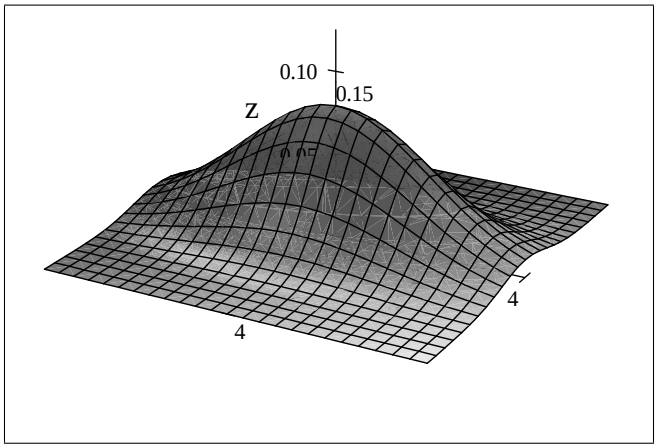
$$X = AZ + b$$

dove A è una matrice e b è un vettore. Se X ha dimensione n, richiediamo che A sia $d \times n$ (cioè $A : \mathbb{R}^d \to \mathbb{R}^n$) e b abbia dimensione n (ma n può essere diverso da d).

Il grafico della densità normale standard in 2 dimensioni è stato tracciato nel paragrafo 1.2.8. Il grafico delle altre densità gaussiane può essere immaginato eseguendo trasformazioni lineari del piano base xy (deformazioni definite da A) e traslazioni (di b). Per esempio, se

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

matrice che amplifica l'asse x di un fattore 2, otteniamo il seguente grafico:

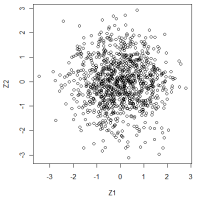


Tuttavia questo tipo di grafico è poco interpretabile, salvo casi facili come quello invariante per rotazione o quello appena disegnato; grazie alle ombreggiature riusciamo ad intuire qualcosa, ma con poca precisione.

Le curve di livello, cioè le curve nello spazio R^k definite dalle equazioni $f(x) = a$ al variare di $a > 0$, dove f è la densità congiunta di X, sono un modo più efficace, a cui siamo abituati da sempre se pensiamo alle cartine geografiche; esso però richiede la conoscenza di f, che preferiamo non usare in questa specifica argomentazione basata sulla definizione 30; quando più tardi avremo la densità, vedremo che le curve di livello sono ellissi concentriche, di centro μ, indipendentemente da A.

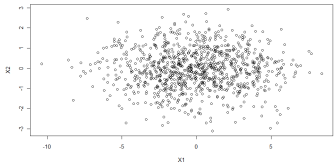
Un altro modo di visualizzare un vettore aleatorio X è di raffigurare un suo campione sperimentale, numeroso: è di nuovo una raffigurazione nello spazio R^k, lo spazio dei valori possibili di X; se i punti sono molti, si riesce ad intuire la struttura delle curve di livello. Tornando alla definizione 30, un modo di avere un campione di numerosità N da X è quello di averlo da Z e trasformarlo. Iniziamo allora osservando che un campione di numerosità $N = 1000$ (per fare un esempio) estratto da Z in dimensione $n = 2$ si ottiene e raffigura coi comandi

Z1<-rnorm(1000); Z2<-rnorm(1000); plot(Z1,Z2)



Possiamo poi osservare l'effetto di una matrice del tipo $A_1 = \begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix}$ con $\lambda \neq 1$, ad esempio $\lambda = 3$:

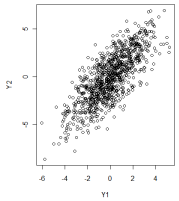
X1<-3*Z1; X2<-Z2; plot(X1,X2)



Infine, possiamo vedere l'effetto di una successiva rotazione (in senso antiorario) di θ radianti, ottenuta componendo ulteriormente con la matrice $A_2 = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ (componiamo le due, cioè applichiamo prima la matrice A_1 poi la A_2); vediamo ad esempio $\theta = 1$:

A11 = cos(1); A12 = -sin(1); A21 = sin(1); A22 = cos(1)

Y1 <- A11*X1+A12*X2; Y2 <- A21*X1+A22*X2; plot(Y1,Y2)



Riprendiamo l'analisi teorica delle gaussiane. Calcoliamo media e covariana di un vettore della forma $X = AZ + b$, con Z di tipo standard. Dagli esercizi 14 e 15 abbiamo:

Proposizione 11 *Il vettore dei valori medi μ e la matrice di covarianza Q di un vettore X della forma precedente sono dati da*

$$\mu = b$$
$$Q = AA^T.$$

Esercizio 17 *Sia $X = (X_1, ..., X_n)$ un vettore gaussiano secondo la definizione 30, B una matrice $n \times m$, c un vettore di $\mathbb{R}^m$. Allora*

$$Y = BX + c$$

è un vettore gaussiano di dimensione m (sempre secondo la definizione 30). La relazione tra medie e covarianze è

$$\mu_Y = B\mu_X + c$$

$$Q_Y = BQ_X B^T.$$

Osservazione 40 *Dall'esercizio vediamo che si può partire da un vettore non-degenere X ed ottenere un vettore degenere Y, se B non è biunivoca. Questo accade sempre se $m > n$.*

Definizione tramite densità

Bisogna premettere il seguente lemma, un po' laborioso. Si consiglia di apprenderne l'enunciato tralasciando la dimostrazione, almeno in un primo momento.

Lemma 2 *Dato un vettore $\mu = (\mu_1, ..., \mu_n)$ ed una matrice $n \times n$ simmetrica definita positiva Q (cioè tale che $v^T Q v > 0$ per ogni $v \neq 0$), si consideri la funzione*

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1} (x-\mu)}{2} \right)$$

dove $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Si noti che la matrice inversa Q^{-1} è ben definita (in quanto Q è definita positiva), il numero $(x-\mu)^T Q^{-1} (x-\mu)$ è non negativo, ed il determinante $\det(Q)$ è positivo. Allora:

i) $f(x)$ è una densità di probabilità;

ii) se $X = (X_1, ..., X_n)$ è un vettore aleatorio con tale densità congiunta, allora μ è il vettore dei valori medi, nel senso che

$$\mu_i = E[X_i]$$

e Q è la matrice di covarianza:

$$Q_{ij} = Cov(X_i, X_j).$$

Proof. Step 1. In questo primo passo spieghiamo il significato dell'espressione che definisce $f(x)$. Abbiamo ricordato sopra che ogni matrice simmetrica Q può essere diagonalizzata, cioè esiste una base ortonormale $e_1, ..., e_n$ di $\mathbb{R}^n$ in cui Q ha la forma

$$Q_e = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}.$$

Inoltre, i valori λ_i sulla diagonale sono gli autovalori di Q, ed i vettori e_i sono i corrispondenti autovettori. Si veda il paragrafo sulla matrice di correlazione per ulteriori dettagli. Sia U la matrice introdotta in quel paragrafo, tale che $U^{-1} = U^T$. Si ricordi la relazione $Q_e = UQU^T$.

Essendo $v^T Q v > 0$ per tutti i vettori $v \neq 0$, vale

$$v^T Q_e v = \left(v^T U\right) Q \left(U^T v\right) > 0$$

per ogi $v \neq 0$ (in quanto $U^T v \neq 0$). Preso in particolare $v = e_i$, troviamo $\lambda_i > 0$.

Se ne deduce che la matrice Q_e è invertibile con inversa

$$Q_e^{-1} = \begin{pmatrix} \lambda_1^{-1} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n^{-1} \end{pmatrix}.$$

Si deduce inoltre che Q, essendo uguale a $U^T Q_e U$ (la relazione $Q = U^T Q_e U$ discende da $Q_e = UQU^T$), è invertibile, con inversa $Q^{-1} = U^T Q_e^{-1} U$. Si deduce allora facilmente $(x-\mu)^T Q^{-1} (x-\mu) > 0$ per ogni $x \neq \mu$. Inoltre, vale

$$\det(Q) = \det\left(U^T\right) \det(Q_e) \det(U) = \lambda_1 \cdots \lambda_n$$

in quanto

$$\det(Q_e) = \lambda_1 \cdots \lambda_n$$

e $\det(U) = \pm 1$. Quest'ultimo fatto discende da

$$1 = \det I = \det\left(U^T U\right) = \det\left(U^T\right) \det(U) = \det(U)^2$$

(che verrà usato nell'esercizio 13). Quindi $\det(Q) > 0$. La formula per $f(x)$ ha senso e definisce una funzione positiva.

Step 2. Proviamo ora che $f(x)$ è una densità. Per il teorema di cambio di variabile negli integrali multidimensionali, col cambio di variabile $x = U^T y$, troviamo

$$\int_{\mathbb{R}^n} f(x)\, dx = \int_{\mathbb{R}^n} f\left(U^T y\right) dy$$

in quanto $\left|\det U^T\right| = 1$ (e la matrice jacobiana di una trasformazione lineare è la matrice stessa). Ora, essendo $UQ^{-1}U^T = Q_e^{-1}$, $f\left(U^T y\right)$ coincide con la seguente funzione:

$$f_e(y) = \frac{1}{\sqrt{(2\pi)^n \det(Q_e)}} \exp\left(-\frac{(y-\mu_e)^T Q_e^{-1} (y-\mu_e)}{2}\right)$$

dove abbiamo posto

$$\mu_e = U\mu.$$

Essendo

$$(y-\mu_e)^T Q_e^{-1} (y-\mu_e) = \sum_{i=1}^n \frac{(y_i - (\mu_e)_i)^2}{\lambda_i}$$

e $\det(Q_e) = \lambda_1 \cdots \lambda_n$, otteniamo

$$f_e(y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{(y_i - (\mu_e)_i)^2}{2\lambda_i}\right).$$

In altre parole, $f_e(y)$ è il prodotto di n densità gaussiane $N\left((\mu_e)_i, \lambda_i\right)$. Sappiamo dalla teoria che il prodotto di densità è la densità congiunta di un vettore fatto di componenti indipendenti. Quindi $f_e(y)$ è una densità di probabilità. Pertanto $\int_{\mathbb{R}^n} f_e(y)\, dy = 1$. Questo dimostra $\int_{\mathbb{R}^n} f(x)\, dx = 1$, ovvero f è una densità di probabilità.

Step 3. Sia $X = (X_1, ..., X_n)$ un vettore aleatorio con densità di probabilità f, se scritto nella base originaria. Sia $Y = UX$. Allora (esercizio 13) Y ha densità $f_Y(y)$ data da $f_Y(y) = f\left(U^T y\right)$. Quindi

$$f_Y(y) = f_e(y) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{(y_i - (\mu_e)_i)^2}{2\lambda_i}\right).$$

In altre parole, le componenti di $(Y_1, ..., Y_n)$ sono v.a. indipendent $N((\mu_e)_i, \lambda_i)$ e quindi

$$E[Y_i] = (\mu_e)_i, \qquad Cov(Y_i, Y_j) = \delta_{ij}\lambda_i.$$

Dagli esercizi 14 e 15 deduciamo che $X = U^T Y$ ha media

$$\mu_X = U^T \mu_Y$$

e covarianza

$$Q_X = U^T Q_Y U.$$

Essendo $\mu_Y = \mu_e$ e $\mu_e = U\mu$ deduciamo $\mu_X = U^T U\mu = \mu$. Ma $Q_Y = Q_e$ e $Q = U^T Q_e U$, per cui $Q_X = Q$. La dimostrazione è completa. ■

Definizione 31 *Dato un vettore $\mu = (\mu_1, ..., \mu_n)$ ed una matrice $n \times n$ simmetrica definita positiva Q, chiamiamo vettore gaussiano di media μ e covarianza Q un vettore aleatorio $X = (X_1, ..., X_n)$ avente densità congiunta*

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1}(x-\mu)}{2}\right)$$

dove $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Scriviamo $X \sim N(\mu, Q)$.

Proposizione 12 *Se $X = (X_1, ..., X_n)$ è un vettore gaussiano $N(\mu, Q)$ secondo questa definizione, B è una matrice invertibile $n \times n$ e $c \in \mathbb{R}^n$, allora $Y = BX + c$ è gaussiano $N\left(B\mu + c, BQB^T\right)$.*

Proof. Per il Corollario 4, Y ha densità congiunta, data da

$$f_Y(y) = \frac{f_X\left(B^{-1}(y-c)\right)}{|\det B|}.$$

Sostituendo la formula di f_X troviamo

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n |\det B| \det(Q) |\det B|}} \exp\left(-\frac{\left(B^{-1}(y-c)-\mu\right)^T Q^{-1}\left(B^{-1}(y-c)-\mu\right)}{2}\right).$$

Da un lato

$$|\det B| \det(Q) |\det B| = \det(BQB^T).$$

Dall'altro,

$$\begin{aligned}
\left(B^{-1}(y-c)-\mu\right)^T Q^{-1}\left(B^{-1}(y-c)-\mu\right) &= \left(B^{-1}(y-c-B\mu)\right)^T Q^{-1}\left(B^{-1}(y-c-B\mu)\right) \\
&= \left\langle Q^{-1}\left(B^{-1}(y-c-B\mu)\right), B^{-1}(y-c-B\mu)\right\rangle \\
&= \left\langle \left(B^{-1}\right)^T Q^{-1} B^{-1}(y-c-B\mu), y-c-B\mu\right\rangle \\
&= (y-c-B\mu)^T (BQB^T)^{-1}(y-c-B\mu).
\end{aligned}$$

Quindi

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det(BQB^T)}} \exp\left(-\frac{(y-c-B\mu)^T (BQB^T)^{-1} (y-c-B\mu)}{2}\right)$$

che è la densità di una $N\left(B\mu + c, BQB^T\right)$. ■

L'unica restrizione della definizione 31 è l'ipotesi che Q sia definita positiva. La definizione 30 non ha questo difetto.

Avendo dato due definizioni diverse di vettore gaussiano, dobbiamo dimostrarne l'equivalenza. Se Q è definita positiva, le due definizioni di vettore gaussiano definiscono lo stesso oggetto, ma se Q è solamente definita non-negativa, abbiamo solamente l'utima definizione, quindi non dobbiamo dimostrare nessuna equivalenza.

Proposizione 13 *Se Q è definita positiva, allora le definizioni 31 e 30 sono equivalenti. Più precisamente, se $X = (X_1, ..., X_n)$ è un vettore aleatorio gaussiano di media μ e covarianza Q nel senso della definizione 31, allora esistono un vettore normale standard $Z = (Z_1, ..., Z_n)$ ed una matrice $n \times n$, A, tali che*

$$X = AZ + \mu.$$

Si può prendere (la scelta di A non è univoca) $A = \sqrt{Q}$, come descritto nella dimostrazione. Viceversa, se $X = (X_1, ..., X_n)$ è un vettore gaussiano nel senso della definizione 30, della forma $X = AZ + \mu$, con Z di dimensione n ed A invertibile (o comunque Z di dimensione $\geq n$ ed AA^T invertibile), allora X è gaussiano nel senso della definizione 31, con media μ e covarianza $Q = AA^T$.

Proof. Dimostriamo la prima affermazione. Ricordiamo che nella sezione 1.5.3 abbiamo definito la radice quadrata di Q come

$$\sqrt{Q} := U\sqrt{Q_e}U^T.$$

Poniamo allora

$$Z = \left(\sqrt{Q}\right)^{-1}(X - \mu)$$

dove osserviamo che $\sqrt{Q}$ è invertibile, in base alla sua definizione ed alla positività stretta dei λ_i. Allora possiamo applicare la Proposizione 12 ed ottenere che Z è gaussiano

$$N\left(\left(\sqrt{Q}\right)^{-1}(\mu - \mu), \left(\sqrt{Q}\right)^{-1} Q \left(\sqrt{Q}^T\right)^{-1}\right)$$

ovvero, semplificando, $N(0, Id)$. Infatti

$$\left(\sqrt{Q}\right)^{-1} Q \left(\sqrt{Q}^T\right)^{-1} = \left(\sqrt{Q}\right)^{-1} \sqrt{Q}\sqrt{Q} \left(\sqrt{Q}\right)^{-1} = Id$$

usando il fatto che $\sqrt{Q}$ è simmetrica ed il suo quadrato è uguale a Q. In conclusione, abbiamo trovato che Z è un vettore normale standard. Dalla definizione di Z troviamo $X = \sqrt{Q}Z + \mu$, così che la prima affermazione è dimostrata.

Viceversa, supponiamo che $X = AZ+\mu$ sia un vettore gaussiano secondo la definizione 30, con Z di dimensione n ed A invertibile. La densità congiunta di Z è $f_Z(z) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{z^T z}{2}\right)$. Per il Corollario 4, X ha densità congiunta, data da

$$f_X(x) = \frac{f_Z\left(A^{-1}(x-\mu)\right)}{|\det A|}.$$

Sostituendo la formula di f_Z troviamo

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{(2\pi)^n |\det A| |\det A|}} \exp\left(-\frac{\left(A^{-1}(x-\mu)\right)^T \left(A^{-1}(x-\mu)\right)}{2}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n |\det (AA^T)|}} \exp\left(-\frac{(x-\mu)^T \left(AA^T\right)^{-1}(x-\mu)}{2}\right) \end{aligned}$$

e quindi X è gaussiano $N\left(\mu, AA^T\right)$ secondo la definizione 31. La dimostrazione è completa. ■

Osservazione 41 *La densità di un vettore gaussiano (quando Q è invertibile) è determinata dal vettore dei valori medi e dalla matrice di covarianza. Questo fatto fondamentale verrà usato più tardi nello studio dei processi stocastici. Usando il concetto di legge di un vettore aleatorio, questo fatto vale anche nel caso degenere, senza densità, in cui si deve usare la definizione 30, però i dettagli sono meno elementari.*

Osservazione 42 *Alcuni dei risultati precedenti sono molto utili se vogliamo generare vettori aleatori secondo una legge gaussiana specificata. Assumiamo di aver prescritto la media μ e la covarianza Q, n-dimensionali, e vogliamo generare un punto casuale $(x_1, ..., x_n)$ dalla $N(\mu, Q)$. Per far questo possiamo generare n numeri casuali indipendenti $z_1, ..., z_n$ dalla normale standard 1-dimensionale e calcolare*

$$\sqrt{Q} z + \mu$$

dove $z = (z_1, ..., z_n)$. Per avere le componenti della matrice $\sqrt{Q}$, se il software non le fornisce automaticamente (alcuni software lo fanno), possiamo usare la formula $\sqrt{Q} = U\sqrt{Q_e}U^T$. La matrice $\sqrt{Q_e}$ è ovvia. Per ottenere la matrice U si ricordi che le sue colonne sono gli autovettori $e_1, ..., e_n$ scritti nella base di partenza. Basta quindi che il software sia in grado di effettuare la decomposizione spettale di Q.

Curve di livello

Come abbiamo già accennato sopra, un modo di visualizzare un grafico in due dimensioni è quello di tracciare le sue curve di livello. Data $f : \mathbb{R}^2 \to \mathbb{R}$, la curva di livello a è il luogo dei punti $x \in \mathbb{R}^2$ tali che $f(x) = a$. Nel caso di una densità f, essendo positiva, ha senso esaminare solo il caso $a > 0$. Nel caso della gaussiana $N(\mu, Q)$, dobbiamo capire l'equazione

$$\frac{1}{\sqrt{(2\pi)^n \det(Q)}} \exp\left(-\frac{(x-\mu)^T Q^{-1}(x-\mu)}{2}\right) = a.$$

Posto $a' = -\frac{1}{2}\log\left(a\sqrt{(2\pi)^n \det(Q)}\right)$, l'equazione diventa

$$(x-\mu)^T Q^{-1}(x-\mu) = a'.$$

Questa è l'equazione di un'ellisse di centro μ. Infatti, usando la solita scomposizione $Q = UQ_eU^T$ dove Q_e è la matrice diagonale

$$Q_e = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

e posto $\widetilde{x} = x - \mu$ (traslazione che porta μ in 0), l'equazione diventa

$$\widetilde{x}^T \left(U^T\right)^{-1} Q_e^{-1} U^{-1} \widetilde{x} = a'$$

e poi, posto $y = U^{-1}\widetilde{x}$ (una rotazione) troviamo

$$y^T Q_e^{-1} y = a'$$

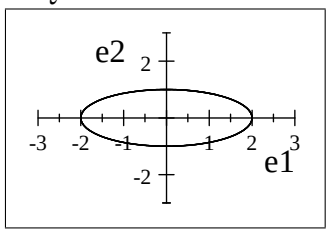
che in coordinate si legge

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = a'$$

ovvero un'ellisse. Le lunghezze lungo gli assi (precisamente le lunghezze dei segmenti che uniscono l'origine ai vertici dell'ellisse) sono pari a $\sqrt{\lambda_1}$ e $\sqrt{\lambda_2}$. L'orientazione degli assi è quella degli autovettori di Q, come si può verificare ragionando più da vicino sul significato della trasformazione U. In conclusione:

Proposizione 14 *Le curve di livello di un vettore gaussiano $N(\mu, Q)$ in due dimensioni sono ellissi di centro μ ed assi dati dagli autovettori di Q, con lunghezze degli assi pari alle radici degli autovalori $\sqrt{\lambda_1}$ e $\sqrt{\lambda_2}$.*

Il seguente disegno raffigura il caso $\frac{y_1^2}{4} + \frac{y_2^2}{1} = 1$. Questa è l'ellisse rispetto agli assi e_1, e_2, base di autovettori di Q.



Se invece vogliamo vedere l'ellisse nella base canonica originaria, quella delle variabili x_i, bisogna eseguire la rotazione U e la traslazione di μ. Non c'è bisogno di sapere con esattezza di che rotazione si tratta, basta sapere come appaiono i vettori e_1, e_2 nella base canonica (cioè avere le loro coordinate), e tracciare l'ellisse con tali assi.

I risultati ora esposti si generalizzano a più di due dimensioni, usando la nozione di ellissoide.

Un'altra definizione

Esistono altre definzioni di vettore gaussiano. Per curiosità enunciamo la seguente, che è forse la più veloce ma può apparire più oscura di altre.

Definizione 32 *Un vettore aleatorio* $X = (X_1, ..., X_n)$ *si dice gaussiano se accade che per ogni vettore di numeri reali* $u = (u_1, ..., u_n)$ *la v.a.*

$$\langle u, X \rangle = \sum_{i=1}^{n} u_i X_i$$

sia gaussiana.

Questa definizione generalizza la nota proprietà che le combinazioni lineari di gaussiane indipedenti sono gaussiane. Con questa definizione è immediato verificare che le trasformazioni lineari di vettori gaussiani sono vettori gaussiani.

La definizione data ha anche una certa interpretazione geometrica. Se u ha lunghezza unitaria, l'epressione $\langle u, X \rangle$ è la proiezione di X su u. La definizione afferma quindi che tutte le proiezioni uni-dimensionali sono gaussiane.

Capitolo 2

Elementi di Statistica

2.1 Introduzione. Stimatori

Gli elementi di statistica esposti in questo capitolo costituiscono solo un breve riassunto e non hanno alcuno scopo di organicità e completezza. Verranno lasciate fuori molte questioni ed argomenti importanti.

I due problemi principali esaminati dalla statistica di base sono:

- la *stima dei parametri* o più in generale la costruzione di modelli probabilistici a partire da dati sperimentali
- i *test di ipotesi*, o più in generale la verifica dei modelli ipotizzati eseguita confrontandoli con dati sperimentali.

Alcune definizioni e considerazioni generali possono essere premesse allo studio di questi due problemi.

Definizione 33 *Data una v.a. X, si chiama campione di numerosità n estratto da X una sequenza di n v.a. $X_1, ..., X_n$ indipendenti e distribuite come X (e definite sullo stesso spazio probabilizzato $(\Omega, \mathcal{F}, P)$).*

Questo concetto è simile a quello di *campione sperimentale* composto da numeri reali $x_1, ..., x_n$ emersi da prove, osservazioni sperimentali. Però sono due concetti diversi: il campione $X_1, ..., X_n$ è fatto di v.a., l'altro, $x_1, ..., x_n$, di numeri; quest'ultimo è per così dire una *realizzazione* del primo. Si può immaginare, per capire la differenza, che dopo aver eseguito gli esperimenti si siano trovati i numeri $x_1, ..., x_n$, mentre prima di eseguirli si possa immaginare che gli esperimenti produrranno dei numeri; in tale fase precedente agli esperimenti, i valori numerici che usciranno dagli esperimenti sono grandezze aleatorie, ad esito incognito, che possiamo descrivere con v.a. $X_1, ..., X_n$.

Definizione 34 *Dato il campione $X_1, ..., X_n$ estratto da X, chiamiamo media aritmetica o*

empirica e varianza empirica le v.a.

$$\overline{X} = \frac{X_1 + ... + X_n}{n}$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Nel caso di un campione sperimentale $x_1, ..., x_n$, si definiscono nello stesso modo i corrispondenti numeri $\overline{x}$ ed s^2.

Proposizione 15 *Se X è $N\left(\mu, \sigma^2\right)$, allora $\overline{X}$ è $N\left(\mu, \frac{\sigma^2}{n}\right)$.*

Più in generale, se X ha media μ e varianza σ^2, allora $\overline{X}$ ha media μ e varianza $\frac{\sigma^2}{n}$, ed è asintoticamente normale nel senso descritto tra breve.

Proof. Diamo solo un cenno. La verifica che $\overline{X}$ ha media μ e varianza $\frac{\sigma^2}{n}$ è elementare, usando le regole dei valori medi. L'asintotica normalità di dimostra col teorema limite centrale. Quando poi X è già gaussiana, lo è anche $\overline{X}$ per il fatto che le combinazioni affini di v.a. gaussiane indipendenti è gaussiana. ■

Le v.a. $\overline{X}$ ed S^2 sono esempi di *stimatori*, cioè di v.a.

$$T = T\left(X_1, ..., X_n\right)$$

che vengono usate per stimare (approssimare) parametri di distribuzioni statistiche. Se la v.a. X ha media μ e varianza σ^2, per varie ragioni si utilizzano $\overline{X}$ ed S^2 come stimatori di μ e σ^2. Vedremo anche un esempio di stimatore della cdf $F(x)$ e, nel capitolo sui processi, esempi di stimatori di grandezze relative a più variabili aleatorie, come la covarianza e la correlazione.

Osservazione 43 *Il fatto che, come dice la proposizione precedente, $\overline{X}$ ha varianza $\frac{\sigma^2}{n}$, è un'indicazione importante del fatto che $\overline{X}$ approssimi μ. Infatti, intanto dal punto di vista grafico, se ad esempio stiamo parlando di v.a. con densità, la densità di $\overline{X}$ è stretta e alta, cioè concentrata attorno alla media, dovendo avere varianza piccola. Dal punto di vista analitico, varianza $\frac{\sigma^2}{n}$ (più media μ) significa*

$$E\left[\left(\overline{X} - \mu\right)^2\right] = \frac{\sigma^2}{n} \overset{n\to\infty}{\to} 0$$

cioè $\overline{X}$ è concentrato vicino a μ in media quadratica. Questo è anche il calcolo tipico della legge dei grandi numeri.

Uno stimatore può avere alcune buone proprietà. Indichiamo con X una v.a., con $X_1, ..., X_n$ un suo campione, con θ un parametro della legge di X, con $T_n = T_n\left(X_1, ..., X_n\right)$ uno stimatore di θ (esplicitiamo la sua dipendenza da n ed ipotizziamo, dove serve, che si possa prendere il campione e lo stimatore per ogni valore di n).

Definizione 35 *Diciamo che T_n è uno stimatore non distorto (o corretto) di θ se*

$$E[T_n] = \theta.$$

Diciamo poi che è uno stimatore consistente se converge in probabilità a θ:

$$T_n \xrightarrow{P} \theta.$$

Diciamo che è uno stimatore asintoticamente normale se

$$\frac{\sqrt{n}}{\sigma}(T_n - \theta) \xrightarrow{\mathcal{L}} N(0,1)$$

(convergenza in legge) per una opportuna costante positiva σ, che viene detta deviazione standard asintotica (σ^2 sarà detta varianza asintotica).

Le proprietà di correttezza e consistenza sono requisiti naturali per credere che uno stimatore stimi abbastanza bene il parametro corrispondente. La proprietà di gaussianità asintotica è invece utile per scrivere intervalli di confidenza asintotici.

Anche se alcune delle affermazioni seguenti valgono sotto ipotesi minori, per semplicità supponiamo per la loro validità che X abbia momento di ordine 4 finito.

Proposizione 16 *$\overline{X}_n = \frac{X_1+\ldots+X_n}{n}$ è uno stimatore corretto, consistente ed asintoticamente normale di $\mu = E[X]$. Lo stesso vale per $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ rispetto a $\sigma^2 = Var[X]$.*

Proof. Non diamo tutta la dimostrazione ma verifichiamo solo, a titolo di esempio, che S_n^2 è stimatore corretto di σ^2. Vale

$$\begin{aligned}
&E\left[\left(X_i - \overline{X}_n\right)^2\right] \\
&= E\left[(X_i-\mu)^2\right] + E\left[\left(\overline{X}_n - \mu\right)^2\right] - 2E\left[(X_i-\mu)\left(\overline{X}_n - \mu\right)\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{1}{n}\sum_{j=1}^n E\left[(X_i-\mu)(X_j-\mu)\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} = \sigma^2\frac{n-1}{n}
\end{aligned}$$

da cui

$$E\left[S_n^2\right] = \frac{1}{n-1}\sum_{i=1}^n E\left[\left(X_i - \overline{X}_n\right)^2\right] = \frac{1}{n-1}\sum_{i=1}^n \sigma^2\frac{n-1}{n} = \sigma^2.$$

L'analoga verifica per $\overline{X}_n$ è banale; la consistenza dei due stimatori si dimostra con la legge dei grandi numeri e l'asintotica normalità con il teorema limite centrale, ogni tanto però con l'aggiunta di un certo numero di considerazioni specifiche, nel caso di S_n^2, a causa della presenza di $\overline{X}_n$ nel termine $\left(X_i - \overline{X}_n\right)^2$. ■

Esercizio 18 *Mostrare la parte della proposizione precedente che riguarda $\overline{X}_n$.*

Esercizio 19 *Mostrare che valgono le affermazioni della proposizione precedente per*

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

rispetto a σ^2.

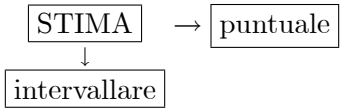
La distorsione si può misurare col numero

$$bias = E[T_n] - \theta.$$

Va detto che in certi problemi può essere utile considerare stimatori distorti, in quanto più semplici o naturali di altri; basta che il bias sia piccolo o meglio che tenda a zero per $n \to \infty$, abbastanza in fretta. Ad esempio, σ^2 si può anche stimare con $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ che è lievemente distorto; uno dei vantaggi è che questa espressione si armonizza meglio con altre nella costruzione di stimatori della covarianza.

2.2 Intervalli di confidenza

Abbiamo detto all'inizio che la *stima dei parametri* è uno dei due problemi principali della statistica di base. La teoria della stima ha due direzioni principali:



La stima puntuale è quella che abbiamo già iniziato a discutere nella sezione precedente, parlando di stimatori. Essi forniscono una stima puntuale dei corrispondenti parametri. Tra le varie cose che ora non affronteremo c'è la ricerca di stimatori tramite il metodo di massima verosimiglianza, tramite il metodo dei momenti, e varie altre cose importanti.

Esaminiamo la stima intervallare. Si tratta di fare affermazioni non solo sul valore T che approssima il parametro θ ma anche sulla bontà di questa approssimazione, sull'errore che si potrebbe commettere.

In analisi numerica, quando si approssima ad es. la soluzione θ di un'equazione con un numero T, si studia l'errore di approssimazione e, se si riesce, si danno risultati del tipo

$$|T - \theta| < \delta$$

(stima dell'errore assoluto) o

$$\left|\frac{T - \theta}{\theta}\right| < \delta$$

(stima dell'errore relativo) dove δ dipenderà da varie cose.

Nei problemi di stima di parametri statistici, è impossibile ottenere esattamente risultati di questo tipo.

Esempio 61 *$\overline{X}_n = \frac{X_1+...+X_n}{n}$ è un buon stimatore di μ. Ma, ad esempio nel caso in cui X sia gaussiana, $\overline{X}_n$ ha una densità positiva su tutto l'asse reale, cioè può assumere (anche se con probabilità piccolissima) valori arbitrariamente grandi (positivi e negativi), quindi arbitrariamente distanti da μ. E' impossibile sperare in un teorema del tipo $\left|\overline{X}_n - \mu\right| < \delta$ (senza ulteriori limitazioni).*

L'esempio precedente però suggerisce la via di uscita: potrebbe valere $\left|\overline{X}_n - \mu\right| < \delta$ *con elevata probabilità*. Questa è la natura dei risultati che possiamo cercare: stime dell'errore corredate di limitazioni sulla loro probabilità di essere valide.

Enunciamo una proposizione sulle gaussiane e vediamone le conseguenze. Ricordiamo che indichiamo con $\Phi(x)$ e q_α la cdf ed il quantile della normale standard, rispettivamente.

Proposizione 17 *Sia X gaussiana, $N\left(\mu, \sigma^2\right)$. Fissato $\delta > 0$, vale*

$$P\left(\left|\overline{X}_n - \mu\right| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1.$$

Viceversa, fissato $\alpha \in (0,1)$, vale

$$P\left(\left|\overline{X}_n - \mu\right| < \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) = 1 - \alpha. \tag{2.1}$$

Proof. Sappiamo che $\overline{X}_n$ è una gaussiana $N\left(\mu, \frac{\sigma^2}{n}\right)$. Allora

$$\begin{aligned} P\left(\mu - \delta < \overline{X}_n < \mu + \delta\right) &= \Phi\left(\frac{(\mu+\delta)-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{(\mu-\delta)-\mu}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(\frac{-\delta\sqrt{n}}{\sigma}\right) \\ &= 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1. \end{aligned}$$

Questo dimostra la prima identità. Fissato $\alpha \in (0,1)$, poniamo

$$2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) - 1 = 1 - \alpha.$$

Si trova

$$\begin{aligned} \Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) &= 1 - \frac{\alpha}{2} \\ \frac{\delta\sqrt{n}}{\sigma} &= q_{1-\frac{\alpha}{2}} \\ \delta &= \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}. \end{aligned}$$

Questo conclude la dimostrazione anche della seconda identità. ■

Possiamo scrivere l'identità (2.1) in due modi più espressivi: fissato $\alpha \in (0,1)$, posto $\delta = \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, vale

$$\mu - \delta < \overline{X}_n < \mu + \delta$$

con probabilità $1-\alpha$; ma anche

$$\overline{X}_n - \delta < \mu < \overline{X}_n + \delta$$

con probabilità $1-\alpha$. Entrambe le scritture sono molto istruttuve. La seconda diventa il nostro esempio fondamentale di *intervallo di confidenza*.

Definizione 36 *Date due v.a.*

$$T_n^- = T_n^-(X_1, ..., X_n) \ \text{ e } T_n^+ = T_n^+(X_1, ..., X_n)$$

diciamo che l'intervallo (aleatorio)

$$\left[T_n^-, T_n^+\right]$$

è un intervallo di confidenza di livello $1-\alpha$ *(a volte si dice livello* α*, ma questo provoca un po' di fraintesi) se*

$$P\left(\theta \in \left[T_n^-, T_n^+\right]\right) \geq 1-\alpha.$$

Il numero $1-\alpha$ *si dice confidenza. A parole diremo che*

$$\theta \in \left[T_n^-, T_n^+\right] \ \textit{con confidenza } 1-\alpha.$$

Corollario 5 *Sia* X *gaussiana,* $N\left(\mu, \sigma^2\right)$*. Allora* $\left[\overline{X}_n - \delta, \overline{X}_n + \delta\right]$ *è intervallo di confidenza di livello* $1-\alpha$ *per* μ*, dove* $\delta = \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$*. Scriveremo anche*

$$\mu = \overline{X}_n \pm \delta \ \textit{ a livello } 1-\alpha.$$

In pratica, ad esempio, si dirà: al 95% vale

$$\mu = \overline{X}_n \pm \delta$$

dove

$$\delta = \frac{\sigma \cdot 1.96}{\sqrt{n}}$$

(essendo $q_{1-\frac{\alpha}{2}} = 1.96$ se $1-\alpha = 95\%$). Oppure: al 90% vale

$$\mu = \overline{X}_n \pm \delta$$

dove

$$\delta = \frac{\sigma \cdot 1.64}{\sqrt{n}}$$

(essendo $q_{1-\frac{\alpha}{2}} = 1.64$ se $1-\alpha = 90\%$).

2.2.1 Esempio

Un'azienda che effettua interventi e riparazioni vuole stimare due grandezze, per poter dimensionare l'organico ed organizzare i turni di lavoro. La prima grandezza è il numero medio μ di ore di lavoro in azienda, giornaliere, necessarie per effettuare tutti i lavori richiesti. La seconda è la probabilità p di dover effettuare interventi esterni. Indichiamo con N il numero di ore di lavoro interne, con X una v.a. di Bernoulli che vale 1 se c'è da effettuare un lavoro esterno (entrambe le variabili sono riferite ad una giornata, generica).

L'azienda si pone le seguenti domande: i) come stimare μ e p? ii) Che errore potremmo aver commesso in tale stima? iii) Quante osservazioni servono per fare tali stime?

Supponendo di avere a che fare con un'azienda di media grandezza, in cui i valori di N siano di varie decine e non di pochissime unità, decidiamo di trattare N come una v.a. continua e per semplicità gaussiana. Invece X è intrinsecamente Bernoulli. Dobbiamo stimare in entrambi i casi il valor medio:

$$\mu = E[N], \qquad p = E[X].$$

La risposta alla domanda (i) in un certo senso è ovvia: si devono effettuare n rilevazioni giornaliere delle due grandezze, chiamiamole

$$N_1, ..., N_n \qquad \text{e} \qquad X_1, ..., X_n$$

(anche numerosità diverse per i due problemi) e poi calcolare gli stimatori

$$\widehat{\mu} = \frac{N_1 + ... + N_n}{n}, \qquad \widehat{p} = \frac{X_1 + ... + X_n}{n}.$$

Detto questo però sorgono tante domande, appunto ad esempio le domande (ii) ed (iii), circa la bontà di queste stime.

Avendo ipotizzato che N è gaussiana, vale

$$\mu = \widehat{\mu} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

con confidenza $1 - \alpha$. Ad esempio, $\mu = \widehat{\mu} \pm \frac{\sigma \cdot 1.96}{\sqrt{n}}$ al 95%. Questo significa che, al 95%, il massimo errore possibile è $\frac{\sigma \cdot 1.96}{\sqrt{n}}$. (In particolare, non c'è un errore massimo possibile certo, ma sempre a meno di una piccola probabilità; c'è sempre una piccola probabilità che l'errore sia ancora più grosso). Questo non significa che l'errore sarà pari a $\frac{\sigma \cdot 1.96}{\sqrt{n}}$, al 95%: *al massimo* sarà $\frac{\sigma \cdot 1.96}{\sqrt{n}}$. Ma se riduciamo la confidenza, esso è minore:

$$\begin{array}{cccc} \text{al } 90\%: & \frac{\sigma \cdot 1.64}{\sqrt{n}} & \text{al } 70\%: & \frac{\sigma \cdot 1.04}{\sqrt{n}} \\ \text{all' } 80\%: & \frac{\sigma \cdot 1.28}{\sqrt{n}} & \text{al } 60\%: & \frac{\sigma \cdot 0.84}{\sqrt{n}} \end{array}$$

e così via. L'idea si vede bene graficamente tracciando la densità gaussiana di $\frac{N_1 + ... + N_n}{n}$ ed osservando come varia l'intervallo attorno a μ quando si varia l'area soprastante. Quindi è molto probabile che l'errore sia molto più piccolo di $\frac{\sigma \cdot 1.96}{\sqrt{n}}$, ad esempio sia la metà. Il numero $\frac{\sigma \cdot 1.96}{\sqrt{n}}$ fornisce l'ordine di grandezza.

Veniamo ora all'aspetto pratico: supponiamo di aver fatto $n = 25$ osservazioni ed aver trovato $\widehat{\mu} = 62.8$. Che possiamo dire, ad esempio al 95%? Che

$$\mu = 62.8 \pm \frac{\sigma \cdot 1.96}{5} = 62.8 \pm 0.39 \cdot \sigma.$$

Ma quanto vale σ? Nessuno ce lo può dire. La cosa più naturale, avendo a disposizione il campione di numerosità $n = 25$, è calcolare S. Supponiamo di farlo ed aver trovato $S = 18.3$. Allora, approssimativamente (S non è σ), possiamo affermare che al 95%

$$\mu = 62.8 \pm 0.39 \cdot 18.3 = 62.8 \pm 7.\,14.$$

In altre parole, al 95%, il valore incognito μ è compreso tra $55.\,66$ e $69.\,94$. Ma, come detto sopra, molto probabilmente è abbastanza più vicino a 62.8. Ad esempio, al 60%, vale circa

$$\mu = 62.8 \pm 3.\,5$$

cioè μ è compreso tra $59.\,3$ e $66.\,3$.

la sostituzione di σ con S ha introdotto un'approssimazione. Un teorema dice che i risultato (cioè l'ampiezza dell'intervallo di confidenza) torna ad essere un risultato esatto se, oltre a sostituire σ con S, si sostituisce il quantile gaussiano standard $q_{1-\frac{\alpha}{2}}$ con il quantile della t di Student a $n-1$ gradi di libertà:

$$\mu = \widehat{\mu} \pm \frac{\sigma t^{(n-1)}_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Nel nostro esempio, usando le tavole, vale $t^{(24)}_{1-\frac{0.05}{2}} = 2.064$ e quindi

$$\mu = 62.8 \pm \frac{18.3 \cdot 2.064}{5} = 62.8 \pm 7.\,55.$$

Il risultato è un po' peggiore di quello approssimato precedente, ma è sicuro. La differenza non è però marcatissima.

La domanda (ii) ha però una variante fondamentale: che si parli di errore relativo invece che assoluto. L'errore assoluto è $|\widehat{\mu} - \mu|$ mentre l'errore relativo è

$$\left| \frac{\widehat{\mu} - \mu}{\mu} \right|.$$

Allora l'errore relativo massimo possibile con confidenza $1 - \alpha$ è

$$\left| \frac{\widehat{\mu} - \mu}{\mu} \right| = \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}\,|\mu|}.$$

Nel nostro esempio, al 95%, usando ad esempio per semplicità i quantili guassiani

$$\left| \frac{\widehat{\mu} - \mu}{\mu} \right| = \frac{18.3 \cdot 1.96}{5 \cdot |\mu|} = \frac{7.\,17}{|\mu|}.$$

naturalmente nessuno ci può dare μ, visto che è la quantità da stimare. Quindi approssimativamente sostiuiamola con $\widehat{\mu}$ che è nota:

$$\left|\frac{\widehat{\mu}-\mu}{\mu}\right| \approx \frac{7.17}{62.8} = 0.114.$$

In sostanza, si commette un errore relativo di un decimo (decente per scopi di commercio non troppo spinti). Ovviamente se si vuole usare la t di Student, viene lievemente più grande (provare).

Sempre relativamente a N, veniamo alla domanda (iii). Il numero di osservazioni da fare non può essere una grandezza assoluta, indipendente da requisiti. Dipende dalla precisione che vogliamo ottenere e dalla confidenza che scegliamo (il rischio che accettiamo di correre). La domanda (iii). Essa è un esempio di DOE (*Design Of Experiments*).

Il numero di osservazioni da fare non può essere una grandezza assoluta, indipendente da requisiti. Dipende dalla precisione che vogliamo ottenere e dalla confidenza che scegliamo (il rischio che accettiamo di correre; rischio di fare una dichiarazione falsa circa l'intervallo in cui cade la media). Supponiamo di correre un rischio del 5%, prendere cioè confidenza 95% e supponiamo di volere un errore (massimo) pari a 5, erroe assoluto. Uguagliando l'errore massimo a 5 abbiamo $\frac{\sigma\cdot 1.96}{\sqrt{n}} = 5$, ovvero

$$n = \left(\frac{\sigma\cdot 1.96}{5}\right)^2 = 0.154\cdot\sigma^2.$$

Con l'uguaglianza si intende in realtà il primo intero $n \geq 0.154\cdot\sigma^2$ (infatti per essere più precisi andrebbe impostata dall'inizio la disuguaglianza $\frac{\sigma\cdot 1.96}{\sqrt{n}} \leq 5$). Resta il grave problema di conoscere σ: se non abbiamo ancora fatto rilevazioni, se non abbiamo dati, σ è incognita. Non ci sono scappatoie generali: o si conosce un valore approssimato di σ sulla base di dati precedenti, oppure si deve ipotizzare l'ordine di grandezza di σ, approssimando ovviamente per eccesso. Senza σ non si può stabilire n in anticipo. Se non si hanno dati precedenti o capacità di stima dell'ordine di grandezza, bisogna iniziare i campionamenti, raccogliere un po' di dati e con essi stimare σ. Questi primi dati concorreranno comunque alla stima finale di μ. Supponiamo di aver raccolto una decina di dati preliminari, dai quali esca la stima

$$S = 20.4$$

Allora troviamo

$$n = 0.154\cdot 20.4^2 = 64.089.$$

Servono circa 65 osservazioni. In realtà, dopo un po' di ulteriori osservazioni conviene ristimare σ per rendere più accurata la previsione del numero di osservazioni da fare.

Se volevamo invece l'errore relativo (massimo) assegnato, es. 10%, dovevamo imporre

$$\frac{\sigma\cdot 1.96}{\sqrt{n}\,|\mu|} = 0.1$$

ovvero

$$n = \left(\frac{\sigma\cdot 1.96}{0.1\cdot|\mu|}\right)^2 = 384.16\cdot\left(\frac{\sigma}{|\mu|}\right)^2.$$

Qui servono addirittura una stima preliminare di σ e μ. Si agisce come sopra. Supponiamo che dopo alcune osservazioni preliminari abbiamo trovato $\overline{x} = 60.5$, $S = 20.4$. Allora

$$n = 384.16 \cdot \left(\frac{20.4}{60.5}\right)^2 = 43.678.$$

Questi esempi numerici mostrano la ragionevolezza dei risultati che si ottengono con questa teoria.

Si noti comunque che questi calcoli producono valori piuttosto alti di n. In certe applicazioni pratiche, molte decine di osservazioni sono davvero costose. C'è un rimedio? Ricordiamo quanto appreso sopra circa l'intervallo di confidenza: esso esprime il risultato più pessimistico. Con buona probabilità, l'intervallo al 95% è pessimistico, la stima è molto migliore, come evidenzia l'intervallo al 60%, ad esempio.

Se accettassimo un rischio molto alto, 40%, i calcoli precedenti darebbero:

$$n^{60\%}_{assoluto} = \left(\frac{\sigma \cdot 0.84}{5}\right)^2 = 0.028 \cdot \sigma^2 \overset{S=20.4}{=} 0.028 \cdot 20.4^2 = 11.652.$$

Naturalmente non possiamo esporci ad un tale rischio, ma questo calcolo ci dice che il 60% delle volte accadrebbe che 12 osservazioni sono sufficienti, invece che 65. Similmente, accettando un rischio del 20%,

$$n^{80\%}_{assoluto} = \left(\frac{\sigma \cdot 1.28}{5}\right)^2 = 0.065 \cdot \sigma^2 \overset{S=20.4}{=} 0.065 \cdot 20.4^2 = 27.05.$$

Insomma, con elevata probabilità, bastano molte meno osservazioni. Che fare? Ovviamente si può decidere di fare poche osservazioni (es. solo 20-30) e sperare che le cose siano andate bene. Si può però tracciare un grafico della stima della media $\widehat{\mu}$ al crescere del numero di prove. Nel senso: dopo aver eseguito n osservazioni, calcoliamo $\widehat{\mu}_n$ ed aggiungiamolo al grafico precedentemente fatto dei valori di $\widehat{\mu}$ in funzione del numero di prove. Al crescere di n questo grafico tenderà ad assestarsi attorno all'asintoto orizzontale μ (asintoto però sconosciuto!). Quando vediamo il grafico diventare sufficientemente orizzontale, abbiamo un forte sintomo che siamo già "arrivati a convergenza", come si suol dire. Non c'è la certezza assoluta, ma è molto difficile che un tale grafico si assesti e poi riprenda a muoversi. Bene, nel 60% dei casi, si assesta molto presto, nell'80%, poco oltre, e così via. Solo in rari casi necessita davvero di valori di n intorno a 65 per assestarsi; è solo il caso più pessimistico, che però è garantito al 95%. A priori, non possiamo sapere se ci capiterà questo caso o quelli più fortunati. Bisogna eseguire le prove sequenzialmente e sperare. Quanto qui espresso è una versione pratica della cosidetta *Sequential Analysis*.

Ripetiamo ora alcuni dei passi precedenti per il problema della stima della proporzione p, altro problema classico e ricorrente. Lo stimatore è $\widehat{p}$, ma ora non vale più la teoria gaussiana dell'intervallo di confidenza. Tuttavia, in modo approssimato essa è ancora vera: vale

$$p = \widehat{p} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \qquad \sigma^2 = Var[X] = p(1-p)$$

con confidenza approssimativamente pari $1-\alpha$. Ciò che è approssimata è la probabilità che p stia nell'intervallo suddetto, non l'intervallo in sé. Tutto deriva dal teorema limite centrale, in quanto

$$P\left(|\widehat{p}-p| \leq \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) = P\left(\left|\frac{X_1+...+X_n}{n}-p\right| \leq \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$
$$= P\left(\left|\frac{X_1+...+X_n-np}{\sqrt{n}\sigma}\right| \leq q_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

dove l'ultima approssimazione è fornita appunto dal TLC.

Facciamo un esempio pratico: supponiamo di aver fatto $n = 25$ osservazioni ed aver trovato $\widehat{p} = 0.21$. Che possiamo dire, ad esempio al 95%? Che con probabilità circa uguale a questa, vale

$$p = 0.21 \pm 0.39 \cdot \sigma$$

un po' come nel caso gaussiano. Resta il problema di conoscere σ.

Qui però c'è un elemento in più, molto particolare: $\sigma^2 = p\,(1-p)$. Il parametro σ è legato alla quantità p che stiamo cercando di stimare. Una prima conclusione quindi è che valga, approssimativamente

$$p = \widehat{p} \pm \frac{\sqrt{\widehat{p}\,(1-\widehat{p})}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Nel nostro esempio,

$$p = 0.21 \pm 0.39 \cdot \sqrt{0.21 \cdot (1-0.21)} = 0.21 \pm 0.16.$$

Vale cioè

$$0.05 \leq p \leq 0.37.$$

Non è un risultato eccellente, in senso relativo. Naturalmente, è abbastanza probabile che l'intervallo sia più piccolo, come abbiamo visto nel caso gaussiano: ad esempio, all'80% vale

$$p = 0.21 \pm \frac{1.28}{5} \cdot \sqrt{0.21 \cdot (1-0.21)} = 0.21 \pm 0.104.$$

cioè p è compreso tra 0.1 e 0.3. Parlando a braccio, la frequenza con cui il negozio deve mandare operatori fuori sede si aggira tra 1/10 e 3/10. Se questa vaghezza di informazione è sufficiente, basta così, altrimenti bisogna campionare di più.

L'errore relativo in astratto è

$$\left|\frac{\widehat{p}-p}{p}\right| \leq \frac{\sqrt{p\,(1-p)}q_{1-\frac{\alpha}{2}}}{p\sqrt{n}} = \frac{\sqrt{\frac{1-p}{p}}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

ed approssimando le espressioni sulla destra diventa

$$\left|\frac{\widehat{p}-p}{p}\right| \leq \frac{\sqrt{\frac{1-\widehat{p}}{\widehat{p}}}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

in questo esempio è (approssimativamente al 95%)

$$\left|\frac{\widehat{p}-p}{p}\right| \leq \frac{\sqrt{\frac{1-0.21}{0.21}}1.96}{5} = 0.76.$$

Per certe applicazioni è davvero troppo grosso, per altre può anche essere accettabile.

Si deve notare che è venuto così grosso perché $\widehat{p}$ è piccolo: se si stima una proporzione piccola, la tendenza è di commettere un errore relativo grosso. Se invece $\widehat{p}$ fosse stato grande, $\sqrt{\frac{1-\widehat{p}}{\widehat{p}}}$ era piccolo e avrebbe contribuito a diminuire l'errore relativo.

Spesso nelle applicazioni si cerca di stimare una proporzione piccola al solo scopo di sapere che è piccola, non di conoscerne con precisione il valore. Sapere che è 0.05 o 0.1 o 0.15 non cambia le nostre azioni successive, anche se questi numeri differiscono di tantissimo in senso relativo. Differiscono poco in senso assoluto. Allora, in problemi di questo genere, basta chiedere che l'errore assoluto sia piccolo. L'errore relativo non serve. In sintesi, in problemi in cui basta scoprire che p è piccolo basta desiderare che l'errore assoluto sia piccolo; e quindi i difetti suddetti dell'errore relativo per $\widehat{p}$ piccolo diventano inessenziali.

In quest'ottica, immaginiamo di voler stimare p con precisione assoluta 0.1 (se $\widehat{p}$ è piccolo, ci basta, p non supererà $\widehat{p}+0.1$; se $\widehat{p}$ è grande, un errore assoluto di 0.1 non è così grave). Dobbiamo imporre

$$\frac{\sqrt{p\left(1-p\right)}\cdot 1.96}{\sqrt{n}} = 0.1$$

ovvero

$$n = \left(\frac{1.96}{0.1}\right)^2 p\left(1-p\right).$$

Serve una stima di p, che in fase di DOE può provenire da campionamenti precedenti, da primi piccoli campionamenti, da ipotesi. Ma in questo caso vale anche la seguente stima universale: siccome l'espressione $p\left(1-p\right)$ può al massimo valere $\frac{1}{4}$, Alla peggio dovremo prendere

$$n = \left(\frac{1.96}{0.1}\right)^2 \frac{1}{4} = 96.04.$$

Ovviamente non è un valore molto incoraggiante, però è universale. E' chiaro che all'80% basta

$$n = \left(\frac{1.28}{0.1}\right)^2 \frac{1}{4} = 40.96$$

ed al 60% addirittura

$$n = \left(\frac{0.84}{0.1}\right)^2 \frac{1}{4} = 17.64.$$

Quindi, eseguendo le cose sequenzialmente e sperando di non essere troppo sfortunati, dovrebbe bastare un numero contenuto di osservazioni.

2.2.2 Soglie, ammissibili ecc.

Citiamo un'applicazione frequentissima dei modelli probabilistici e della teoria degli intervalli di confidenza: il calcolo di soglie, ammissibili, scorte di sicurezza, valori minimi o massimi a meno di probabilità prefissate, e loro correzione in caso di incertezza sui parametri. Mostriamo un esempio relativo al problema degli ammissibili di progetto in un problema di resistenza di strutture; ma il ragionamento sarebbe identico per la soglia di traffico telefonico oltre la quale una stazione smette di servire tutte le comunicazioni, per il valore della scorta di sicurezza che serve a soddisfare tutta la clientela, e per tanti altri problemi della stessa natura matematica, in cui si cerca un valore minimo o massimo di una grandezza.

La prima caratteristica cruciale di questi problemi è che la grandezza in questione non ha minimo o massimo, o se anche li l'ha sono valori irraggiungibili in pratica (es. una quantità di scorte esagerata, la necessità di una potenza esagerata della stazione telefonica ecc.). Allora si deve accettare una piccola probabilità α di errore, mal funzionamento, esposizione a pericolo ecc. e, relativamente ad α fissato si deve trovare quella soglia λ che viene superata (nel senso negativo per il problema in questione) solamente con probabilità α. Non si tratta quindi di trovare il minimo assoluto o il massimo assoluto della grandezza (improponibili per gli scopi pratici o addirittura infiniti), ma di trovare minimo o massimo a meno di una certa probabilità, un certo rischio.

Il numero λ che cerchiamo è un quantile della distribuzione di probabilità della grandezza in questione. Bisogna quindi saper calcolare quantili. Per le gaussiane ci sono formule generali che riportano i quantili di una gaussiana generica a quelli della normale standard. Per altre distribuzioni si può ad esempio usare il software. In R si calcolano i quantili coi comandi del tipo `qnorm`, `qweibull` ecc.

La seconda caratteristica cruciale di questi problemi è che di solito la distribuzione di probabilità della grandezza in questione non è nota con precisione ma, pur supponendo noto il tipo di distribuzione (gaussiana, Weibull ecc.), c'è incertezza sui parametri. Bisogna allora propagare questa incertezza sul calcolo della soglia. Essa sarà quindi una soglia definita a meno di due valori di probabilità α ed α': α è la probabilità che la soglia venga superata (a causa della variabilità della grandezza), α' è la probabilità che i parametri utilizzati per il calcolo (più pessimistico possibile) della soglia siano sbagliati, quindi la soglia sia semplicemente sbagliata. Vediamo un esempio.

Sappiamo che una struttura cede oltre una certa soglia di carico, che però è un po' aleatoria a causa delle imperfezioni (sconosciute) del materiale e della costruzione. Relativamente ad una generica struttura di quel tipo, sia S la soglia di rottura, che stiamo trattando come una variabile aleatoria.

Problema: determinare l'ammissibile al 99%: quel valore S^* tale che $P(S > S^*) = 0.99$. Questo significa che, mediamente, solo una struttura su 100 ha la propria soglia di rottura inferiore ad S^* e quindi, sottoponendo la struttura ad un carico uguale (o inferiore) a S^*, mediamente 99 strutture su 100 resistono.

Si tratta di calcolare un quantile della distribuzione di S: S^* è dato dalla relazione $P(S < S^*) = 0.01$ e quindi, per definizione di quantile q_α^S della distribuzione di S, vale

$$S^* = q_{0.01}^S.$$

Il problema non contiene alcuna difficoltà se si conosce bene la distribuzione di S. Se ad

esempio è una $N(10, 0.64)$, vale

$$S^* = 8.14$$

in quanto

$$q^S_{0.01} = 10 - 0.8 \cdot q_{0.99} = 10 - 0.8 \cdot 2.326 = 8.14$$

dove q_α è il quantile della normale standard. Se S ha una distribuzione più complessa della gaussiana ma comunque di classe abbastanza nota, i valori dei quantili sono reperibili nei software o in tavole o tramite integrazione numerica della densità. Infine, se non disponiamo di questi strumenti ma solo della possibilità di generare numeri aleatori con distribuzione S, possiamo trovare i quantili con un metodo tipo Monte Carlo.

Che dire però se di S si conosce solo un campione e non la distribuzione precisa? Se la numerosità del campione fosse molto elevata, potremmo calcolare un'approssimazione del quantile con una semplice proporzione sui dati. Questo caso però è molto raro nelle applicazioni.

Ipotizzando un tipo di distribuzione per S, possiamo stimare i parametri tramite il campione. Così però si commette un errore, che può influire parecchio sul risultato: se ad esempio i valori stimati dessero una $N(10, 0.64)$, ma con un errore del 10% su media e deviazione standard, in realtà la vera distribuzione potrebbe essere una $N(9, 0.88^2)$, quindi sarebbe

$$S^* = 9 - 0.88 \cdot 2.326 = 6.95.$$

La differenza è notevole: non possiamo trascurarla nel dichiarare l'ammissibile, visti i rischi in gioco. La teoria degli intervalli di confidenza sviluppata sopra permette allora di risolvere questo problema.

Supponiamo che la varianza sia nota, per semplificare un po', mentre che la media sia stata stimata con un campione di numerosità 20. Supponiamo che la σ nota valga 0.8 mentre la stima $\overline{x}$ della μ vera abbia dato il valore 10. Sappiamo allora che

$$\mu = 10 \pm \delta \text{ al } 95\%$$

dove

$$\delta = \frac{0.8 \cdot 1.96}{\sqrt{20}} = 0.35.$$

Questo significa che, al 95%, non possiamo escludere un valore pessimistico della soglia media di rottura pari a $10 - 0.35 = 9.65$. Se questo fosse il valore vero della media, la soglia sarebbe

$$S^* = 9.65 - 0.8 \cdot 2.326 = 7.789.$$

L'affermazione finale allora è: il 99% degli esemplari di quella struttura ha una soglia di rottura maggiore di 7.789:

$$S^* \geq 7.789$$

e questo valore ha un grado di affidabilità (grado di fiducia) del 95%.

Volendo esplicitare in senso frequenziale l'interpretazione di questo "grado di affidabilità" potremmo dire che se i parametri della distribuzione statistica, qui μ, venissero stimati 100

volte tramite campioni di numerosità 20, nel 95% dei casi il loro valore di $\overline{x}$ disterebbe dal valore vero μ meno di 0.35, quindi la disuguaglianza

$$\mu \geq \overline{x} - 0.35$$

su cui abbiamo basato il calcolo di S^* sarebbe vera in 95 su 100 dei casi. Noi però stimiamo una volta sola μ, quindi non c'è una realtà frequenziale dietro questo ragionamento (come invece c'è nel dire che 99 su 100 delle strutture ha soglia $>S^*$). Quindi il ragionamento frequenziale appena fatto deve tradursi in una dichiarazione di *fiducia* nel risultato. Ci fidiamo al 95% che sia $\mu \geq 10 - 0.35$ e quindi attribuiamo la stessa fiducia al risultato $S^* \geq 7.789$.

2.3 Test statistici

2.3.1 Un esempio prima della teoria

Una compagnia ferroviaria dichiara che il servizio lungo una certa tratta critica è stato migliorato ed il ritardo medio è ora di $\mu_0 = 5$ Min.

Per 10 giorni misuriamo i ritardi ed osserviamo i valori:

$$5, 7, 4, 10, 6, 5, 8, 2, 8, 6.$$

La compagnia ha ragione?

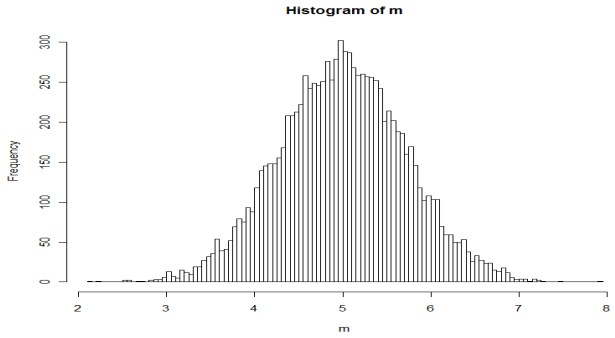
La media empirica è $\overline{x} = 6.1$.

Naturalmente è diversa da 5 (è impossibile osservare $\overline{x} = \mu_0$). Ci sono sempre fluttuazioni casuali nei dati reali. La domanda allora è: la media empirica è troppo diversa da quella teorica oppure può essere considerata una fluttuazione casuale?

Per semplicità, supponiamo che la distribuzione statistica dei ritardi sia gaussiana. Come varianza di questa distribuzione prendiamo pragmaticamente il valore che proviene dai dati sperimantali (non abbiamo altro): $sd = 2.28$.

Generiamo 10000 campioni di numerosità 10 da una $N\left(5, 2.28^2\right)$ e vediamo quanto naturale o viceversa estremo è il valore $\overline{x} = 6.1$ trovato per i nostri dati.

```
N<-10000; m<-1:N
for (i in 1:N) {m[i]<-mean(rnorm(10,5,2.28))}
hist(m,100)
```



Il valore $\overline{x} = 6.1$ è abbastanza estremo. Potremmo calcolare la *probabilità che un valore di* $\overline{x}$ *sia più estremo di* 6.1. Questo numero verrà chiamato *p-value*, o *valore p*. Risulta (vedremo tra un momento come)

$$p\text{-value} = 0.064.$$

E' abbastanza piccolo. Tuttavia, è maggiore di 0.05, una delle soglie usuali in statistica per giudicare la piccolezza di una probabilità. Viene quindi demandato a noi decidere se il campione è naturale o no, se 6.1 è un valore naturale o no. Demandato a noi, ma con l'ausilio della conoscenza del p-value$= 0.065$.

Le componenti di questo esempio sono:

- un campione
- un'ipotesi (es. $\mu_0 = 5$)
- un riassunto $\overline{x}$ del campione, chiamata *statistica del test* (*test statistic*) utile per eseguire un confronto tra campione e ipotesi
- la distribuzione del test statistic
- il p-value, cioè la probabilità che il test statistic sia più estremo del valore osservato.

Da un punto di vista pratico ed operativo potremmo dire che questa è la sostanza di tutti i test statistici: si vuole capire la compatibilità o meno di un campione sperimentale rispetto ad un'ipotesi, e la si valuta calcolando una grandezza statistica (la statistica del test) che, se c'è compatibilità dovrebbe cadere in un range "normale", mentre se non c'è compatibilità (e quindi l'ipotesi va rifiutata) essa cade in una regione un po' estrema; infine, il grado di anomalia della grandezza statistica rispetto alla condizione normale viene valutato tramite il calcolo del p-value.

In resto della sezione sviluppa alcuni elementi teorici e concettuali in modo più organico, ma la sostanza è quella già esposta.

2.3.2 Calcolo analitico del p-value nel precedente test per la media

Sappiamo che la media aritmetica $\overline{X}$ di un campione gaussiano $N\left(\mu_0, \sigma^2\right)$ di numerosità n ha distribuzione $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. Il p-value relativo ad un certo valore sperimentale $\overline{x}$ è definito da

$$p = P\left(\overline{X} > \overline{x}\right)$$

quindi vale

$$p = 1 - P\left(\overline{X} \leq \overline{x}\right) = 1 - \Phi\left(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

usando le formule che trasformano la cdf di una gaussiana qualsiasi in quella standard.

Nel nostro esempio allora $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{6.1-5}{2.28/\sqrt{10}} = 1.525\,7$, $\Phi\left(1.5257\right)$ (calcolabile in R con `pnorm(1.5257)`) vale 0.936, quindi

$$p = 1 - 0.936 = 0.064.$$

Il p-value appena calcolato è il cosidetto p-value unilaterale. Si potrebbe anche calcolare il p-value bilaterale, cioè quello in cui la frase "valori più estremi di quello sperimentale" (che compare nella definizione di p-value) viene intesa bilateralmente rispetto alla media, o coi valori assoluti, per così dire. Secondo questa accezione dobbiamo calcolare

$$p = P\left(\left|\overline{X} - \mu_0\right| > \left|\overline{x} - \mu_0\right|\right).$$

Quindi, standardizzando, vale

$$p = P\left(\left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) = P\left(|Z| > |z|\right)$$

dove Z è una v.a. $N(0,1)$ e z è il numero sperimentale $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$. Quindi (come si vede facilmente tracciando il grafico di una $N(0,1)$ e raffigurando le aree delle due code che dobbiamo calcolare)

$$p = 2 - 2\Phi\left(|z|\right)$$

dove Φ è la cdf normale standard. Come potevamo intuire sin da subito da un disegno della densità della v.a. $\overline{X}$, questo p-value è il doppio di quello unilaterale. Se per distinguerli indichiamo quello unilaterale con p^U e quello bilaterale con p^B, vale

$$p^B = 2p^U.$$

Nel nostro esempio quindi $p^B = 0.128$.

2.3.3 Ipotesi nulla

Partiamo da un campione. Su di esso si fa un'ipotesi, del tipo: proviene da una distribuzione di media 5, proviene da una Weibull con certi parametri, e così via. Scopo del test: rigettare questa ipotesi.

Il primo elemento di un test è quindi l'ipotesi, che verrà detta *ipotesi nulla*, indicata con $\mathcal{H}_0$.

Al termine del test, o avremo rifiutato l'ipotesi, oppure non l'avremo rifiutata (che non equivale a dire che l'abbiamo confermata, ma solo che non abbiamo trovato nessuna contraddizione tra il campione sperimentale e l'ipotesi).

Esempio 62 *Esempio di $\mathcal{H}_0$: il ritardo medio μ è maggiore di 5.*

Avendo introdotto il simbolo $\mathcal{H}_0$ per l'ipotesi nulla, riscriviamo la definizione di valore p in modo più enfatico:

$$p = P_{\mathcal{H}_0}\left(\overline{X} > \overline{x}\right).$$

Ipotesi alternativa. Differenza rispetto alla teoria delle decisioni

La teoria rigorosa dei test statistici richiede anche il concetto di *ipotesi alternativa* $\mathcal{H}_1$. Siccome non enunciamo e dimostriamo teoremi sui test, il suo ruolo sarà abbastanza nascosto.

Esempio 63 *Esempio di $\mathcal{H}_1$: il ritardo medio è $\mu > 5$.*

Esempio 64 *Altro esempio di $\mathcal{H}_1$: il ritardo medio è $\mu \neq 5$.*

Lo schema matematico formato dalle due ipotesi complementari (rispetto ad un certo universo di possibilità) $\mathcal{H}_0$ e $\mathcal{H}_1$ appare simile a quello della *teoria delle decisioni*: sulla base di alcune osservazioni sperimentali, dobbiamo decidere se vale $\mathcal{H}_0$ oppure $\mathcal{H}_1$.

Tuttavia, nella teoria delle decisioni le due ipotesi vengono considerate allo stesso livello, in modo simmetrico, e la decisione di conclude con una scelta tra le due: o vale l'una o vale l'altra.

Invece, nella teoria dei test statistici, il ruolo di $\mathcal{H}_0$ ed $\mathcal{H}_1$ è asimmetrico. $\mathcal{H}_0$ può solo essere rifiutata o non rifiutata, non possiamo arrivare ad una conclusione del tipo: "$\mathcal{H}_0$ è vera".

Per capire meglio in che senso c'è simmetria nella teoria delle decisioni, ricordiamone alcuni elementi nel caso della *teoria bayesiana*. Si ha un universo Ω, una partizione (B_k) e si deve prendere una decisione circa quale degli eventi B_k sia vero.

Supponiamo che gli eventi B_k influenzino qualcosa che possiamo osservare, diciamo l'evento A. Si pensi per esempio che B_i siano le possibili cause, A e A^c le possili conseguenze.

La regola di decisione bayesiana è semplicemente: si sceglie la causa più probabile, condizionata al fatto che A si è avverato.

Per la formula di Bayes

$$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum_k P(A|B_k)\,P(B_k)}.$$

Il demoninatore è uguale per tutti, quindi basta massimizzare il numeratore:

$$B_i^{opt} := \arg\max_{B_i} P(A|B_i)\,P(B_i)\,.$$

Se decidiamo che a priori le diverse possibilità B_i sono equiprobabili (assenza di pregiudizi) troviamo semplicemente

$$B_i^{opt} := \arg\max_{B_i} P(A|B_i)\,.$$

Le probabilità $P(A|B_i)$ sono simili a dei p-values, se si prende come evento A l'evento "la test statistic assume valori più estremi di quello osservato", e come B_i le due diverse ipotesi del test. Allora

$$p\text{-value} = P(A|\mathcal{H}_0)\,.$$

Ma mentre in teoria delle decisioni calcoleremmo anche $P(A|\mathcal{H}_1)$ e sceglieremmo tra le due alternative sulla base del valore più grande delle due probabilità, nella teoria dei test calcoliamo solo $P(A|\mathcal{H}_0)$ e rifiutiamo l'ipotesi $\mathcal{H}_0$ se questa probabilità (il p-value) è molto piccola, in genere più piccola di 0.05. Se non è piccola, non confrontiamo con 0.5 come si farebbe in teoria delle decisioni; concludiamo semplicemente che non c'è evidenza per rifiutare $\mathcal{H}_0$.

Precisazioni sulla statistica del test

Un test è un algoritmo. L'input è il campione sperimentale e l'ipotesi $\mathcal{H}_0$. l'output è il valore della statistica del test, o un passo oltre il p-value. Indichiamo genericamente con z il valore della statistica del test (era $\overline{x}$ nell'esempio).

Esempio 65 *Un politico afferma che il 65% della popolazione è con lui, preferisce cioè l'alternativa A alla B. Sospettiamo che abbia torto. Chiediamo allora a 100 persone ed osserviamo che solo 47 preferiscono A a B. Dobbiamo confrontare l'ipotesi nulla* $\mathcal{H}_0$ *="il 65% preferisce A a B" col campione. Abbiamo bisogno di un algoritmo che, presi i numeri 65, 47, 100 restituisca un risultato, la statistica del test, che indichiamo con z. Un esempio banale potrebbe essere l'errore relativo*

$$z = \left| \frac{65 - 47}{65} \right|$$

che però non tiene conto della numerosità del campione (è certo diverso chiedere a 10, a 100 o a 1000 persone).

Possiamo pensare che z sia aleatoria (si pensi a ripetere il campionamento), per cui sarebbe meglio usare la notazione Z. La v.a. Z è più propriamente chiamata statistica del test, ed ha una sua distribuzione di probabilità. Supponiamola descritta da una densità $f(z)$.

Più precisamente, *se* $\mathcal{H}_0$ *vale*, allora Z ha densità $f_{\mathcal{H}_0}(z)$. Se invece vale una certa ipotesi alternativa $\mathcal{H}_1'$, Z avrà un'altra densità $f_{\mathcal{H}_1'}(z)$. Queste frasi spiegano l'idea ma non sono molto precise, o almeno non universali. Dipende se le ipotesi sono così precise da identificare una densità oppure sono vaghe. Un'ipotesi del tipo: la v.a. è gaussiana di media 3 e varianza 24 è precisa; mentre un'ipotesi del tipo: la v.a. è gaussiana di media $\mu \neq 3$ e varianza 24 non identifica una densità. In questo caso, se abbiamo bisogno di una densità anche per l'ipotesi alternativa, questa va frammentata in sottoipotesi precise.

Introdotti questi concetti preliminari, il problema è: come si conclude, sulla base del valore sperimentale z, se $\mathcal{H}_0$ è falsa?

Basta calcolare la probabilità che Z assuma valori più estremi di z, probabilità secondo la densità $f_{\mathcal{H}_0}(z)$ (il p-value).

C'è però un'alternativa: prescrivere a priori un valore piccolo di probabilità α, es. 5%, che identifica una coda (o due code) con tale probabilità; fatto questo, si va a vedere se z cade nella coda o no. Se cade nella coda, si rifiuta $\mathcal{H}_0$.

Se z cade nella coda, significa che sperimentalmente abbiamo osservato un accadimento che aveva probabilità molto piccola di accadere, secondo $f_{\mathcal{H}_0}(z)$. Siccome questo era improbabile (relativamente al valore piccolo di probabilità α prefissato), riteniamo che il campione che ha prodotto quel valore di z non potesse provenire dalla v.a. di densità $f_{\mathcal{H}_0}(z)$.

Possiamo quindi o calcolare la probabilità della coda identificata da z oppure decidere la coda a priori e vedere se z ci cade dentro.

2.3.4 Errori di prima e seconda specie; significatività e potenza di un test

Supponiamo di aver sviluppato un test per valutare la validità di un'ipotesi. Il test non è infallibile. Rifiutiamo l'ipotesi $\mathcal{H}_0$ se z cade nelle code; ma, quando l'ipotesi $\mathcal{H}_0$ è valida, z

può cadere nelle code, solo che ciò è molto improbabile. C'è quindi una piccola probabilità di rifiutare $\mathcal{H}_0$ quando invece è valida (ciò avviene quando il campione che produce z, pur essendo in perfetto accordo con l'ipotesi, è un campione un po' anomalo, cosa rara ma possibile per puro caso).

Il primo errore possibile, quindi, nella teoria dei test è la possibilità di rifiutare $\mathcal{H}_0$ quando invece è vera. Viene detto *errore di prima specie*. La sua probabilità è α:

$$\begin{aligned}\alpha &= P\left(\text{errore di prima specie}\right)\\ &= P\left(\text{rifiutare } \mathcal{H}_0 \text{ quando è vera}\right).\end{aligned}$$

Il numero α è anche chiamato *significatività* del test. Il suo valore viene di solito paragonato a 0.05: se lo si prefissa, lo si prende pari o minore a 0.05; se lo si calcola a posteriori dev'essere ≤ 0.05. (A volte è più spontaneo dire che la significatività è 95%, invece che 5%, che pur essendo la frase canonica, è però un po' opposta al senso comune).

Osservazione 44 *Nella teoria delle decisioni si calcolano invece due probabilità di errore simmetriche tra loro.*

Esiste poi un secondo errore che si può commettere facendo un test. Può accadere che sia valida l'ipotesi alternativa $\mathcal{H}_1$ ma il test non se ne accorge, non trova nulla di contraddittorio nel campione rispetto ad $\mathcal{H}_0$. In pratica questo è possibilissimo: si pensi al solito esempio, si supponga che la media vera sia un poco diversa da quella ipotizzata da $\mathcal{H}_0$, ma non troppo diversa; se estraiamo un campione sperimentale, questo non sarà così diverso da un generico campione estratto secondo $\mathcal{H}_0$; come può z cadere nelle code relative al 5% di probabilità?

L'errore che si commette quando non si rifiuta $\mathcal{H}_0$ mentre era falsa, viene detto *errore di seconda specie*.

La sua probabilità però non è ben definita, perché non lo è la densità sotto l'ipotesi troppo generica $\mathcal{H}_1$. Bisogna specificare meglio $\mathcal{H}_1$, cioè formulare delle ipotesi alternative $\mathcal{H}_1'$ più specifiche, che identifichino una sola densità, da cui sia calcolabile la probabilità dell'errore di seconda specie. Le ipotesi $\mathcal{H}_1'$ saranno descritte da un parametro d, ad esempio

$$d = \frac{\mu - \mu_0}{\sqrt{n}}\sigma$$

nel caso in cui μ_0 sia la media dell'ipotesi $\mathcal{H}_0$ mentre μ sia la media dell'ipotesi $\mathcal{H}_1'$. Se indichiamo con β la probabilità dell'errore di seconda specie, esso sarà funzione di d:

$$\begin{aligned}\beta\left(d\right) &= P\left(\text{errore di seconda specie relativo a } d\right)\\ &= P\left(\text{non riconoscere che vale } \mathcal{H}_1^d\right)\end{aligned}$$

dove $\mathcal{H}_1^d$ è l'ipotesi alternativa ditipo specifico con parametro d.

La quantità

$$1 - \beta\left(d\right)$$

è detta potenza del test, relativa alla particolare ipotesi alternativa $\mathcal{H}_1^d$. Come complementare dell'altra, il suo significato è quello di probabilità di accorgersi che $\mathcal{H}_0$ è falsa, quando lo è

(nel senso preciso che vale $\mathcal{H}_1^d$). Quanto è capace il test di riconoscere che $\mathcal{H}_0$ è falsa, quando la verità è $\mathcal{H}_1^d$? La potenza quantifica questa capacità:

$$1-\beta\left(d\right)=P\left(\text{riconoscere che } \mathcal{H}_0 \text{ è falsa, se vale } \mathcal{H}_1^d\right).$$

2.3.5 Struttura diretta della procedura di test

- Specificare $\mathcal{H}_0$
- specificare $\mathcal{H}_1$ (a volte nella pratica resta nascosta)
- scegliere α
- calcolare z dal campione sperimentale.
- Vedere se z cade nelle code di probabilità α. Se sì, rifiutare $\mathcal{H}_0$, altrimenti dichiarare che il campione non contraddice $\mathcal{H}_0$.

Si può anche immaginare la seguente struttura più elaborata, a priori, in fase di Design Of Experiments (DOE):

- specificare $\mathcal{H}_0$ e scegliere α
- specificare anche una particolare $\mathcal{H}_1'$
- scegliere β, calcolare il numero di elementi del campione che garantisca potenza $1-\beta$
- eseguire il test.

Questa è la struttura logica in fase di DOE. Per soddisfare due valori scelti a priori di significatività e potenza, l'unico parametro su cui possiamo giocare è la numerosità del campione.

2.3.6 p-value (struttura indiretta)

La maggior parte dei software, quando esegue un test, non chiede di assegnare α. Chiede solo il campione ed alcuni dettagli di $\mathcal{H}_0$.

L'ouput prodotto dal software è un numero, il *p-value*, introdotto all'inizio della sezione.

Che relazione c'è tra p-value e significatività α scelta a priori?

Per capire, si noti che per certi $\alpha \in (0,1)$ il test produrrà rifiuto di $\mathcal{H}_0$, per altri no. Se α è grande, la coda (o la coppia di code) ha probabilità grande, quindi è facile che z ci cada dentro. Se α è piccolo è difficile. Quindi un test tende a rifiutare per α grande, non rifiutare per α piccolo. L'intervallo $(0,1)$ si divide in due parti

$$\underset{\text{no rejection}}{(0,p)} \quad \text{——} \quad \underset{\text{rejection}}{(p,1)}$$

Per tutti gli $\alpha \in (p,1)$ il test rifiuta $\mathcal{H}_0$. Per tutti gli $\alpha \in (0,p)$ no. Il numero p di separazione è il p-value. Quindi il p-value è il miglior livello di significatività che porta al rifiuto.

Da qui si vede che, usando un software, se si vuole eseguire un test nel modo diretto, cioè pre-assengando α, lo si può fare: basta deciderlo, poi far eseguire il test al software che produrrà un p-value, ed infine confrontare α col p-value.

Siccome questo può essere fatto per ogni scelta di α, è giusto che il software dia il p-value. Così non solo risponde ad ogni possibile pre-assegnazione di α ma ci dice anche quale era il valore più piccolo possibile di α che avrebbe portato al rifiuto.

Di fatto molto spesso in pratica non prefissiamo α ma calcoliamo direttamente il p-value e lo giudichiamo a posteriori: se è piccolo, significa che anche con α piccoli si sarebbe arrivati al rifiuto. Comunque, il valore 0.05 viene spesso preso come riferimento di piccolezza.

2.3.7 Test gaussiano per la media unilaterale e bilaterale, varianza nota

Test unilaterale destro

Riassumiamo il test unilaterale per la media appreso fino ad ora. Come ipotesi non messe in discussione c'è la gaussianità della grandezza esaminata ed il valore noto di σ. L'ipotesi nulla è una dichiarazione circa il valore della media, che indicheremo con μ_0. Il test in forma diretta allora procede così:

- si fissa la significatività α (es. 0.05) e si calcola il relativo quantile $q_{1-\alpha}$ (1.64 se $\alpha = 0.05$)
- si calcola $\overline{x}$ dal campione e poi $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$.
- se $z > q_{1-\alpha}$ si rifiuta l'ipotesi, cioè si afferma che la media non è μ_0.

Nel Paragrafo 2.3.1, come statistica del test avevamo calcolato $\overline{x}$ e da quello il valore p. Ma volendo svolgere il test in modo diretto, fissato cioè α, dovremmo vedere se $\overline{x} > \mu_0 + \frac{\sigma}{\sqrt{n}}q_{1-\alpha}$ (questo è il numero che identifica la coda destra di area α). Con facili passaggi si riconosce che questo confronto numerico equivale a $|z| > q_{1-\alpha}$, dopo aver introdotto la statistica standardizzata z al posto di $\overline{x}$. Si può notare anche il fatto che la formula per il p-value del Paragrafo 2.3.2 è espressa in termini di z:

$$p^{(U)} = 1 - \Phi(z).$$

Calcoliamo inoltre $\beta(d) = P(\text{errore di seconda specie relativo a } d)$, dove d è un parametro che serve a specificare meglio la condizione alternativa ad $\mathcal{H}_0$. Il parametro naturale per questo problema sarebbe la media vera μ, diversa da μ_0. Ma come abbiamo già fatto poco fa, conviene "standardizzarlo", introducendo il nuovo parametro

$$d = \frac{\mu - \mu_0}{\sigma}\sqrt{n}.$$

Per definizione, vale

$$\beta(d) = P_{\mathcal{H}_1^d}(Z \leq q_{1-\alpha})$$

dove $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$. Ma $\overline{X}$ ora è una gaussiana $N\left(\mu, \frac{\sigma^2}{n}\right)$, non $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. Quindi Z non è $N(0,1)$, ma $N(d,1)$ (come si verifica facilmente). Quindi $Z - d$ è $N(0,1)$. Allora

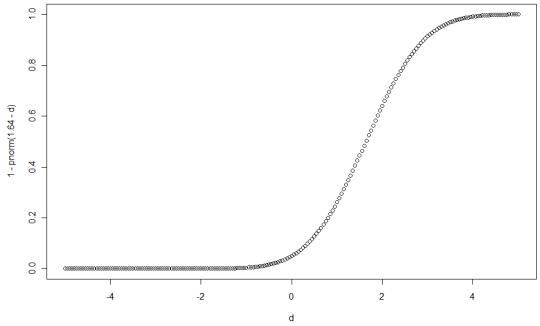
$$\beta(d) = P_{\mathcal{H}_1^d}(Z - d \leq q_{1-\alpha} - d) = \Phi(q_{1-\alpha} - d).$$

La potenza del test, cioè la probabilità di riconoscere che $\mathcal{H}_0$ è falsa, quando vale $\mathcal{H}_1^d$, vale quindi

$$\text{potenza}(d) = 1 - \Phi\left(q_{1-\alpha} - d\right).$$

Osserviamo il grafico della potenza, per $\alpha = 0.05$, al variare di d:

```
d<-(-100:100)/20
plot(d,1-pnorm(1.64-d))
```



Vediamo che la potenza è buona per valori di d intorno a 2.5 ed oltre, cioè per $\frac{\mu-\mu_0}{\sigma}\sqrt{n} > 2.5$ ovvero

$$\mu > \mu_0 + \frac{\sigma \cdot 2.5}{\sqrt{n}}.$$

Invece è bassa per valori di μ più vicini a μ_0: è difficile accorgersi che l'ipotesi nulla è falsa se la media vera è vicina a quella di tale ipotesi.

Una cosa però importante è che la potenza è terribilmente bassa se la media vera μ differisce da μ_0 considerevolmente ma alla sua sinistra, cioè quando $d < 0$. Questo test, che dovrebbe accorgersi che μ_0 è falsa, non ci riesce affatto quando μ è minore di μ_0. Qui sta il succo del discorso unilaterale-bilaterale, che ora specificheremo meglio. Se abbiamo ragione di credere che o valga l'ipotesi nulla (media μ_0) oppure che la media vera sia una certa $\mu > \mu_0$, il test visto fino ad ora, detto test unilaterale destro, è ottimo, diventa molto potente anche per scarti piccoli tra μ e μ_0. Ma se non sapessimo che le deviazioni da μ_0, se ci sono, sono alla sua destra, cioè se fosse altrettanto possibile che la media vera sia minore di μ_0, allora il test precedente ha potenza praticamente nulla.

Oppure: il punto è se per noi è importante accorgerci solo delle variazioni a destra, mentre quelle a sinistra, anche se ci sono, non importa rilevarle. L'esempio del treno è in questa direzione. Serve sapere se i dati reali confermano ritardo medio al massimo pari a 5; se il ritardo medio reale fosse 3, non serve che il test rifiuti l'ipotesi.

Test unilaterale sinistro

Ovviamente ci saranno esempi in cui le potenziali variazioni ripsetto a μ_0, o comunque quelle che vorremmo mettere in evidenza col test, sono alla sinistra di μ_0. Basta usare la variante unilaterale sinistra:

- si fissa la significatività α (es. 0.05) e si calcola il relativo quantile $q_{1-\alpha}$ (1.64 se $\alpha = 0.05$)
- si calcola $\overline{x}$ dal campione e poi $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$.
- se $z < -q_{1-\alpha}$ si rifiuta l'ipotesi, cioè si afferma che la media non è μ_0.

Esercizio 20 *Calcolare p-value e potenza di questo test. Tracciare il grafico della potenza, con R.*

Test bilaterale

Ed infine ci saranno problemi in cui ci interessa scoprire sia le variazioni in positivo che quelle in negativo. Il test è allora:

- si fissa la significatività α (es. 0.05) e si calcola il relativo quantile $q_{1-\frac{\alpha}{2}}$ (1.96 se $\alpha = 0.05$)
- si calcola $\overline{x}$ dal campione e poi $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$.
- se $|z| > q_{1-\frac{\alpha}{2}}$ si rifiuta l'ipotesi, cioè si afferma che la media non è μ_0.

Esercizio 21 *Un certo sistema di servizio (si pensi ad esempio agli sportelli di una banca) è ben dimensionato se ci sono in media 100 richieste al giorno (se sono di più bisogna aumentarlo, se sono di meno si stanno sprecando risorse). Forse il mercato è cambiato e le richieste non sono più 100 in media. Si registra un campione per 9 giorni:*

$$98, 112, 103, 96, 108, 115, 102, 99, 109.$$

Al 95%, il servizio è ben dimensionato? Si supponga, sulla base di esperienze passate, che sia $\sigma = 4$. Sol:

$$z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{104.2 - 100}{4}\sqrt{9} = 3.15$$

maggiore (in valore assoluto) di $q_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}} = 1.96$ (vale $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $1 - \frac{\alpha}{2} = 0.975$, $q_{0.975} = 1.96$). Il sistema non è ben dimensionato.

Il p-value è stato già calcolato mel Paragrafo 2.3.2:

$$p^{(B)} = 2 - 2\Phi(z).$$

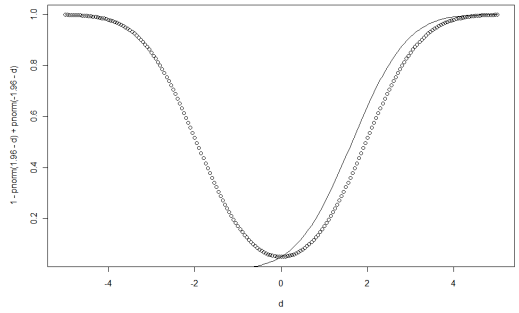
Esercizio 22 *Verificare che la potenza del test bilaterale vale*

$$potenza(d) = 1 - \Phi\left(q_{1-\frac{\alpha}{2}} - d\right) + \Phi\left(-q_{1-\frac{\alpha}{2}} - d\right).$$

Osserviamo il grafico di questa funzione, per $\alpha = 0.05$:

```
d<-(-100:100)/20
plot(d,1-pnorm(1.96-d)+pnorm(-1.96-d))
```

Vediamo che la potenza è alta sia per scostamenti positivi sia negativi, da μ_0 (cioè scostamenti di d da zero). Se però sovrapponiamo i grafici vediamo che la potenza del test unilaterale destro è migliore, a destra:



Quindi conviene usare i test unilaterali quando vogliamo vedere variazioni in una direzione e non ci importa se ci sono o meno nell'altra.

2.3.8 Curve OC e DOE nei test

Le curve di potenza appena tracciate vengono anche dette Curve Caratteristiche Operative (o curve OC). Esse sono usate in fase di DOE, cioè in fase di progettazione degli esperimenti. Lo scopo è determinare la numerosità n degli esperimenti che porterà ad avere una certa potenza, relativamente ad una significatività scelta ed a un valore ipotetico di μ (la potenza è definita in funzione di μ).

Ad esempio, nel caso unilaterale destro, se vogliamo potenza 0.9, dobbiamo risolvere l'equazione

$$\Phi\left(q_{1-\alpha} - d\right) = 0.1$$

da cui $q_{1-\alpha} - d = q_{0.1}$, $d = q_{0.9} + q_{1-\alpha}$, da cui si trova n ricordando che $d = \frac{\mu-\mu_0}{\sigma}\sqrt{n}$. Qui non serve alcuna curva, visto che il calcolo è esplicito e semplice, ma nel caso bilaterale il calcolo esplicito non è possibile e quindi si può usare il disegno.

Ma non è solo questa la ragione per cui si usano curve invece che formule, a volte. Le curve hanno il pregio di far ragionare con buon senso, circa il desiderio di miglioramento in rapporto al costo. Un ragionamento può essere: fino a circa $d = 3$ ogni aumento di d provoca un miglioramento netto della potenza, ma da lì in poi il tasso di miglioramento cala, serve un incremento sempre più ampio di d per ottenere piccoli miglioramenti della potenza. Anche se è un concetto, vago, è come se la curva avesse un "gomito". Allora accontentiamoci di $d = 3$. [Questo valore non è universale: è relativo alla curva OC per $\alpha = 0.05$, e comunque è una nostra intuizione ad occhio.] Quindi $\frac{\mu-\mu_0}{\sigma}\sqrt{n} = 3$, da cui si trova n. La scelta $d = 3$ corrisponde alla potenza $1 - \beta = 1 - \Phi\left(1.64 - 3\right) = 1 - 0.08 = 0.92$.

A livello pratico, ci sono molte scelte da fare, prima di calcolare n. Una delle più critiche è μ. Un'idea possibile, anche se vaga, è che μ sia il primo valore critico diverso da μ_0, cioè il primo valore che, se realizzato, provoca delle conseguenze rilevanti, e che quindi deve essere rilevato dal test. L'esempio seguente aiuta a capire questo concetto.

Carte di controllo

Quando si eseguono test nelle aziende? Ad esempio quando si fa monitoraggio, ad esempio con le *carte di controllo*. Si veda un testo o internet per una descrizione anche molto sommaria

delle carte di controllo, le loro bande, il campionamento a tempi regolari, l'allarme quando si esce dalle bande, l'analogia con l'eseguire un test ad ogni istante di controllo. Nelle carte per tenere sotto controllo la media, vengono calcolate le bande con la formula $\mu_0 \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, come per l'intervallo di confidenza (ma lo scopo è un test, non la stima della media).

Operativamente, pensando ad un esempio concreto, come si realizza una carta? Vanno scelti α ed n, fissati una volta per tutte. Per sceglierli si deve aver chiaro il significato di ogni elemento della teoria del test. α è la probabilità di uscire dalle bande quando invece la media è rimasta μ_0. In molti casi questo non è così grave: quando accade basta rifare il campionamento, per controllare meglio. n va scelto per avere una certa potenza, relativamente ad un certo μ. Si deve sapere, ad es. dagli esperti delle cose prodotte, quali deviazioni da μ_0 rendono inservibili o pericolose le cose prodotte. μ corrisponde a tali valori critici. A quel punto va scelto $\beta(\mu)$. E' la probabilità di non accorgersi di un cambiamento di μ_0, quando questo è avvenuto. Questo sì che può essere pericoloso: vendere cose fuori norma senza saperlo. Allora $\beta(\mu)$ va preso molto piccolo, e trovato n in corrispondenza.

Esempio 66 *Consideriamo un'azienda che produce filato. Il filato prodotto correttamente ha le seguenti caratteristiche:* $\mu_0 = 0.2$ *mm,* $\sigma_0 = 0.02$ *mm. Spessore da evitare (altrimenti diventa visibile ad occhio nelle tessiture):* 0.3 *mm, o* 0.1 *mm. Vorremmo allora creare una carta di controllo.*

Osservazione 1: si tratta di un processo ad alta precisione, cioè con σ_0 *molto piccola. Un campione ha probabilità piccolissima di superare* 0.3 *mm per caso (servono 5 sigma, quasi impossibile).*

Osservazione 2: ha quindi senso tenere sotto controllo la media, invece che il singolo esemplare. Infatti il singolo esemplare è improbabile che superi 0.3 *mm per caso, se la media resta quella. Il pericolo non è nella causalità, ma in un peggioramento sistematico della media.*

Osservazione 3: oppure il pericolo è in un peggioramento della varianza: se fosse ad es. $\sigma = 0.05$ *mm., basta arrivare a* 2σ *per raggiungere* 0.3 *mm per caso in un esemplare. Questo sarebbe frequente.*

Conclusione: vanno tenute sotto controllo media e varianza. Noi qui studiamo solo la media. Si crea una carta di controllo per la media, in cui $UCL = \mu_0 + \frac{\sigma_0 q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ *e* $LCL = \mu_0 - \frac{\sigma_0 q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. *Vanno scelti* α *ed* n. *Come già detto sopra,* α *forse può essere scelto non troppo severo, es. 0.05. Tuttavia, immaginiamo di costruire un sistema di controllo automatico, che suona quando si superano le soglie: non vorremo che suoni per caso 5 volte su 100. Prendiamo allora* $\alpha = 0.001$, *ad esempio (*$q_{1-\frac{\alpha}{2}} = 3.29$*).*

Invece n *va scelto con lo scopo di avere una certa potenza. Va identificato un valore* μ *che non si vuole raggiungere. Va evitato che lo spessore sia* 0.3 *mm, o* 0.1 *mm. Trascurando l'influsso delle piccole fluttuazioni del singolo esemplare, va evitato che la media raggiunga questi valori. Quindi* $\mu = 0.3$ *oppure* $\mu = 0.1$ *sono i valori di riferimento rispetto a cui calcolare la potenza: essa misura la capacità della carta di controllo di accorgersi che lo spessore ha raggiunto quei livelli inaccettabili. Allora, scelta una potenza, es. 0.9999, cioè* $\beta = 0.0001$, *si impone l'equazione*

$$\Phi(3.29 - d) - \Phi(-3.29 - d) = 0.0001$$

dove $d = \frac{0.1}{0.02}\sqrt{n} = 5\sqrt{n}$. *Trascuriamo* $\Phi(-3.29 - d)$ *che plausibilmente è molto piccolo, risolviamo* $\Phi(3.29 - d) = 0.0001$, *cioè* $3.29 - d = q_{0.0001} = -3.72$, $d = 3.29 + 3.72 = 7.01$, *da cui* $5\sqrt{n} = 7$, $\sqrt{n} = \frac{7}{5} = 1.4$, $n = 2$. *Questo risultato è un po' troppo incoraggiante rispetto a molti esempi applicativi, ed è dovuto al fatto che ci vogliono* 5σ *per passare causalmente da 0.2 a 0.3; cioè si tratta di un esempio con un grado di precisione già elevatissimo.*

Esercizio 23 *Proponiamo un esercizio piuttosto lungo. Anche se non lo si vuole svolgere, è bene leggerlo e soffermarsi a riflettere sulla struttura generale, l'interesse applicativo ed il legame con i vari elementi studiati.*

Un'azienda produce delle componenti di metallo da assemblare in macchine complesse.

1) Per un corretto utilizzo, le componenti devono avere lunghezza 20 cm e risulta accettabile uno scarto di 0.1 mm. La produzione è inizialmente piuttosto precisa per cui, statisticamente, si rileva (dopo un campionamento molto numeroso) una deviazione standard campionaria di 0.02 mm.

i) Si mette in funzione l'impianto e si vuole verificare con un test che la lunghezza media sia in regola. Discutere che test svolgereste e progettarlo, nel senso di stabilire la numerosità campionaria sulla base di un criterio ragionevole.

ii) Appurato che inizialmente l'impianto funziona correttamente, e supponendo che la deviazione standard della produzione non cambi ma che possa esserci una lenta deriva nella lunghezza, progettare una carta di controllo per la media da utilizzarsi durante la produzione. Descrivere come avverrà l'utilizzo della carta, esemplificando anche numericamente i possibili scenari (si scelgano a piacere dei potenziali dati che potrebbero emergere dai campionamenti).

iii) Per la determinazione della carta di controllo vanno scelti alcuni parametri: se ne discuta la rilevanza ed i motivi per effettuare una scelta piuttosto che un'altra. Eventualmente, a questo scopo, fare riferimento al concetto di curva caratteristica operativa.

iv) Considerando i valori sperimentali, registrati sulla carta di controllo, come una serie storica, pur nel caso in cui questi valori siano entro i limiti, descrivere con quali tecniche si potrebbe individuare un trend e prevedere una deriva prima che questa accada. [Questa parte dell'esercizio richiede elementi del capitolo sulle serie storiche.]

2) In un periodo successivo, viene chiesto all'azienda se sia in grado di produrre componenti metalliche di quel tipo, ma con nuove caratteristiche di robustezza. L'azienda riesce ad ideare una lega ed un processo produttivo che migliorano la robustezza e deve ora caratterizzarla, tramite esperimenti.

i) Deve decidere quanti provini testare per caratterizzare la robustezza media con una precisione del 5%. Come ragionereste?

ii) Ipotizzata per semplicità la gaussianità della robustezza, una volta caratterizzati i parametri, con la loro incertezza, come calcolereste la robustezza minima al 99%, cioè quel valore della robustezza che viene superato in negativo solo da una componente su cento (in media)? Esemplificare numericamente tramite dati scelti a piacere.

iii) In fase progettuale sono state individuate due componenti della lega metallica ed un trattamento del processo produttivo potenzialmente rilevanti per aumentare la robustezza. Con quali tecniche statistiche si può esplorare l'effettivo impatto di queste variabili e cercare di ottenere il miglior prodotto a parità di costi? [Questa parte dell'esercizio richiede elementi del capitolo sulla statistica multivariata.]

2.3.9 Test di "adattamento"

Con questo termine si intendono i test che cercano di capire se va rifiutata un'ipotesi sulla distribuzione di probabilità di una v.a., invece che sul solo valor medio o sulla varianza. L'ipotesi nulla potrebbe ad esempio avere la forma: la v.a. è $N(3,5)$. Avendo un campione sperimentale che forse proviene da tale distribuzione o forse no, si esegue un test per capire la compatibilità tra campione e densità ipotizzata. Nel test per la media, invece, la gaussianità non veniva messa in dubbio, e neppure la varianza, ma solo la media era in discussione. Ora invece è l'intera densità $N(3,5)$ che viene messa in discussione.

Illustriamo un po' in dettaglio il test chi-quadro, per il quale servono alcune premesse sulle distribuzioni chi-quadro; al termine accenneremo al test di Kolmogorov-Smirnov.

Distribuzione chi-quadro a k gradi di libertà

Definizione 37 *Date delle v.a. $Z_1, ..., Z_k$ gaussiane standard indipendenti, la v.a.*

$$X^{(k)} := Z_1^2 + ... + Z_k^2$$

è detta chi-quadro a k gradi di libertà.

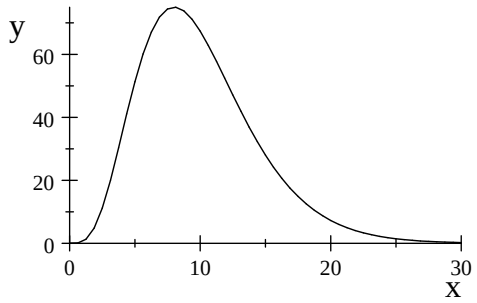
Si può dimostrare che:

Teorema 23 *La densità di probabilità di una v.a. chi-quadro a k gradi di libertà è*

$$f(x) = Cx^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) \quad \text{per } x > 0$$

e zero per $x < 0$; $C^{-1} = 2^{k/2}\Gamma(k/2)$. Quindi è un caso particolare di densità Gamma, con

$$shape = \frac{k}{2}, \quad scale = 2.$$



Chi quadro per $k = 10$

Dalle regole per media e varianza abbiamo

$$E\left[X^{(k)}\right] = kE\left[Z_1^2\right] = k$$
$$Var\left[X^{(k)}\right] = kVar\left[Z_1^2\right] = k\left(E\left[Z_1^4\right] - E\left[Z_1^2\right]^2\right) = 2k$$

in quanto $E\left[Z_1^4\right] = 3$. Quindi

$$\frac{X^{(k)}}{k} \text{ ha media 1 e dev.st. } \frac{\sqrt{2}}{\sqrt{k}}.$$

I valori di $\frac{X^{(k)}}{k}$ si trovano pertanto vicino ad 1, con elevata probabilità, se k è elevato. Per esempio,

$$P\left(\frac{X^{(10)}}{10} > 1.83\right) = 0.05$$
$$P\left(\frac{X^{(100)}}{100} > 1.24\right) = 0.05.$$

Questo fatto sarà la base del test chi-quadro.

Asintoticamente chi-quadro

Supponiamo che X sia una v.a. discreta che assume i valori $1, 2, ..., n_{class}$ con probabilitià

$$P\left(X = k\right) = p_k$$

dove $p_k \in [0,1]$, $\sum_{i=1}^{n_{class}} p_k = 1$.

Supponiamo che $x_1, ..., x_n$ sia un campione sperimentale fatto di numeri che appartengono all'insieme $\{1, 2, ..., n_{class}\}$. Per ogni $k \in \{1, 2, ..., n_{class}\}$, indichiamo con $\widehat{n}_k$ il numero di elementi tra $x_1, ..., x_n$ che valgono k ($n_k = \sum_{i=1}^{n} \delta\left(x_i - k\right)$) e con $\widehat{p}_k$ la frequenza relativa con cui si osserva k nel campione, cioè $\widehat{p}_k = \frac{\widehat{n}_k}{n}$.

Esempio 67 *Nel problema delle preferenze tra A e B, codifichiamoli con A=1, B=2; allora* $n_{class} = 2$, $p_1 = 65\%$, $p_2 = 35\%$ *(quelle ipotizzate nell'ipotesi nulla),* $\widehat{p}_1 = 47/100$, $\widehat{p}_2 = 53/100$.

Calcoliamo la grandezza

$$\chi^2 = \sum_{k=1}^{n_{class}} \frac{\left(\widehat{p}_k - p_k\right)^2}{p_k}.$$

Questo è un numero relae positivo se con $\widehat{p}_k$ intendiamo le frequenze empiriche di un vero campione sperimentale, altrimenti è una v.a. se intendiamo le frequenze empiriche di un ipotetico campione estratto da X, campione aleatorio.

Teorema 24 *La distribuzione di probabilità della v.a.* χ^2 *converge in legge, per* $n \to \infty$, *ad una chi-quadro con* $n_{class} - 1$ *gradi di libertà.*

Significa che

$$P\left(\chi^2 \in [a,b]\right) \sim \int_a^b f_{n_{class}-1}\left(x\right) dx$$

dove $f_{n_{class}-1}\left(x\right)$ è la densità chi-quadro con $n_{class} - 1$ gradi di libertà.

Test chi-quadro

L'idea è: se la distribuzione *empirica* ($\widehat{p}_k$) è troppo diversa da quella *teorica* (p_k), χ^2 assume valori troppo grossi rispetto al valore medio $n_{class} - 1$.

Formalizziamo il test:

- $\mathcal{H}_0$ = "il campione proviene dalla distribuzione (p_k)". Sotto questa ipotesi, la v.a. χ^2 è approssimativamente chi-quadro con $n_{class} - 1$ gradi di libertà;
- scegliamo α, per esempio $\alpha = 0.05$;
- identifichiamo la coda destra della densità $f_{n_{class}-1}(x)$ avente area α (scegliamo la coda destra e solo quella per via dell'idea intuitiva esposta sopra); questo identifica il quantile $\lambda_{\alpha,n_{class}}$;
- calcoliamo la *test statistic* χ^2 dal campione e confrontiamola con $\lambda_{\alpha,n_{class}}$; se $\chi^2 > \lambda_{\alpha,n_{class}}$, rifiutiamo $\mathcal{H}_0$.

Per esempio,

$$\text{se } n_{class} = 11\text{: rifiutiamo quando } \frac{\chi^2}{n_{class} - 1} > 1.83$$
$$\text{se } n_{class} = 111\text{: rifiutiamo quando } \frac{\chi^2}{n_{class} - 1} > 1.24$$

a livello $\alpha = 0.05$.

Questo test ha diverse applicazioni, per esempio alle cosidette tavole di contingenza, che non discutiamo. Una delle applicazioni è all'esame di ipotesi su distribuzioni di probabilità (test di adattamento). Ovviamente il test è in origine un test di adattamento: l'ipotesi nulla è un'ipotesi su una distribuzione di probabilità. Solo che questa è discreta, e se fosse solo così l'ambito applciativo sarebbe molto ristretto. Ecco come si opera più in generale.

Test chi-quadro per il fit di una densità

- $\mathcal{H}_0$ = "il campione proviene dalla densità $f(x)$".
- Suddividiamo la parte rilevante del dominio di $f(x)$ in intervalli (una partizione). Questo passo è soggettivo ed il risultato del test dipenderà in qualche misura da questa scelta. In questa scelta possiamo includere preferenze del tipo: dare più importanza alle code, o meno.
- Siano $I_1, ..., I_{n_{class}}$ gli intervalli scelti, le cosidette *classi*. Calcoliamo

$$p_k = \int_{I_k} f(x)\, dx, \quad k = 1, 2, ..., n_{class}.$$

Queste saranno le probabilità teoriche.

- Calcoliamo anche le probabilità empiriche $\widehat{p}_k$, come si fa in un istogramma: $\widehat{p}_k$ è la frequenza relativa con cui il campione sperimentale cade in I_k.
- Eseguiamo il test chi-quadro descritto al paragrafo precedente, test di confronto tra le frequenze empriche $(\widehat{p}_k)$ e quelle teoriche (p_k).

Se si usa `R`, questo fornisce il p-value.

Osservazione 45 *Contrariamente alle usuali situazioni in cui si esegue un test, qui di solito si vorrebbe confermare una distribuzione ipotizzata, non rifiutarla: di solito si svolge questo studio quando si sta tentando di scoprire una buon modello (una buona distribuzione) che descriva dei dati sperimentali. Quindi saremmo più contenti di un non-rifiuto ma sappiamo che il non rifiuto non è una conferma, e varie ipotesi simili ma un po' diverse possono avere lo stesso esito di non rifiuto (o di rifiuto), quindi non essere distinguibili solo sulla base del rifiuto o meno. Allora spesso il test non viene eseguito per rifiutare o non rifiutare, ma per confrontare i p-value relativi a diverse desità ipotizzate. Si sceglie la densità che ha il p-value più grande, quello più lontano dal rifiuto.*

Il test di Kolmogorov-Smirnov

Esso esegue un confronto tra cdf, precisamente tra quella teorica ipotizzata e quella empirica. La test statistic viene di solito indicata con D. L'idea è di calcolare il massimo (o l'estremo superiore) della differenza tra le due cdf.

La teoria dietro questo test è più elaborata della teoria chi-quadro, ma l'implementazione con R è più semplice. Basta usare il comando

`ks.test(Dati, ''pweibull'', a, s)`.

L'output è il valore D della Kolmogorov-Smirnov test statistics, ed il p-value. L'ipotesi nulla $\mathcal{H}_0$ della riga di R appena scritta è che la distribuzione sia una Weibull di parametri a, s; le modifiche per altre distribuzioni sono ovvie, ma si controlli l'help di R per sicurezza.

Capitolo 3

Processi Stocastici

3.1 Processi a tempo discreto

Definizione 38 *Chiamiamo processo stocastico a tempo discreto una successione*

$$X_0, X_1, X_2, ..., X_n, ...$$

di variabili aleatorie definte su uno stesso spazio probabilizzato (Ω, F, P)*, a valori in* $\mathbb{R}$*. Fissato* $\omega \in \Omega$*, la successione numerica*

$$X_0(\omega), X_1(\omega), ..., X_n(\omega), ...$$

verrà detta realizzazione del processo stocastico.

Questa definizione non è rigida e può essere modificata rispetto ad alcuni dettagli: lo stesso nome si usa per sequenze che partono dal tempo 1, ad esempio, $X_1, X_2, ..., X_n, ...$, o al caso in cui le v.a. X_n prendono valori in spazi diversi da $\mathbb{R}$, ad esempio ciascuna di esse è un vettore aleatorio. Nel seguito considereremo anche il caso in cui l'insieme degli indici (il "tempo") è l'insieme dei numeri interi relativi (tempo che va da $-\infty$ a $+\infty$).

Gli oggetti principali associati ad una v.a. sono la sua legge (ad esempio la densità, se c'è) ed i suoi momenti di ordine uno e due (ed a volte i superiori, o la funzione generatrice, e la cdf). Si fa la stessa cosa per un processo $(X_n)_{n\geq 0}$: la densità di probabilità della v.a. X_n, quando esiste, verrà indicata con $f_n(x)$, la media con μ_n, la deviazione standard con σ_n. Spesso scriveremo t al posto di n, pur essendo un intero. Quindi, i nostri primi concetti sono:

i) *la funzione valor medio e la funzione varianza*:

$$\mu_t = E[X_t], \qquad \sigma_t^2 = Var[X_t], \qquad t = 0, 1, 2, ...$$

Oltre a questo, della massima importanza è la correlazione temporale, come dipende o è collegato il processo ad un certo istante ad un altro. Introduciamo tre funzioni:

ii) la *funzione di autocovarianza* $C(t, s)$, $t, s = 0, 1, 2, ...$:

$$C(t, s) = Cov(X_t, X_s) = E[(X_t - \mu_t)(X_s - \mu_s)]$$

e la funzione

$$R(t,s) = E[X_t X_s]$$

(il suo nome verrà discusso sotto). Esse sono simmetriche ($R(t,s) = R(s,t)$ e lo stesso per $C(t,s)$) quindi è sufficiente conoscerle per $t \geq s$. Vale

$$C(t,s) = R(t,s) - \mu_t \mu_s, \qquad C(t,t) = \sigma_t^2.$$

In particolare, quando $\mu_t \equiv 0$ (che accade spesso), $C(t,s) = R(t,s)$. Le più importanti, tra tutte queste funzioni sono considerate μ_t e $R(t,s)$. Infine, introduciamo:

iii) la *funzione di autocorrelazione*

$$\rho(t,s) = Corr(X_t, X_s) = \frac{C(t,s)}{\sigma_t \sigma_s}$$

Vale

$$\rho(t,t) = 1, \qquad |\rho(t,s)| \leq 1.$$

Le funzioni $C(t,s)$, $R(t,s)$, $\rho(t,s)$ vengono usate per identificare ripetizioni (in senso vago, sporcato da errore) nel processo, somiglianze tra il processo ed una sua traslazione temporale. Per esempio, se $(X_n)_{n\geq 0}$ è vagamente periodica di periodo P, $\rho(t+P,t)$ sarà significativamente più alto degli altri valori di $\rho(t,s)$ (ad eccezione di $\rho(t,t)$ che sempre uguale ad 1).

Esempio 68 *Supponiamo che X_n rappresenti le vendite mensili di un certo prodotto, soggetto a stagionalità, cioè più venduto in una certa stagione piuttosto che un'altra (vestiti di un certo tipo, ad esempio). Allora il processo $X_1, X_2, ...$ ed il processo $X_{12+1}, X_{12+2}, ...$ saranno simili, anche se non identici per via delle differenze che intercorrono sempre tra un'annata ed un'altra (concorrenza diversa, situazione economica diversa del luogo di vendita ecc.). Se fossero identici, cioè se fosse $X_{12+1} = X_1$, $X_{12+2} = X_2$ ecc., allora avremmo*

$$\rho(t+12,t) = \frac{C(t+12,t)}{\sigma_{t+12}\sigma_t} = \frac{\sigma_t^2}{\sigma_t \sigma_t} = 1$$

dove abbiamo usato il fatto che $\sigma_{t+12} = \sigma_t$ (essendo $X_{12+t} = X_t$) e

$$C(t+12,t) = Cov(X_{t+12}, X_t) = Cov(X_t, X_t) = \sigma_t^2.$$

Quindi $\rho(t+12,t) = 1$, il valore massimo possibile. Se non vale esattamente $X_{12+t} = X_t$ ma solo approssimativamente, $\rho(t+12,t)$ non sarà esattamente 1 ma emergerà comunque rispetto agli altri valori.

Esempio 69 *Precisamente, per fare un esempio numerico, supponiamo che il legame tra tempi a distanza di un anno sia*

$$X_{12+t} = X_t + \varepsilon_t$$

dove per semplicità supponiamo che ε_t sia indipendente da X_t e sia piccolo nel senso che la deviazione standard di ε_t sia un decimo di quella di X_t:

$$Var[\varepsilon_t] = \frac{1}{100} Var[X_t].$$

Da queste ipotesi discende che

$$Var\left[X_{12+t}\right] = Var\left[X_t\right] + \frac{1}{100}Var\left[X_t\right] = \frac{101}{100}Var\left[X_t\right]$$

ovvero

$$\sigma_{t+12} = \sqrt{\frac{101}{100}}\sigma_t = 1.005 \cdot \sigma_t$$

$$Cov\left(X_{t+12}, X_t\right) = Cov\left(X_t, X_t\right) + Cov\left(\varepsilon_t, X_t\right) = Cov\left(X_t, X_t\right) = \sigma_t^2$$

da cui

$$\rho\left(t+12, t\right) = \frac{C\left(t+12, t\right)}{\sigma_{t+12}\sigma_t} = \frac{1}{1.005}\frac{\sigma_t^2}{\sigma_t\sigma_t} = 0.995.$$

Anche un trend è una forma di ripetizione temporale. Qui però la teoria si fa più difficile ed è meglio capirla più avanti. Anticipando un po' le cose, bisogna far distinzione tra processi stocastici e serie storiche (che tratteremo nel prossimo capitolo). Una serie storica è una sequenza di numeri, non di variabili aleatorie. Può però essere una realizzazione sperimentale di un processo (così come un singolo numero può essere il risultato sperimentale associato ad una v.a.). Allora si tende a confondere i due concetti, processo stocastico e serie storica. Il teorema ergodico che vedremo tra poco rende ancor più stretto questo legame. Tuttavia, i due concetti sono diversi. Ora, tornando al trend, un conto è un processo con trend, un altro una serie storica con trend. Le due cose hanno riflessi diversi sull'autocorrelazione. Mentre per le serie storiche vedremo che l'autocorrelazione di una serie con trend ha valori tutti abbastanza alti (da ciò si può ad esempio dedurre che c'è un trend se non fosse visibile ad occhio), per un processo stocastico con trend la funzione $\rho\left(t, s\right)$ potrebbe non manifestare nulla di sigificativo. Questa differenza nell'autocorrelazione di processi e serie con trend non contraddice il teorema ergodico (quello che rigorosamente lega i due concetti) perché esso vale sotto ipotesi di stazionarietà del processo, ipotesi che sono appunto violate quando c'è un trend. In altre parole, quando c'è un trend, la teoria dei processi e quella delle serie storiche presenta delle divergenze.

Esempio 70 *Sia* $\left(Z_n\right)_{n\geq 0}$ *una successione di v.a. indipedenti di media zero e varianza 1 e sia* $\left(X_n\right)_{n\geq 0}$ *definito da*

$$X_n = a \cdot n + b + \varepsilon \cdot Z_n.$$

X_n *è un processo con trend, se* ε *è piccolo rispetto ad* a*: il grafico di una realizzazione di* X_n *è la retta* $a \cdot n + b$ *sporcata dalle piccole variazioni casuali* $\varepsilon \cdot Z_n$*. Vale*

$$Var\left[X_n\right] = Var\left[\varepsilon \cdot Z_n\right] = \varepsilon^2$$

ovvero $\sigma_t = \varepsilon$*, e ricordando che nella covarianza le costanti additive si possono cancellare,*

$$\begin{aligned} Cov\left(X_t, X_s\right) &= Cov\left(at + b + \varepsilon Z_t, as + b + \varepsilon Z_s\right) \\ &= Cov\left(\varepsilon Z_t, \varepsilon Z_s\right) = \varepsilon^2\delta\left(t - s\right) \end{aligned}$$

dove il simbolo $\delta\left(t - s\right)$ *(delta di Dirac) vale 0 per* $t \neq s$*, 1 per* $t = s$*. Quindi*

$$\rho\left(t, s\right) = \frac{C\left(t, s\right)}{\sigma_t\sigma_s} = \frac{\varepsilon^2\delta\left(t - s\right)}{\varepsilon^2} = \delta\left(t - s\right).$$

In altre parole, $\rho(t,s)$ è 1 per $t=s$, zero altrimenti (quest'ultima cosa è l'opposto di ciò che si osserva per una serie con trend).

Altri oggetti (se definiti) collegati alla struttura temporale sono:
iv) la *densità di probabilità congiunta*

$$f_{t_1,...,t_n}(x_1,...,x_n),, \qquad t_n \geq ... \geq t_1$$

del vettore $(X_{t_1},...,X_{t_n})$, nel caso continuo, oppure le probabilità marginali

$$P(X_{t_1}=x_1,...,X_{t_n}=x_n)$$

nel caso discreto
v) la *desità condizionale*

$$f_{t|s}(x|y)=\frac{f_{t,s}(x,y)}{f_s(y)}, \qquad t>s$$

nel caso continuo, oppure le probabilità condizionali

$$P(X_t=y|X_s=x)=\frac{P(X_t=y,X_s=x)}{P(X_s=x)}, \qquad t>s$$

nel caso discreto. Vedremo ad esempio all'opera quest'ultimo concetto nelle catene di Markov, in capitolo successivo.

Ora, un'osservazione circa il nome di $R(t,s)$. In Statistica e nella Time Series Analysis, si usa il nome *funzione di autocorrelazione* per la funzione $\rho(t,s)$, come abbiamo fatto sopra. Ma in altre discipline legate al signal processing, questo nome è dato alla funzione $R(t,s)$. Non ci sono motivi particolari per tale scelta se non il fatto che $R(t,s)$ è la quantità fondamentale da capire e studiare, mentre le altre ($C(t,s)$ e $\rho(t,s)$) sono semplici trasformazioni di $R(t,s)$. Quindi ad $R(t,s)$ viene dato quel nome che maggiormente ricorda il concetto di auto-relazione tra valori a tempi differenti. Nel seguito useremo entrambi i termini ed a volte, come si fa nel signal processing, chiameremo $\rho(t,s)$ il *coefficiente di autocorrelazione.*

L'ultimo oggetto che introduciamo si riferisce a *due processi* contemporaneamente: $(X_n)_{n\geq 0}$ ed $(Y_n)_{n\geq 0}$. E' chiamato:
vi) *funzione di mutua-correlazione (cross-correlation function)*

$$C_{X,Y}(t,s)=E\left[\left(X_t-E[X_t]\right)\left(Y_s-E[Y_s]\right)\right].$$

Questa funzione misura la somiglianza tra i due processi traslati nel tempo. Per esempio, può essere usata col seguente scopo: uno dei due processi, diciamo Y, è noto, le sue realizzazioni hanno ad esempio una forma per noi nota ed importante, l'altro processo, X, è il processo meno noto che stiamo esaminando, in cui vorremmo scoprire se ci sono porzioni, finestre, con una forma simile a quella di Y. Per esempio, nella trasmissione dei segnali o nella ricezione radar, Y è una forma nota di riferimento, X è il sganale ricevuto che stiamo esaminando, in cui vorremmo scoprire se c'è quella forma (nascosta dal rumore e dalla traslazione temporale, visto che non sappiamo a che istante si manifesti la forma).

Un'altro ambito applicativo importante può essere quello economico-gestionale. Qui X ed Y possono raprresentare i dati (mensili ad esempio) di due grandezze economiche, per esempio Y le vendite ed X la pubblicità. E si vuole scoprire se X infuisce su Y e quando, dopo quanto tempo. Tramite la cross-correlation vediamo se c'è un legame, una somiglianza, tra i due processi e lo vediamo a seconda della traslazione temporale. Se $C_{X,Y}(t, t+2)$ è più grande degli altri valori, questa può essere l'indicazione che dopo due mesi la pubblicità ha effetto, più di quanto non abbia dopo un mese o a distanza maggiore di tempo.

Esempio 71 *Supponiamo che sia*

$$\begin{aligned} &X_n = 0 \text{ per } n \neq 3 \\ &X_3 \sim N(\mu_X, 1) \end{aligned}$$

$$\begin{aligned} &Y_n = \varepsilon_n \text{ per } n \neq 5 \\ &Y_5 \sim \varepsilon_5 + \lambda X_3 \end{aligned}$$

dove ε_n è come nell'esempio 69. Supponiamo inoltre che sia X_3 indipendente dalle ε_n. Allora

$$C_{X,Y}(n, n+k) = 0 \text{ per } n \neq 3$$

in quanto le costanti additive si possono cancellare, mentre

$$\begin{aligned} C_{X,Y}(3, n) &= Cov(X_3, \varepsilon_n) = 0_n \text{ per } n \neq 5 \\ C_{X,Y}(3, 5) &= Cov(X_3, \varepsilon_5 + \lambda X_3) = \lambda. \end{aligned}$$

Vediamo quindi che $C_{X,Y}(t, s)$ è nulla per tutti i valori di t, s salvo per la combinazione $C_{X,Y}(3, 5)$, che mostra quindi un legame. Ovviamente se il processo è noto il legame lo vedevamo dalla sua definizione; in quest'ottica l'esempio mostra solo la coerenza del concetto di $C_{X,Y}(t, s)$ con le aspettative. Un diverso spirito con cui si può invece giudicare il risultato di questo esempio è: se il processo non ci è noto e possiamo osservare solo la funzione $C_{X,Y}(t, s)$, vedendo che essa è nulla salvo per $C_{X,Y}(3, 5)$, capiamo una caratteristica del processo.

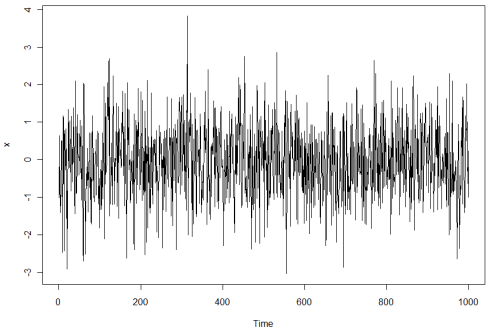
Quando si studiano più processi, può essere conveniente scrivere $R_X(t, s)$, $C_X(t, s)$ per le quantità associate al processo X. Sviluppiamo in dettaglio i calcoli relativi ad alcuni altri esempi. Essi, il white noise, la random walk ed i processi con ritardo 1, sono gli esempi fondamentali di carattere generale.

Esempio 72 (white noise) *Il white noise (rumore bianco, WN) di intensità σ^2 è il processo $(X_n)_{n \geq 0}$ avente le seguenti proprietà:*

i) $X_0, X_1, X_2, ..., X_n, ...$ sono v.a. indipendenti

ii) $X_n \sim N(0, \sigma^2)$.

Si tratta di un processo molto elementare, con una struttura temporale banale, ma che viene usato come mattone di costruzione per altri processi, dove di solito è indicato con ε_n (si vedano gli esempi precedenti). Viene anche usato come termine di paragone per capire le proprietà di esempi più complessi. La seguente figura è stata ottenuta col software R tramite il comando `x<-rnorm(1000); ts.plot(x)`.



Esempio 73 (continuazione sul WN) *Calcoliamo le quantità fondamentali associate al WN (le verifiche sono lasciate per esercizio):*

$$\mu_t = 0 \qquad \sigma_t^2 = \sigma^2$$

$$R(t,s) = C(t,s) = \sigma^2 \cdot \delta(t-s)$$

dove il simbolo $\delta(t-s)$ *della delta di Dirac è stato già usato nell'esempio (70), quindi*

$$\rho(t,s) = \delta(t-s)$$

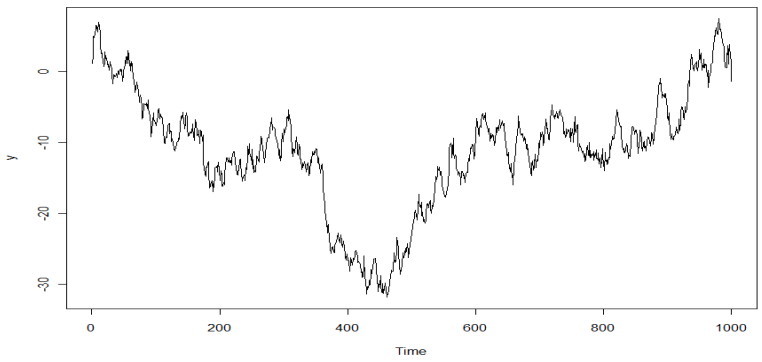
$$f_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i) \qquad \text{where } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$f_{t|s}(x|y) = p(x).$$

Esempio 74 (random walk) *Sia* $(W_n)_{n\geq 0}$ *un white noise. Poniamo*

$$\begin{aligned} X_0 &= 0 \\ X_{n+1} &= X_n + W_n, \qquad n \geq 0. \end{aligned}$$

Questa è una random walk (passeggiata casuale, RW). Il white noise è stato utilizzato come mattone da costruzione: la RW $(X_n)_{n\geq 0}$ *è soluzione di un'equazione per ricorrenza lineare, forzata da un white noise vedremo esempi più generali tra un momento). La seguente figura è stata ottenuta col software R tramite il comando* `x<-rnorm(1000); y<-cumsum(x); ts.plot(y)`.



Esempio 75 (continuazione sulla RW) *Le variabili X_n non sono indipendenti (X_{n+1} dipende in modo ovvio da X_n). Usando la ricorrenza si verifica subito che*

$$X_{n+1} = \sum_{i=0}^{n} W_i.$$

Inoltre, si verifica subito che

$$\begin{aligned} \mu_0 &= 0 \\ \mu_{n+1} &= \mu_n, \qquad n \geq 0 \end{aligned}$$

quindi $\mu_n = 0$ per ogni $n \geq 0$. Infine, verificare per esercizio che, se σ^2 è l'intensità del WN e σ_n è la deviazione standard della RW, vale

$$\sigma_n = \sqrt{n}\sigma, \qquad n \geq 0.$$

L'interpretazione intuitiva di questo risultato è che X_n "cresce" (pur fluttuando tra positivi e negativi, si veda la figura sopra) come $\sqrt{n}$, grossolanamente parlando.

Esempio 76 (continuazione sulla RW) *Per quanto riguarda la dipendenza temporale, intanto vale $C(t,s) = R(t,s)$. Vale poi*

$$R(m,n) = E\left[\sum_{i=0}^{n} W_i \cdot \sum_{j=0}^{m} W_j\right] = \sum_{i=0}^{n}\sum_{j=0}^{m} E[W_i W_j] = \sum_{i=0}^{n}\sum_{j=0}^{m} \delta(i-j)\,\sigma^2.$$

Se $m \geq n$, troviamo $R(m,n) = n\sigma^2$, quindi in generale $R(m,n) = (n \wedge m)\,\sigma^2$. Si può poi vrificare, per $m \geq n$, che

$$\rho(m,n) = \sqrt{\frac{n}{m}}.$$

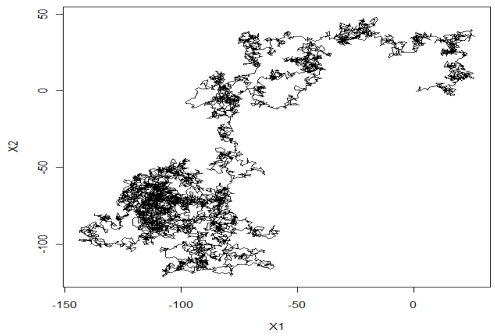
Questo implica in particolare

$$\rho(m,1) \to 0 \text{ as } m \to \infty.$$

Possiamo interpretare questo risultato dicendo che la RW perde memoria della posizione iniziale.

Esempio 77 *Per curiosità, vediamo almeno graficamente la random walk a valori nel piano (dim=2):*

```
N<-10000
W1<-rnorm(N,0,1)
W2<-rnorm(N,0,1)
X1<-1:N
X2<-1:N
X1[1]<-0
X2[1]<-0
X1<-cumsum(W1)
X2<-cumsum(W2)
plot(X1,X2, type=''l'')
```



3.1.1 Legame tra v.a. esponenziali e di Poisson

Variabili aleatorie di Erlang

Capita spesso di considerare la somma di n v.a. esponenziali indipendenti con lo stesso parametro λ: ad esempio è l'istante in cui arriva in coda l'n-esimo cliente, se tra un arrivo e l'altro passa un tempo esponenziale, e tali *intertempi* sono indipendenti ed ugualmente distribuiti.

Date $T_1, ..., T_n$ v.a. indipendenti con distribuzione esponenziale di parametro $\lambda > 0$, diciamo che $S_n = T_1 + ... + T_n$ è una *v.a. di Erlang* di numerosità n e parametro λ.

Lemma 3 *La sua densità g_n e la sua funzione di distribuzione G_n sono date da*

$$g_n(x) = \lambda \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x} \quad per\ x > 0$$

$$G_n(x) = 1 - e^{-\lambda x}\left(1 + \frac{\lambda x}{1!} + ... + \frac{(\lambda x)^{n-1}}{(n-1)!}\right) \quad per\ x > 0.$$

Proof. Dimostriamo il lemma per induzione. Per $n = 1$ la g_1 è la densità esponenziale, quindi l'affermazione è vera. Supponiamo che l'affermazione del lemma sia vera per n, dimostriamola per $n + 1$. Consideriamo la somma

$$S_{n+1} = T_1 + ... + T_n + T_{n+1} = S_n + T_{n+1},$$

dove $S_n = T_1 + ... + T_n$ ha $g_n(x) = \lambda \frac{(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}$ come densità. Abbiamo ricordato sopra che la densità della somma di due v.a. indipendenti è la convoluzione delle densità. Quindi

$$g_{n+1}(x) = \int_{-\infty}^{+\infty} g_n(x-y) f_{T_{n+1}}(y)\, dy$$

dove abbiamo indicato con $f_{T_{n+1}}(y)$ la densità di T_{n+1}. Per $x > 0$ vale allora (si deve prestare un attimo di attenzione agli estremi di integrazione, motivati dal fatto che le due densità integrande sono nulle per i loro argomenti negativi)

$$\begin{aligned} g_{n+1}(x) &= \int_0^x \lambda \frac{(\lambda(x-y))^{n-1}}{(n-1)!} e^{-\lambda(x-y)} \lambda e^{-\lambda y} dy \\ &= \frac{\lambda^2 e^{-\lambda x}}{(n-1)!} \lambda^{n-1} \int_0^x (x-y)^{n-1} dy = \frac{\lambda^2 e^{-\lambda x}}{(n-1)!} \lambda^{n-1} \int_0^x t^{n-1} dt \\ &= \frac{\lambda^2 e^{-\lambda x}}{(n-1)!} \lambda^{n-1} \frac{x^n}{n} = \lambda \frac{(\lambda x)^n}{n!} e^{-\lambda x}. \end{aligned}$$

La dimostrazione per induzione che g_n è la densità è completa. Per dimostrare che G_n è la funzione di distribuzione si può eseguire l'integrale di g_n, o conoscendo già l'espressione data sopra per G_n basta far vedere che la sua derivata è g_n (derivando, i vari termini si cancellano a due a due, escluso uno) e che $G_n(0) = 0$. La dimostrazione è completa. ■

Osservazione 46 *Il valore medio e la varianza di S_n sono pari rispettivamente a $\frac{n}{\lambda}$ e $\frac{n}{\lambda^2}$ (segue subito dalla definizione e dalle proprietà del valor medio).*

Il legame

Supponiamo che, sull'asse dei tempi $[0, \infty)$, accadano degli eventi ad istanti aleatori successivi (es. gli arrivi di clienti ad una coda). Indichiamo con $T_1, ..., T_n$ gli *interarrivi* tra un evento e l'altro (T_1 è l'istante in cui accade il primo evento, $T_1 + T_2$ l'istante del secondo, mentre T_2 è il tempo che intercorre tra il primo evento ed il secondo, e così via). Fissato un tempo $t > 0$ deterministico, ci chiediamo quanti eventi sono accaduti entro t, ovvero nell'intervallo di tempo $[0, t]$. Indichiamo con N_t questo numero aleatorio (ad esempio il numero di clienti arrivati entro il tempo t).

Teorema 25 *Se le T_i sono esponenziali $Exp(\lambda)$ indipendenti, allora N_t è una v.a. di Poisson di parametro λt.*

Proof. Ricordiamo che $S_n = T_1 + ... + T_n$ è una v.a. di Erlang. N_t e la famiglia $(S_n)_{n\geq 0}$ sono legate dal questa relazione logica:

$$N_t \leq k \Leftrightarrow S_{k+1} > t.$$

Questa darà la chiave di tutto. Dimostriamola mostrando la validità delle due implicazioni separatamente. Se $N_t \leq k$, ovvero se entro t arrivano al più k chiamate, allora non ne possono arrivare $k+1$, quindi la $k+1$-esima chiamata arriva dopo il tempo t, quindi $S_{k+1} > t$. Viceversa, se $S_{k+1} > t$, cioè se la $k+1$-esima chiamata arriva dopo il tempo t, allora entro il tempo t sono arrivate meno di $k+1$ chiamate, quindi al più k, quindi $N_t \leq k$.

Allora

$$\begin{aligned} P(N_t \leq k) &= P(S_{k+1} > t) = 1 - G_{k+1}(t) \\ &= e^{-\lambda t}\left(1 + \frac{\lambda t}{1!} + ... + \frac{(\lambda t)^k}{k!}\right). \end{aligned}$$

Abbiamo così trovato la funzione di distribuzione di N_t. Vale quindi

$$P(N_t = k) = P(N_t \leq k) - P(N_t \leq k-1) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

come volevasi dimostrare. ■

Esempio 78 *Durante le 8 ore di apertura di un buon negozio di bici, l'intertempo di vendita, cioè il tempo tra due vendite successive, è aleatorio ed ha media pari a 2 ore. Supponiamo che sia distribuito secondo una legge esponenziale. Supponiamo che gli intertempi tra le vendite successive della singola giornata siano indipendenti. Calcoliamo alcune cose di interesse.*

La probabilità che nella singola giornata non si abbia nessuna vendita è

$$P(T > 8) = e^{-\frac{1}{2}8} = 0.018.$$

Abbiamo indicato con T il tempo tra una vendita e l'altra; il suo parametro è

$$\lambda = \frac{1}{E[T]} = \frac{1}{2}$$

(misurando il tempo in ore).

La probabilità che in un giorno si effettuino almeno 3 vendite è

$$\begin{aligned} P(N_8 \geq 3) &= 1 - P(N_8 < 3) = 1 - \sum_{k=0}^{2} P(N_8 = k) \\ &= 1 - \sum_{k=0}^{2} e^{-\frac{1}{2}8}\frac{\left(\frac{1}{2}8\right)^k}{k!} = 0.7619. \end{aligned}$$

dove N_8 indica il numero di vendite in 8 ore; abbiamo usato il teorema di legame con le v.a. di Poisson.

Il numero medio di bici vendute in un giorno è

$$E[N_8] = \frac{1}{2}8 = 4.$$

Almeno questo si poteva immaginare anche senza la teoria.

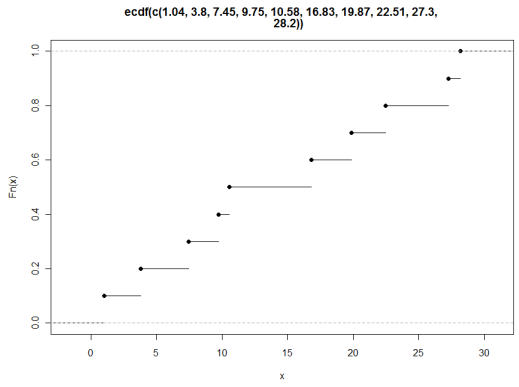
A prima vista l'enunciato può essere di non facile lettura o interpretazione. Raffiguriamolo. Generiamo 10 numeri esponenziali di parametro $\lambda = 1/2$, indicati con `T`, e cumuliamoli in `S`:

```
T <- rweibull(10,1,2)
S <- cumsum(T)
S
```

1.04 3.80 7.45 9.75 10.58 16.83 19.87 22.51 27.30 28.20

Consideriamo la funzione costante a tratti che aumenta di un'unità ad ogni istante S_n (la si può disegnare usando l'espediente della ecdf):

```
plot(ecdf(c(1.04,3.80,7.45,9.75,10.58,16.83,19.87,22.51,27.30,28.20)))
```



Se abbiamo precedentemente fissato, ad esempio, il tempo $t = 15$, il valore di questa funzione all'istante t è N_t. Relativamente a questa simulazione, $N_t = 5$, ma naturalmente dipende dalla simulazione. Il teorema afferma che N_t è una v.a. di Poisson.

Processo di Poisson

Appoggiamoci al teorema precedente ed alla figura. Per ogni $t \geq 0$, è definita una v.a. N_t, il *numero di salti* entro il tempo t, che per il teorema è una Poisson di parametro λt. Vedendo questa variabile al trascorrere di t, cioè come processo stocastico $(N_t)_{t\geq 0}$, abbiamo costruito il cosidetto *processo di Poisson* $(N_t)_{t\geq 0}$ *di intensità* $\lambda > 0$.

Esso può essere definito o nel modo precedente (costruttivo) oppure tramite le seguenti proprietà (in modo analogo al MB):

- $N_0 = 0$

- $N_t - N_s$, per ogni $t \geq s \geq 0$, è una v.a. di Poisson di parametro $\lambda\,(t-s)$
- gli incrementi $N_{t_n} - N_{t_{n-1}}$, ... , $N_{t_1} - N_{t_0}$ sono indipendenti, per ogni $n \geq 1$ e $0 \leq t_0 < t_1 < ... < t_n$.

L'equivalenza tra le due definizioni richiede del lavoro, ma possiamo ad esempio sservare che dalla definizione più astratta discende che: i) se si prende $s = 0$, scopriamo che N_t è una v.a. di Poisson di parametro λt; ii) quindi i valori del processo di Poisson $(N_t)_{t\geq 0}$ sono i numeri interi non negativi: questo significa che se tracciamo il grafico di una generica realizzazione, tale grafico è una curva costante a tratti; iii) ed inoltre è non decrescente in quanto gli incrementi sono anch'essi di Poisson; quindi il grafico è una sorta di scalinata ascendente, con salti posizionati irregolarmente. Ci avviciniamo quindi alla visione costruttiva, pur mancando la dimostrazione del fatto che gli intertempi tra un salto ed il successivo sono v.a. esponenziali di parametro λ indipendenti.

Il processo di Poisson serve ad esempio a descrivere il numero di arrivi ad un sistema di servizio, se si può assumere che gli intertempi di arrivo siano esponenziali.

Generalizzazioni di Poisson

Ci sono varie generalizzazioni interessanti. Data una densità di probabilità $f(x)$ (non è nemmeno necessario che sia normalizzata ad 1) si può usare questa al posto di λ nel seguente modo: nella seconda proprietà del processo di Poisson si richiede che

- $N_t - N_s$, per $t \geq s \geq 0$, è una v.a. di Poisson di parametro $\int_s^t f\,(x)\,dx$.

Ne deriva il cosidetto processo di Poisson *non omogeneo*, avente funzione di intensità f. Serve ad esempio, per gli arrivi ad un servizio, a distinguere le ore di punta.

Un'altra generalizzazione è quella al caso bidimensionale (o multidimensionale): *punti di Poisson nel piano*. Però serve un cambio di punto di vista: bisogna introdurre un processo $(N_A)_{A\subset\mathbb{R}^2}$ indicizzato dagli insiemi A del piano e richiedere:

- $N_A \sim \mathcal{P}\,(\lambda\,|A|)$
- N_{A_1}, ... , N_{A_k} indipendenti se A_1, ... , A_k sono insiemi disgiunti.

Qui $|A|$ indica l'area di A. Oppure si può generalizzare chiedendo

$$N_A \sim \mathcal{P}\left(\lambda \int_A f(x)dx\right).$$

Tali processi di Poisson, detti anche processi di punto (di tipo Poisson) nel piano, descrivono vari problemi. Possono indicare il numero di individui colpiti da una certa epidemia, al variare della zona in esame. Oppure, possono descrivere le posizioni aleatorie in cui avvengono determinati avvenimenti. Infatti, in un certo senso opportuno, è come se dietro un processo di Poisson del piano ci fosse una famiglia di punti aleatori (P_i), $P_i \in \mathbb{R}^2$, e la generica v.a. N_A fosse definita dal numero di punti che cadono in A (si pensi al caso unidimensionale e

gli istanti legati ai tempi T_i). Il software `R` permette una qualche analisi di questi processi, tramite il package "`spatial`".

Un modo approssimato di produrre dei punti di Poisson nel piano è quello di produrre dei punti distribuiti in modo uniforme. Però la densità uniforme ha senso solo su insiemi limitati, per questo è un'approssimazione. Si può quindi prendere un grande intervallo $[-L, L]$, generare delle coppie (X, Y) di coordinate indipendenti, ciascuna uniforme in $[-L, L]$.

3.2 Processi stazionari

Un processo stocastico si dice *stazionario in senso lato* se μ_t e $R(t+n, t)$ sono indipendenti da t.

Ne segue che anche σ_t, $C(t+n, t)$ e $\rho(t+n, t)$ sono indipendenti da t. Quindi possiamo parlare di:

i) media μ del processo

ii) deviazione standard σ

iii) funzione di covarianza $C(n) := C(n, 0)$

iv) funzione di autocorrelazione (nel senso improprio descritto sopra)

$$R(n) := R(n, 0)$$

v) coefficiente di autocorrelazione (o anche funzione di autocorrelazione, nel linguaggio della Statistica)

$$\rho(n) := \rho(n, 0).$$

Si noti che è sparita una variabile temporale da ciascuna delle precedenti quantità. Le funzioni di autocorrelazione ecc. restano funzioni, ma solo di una variabile, non più di due.

Un processo stocastico si dice *stazionario in senso stretto* se la legge di un generico vettore $(X_{n_1+t}, ..., X_{n_k+t})$ è indipendente da t. Questa condizione implica la stazionarietà in senso debole. Il viceversa non vale in generale ma vale almeno per i processi gaussiani (si veda sotto).

Esempio 79 (WN) *Abbiamo*

$$R(t, s) = \sigma^2 \cdot \delta(t - s)$$

quindi

$$R(n) = \sigma^2 \cdot \delta(n).$$

Osservazione 47 (RW) *La RW non è stazionaria, come si vede ad esempio dalla formula* $\sigma_n = \sqrt{n}\sigma$.

Esempio 80 (equazione lineare con smorzamento) *Si consideri la seguente variante con smorzamento della RW:*

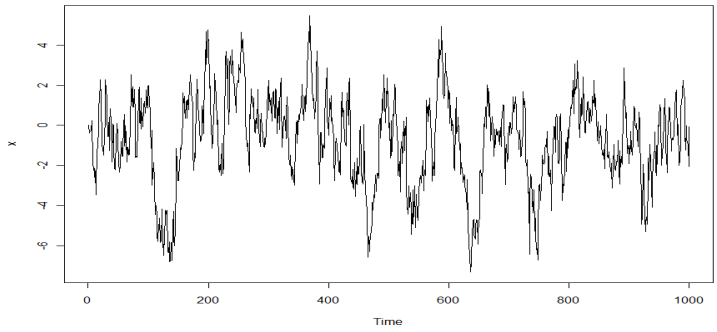
$$X_{n+1} = \alpha X_n + W_n, \qquad n \geq 0$$

dove $(W_n)_{n\geq 0}$ *è un white noise di intensità* σ^2 *e*

$$\alpha \in (-1, 1).$$

La seguente figura è stata ottenuta col software R tramite i comandi ($\alpha = 0.9$, $X_0 = 0$)*:*

```
w <- rnorm(1000)
x <- rnorm(1000)
x[1]=0
for (i in 1:999) {
x[i+1] <- 0.9*x[i] + w[i]
}
ts.plot(x)
```



Esercizio 24 (continuazione) *Il disegno ha alcuni aspetti simii a quello del WN, ma è meno aleatorio, più persistente nelle direzioni in cui si muove. E' una sorta di RW riportata sistematicamente verso l'origine (questo è l'effetto del termine di smorzamento).*

Supponiamo di prendere X_0 indipendente dal WN, di media 0 e varianza σ_0^2. Mostriamo che $(X_n)_{n\geq 0}$ è stazionario (in senso lato) se σ_0^2 è scelto opportunamente rispetto a σ^2. Abbiamo in primo luogo

$$\mu_0 = 0$$
$$\mu_{n+1} = \alpha\mu_n, \qquad n \geq 0$$

come si vede subito dall'equazione per ricorrenza. Quindi $\mu_n = 0$ per ogni $n \geq 0$. La funzione valor medio è costante, prima verifica della stazionarietà.

Come calcolo preliminare, vediamo quando la varianza è constate. Dall'equazione per ricorrenza discende (usando l'indipendenza tra X_n e W_n, valida per ogni n fissato, che si riconosce per induzione)

$$\sigma_{n+1}^2 = \alpha^2\sigma_n^2 + \sigma^2, \qquad n \geq 0.$$

Se vogliamo $\sigma_{n+1}^2 = \sigma_n^2$ per ogni $n \geq 0$ cioè

$$\sigma_n^2 = \alpha^2\sigma_n^2 + \sigma^2, \qquad n \geq 0$$

ovvero

$$\sigma_n^2 = \frac{\sigma^2}{1-\alpha^2}, \qquad n \geq 0.$$

dobbiamo prendere

$$\sigma_0^2 = \frac{\sigma^2}{1-\alpha^2}.$$

Qui, tra l'altro, vediamo per la prima volta il motivo dell'ipotesi $|\alpha| < 1$ *fatta all'inizio. Supponendo* $\sigma_0^2 = \frac{\sigma^2}{1-\alpha^2}$ *troviamo*

$$\sigma_1^2 = \alpha^2 \frac{\sigma^2}{1-\alpha^2} + \sigma^2 = \frac{\sigma^2}{1-\alpha^2} = \sigma_0^2$$

e così via, $\sigma_{n+1}^2 = \sigma_n^2$ *per ogni* $n \geq 0$*, cioè la funzione varianza è constante. Ricordiamo che questo è solo un sintomo della stazionarietà. La definizione di stazionarietà si riferisce alla media ed alla funzione* $R(t,s)$*.*

Esercizio 25 (continuazione) *Verifichiamo fialmente che* $R(t+n,t)$ *non dipende da* t*, imponendo ovviamente la condizione* $\sigma_0^2 = \frac{\sigma^2}{1-\alpha^2}$*, altrimenti non c'è speranza di avere la stazionarietà (in quanto essa implica che la varianza deve essere costante). Vale*

$$R(t+1,t) = E\left[(\alpha X_t + W_t) X_t\right] = \alpha\sigma_n^2 = \frac{\alpha\sigma^2}{1-\alpha^2}$$

che è indipedente da t*;*

$$R(t+2,t) = E\left[(\alpha X_{t+1} + W_{t+1}) X_t\right] = \alpha R(t+1,t) = \frac{\alpha^2\sigma^2}{1-\alpha^2}$$

e così via,

$$\begin{aligned} R(t+n,t) &= E\left[(\alpha X_{t+n-1} + W_{t+n-1}) X_t\right] = \alpha R(t+n-1,t) \\ &= ... = \alpha^n R(t,t) = \frac{\alpha^n\sigma^2}{1-\alpha^2} \end{aligned}$$

che è indipedente da t*. Quindi il processo è stazionario ed abbiamo anche scoperto che*

$$R(n) = \frac{\alpha^n\sigma^2}{1-\alpha^2}.$$

Inoltre

$$\rho(n) = \alpha^n.$$

Il coefficiente di autocorrelazione decade esponenzialmente in t*.*

3.2.1 Processi definiti anche per tempi negativi

Possiamo estendere un po' le precedenti definizioni e chiamare processo a tempo discreto anche una successione bilaterale di v.a. $(X_n)_{n\in\mathbb{Z}}$, definita per tempi anche negativi. L'idea intuitiva è quella che il processo (fisico, economico ecc.) che stiamo esaminando non inizia ora, al presente, ad esistere ma è attivo da molto tempo, nel lontano passato.

Questa nozione risulta particolarmente naturale per i processi stazionari. In tal caso la funzione $R(n)$ (e così $C(n)$ e $\rho(n)$) è definita anche per n negativi:

$$R(n) = E\left[X_n X_0\right], \qquad n \in \mathbb{Z}.$$

Per la stazionarietà,

$$R(-n) = R(n)$$

in quanto $R(-n) = E\left[X_{-n}X_0\right] = E\left[X_{-n+n}X_{0+n}\right] = E\left[X_0X_n\right] = R(n)$. Vediamo quindi che questa estensione non contiene molta informazione nuova; tuttavia ogni tanto è utile e semplifica alcuni calcoli.

3.2.2 Serie temporli e grandezze empiriche

Una serie temporale è una sequenza finita di numeri reali $x_1, ..., x_n$, dove di solito l'indice i ha il significato di tempo. Sono serie temporali i dati mensili o annuali di tipo economico-sociale rintracciabili su siti come Istat o Eurostat. Sono serie temporali anche le realizzazioni di processi stocastici, accettando eventualmente in questa accezione che la sequenza di numeri sia infinita.

Idealmente, quando osserviamo una serie temporale $x_1, ..., x_n$ del mondo reale, immaginiamo che alle sue spalle ci sia un processo stocastico $(X_k)_{k\in\mathbb{N}}$ di cui essa sia una realizzazione. Con questa immaginazione, applichiamo la teoria dei processi stocastici all'analisi della serie temporale.

Il procedimento è del tutto analogo a ciò che si fa in statistica elementare quando, a fronte di valori sperimentali di una grandezza fisica, economica ecc., caratterizzata da na certa casualità, imprevedibilità del suo risultato, si immagina che alle spalle di questi numeri sperimentali ci sia una variabile aleatoria X, si cui essi siano valori possibili $X(\omega)$ per qualche ω dello spazio Ω su cui la v.a. è definita.

C'è qui somiglianza ma anche una differenza essenziale rispetto alla statistica elementare. Se vogliamo stimare la media di una v.a. X, sappiamo che dobbiamo raccogliere un campione $x_1, ..., x_n$ da X e calcolare $\overline{x} = \frac{1}{n}(x_1 + ... + x_n)$. Se possediamo solo un valore sperimentale x_1 estratto da X, la stima della media è troppo povera. Ora, nell'ambito dei processi stocastici, se vogliamo stimare grandezze medie associate al processo (es. R) dovremmo, in analogia col caso di una v.a. X, possedere un campione estratto dal processo, cioè n sue realizzazioni; ogni realizzazione è una stringa di numeri, quindi dovremmo possedere n stringhe di numeri, n serie storiche.

La novità è che, in ipotesi di stazionarietà, basta una realizzazione, una serie storica, per stimare le grandezze medie. Sarebbe come sperare di stimare media e varianza di una v.a. X possedendo solo un suo valore sperimentale x_1. Da questo punto di vista quindi può sembrare sorprendente ma la chiave è la stazionarietà, che rende simili, per certi versi, le v.a. del processo. Una singola realizzazione è fatta di un valore per ciascuna v.a. del processo, e siccome le v.a. del processo sono per certi versi come tante copie di una X, una sola serie storica somiglia in un certo senso ad un campione estratto da una singola X.

In altre parole, si sta sostituendo l'indipendenza ed equidistribuzione (ipotesi alla base del concetto di campione estratto da una v.a. X) con la stazionarietà.

Per essere più precisi, come vedremo tra poco, serve anche un'ipotesi aggiuntiva chiamata ergodicità. Questo si può capire intuitivamente. L'indipendenza ed equidistribuzione è una doppia ipotesi, appunto. La stazionarietà è un rimpiazzo dell'equidistribuzione, ma non dell'indipendenza. Anzi, il processo stazionario più semplice possibile è quello fatto da ua ripetizione della stessa identica v.a. X, processo in cui c'è dipendenza completa. L'ipotesi di ergodicità sarà quella che rimpiazza l'indipendenza (sarà una sorta di indipendenza asintotica, cioè tra X_n ed X_1 per n grande).

Si consideri allora la serie temporale $x_1, ..., x_n$. Ipotizziamo che provenga da un processo stazionario (ed ergodico, concetto che definiremo più avanti). Allora

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

vengono prese come approssimazioni di μ e σ^2. Nel paragrafo sul teorema ergodico daremo risultati rigorosi di approssimazione. Più delicata è l'approssimazione di $R(k), C(k), \rho(k)$. La quantità

$$\widehat{R}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i x_{i+k}$$

viene presa come approssimazione di $R(k)$. Si noti però che l'approssimazione sarà ragionevole solamente per k sensibilmente minore di n, altrimenti la numerosità $n-k$ del campione che si sta usando è troppo piccolo. Per prudenza i software iniziano calcolando $\widehat{R}(k)$ solo per k dell'ordine di $\log n$; però si può chiedere di dare i valori anche per k più grandi, mantenendo qualche dubbio sulla bntà del risultato.

Sarebbe ora naturale definire $\widehat{C}(k)$ come $\widehat{R}(k) - \widehat{\mu}^2$ ma c'è un problema: $\widehat{R}(k)$ è calcolato usando le sequenze

$$x_1, ..., x_{n-k}$$
$$x_{k+1}, ..., x_n$$

mentre $\widehat{\mu}$ usando tutta la sequenza $x_1, ..., x_n$. Quindi, pur essendo lecito prendere $\widehat{R}(k) - \widehat{\mu}^2$ come stimatore (per k basso e n alto va benissimo), forse è preferibile l'espressione più complicata ma più simmetrica

$$\widehat{C}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} (x_i - \widehat{\mu}_0)(x_{i+k} - \widehat{\mu}_k)$$
$$\text{dove } \widehat{\mu}_0 = \frac{1}{n-k}\sum_{i=1}^{n-k} x_i, \qquad \widehat{\mu}_k = \frac{1}{n-k}\sum_{i=1}^{n-k} x_{i+k}.$$

Se ora si pone

$$\widehat{\sigma}_0^2 = \frac{1}{n-k}\sum_{i=1}^{n-k} (x_i - \widehat{\mu}_0)^2, \qquad \widehat{\sigma}_k^2 = \frac{1}{n-k}\sum_{i=1}^{n-k} (x_{i+k} - \widehat{\mu}_k)^2$$

vale la disuguaglianza

$$\left|\widehat{C}(k)\right| \leq \widehat{\sigma}_0 \widehat{\sigma}_k^2$$

che è alla base della seguente definizione:

$$\widehat{\rho}(k) = \frac{\widehat{C}(k)}{\widehat{\sigma}_0 \widehat{\sigma}_k} = \frac{\sum_{i=1}^{n-k} (x_i - \widehat{\mu}_0)(x_{i+k} - \widehat{\mu}_k)}{\sqrt{\sum_{i=1}^{n-k} (x_i - \widehat{\mu}_0)^2 \sum_{i=1}^{n-k} (x_{i+k} - \widehat{\mu}_k)^2}} \tag{3.1}$$

Questa è l'pprossimazione più coerente di $\rho(k)$. Infatti, oltre ad esserne una buona approssimazione, vale

$$|\widehat{\rho}(k)| \leq 1$$

come per $\rho(k)$. Se avessimo preso altre espressioni come definizione di $\widehat{\rho}(k)$, come ad esempio $\frac{\widehat{C}(k)}{\widehat{\sigma}^2}$ o addirittura $\frac{\widehat{R}(k)-\widehat{\mu}^2}{\widehat{\sigma}^2}$, che per certi versi potevano essere più semplici, avremmo poi potuto trovare valori di $|\widehat{\rho}(k)|$ eccedenti l'unità, cosa assurda ma anche assai spiacevole per gli usi pratici, visto che la proprietà $|\rho(k)| \leq 1$ è proprio la base, adimensionale, universale, che permette di giudicare la presenza o meno di correlazione. Di fronte ad un valore $\widehat{\rho}(k) = 0.95$ non avremmo saputo dire se era alto o meno, visto che il valore massimo non era più 1 ma chissà quale, a seconda delle patologie create dalle espressioni usate per $\widehat{\rho}(k)$.

Osservazione 48 *Il comando* `acf` *del software R calcola* $\widehat{\rho}(k)$.

Osservazione 49 *Il calcolo di* $\widehat{\rho}(k)$ *basato sulla formula (3.1) ha senso anche senza stazionarietà del processo (lo stesso vale per* $\widehat{C}(k)$*, ma non per* $\widehat{\mu}$*,* $\widehat{\sigma}^2$*,* $\widehat{R}(k)$*). Si sta calcolando il coefficiente di correlazione tra due sequenze di numeri*

$$x_1, ..., x_{n-k}$$
$$x_{k+1}, ..., x_n$$

cosa che ha senso e può essere utile qualunque sia la natura di queste sequenze. Se il coefficiente di correlazione è elevato, nel senso che è prossimo ad 1, vuol dire che le due sequenze sono linearmente legate, in modo approssimato. Questo fatto sarà noto dalla statistica elementare, dalla regressione lineare, oppure può essere visto ad occhio pensando che l'espressione

$$\sum_{i=1}^{n-k} (x_i - \widehat{\mu}_0)(x_{i+k} - \widehat{\mu}_k)$$

è grande (non si pensi all'unità di misura ed allo sparpagliamento dei numeri stessi, eliminato dalla divisione per $\widehat{\sigma}_0\widehat{\sigma}_k$*) se i numeri* x_i *e* x_{i+k} *si trovano, al variare di* i*, concordemente dalla stessa parte rispetto a* $\widehat{\mu}_0$ *e* $\widehat{\mu}_k$*, quindi le coppie* (x_i, x_{i+k}) *stanno approssimativamente su una retta a coefficiente angolare positivo.*

Osservazione 50 *Intanto, l'osservazione precedente mostra che se la serie storica è approssimativamente periodica di periodo* P*, allora quando si trasla proprio di* P*, cioè si considerano le sequenze*

$$x_1, ..., x_{n-P}$$
$$x_{P+1}, ..., x_n$$

esse sono approssimativamente uguali e quindi si trova elevata correlazione. Il numero $\widehat{\rho}(P)$ *è alto, prossimo a 1.*

Esempio 81 *Vediamo cosa accade se c'è un trend. Immaginiamo ad esempio che la serie storica abbia la forma*

$$x_i = a \cdot i + b + \varepsilon_i, \qquad i = 1, 2, ..., n$$

con $a > 0$, dove i numeri ε_i sono piccoli (rispetto ad a, soprattutto). Allora calcoliamo, per un qualsiasi (basso), la correlazione delle due sequenze

$$a \cdot 1 + b + \varepsilon_1, ..., a \cdot (n-k) + b + \varepsilon_{n-k}$$
$$a \cdot (k+1) + b + \varepsilon_{k+1}, ..., a \cdot n + b + \varepsilon_n.$$

La seconda sequenza si può riscrivere

$$ak + a \cdot 1 + b + \varepsilon_{k+1}, ..., ak + a \cdot (n-k) + b + \varepsilon_n$$

cioè è data dalla costante ak più ua sequenza simile alla prima (visto che i numeri ε_i sono piccoli). Due sequenze che differiscono approssimativamente per una costante hanno correlazione elevata. Si può arrivare a questo risultato osservando anche che la seconda sequenza è approssimativamente la trasformazione lineare della prima:

$$(a \cdot (k+1) + b + \varepsilon_{k+1})$$
$$= (a \cdot 1 + b + \varepsilon_1) + ak + (\varepsilon_{k+1} - \varepsilon_1)$$

e così via per i termini successivi, per cui i punti

$$(a \cdot i + b + \varepsilon_i, a \cdot (k+i) + b + \varepsilon_{k+i})$$

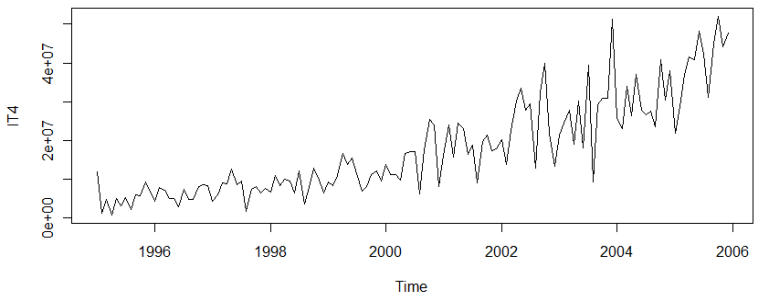
stanno approssimtivamente (cioè a meno dei numeri $\varepsilon_{k+i} - \varepsilon_i$) sulla retta di equazione

$$y = x + ak.$$

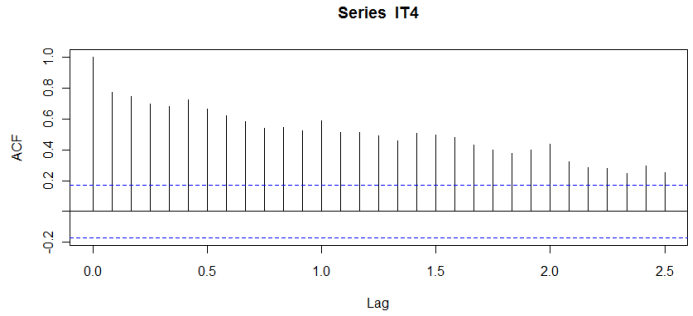
In conclusione, i valori di $\widehat{\rho}(k)$ sono tutti elevati. Abbiamo verificato in un esempio che l'autocorrelazione di una serie con trend positivo ha valori elevati per tutti i k (in pratica i valori di $\widehat{\rho}(k)$ decrescono un po', ma lentamente).

Esercizio 26 *Cosa accade ad una serie storica con trend negativo?*

Esempio 82 *Nella sezione degli esercizi considereremo la seguente serie storica di ambito economico:*



La sua autocorrelazione empirica $\widehat{\rho}(t)$ *è data da*



Vediamo che assume valori elevati uniformemente. La ragione è il trend..

3.3 Processi gaussiani

Definizione 39 *Un processo* $(X_n)_{n\in\mathbb{N}}$ *si dice gaussiano se ogni suo vettore marginale* $(X_{t_1},...,X_{t_n})$ *è congiuntamente gaussiano.*

Per un processo $(X_n)_{n\in\mathbb{N}}$ è equivalente richiedere che $(X_1,...,X_n)$ sia congiuntamente gaussiano per ogni n. La struttura più generale $(X_{t_1},...,X_{t_n})$ sia adatta meglio a insiemi di indici diversi da $\mathbb{N}$.

La densità di un vettore gaussiano $(X_{t_1},...,X_{t_n})$ è individuata dal vettore dei valori medi e dalla matrice di covarianza:

$$(E[X_{t_1}],...,E[X_{t_n}]) = (\mu_{t_1},...,\mu_{t_n})$$
$$\left(Cov\left(X_{t_i},X_{t_j}\right)\right)_{i,j=1,...,n} = (C(t_i,t_j))_{i,j=1,...,n}$$

quindi le funzioni μ_t e $C(s,t)$ determinano le densità di tutte le marginali del processo. Tutte le probabilità associate al processo sono univocamente determinate da μ_t e $C(s,t)$. Tra le conseguenze di questo fatto c'è l'equivalenza dei due concetti di stazionarietà:

Proposizione 18 *Per un processo gaussiano, stazionarietà in senso debole e forte coincidono.*

Proof. Basta ovviamente verificare che la stazionarietà in senso debole implica quella in senso forte. Supponiamo che il processo sia stazionario in senso debole. Consideriamo il generico vettore $(X_{t_1+s},...,X_{t_n+s})$. Dobbiamo verificare che la sua densità non dipende da s. La sua densità è gaussiana. Il vettore delle medie è

$$(E[X_{t_1+s}],...,E[X_{t_n+s}]) = (\mu,...,\mu)$$

indipendente da s e la matrice di covarianza è

$$\begin{aligned}\left(Cov\left(X_{t_i+s},X_{t_j+s}\right)\right)_{i,j=1,...,n} &= (C(t_i+s,t_j+s))_{i,j=1,...,n}\\ &= (C(t_i,t_j))_{i,j=1,...,n}\end{aligned}$$

in quanto $C(u,v)$ dipende solo da $u-v$ (quindi $C(u,v)=C(u+s,v+s)$). Anche la matrice di covarianza non dipende da s, quindi la densità non dipende da s. La dimostrazione è completa. ■

Molto utile è:

Proposizione 19 *Sia $(X_n)_{n\in\mathbb{N}}$ un processo gaussiano e sia $(Y_n)_{n\in\mathbb{N}}$ un processo tale che*

$$Y_n=\sum_{j=1}^{\infty}a_{nj}X_j+b_n$$

cioè è trasformazione lineare di $(X_n)_{n\in\mathbb{N}}$. Sui coefficienti a_{nj} supponiamo che, per ogni n, siano non nulli solo per un numero finito di indici j (ma il risultato finale resta vero anche per somme infinite, sotto opportune condizioni di sommabilità). Allora anche $(Y_n)_{n\in\mathbb{N}}$ è gaussiano.

Proof. Fissiamo n. Consideriamo il vettore

$$Y=(Y_1,...,Y_n).$$

Esso è, per ipotesi trasformazione lineare (o meglio affine) di una stringa finita

$$X=(X_1,...,X_{N_n})$$

cioè esiste N_n, una matrice A ed un vettore b tale che

$$Y=AX+b.$$

Il vettore X è gaussiano per ipotesi. Quindi anche Y lo è, per una proprietà che abbiamo dimostrato sui vettori gaussiani.

Avendo dimostrato che $(Y_1,...,Y_n)$ è gaussiano per ogni n, abbiamo che il processo $(Y_n)_{n\in\mathbb{N}}$ è gaussiano. ■

Molti modelli trattati in queste note sono trasformazioni lineari del white noise, che è un processo gaussiano, quindi tali modelli definiscono processi gaussiani. Quando essi sono stazionari in senso lato, lo sono anche in senso stretto.

3.4 Un teorema ergodico

Dal capitolo sulle convergenze di v.a. ed i teoremi limite, ricordiamo che una successione Y_n converge a Y in media quadratica se

$$\lim_{n\to\infty}E\left[|Y_n-Y|^2\right]=0,$$

in probabilità se per ogni $\varepsilon>0$

$$\lim_{n\to\infty}P\left(|Y_n-Y|>\varepsilon\right)=0$$

e tra i due concetti vale il legame:

Lemma 4 *La convergenza di in media quadratica implica quella in probabilità.*

Proof. Per la disuguaglianza di Chebyshev, abbiamo

$$P\left(|Y_n - Y| > \varepsilon\right) \leq \frac{E\left[|Y_n - Y|^2\right]}{\varepsilon^2}$$

per ogni $\varepsilon > 0$. Pertanto, se $E\left[|Y_n - Y|^2\right]$ tende a zero, allora anche $P\left(|Y_n - Y| > \varepsilon\right)$ tende a zero. ■

Ai vari teoremi limite premettiamo il seguente:

Lemma 5 *Sia $(X_n)_{n\geq 1}$ una successione di v.a. con momenti secondi finiti ed ugual media μ. Se*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i,j=1}^{n} Cov\left(X_i, X_j\right) = 0 \tag{3.2}$$

allora $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge a μ in media quadratica ed in probabilità.

Proof. Vale

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

quindi

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} \left(X_i - \mu\right)\left(X_j - \mu\right)$$

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|^2\right] = \frac{1}{n^2}\sum_{i,j=1}^{n} E\left[\left(X_i - \mu\right)\left(X_j - \mu\right)\right] = \frac{1}{n^2}\sum_{i,j=1}^{n} Cov\left(X_i, X_j\right).$$

Per ipotesi quest'ultima quantità tende a zero, quindi la convergenza in media quadratica è dimostrata. Da questa poi discende la convergenza in probabilità. ■

Ci sono varie versioni di teoremi ergodici. La più semplice è la Legge dei Grandi Numeri, che ricordiamo in una delle sue versioni cosidette deboli, cioè relative alla convergenza in media quadratica ed in probabilità.

Teorema 26 (Legge debole dei Grandi Numeri) *Se $(X_n)_{n\geq 1}$ è una successione di v.a. scorrelate ($Cov\left(X_i, X_j\right) = 0$ per $i \neq j$), con uguali media μ e varianza σ^2 finite, allora $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge a μ in media quadratica*

$$\lim_{n\to\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|^2\right] = 0$$

ed in probabilità: per ogni $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) = 0.$$

Proof. Per la scorrelazione,

$$Cov\left(X_i, X_j\right) = \sigma^2 \delta_{ij}$$

quindi

$$\frac{1}{n^2} \sum_{i,j=1}^{n} Cov\left(X_i, X_j\right) = \frac{1}{n^2} \sum_{i,j=1}^{n} \sigma^2 \delta_{ij} = \frac{\sigma^2}{n} \to 0.$$

Per il Lemma 5, si ottengono i risultati desiderati. ■

Riunendo le precedenti dimostrazioni abbiamo anche ottenuto due stime quantitative interessanti di per sé:

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

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2 \cdot n}.$$

Osservazione 51 *Spesso questo teorema si enuncia nel caso particolare di v.a. indipendenti ed identicamente distribuite, con varianza finita. La dimostrazione però non è più semplice di quella vista ora.*

Osservazione 52 *Il teorema può essere immediatamente generalizzato alle seguenti ipotesi:* $(X_n)_{n\geq 1}$ *è una successione di v.a. scorrelate; i momenti* $\mu_n = E\left[X_n\right]$ *and* $\sigma_n^2 = Var\left[X_n\right]$ *soddisfano*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_i = \mu, \qquad \sigma_n^2 \leq \sigma^2 \text{ for every } n \in \mathbb{N}$$

per qualche costante μ *e* σ^2 *finite. Sotto queste ipotesi, vale lo stesso teorema, con una dimostrazione solo di poco più complicata.*

Vediamo ora un vero e proprio teorema ergodico. In linea generale, un tale teorema afferma che, se un processo è stazionario e soddisfa una proprietà aggiuntiva di ergodicità, allora le sue medie temporali convergono alla media μ. Le diverse versioni di questo teorema dfferiscono sia per dettagli sulle convergenze sia per l'ipotesi di ergodicità: questo non è un concetto univoco, ma ci sono vari gradi di ergodicità. Comuque tutti sono delle generalizzazioni della scorrelazione o indipendenza. Il teorema che segue ha come ipotesi ergodica la *scorrelazione asintotica*.

Teorema 27 (teorema ergodico) *Supponimo che* $(X_n)_{n\geq 1}$ *sia un processo stocastico stazionario in senso lato. Se*

$$\lim_{n\to\infty} C\left(n\right) = 0 \tag{3.3}$$

allora $\frac{1}{n}\sum_{i=1}^{n} X_i$ *converge a* μ *in media quadratica ed in probabilità. Il risultato resta vero sotto l'ipotesi più generale:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} \left|C\left(k\right)\right| = 0. \tag{3.4}$$

Proof. Passo 1. Prima di tutto accertiamoci che la condizione (3.4) sia davvero più generale della (3.3), cioè che (3.3) implichi (3.4). Questo forse è un fatto noto dai corsi di analisi (la convergenza (3.4) viene detta convergenza di Cesaro ed è più debole della convergenza tradizionale, la (3.3)), comunque lo ridimostriamo.

Siccome vale (3.3), per ogni $\varepsilon > 0$ esiste un n_0 tale che per ogni $n \geq n_0$ vale $|C(n)| \leq \varepsilon$. Quindi, per $n \geq n_0$,

$$\frac{1}{n}\sum_{k=0}^{n-1}|C(k)| \leq \frac{1}{n}\sum_{k=0}^{n_0-1}|C(k)| + \frac{1}{n}\sum_{k=n_0}^{n-1}\varepsilon \leq \frac{1}{n}\sum_{k=0}^{n_0-1}|C(k)| + \varepsilon.$$

Il numero $\sum_{k=0}^{n_0-1}|C(k)|$ è indipendente da n, è una costante, per cui esiste $n_1 \geq n_0$ tale che per ogni $n \geq n_1$

$$\frac{1}{n}\sum_{k=0}^{n_0-1}|C(k)| \leq \varepsilon.$$

Quindi, per $n \geq n_1$,

$$\frac{1}{n}\sum_{k=0}^{n-1}|C(k)| \leq 2\varepsilon.$$

Questo significa che $\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}|C(k)| = 0$.

Passo 2. In base al passo 1, se dimostriamo che la condizione (3.4) implica la condizione (3.2) del Lemma 5, abbiamo concluso la dimostrazione del teorema. Questa implicazione è vera in quanto

$$\left|\frac{1}{n^2}\sum_{i,j=1}^{n}Cov(X_i, X_j)\right| \leq \frac{2}{n}\sum_{k=0}^{n-1}|C(k)|. \tag{3.5}$$

Mostriamo questa disuguaglianza. Per la disuguaglianza triangolare ed essendo $Cov(X_i, X_j) = Cov(X_j, X_i) = C(i-j)$, abbiamo

$$\begin{aligned}\left|\sum_{i,j=1}^{n}Cov(X_i, X_j)\right| &\leq \sum_{i,j=1}^{n}|Cov(X_i, X_j)| \\ &\leq 2\sum_{i=1}^{n}\sum_{j=1}^{i}|Cov(X_i, X_j)| \\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{i}|C(i-j)|.\end{aligned}$$

Riscriviamo opportunamente questa doppia somma. Per ogni $i = 1, ..., n$ vale, ponendo $k = i - j$

$$\sum_{j=1}^{i}|C(i-j)| = \sum_{k=0}^{i-1}|C(k)|$$

quindi

$$\sum_{i=1}^{n}\sum_{j=1}^{i}|C(i-j)| = \sum_{i=1}^{n}\sum_{k=0}^{i-1}|C(k)|.$$

Scriviamo esplicitamente quest'ultima doppia somma:

$$\begin{aligned}
&= |C(0)| + (|C(0)| + |C(1)|) + (|C(0)| + |C(1)| + |C(2)|) + ... + (|C(0)| + ... + |C(n-1)|) \\
&= n|C(0)| + (n-1)|C(1)| + (n-2)|C(2)| + ... + |C(n-1)| \\
&= \sum_{k=0}^{n-1}(n-k)|C(k)| \leq \sum_{k=0}^{n-1} n|C(k)| = n\sum_{k=0}^{n-1}|C(k)|.
\end{aligned}$$

Sostituendo questa riscrittura della doppia somma $\sum_{i=1}^{n}\sum_{j=1}^{i}|C(i-j)|$ nella disuguaglianza precedente si ottiene

$$\left|\sum_{i,j=1}^{n} Cov(X_i, X_j)\right| \leq 2n\sum_{k=0}^{n-1}|C(k)|$$

da cui discende la disuguaglianza (3.5), che implica la tesi. La dimostrazione è completa. ■

3.4.1 Tasso di convergenza

Per quanto riguarda il tasso di convergenza, ricordiamo dalla dimostrazione della legge dei grandi numeri che

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right|^2\right] \leq \frac{\sigma^2}{n}.$$

Sotto opportune ipotesi, possiamo dimostrare lo stesso risultato nel caso del teorema ergodico.

Proposizione 20 *Se $(X_n)_{n\geq 1}$ è un processo stazionario in senso lato tale che*

$$\alpha := \sum_{k=0}^{\infty}|C(k)| < \infty$$

(questo implica $\lim_{n\to\infty} C(n) = 0$) allora

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right|^2\right] \leq \frac{2\alpha}{n}.$$

Proof. E' sufficiente mettere insieme alcuni frammenti della dimostrazione precedente:

$$\begin{aligned}
E\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right|^2\right] &= \frac{1}{n^2}\sum_{i,j=1}^{n} Cov(X_i, X_j) \leq \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{i}|Cov(X_i, X_j)| \\
&\leq \frac{2}{n}\sum_{k=0}^{n-1}|C(k)| \leq \frac{2\alpha}{n}.
\end{aligned}$$

■

Si noti che le ipotesi di questi due teoremi ergodici sono molto generali e si potrebbe dimostrare che valgono per tutti gli esempi stazionari studiati in questo corso.

3.4.2 Funzione di autocorrelazione empirica

Spesso abbiamo bisogno della convergenza delle medie temporali di certe *funzioni* del porcesso, e non solo del processo stesso:

$$\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right) \to \eta_g$$

Dobbiamo allora verificare le ipotesi del teorema ergodico per il nuovo processo stocastico $\left(g\left(X_n\right)\right)_{n\geq 1}$. Ecco un esempio semplice.

Proposizione 21 *Sia $\left(X_n\right)_{n\geq 0}$ un processo stazionario in senso lato, con momento quarto finito, tale che il valor medio $E\left[X_n^2 X_{n+k}^2\right]$ sia indipendente da n e*

$$\lim_{k\to\infty} E\left[X_0^2 X_k^2\right] = E\left[X_0^2\right]^2.$$

In altre parole, assumiamo che

$$\lim_{k\to\infty} Cov\left(X_0^2, X_k^2\right) = 0.$$

Allora $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ converge a $E\left[X_1^2\right]$ in media quadratica ed in probabilità.

Proof. Si consideri il processo $Y_n = X_n^2$. La funzione valor medio di (Y_n) è $E\left[X_n^2\right]$ che è indipendente da n per la stazionarietà di (X_n). Per la funzione di autocorrelazione abbiamo poi

$$C\left(n, n+k\right) = E\left[Y_n Y_{n+k}\right] - E\left[X_n^2\right]^2 = E\left[X_n^2 X_{n+k}^2\right] - E\left[X_n^2\right]^2$$

e qui abbiamo bisogno delle nuove ipotesi della proposizione. Quindi (Y_n) è stazionario in senso lato. Infine, grazie all'ipotesi $\lim_{k\to\infty} E\left[X_0^2 X_k^2\right] = E\left[X_0^2\right]^2$, che significa $\lim_{k\to\infty} C_Y\left(k\right) = 0$ dove $C_Y\left(k\right)$ è la funzione di autocorrelazione di (Y_n), possiamo applicare il teorema ergodico. La dimostrazione è completa. ■

Ancor più interessante è il seguente risultato, legato alla stima empirica della funzione di autocorrelazione $R\left(n\right)$. Dato un processo $\left(X_n\right)_{n\geq 1}$, chiamiamo funzione di autocorrelazione empirica l'espressione

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_{i+k}.$$

Teorema 28 *Sia $\left(X_n\right)_{n\geq 0}$ un processo stazionario in senso lato, con momento quarto finito, tale che $E\left[X_n X_{n+k} X_{n+j} X_{n+j+k}\right]$ sia indipendente da n e*

$$\lim_{j\to\infty} E\left[X_0 X_k X_j X_{j+k}\right] = E\left[X_0 X_k\right]^2 \qquad \text{per ogni } k = 0, 1, ...$$

In altre parole, assumiamo che

$$\lim_{j\to\infty} Cov\left(X_0X_k, X_jX_{j+k}\right) = 0.$$

Allora la funzione di autocorrelazione empirica $\frac{1}{n}\sum_{i=1}^{n} X_iX_{i+k}$ *converge a* $R(k)$ *per* $n \to \infty$ *in media quadratica ed in probabilità. Precisamente, per ogni* $k \in \mathbb{N}$*, abbiamo*

$$\lim_{n\to\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_iX_{i+k} - R(k)\right|^2\right] = 0$$

ed analogamente per la convergenza in probabilità.

Proof. Dato $k \in \mathbb{N}$, si consideri il nuovo processo $Y_n = X_nX_{n+k}$. La sua funzione valor medio è costante in n per la stazionarietà di (X_n). Per la funzione di autocorrelazione abbiamo

$$R_Y(n, n+j) = E[Y_nY_{n+j}] = E[X_nX_{n+k}X_{n+j}X_{n+j+k}]$$

che è indipendente da n per ipotesi. Inoltre, $C_Y(j)$ converge a zero. Quindi è sufficiente applicare il teorema ergodico, osservando che $E[Y_0] = R(k)$. La dimostrazione è completa. ■

Con dimostrazioni simili si possono ottenere vari risultati di questo tipo per altre grandezze di interesse pratico.

Circa le ipotesi aggiuntive delle proposizioni precedenti ed altre simili, vale:

Proposizione 22 *Se il processo* $(X_n)_{n\geq 0}$ *è stazionario in senso stretto, allora* $E\left[X_n^2X_{n+k}^2\right]$ *e* $E[X_nX_{n+k}X_{n+j}X_{n+j+k}]$ *sono indipendenti da* n*.*

La dimostrazione è ovvia. Questo aggiunge importanza al concetto di stazionarietà in senso stretto ed alla gaussianità dei processi (per i quali le due nozioni di stazionarietà sono equivalenti).

Osservazione 53 *Non c'è modo di verificare le ipotesi di ergodicità (scorrelazione asintotica, nei nostri enunciati) negli esempi pratici, cioè su una serie storica. Ha senso chiedersi se un processo sia ergodico, non una serie storica. Quindi, quando applichiamo i criteri esposti in questo paragrafo a serie storiche, facciamo un atto di fiducia. Se pensiamo che il processo in esame perda memoria della sua situazione iniziale all'avanzare del tempo, questa fiducia è ragionevole.*

3.5 Analisi di Fourier dei processi stocastici

3.5.1 Premesse

In questa sezione conviene considerare anche processi stocastici a valori complessi, intendendo con questo successioni $(X_n)_{n\in\mathbb{Z}}$ di v.a. $X_n : \Omega \longrightarrow \mathbb{C}$. Le realizzazioni $x_n = X_n(\omega)$, $n \in \mathbb{Z}$, $\omega \in \Omega$, verranno anche chiamate serie temporali, nome che daremo anche a generiche successioni

di numeri reali o complessi $x = (x_n)_{n\in\mathbb{Z}}$. Scriveremo anche $x(n)$ al posto di x_n quando sarà conveniente. Il tempo sarà sempre bilaterale.

Conviene introdurre lo spazio vettoriale l_2 di tutte le serie temporali $x = (x_n)_{n\in\mathbb{Z}}$ tali che

$$\sum_{n\in\mathbb{Z}} |x_n|^2 < \infty.$$

Il numero $\sum_{n\in\mathbb{Z}} |x_n|^2$ viene a volte interpretato come una forma di energia. Le serie temporali che appartengono a l_2 sono dette serie a energia finita.

Un altro spazio importante è l_1 delle serie temporali tali che

$$\sum_{n\in\mathbb{Z}} |x_n| < \infty.$$

Si noti che quest'ipotesi implica energia finita in quanto

$$\sum_{n\in\mathbb{Z}} |x_n|^2 \leq \sup_{n\in\mathbb{Z}} |x_n| \sum_{n\in\mathbb{Z}} |x_n|$$

e $\sup_{n\in\mathbb{Z}} |x_n|$ è limitato perché i termini della serie convergente $\sum_{n\in\mathbb{Z}} |x_n|$ sono infinitesimi, quindi limitati.

Date due serie temporali $f(n)$ e $g(n)$, definiamo la loro *convoluzione* (una nuova serie storica $h(n)$)

$$h(n) = (f * g)(n) = \sum_{k\in\mathbb{Z}} f(n-k)\, g(k).$$

La definzione ha senso quando la serie converge, cosa che accade ad esempio se f e g hanno energia finita. Infatti, per la disuguaglianza di Schwartz,

$$\begin{aligned}\left|\sum_{k\in\mathbb{Z}} f(n-k)\, g(k)\right| &\leq \sum_{k\in\mathbb{Z}} |f(n-k)|^2 \sum_{k\in\mathbb{Z}} |g(k)|^2 \\ &= \sum_{j\in\mathbb{Z}} |f(j)|^2 \sum_{k\in\mathbb{Z}} |g(k)|^2 < \infty.\end{aligned}$$

3.5.2 Trasformata di Fourier a tempo discreto

Data una serie storica $x = (x_n)_{n\in\mathbb{Z}} \in l_2$, introduciamo la *discrete time Fourier transform* (DTFT), che indicheremo con la notaione $\widehat{x}(\omega)$ o con $\mathcal{F}[x](\omega)$, definita da

$$\widehat{x}(\omega) = \mathcal{F}[x](\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-i\omega n} x_n, \qquad \omega \in [0, 2\pi] \ .$$

La convergenza della serie verrà discussa nel paragrafo successivo. C'è purtroppo una sovrapposizione di simboli. Usualmente nel calcolo delle probabilità ω è riservato per l'elemento dello spazio Ω, l'evento casuale elementare. Qui storicamente invece $\omega \in [0, 2\pi]$ indica una frequenza (angolare). Siccome non si scrive praticamente mai esplicitamente il simbolo $\omega \in \Omega$ (questa variabile c'è sempre ma è sottaciuta), nel seguito di questa sezione ω sarà sempre la

frequenza $\omega \in [0, 2\pi]$, salvo venga detto il contrario esplicitamente. Anche il simbolo $\widehat{\ }$ è stato usato in precedenza con altro significato, quello di stima empirica di un parametro statistico; di nuovo, in questa sezione, esso indicherà la trasformata di Fourier.

La successione x_n può essere ricostruita dalla sua DTFT per mezzo della trasformata di Fourier inversa

$$x_n = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{i\omega n} \widehat{x}\,(\omega)\, d\omega.$$

Infatti

$$\begin{aligned}\frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{i\omega n} \widehat{x}\,(\omega)\, d\omega &= \frac{1}{2\pi} \int_0^{2\pi} e^{i\omega n} \sum_{k\in\mathbb{Z}} x_k e^{-i\omega k} d\omega = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} x_k \int_0^{2\pi} e^{i\omega(n-k)} d\omega \\ &= \sum_{k\in\mathbb{Z}} x_k 2\pi\delta(n-k) = x_n\end{aligned}$$

Un passaggio richiede un teorema limite opportuno per poter scambiare serie e integrale.

Osservazione 54 *Le serie temporali (tempo discreto) possono derivare da operazioni di campionamento discreto di segnali fisici a tempo continuo. Se l'incremento temporale tra due campioni consecutivi è* 1*, allora si usa* $-i\omega n$ *all'esponente delle trasformate. Invece, se l'incremento temporale tra due campioni consecutivi è* Δt*, conviene usare* $-i\omega n\Delta t$*. La quantità* $\frac{1}{\Delta t}$ *è chimata frequenza di campionamento.*

Osservazione 55 *La funzione* $\widehat{x}\,(\omega)$ *si può considerare definita per ogni* $\omega \in \mathbb{R}$*, ma è* 2π*-periodica (o* $\frac{2\pi}{\Delta t}$*-periodica).* ω *è detta frequenza angolare.*

Osservazione 56 *A volte (in particolare in fisica), la DTFT è definita senza il segno* $-$ *all'esponente; in tal caso il segno* $-$ *dev'essere presente nella trasformata inversa..*

Osservazione 57 *A volte si omette il fattore* $\frac{1}{\sqrt{2\pi}}$ *nella definizione; noi lo abbiamo incluso per simmetria con la trasformata inversa o per semplicità della formula di Plancherel (senza* $\frac{1}{\sqrt{2\pi}}$*, il fattore* $\frac{1}{2\pi}$ *appare in esse).*

Osservazione 58 *A volte si usa la seguente variante della DTFT:*

$$\widehat{x}\,(f) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-2\pi i f n} x_n, \qquad f \in [0, 1]\,.$$

dove $f = \frac{\omega}{2\pi}$*.*

Dovendo comunque fare una scelta tra tutte queste varianti, usiamo l definizione di DTFT scritta all'inizio, mettendo in guardia il lettore che il confronto con altri testi va fatto modulo determinate modifiche.

Introduciamo infine il concetto di troncamento di una serie storica $x = (x_n)_{n\in\mathbb{Z}}$. Fissata una finestra di ampiezza $2N$, contenente i $2N+1$ punti da $-N$ a N inclusi, usando la funzione indicatrice

$$1_{[-N,N]}(n) = \begin{cases} 1 \text{ se } -N \le n \le N \\ 0 \text{ altrimenti} \end{cases}$$

si introduce il troncamento $x_{2N}(n)$ della serie storica x_n definito da

$$x_{2N}(n) = x_n \cdot 1_{[-N,N]}(n) = \begin{cases} x_n \text{ se } -N \leq n \leq N \\ 0 \text{ altrimenti} \end{cases}$$

e la sua DTFT definita da

$$\widehat{x}_{2N}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{|n| \leq N} e^{-i\omega n} x_n.$$

Questo concetto non necessita dell'ipotesi $x \in l_2$ fatta all'inizio del paragrafo, utilizzata per definire la DTFT. In altre parole, mentre la definizione della DTFT richiede ipotesi di sommabilità della serie storica (noi abbiamo usato l'ipotesi $x \in l_2$), cioè un opportuno decadimento all'infinito dei valori x_n, la definizione di $\widehat{x}_{2N}(\omega)$ non richiede alcuna ipotesi di questo tipo, è applicabile a qualsiasi successione di numeri reali o complessi.

Osservazione 59 *Se $(X_n)_{n\in\mathbb{Z}}$ è un processo stocastico stazionario (non nullo) e $(x_n)_{n\in\mathbb{Z}}$ è una sua tipica realizzazione, x_n non tende a zero per $n \to \infty$. Questo sarebbe incompatibile con la stazionarietà (nella prossima osservazione diamo delle giustificazioni per quest'affermazione). Quindi non possiamo calcolare $\widehat{x}(\omega)$ per le realizzazioni dei processi stazionari. Ma possiamo calcolare $\widehat{x}_{2N}(\omega)$. Su questa operazione si fonda il teorema di Wiener-Khinchine che vedremo tra poco.*

Osservazione 60 *Intuitivamente, si capisce che le realizzazioni di un processo stazionario non nullo non tendono a zero, osservando che le v.a. X_n hanno tutte la stessa varianza (non nulla), quindi hanno valori distribuiti in un certo range, e non è naturale pensare che al crescere di n questi valori tendano a zero, per le singole realizzazioni, quando la loro varianza rimane costante. Rigorosamente, sotto opportune ipotesi di ergodicità, si può fare il seguente ragionamento. Sia $\lambda > 0$ un numero tale che $P(|X_n| \geq \lambda) > 0$. Un tale λ esiste altrimenti la v.a. X_n sarebbe identicamente nulla. Consideriamo la funzione indicatrice dell'insieme $A_\lambda := \{x : |x| \geq \lambda\}$*

$$1_{A_\lambda}(x) = \begin{cases} 1 \text{ se } |x| \geq \lambda \\ 0 \text{ altrimenti} \end{cases}.$$

Consideriamo il processo $Y_n = 1_{A_\lambda}(X_n)$. La v.a. Y_n vale 1 se $|X_n| \geq \lambda$, zero altrimenti, quindi Y_n è una Bernoulli di parametro $p = P(|X_n| \geq \lambda)$. Il processo $(Y_n)_{n\in\mathbb{Z}}$ somiglia quindi al processo di Bernoulli salvo per il fatto che le v.a. non sono indipendenti. E' però stazionario, se $(X_n)_{n\in\mathbb{Z}}$ era stazionario in senso stretto (ogni trasformazione di un processo stazionario in senso stretto è stazionario in senso stretto, se la trasformazione ha una certa proprietà di "misurabilità"; non entramio in questo dettaglio). Se il processo $(Y_n)_{n\in\mathbb{Z}}$ è anche ergodico, vale

$$\frac{1}{n}(Y_1 + ... + Y_n) \to E[Y_1] = p$$

ovvero

$$\frac{1}{n}(1_{A_\lambda}(X_1) + ... + 1_{A_\lambda}(X_n)) \to E[Y_1] = p$$

Purtroppo noi conosciamo questo risultato solo nel senso della convergenza in media quadratica o in probabilità, non nel senso della convergenza lungo le tipiche realizzazioni. Se chiudiamo un occhio su questo (difficile) dettaglio, e supponiamo che, presa una realizzazione $(x_n)_{n\in\mathbb{Z}}$*, in base al ragionamento precedente valga*

$$\frac{1}{n}\left(1_{A_\lambda}(x_1)+...+1_{A_\lambda}(x_n)\right)\to p$$

vediamo che per infiniti indici n *deve essere* $1_{A_\lambda}(x_n)=1$ *(altrimenti, se esistesse* n_0 *tale che per ogni* $n\geq n_0$ *vale* $1_{A_\lambda}(x_n)=0$*, avremmo* $\frac{1}{n}\left(1_{A_\lambda}(x_1)+...+1_{A_\lambda}(x_n)\right)\to 0$*). Quindi per infiniti indici* n *deve valere* $|x_n|\geq\lambda$*. Questo impedisce che sia* $\lim_{n\to\infty}x_n=0$ *e quindi ad esempio che sia* $x\in l_2$*.*

3.5.3 Proprietà della DTFT

1) Se $x\in l_1$, la serie

$$\frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}}e^{-i\omega n}x_n$$

converge assolutamente per ogni $\omega\in[0,2\pi]$. Infatti

$$\left|\sum_{n\in\mathbb{Z}}e^{-i\omega n}x_n\right|\leq\sum_{n\in\mathbb{Z}}|x_n|<\infty$$

(ricordiamo che $\left|e^{-i\omega n}\right|=1$). Anzi, converge uniformemente in ω e quindi definisce una funzione continua di ω:

$$\sum_{n\in\mathbb{Z}}\sup_{\omega\in[0,2\pi]}\left|e^{-i\omega n}x_n\right|=\sum_{n\in\mathbb{Z}}\sup_{\omega\in[0,2\pi]}\left|e^{-i\omega n}\right||x_n|=\sum_{n\in\mathbb{Z}}|x_n|<\infty.$$

Ricordiamo però che $l_1\subset l_2$, quindi questa prima osservzione non garantisce un significato a $\widehat{x}(\omega)$ quando $x\in l_2$.

2) La teoria L^2 delle serie di Fourier garantisce che, se $x\in l_2$, la serie $\sum_{n\in\mathbb{Z}}e^{-i\omega n}x_n$ converge in media quadratica rispetto ad ω, cioè esiste una funzione $\widehat{x}(\omega)$ di quadrato integrabile tale che

$$\lim_{N\to\infty}\int_0^{2\pi}\left|\widehat{x}_{2N}(\omega)-\widehat{x}(\omega)\right|^2d\omega=0.$$

In generale non è più vero che la serie $\frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}}e^{-i\omega n}x_n$ converga per i singoli ω fissati.

3) Vale, nella teoria L^2, la formula di Plancherel

$$\sum_{n\in\mathbb{Z}}|x_n|^2=\int_0^{2\pi}\left|\widehat{x}(\omega)\right|^2d\omega.$$

Proof.

$$\int_0^{2\pi}\left|\widehat{x}(\omega)\right|^2d\omega=\int_0^{2\pi}\widehat{x}(\omega)\left(\widehat{x}(\omega)\right)^*d\omega$$

$$
\begin{aligned}
&= \frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{n\in\mathbb{Z}} e^{-i\omega n}x_n\right)\left(\sum_{m\in\mathbb{Z}} e^{-i\omega m}x_m\right)^* d\omega \\
&= \frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{n,m\in\mathbb{Z}} x_n x_m^* e^{-i\omega n}e^{i\omega m}\right) d\omega \\
&= \frac{1}{2\pi}\sum_{n,m\in\mathbb{Z}} x_n x_m^* \int_0^{2\pi} e^{-i\omega(n-m)} d\omega
\end{aligned}
$$

usando un opportuno teorema di scambio tra serie ed integrali,

$$
= \frac{1}{2\pi}\sum_{n,m\in\mathbb{Z}} x_n x_m^* 2\pi\delta(n-m) = \sum_{n\in\mathbb{Z}} x_n x_n^* = \sum_{n\in\mathbb{Z}} |x_n|^2 .
$$

■

Il significato della formula di Plancherel è che l'energia "contenuta" in una serie temporale e l'energia "contenuta" nella sua DTFT coincdono. Un fattore 2π appare in uno dei due membri se si usano definizioni diverse di DTFT.

4) Se $g(n)$ è a valori reali, allora $\widehat{g}(-\omega) = \widehat{g}^*(\omega)$.

Proof.

$$
\begin{aligned}
\widehat{g}(-\omega) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} g(n)e^{-i(-\omega)n} = \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} g(n)\left(e^{-i\omega n}\right)^* \\
&= \left(\frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}} g(n)e^{-i\omega n}\right)^* = g^*(\omega) .
\end{aligned}
$$

■

5) La DTFT della convoluzione di due serie temporali corrisponde (a meno di un fattore, per altre definizioni della DTFT) al prodotto delle DTFT delle due serie di partenza:

$$
\mathcal{F}[f*g](\omega) = \sqrt{2\pi}\widehat{f}(\omega)\widehat{g}(\omega) .
$$

Si noti il fattore $\sqrt{2\pi}$, che non sarebbe stato presente se nella definizione di DTFT avessimo omesso $\frac{1}{\sqrt{2\pi}}$.

Proof.

$$
\begin{aligned}
\mathcal{F}[f*g](\omega) &= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}}(f*g)(n)e^{-i\omega n} \\
&= \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}}\left(\sum_{k\in\mathbb{Z}} f(n-k)g(k)\right)e^{-i\omega n} \\
&= \frac{1}{\sqrt{2\pi}}\sum_{k\in\mathbb{Z}} g(k)e^{-i\omega k}\sum_{n\in\mathbb{Z}} f(n-k)e^{-i\omega(n-k)} \\
&= \frac{1}{\sqrt{2\pi}}\sum_{k\in\mathbb{Z}} g(k)e^{-i\omega k}\sum_{m\in\mathbb{Z}} f(m)e^{-i\omega m} \\
&\sqrt{2\pi}\widehat{f}(\omega)\widehat{g}(\omega)
\end{aligned}
$$

usando un opportuno teorema sullo scambio di serie. ■

6) Combinando le proprietà precedenti, se g è a valori reali, abbiamo

$$\mathcal{F}\left[\sum_{k\in\mathbb{Z}} f(n+k)\, g(k)\right](\omega) = \mathcal{F}\left[\sum_{k\in\mathbb{Z}} f(n-k)\, g(-k)\right](\omega) = \widehat{f}(\omega)\,\widehat{g}(-\omega) = \sqrt{2\pi}\,\widehat{f}(\omega)\,\widehat{g}^{*}(\omega).$$

Questa proprietà verrà usata nel calcolo della DTFT della funzione di autocorrelazione.

3.5.4 DTFT generalizzata

In casi particolari si può definire la DTFT anche per serie temporali che non soddisfano la condizione $\sum_{n\in\mathbb{Z}} |x_n|^2 < \infty$. Il metodo è usare la definizione

$$\widehat{x}(\omega) = \lim_{N\to\infty} \widehat{x}_{2N}(\omega)$$

se tale limite esiste (in qualche senso ragionevole). Se $x \in l_1$ il limite esiste uniformemente in ω. Se $x \in l_2$ il limite esiste in media quadratica. Ci accontenteremo in questo paragrafo che esista il limite

$$\lim_{N\to\infty} \int_0^{2\pi} \widehat{x}_{2N}(\omega)\, f(\omega)\, d\omega$$

per ogni funzione f continua. Con questa nozione molto debole di limite (è una versione abbreviata del concetto di limite nel senso delle distribuzioni), possiamo definire $\widehat{x}$ anche per certe $x \notin l_2$.

Consideriamo ad esempio la successione

$$x_n = a \sin(\omega_1 n).$$

Calcoliamo la DFTT della successione troncata:

$$\widehat{x}_{2N}(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{|n|\le N} e^{-i\omega n} a \sin(\omega_1 n).$$

Ricordando che

$$\sin t = \frac{e^{it} - e^{-it}}{2i}$$

quindi che $\sin(\omega_1 n) = \frac{e^{i\omega_1 n} - e^{-i\omega_1 n}}{2i}$, vale

$$\sum_{|n|\le N} e^{-i\omega n} a \sin(\omega_1 n) = \frac{1}{2i} \sum_{|n|\le N} e^{-i(\omega-\omega_1)n} - \frac{1}{2i} \sum_{|n|\le N} e^{-i(\omega+\omega_1)n}.$$

Vederemo tra un attimo che questa succesione converge, per $N \to \infty$, nel senso detto sopra.

Siamo costretti, per proseguire, ad usare il concetto di funzione generalizzata, o distribuzione, che è fuori dagi scopi di questo corso, ma che fa parte del bagaglio almeno intutivo di alcuni percorsi di studio in ingegeria. Utilizzeremo la funzione generalizzata chamata $\delta(t)$

delta Dirac (da non confondersi con la semplice delta Dirac $\delta(n)$ nel discreto, che abbiamo usato in precedenza). Essa è caratterizzata dalla proprietà

$$\int_{-\infty}^{\infty} \delta(t) f(t)\, dt = f(0) \tag{3.6}$$

per ogni funzione continua f. Nessuna funzione nel senso usuale del termine ha questa proprietà. Un modo per farsi un'idea intuitiva è il seguente. Consideriamo una funzione, che indichiamo con $\delta_n(t)$, uguale a zero per t fuori da $\left[-\frac{1}{2n}, \frac{1}{2n}\right]$, intervallo di ampiezza $\frac{1}{n}$ intorno all'origine; ed uguale a n in $\left[-\frac{1}{2n}, \frac{1}{2n}\right]$. Abbiamo

$$\int_{-\infty}^{\infty} \delta_n(t)\, dt = 1.$$

Ora,

$$\int_{-\infty}^{\infty} \delta_n(t) f(t)\, dt = n \int_{-\frac{1}{2n}}^{\frac{1}{2n}} f(t)\, dt$$

che è la media integrale di f attorno a 0. Per $n \to \infty$, questa media converge a $f(0)$ se f è continua. In altre parole, vale

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \delta_n(t) f(t)\, dt = f(0)$$

che può essere presa come una sorta di definizione rigorosa dell'identità (3.6), espressa mediante concetti tradizionali. E' inoltre analoga al concetto di limite descritto all'inizio del paragrafo. In un certo senso, quindi, la funzione generalizzata $\delta(t)$ è il limite delle funzioni tradizionali $\delta_n(t)$, ma usando un concetto di limite nuovo. Se si usassero le nozioni usuali di limite vedremmo che $\delta_n(t)$ converge a zero per ogni $t \neq 0$, e a $+\infty$ per $t = 0$. In questo senso molto vago, $\delta(t)$ è zero per $t \neq 0$, $+\infty$ per $t = 0$; ma questa è un'informazione povera, perché non permette di dedurre l'identità (3.6).

Lemma 6 *Vale*

$$\lim_{N\to\infty} \frac{1}{2\pi} \sum_{|n|\leq N} e^{-itn} = \delta(t)$$

nel senso che

$$\lim_{N\to\infty} \int_0^{2\pi} \frac{1}{2\pi} \sum_{|n|\leq N} e^{-itn} f(t)\, dt = f(0)$$

per ogni funzione continua f.

Da questo lemma discende che

$$\lim_{N\to\infty} \sum_{|n|\leq N} e^{-i\omega n} a \sin(\omega_1 n) = \frac{\pi}{i}\delta(\omega - \omega_1) - \frac{\pi}{i}\delta(\omega + \omega_1).$$

In altre parole:

Corollario 6 *La successione*

$$x_n = a \sin(\omega_1 n)$$

ha una DTFT generalizzata, nel senso che esiste il seguente limite

$$\widehat{x}(\omega) = \lim_{N\to\infty} \widehat{x}_{2N}(\omega) = \frac{\sqrt{\pi}}{\sqrt{2}i} \left(\delta(\omega - \omega_1) - \delta(\omega + \omega_1)\right)$$

secondo il significato di limite descritto sopra.

Questo è solo un esempio specifico della possibilità di estendere la DTFT fuori da l_2. Questo esempio ha un'interessante interpretazione. Se il segnale x_n ha una coonente periodica (si noti che la DTFT è lineare, quindi la DTFT della somma di componenti è la somma delle DTFT delle componenti) con frequenza angolare ω_1, allora la sua DTFT due picchi simmetrici (due componenti delta di Dirac) a $\pm\omega_1$. In altre parole, la DTFT rivela le componenti periodiche dei segnali tramite picchi.

Esercizio 27 *Dimostrare che la successione*

$$x_n = a \cos(\omega_1 n)$$

ha DTFT generalizzata

$$\widehat{x}(\omega) = \lim_{N\to\infty} \widehat{x}_{2N}(\omega) = \frac{\sqrt{\pi}}{\sqrt{2}} \left(\delta(\omega - \omega_1) + \delta(\omega + \omega_1)\right).$$

3.6 Densità spettrale di potenza

Definizione 40 *Dato un processo stazionario $(X_n)_{n\in\mathbb{Z}}$ con funzione di autocorrelazione $R(n) = E[X_n X_0]$, $n \in \mathbb{Z}$, chiamiamo densità spettrale di potenza (power spectral density, PSD) la funzione*

$$S(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-i\omega n} R(n), \qquad \omega \in [0, 2\pi]$$

quando la serie converge.

In alternativa, si può usare la definizione

$$S(f) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-2\pi i f n} R(n), \qquad f \in [0, 1]$$

che produce visualizzazioni più semplici in quanto è più semplice vedere ad occhio le frazioni dell'intervallo $[0, 1]$.

Osservazione 61 *Se $\sum_{n\in\mathbb{Z}} |R(n)| < \infty$, la PSD converge uniformemente, ad una funzione continua. Se $\sum_{n\in\mathbb{Z}} |R(n)|^2 < \infty$, la PSD converge in media quadratica. Esistono poi dei casi ulteriori in cui la serie converge in qualche senso generalizzato o grazie a cancelazioni particolari. Dal punto di vista pratico, può essere comunque utile considerare una variante a tempo finito, come $\sum_{|n|\leq N} e^{-i\omega n} R(n)$.*

La funzione $S(f)$ ha alcune proprietà non immediatamene visibili dalla definizione. In particolare essa assume valori reali non negativi. Noi lo vedremo rigorosamente attraverso il teorema di Wiener-Khinchin, ma è bene sapere che c'è un'altra via per stabilirlo. La funzione $R(n)$ è *definita non-negativa*, nel senso che $\sum_{i,j=1}^{n} R(t_i - t_j)\, a_i a_j \geq 0$ per ogni $t_1, ..., t_n$ e $a_1, ..., a_n$. La verifica è facile (come per il fatto che la matrice di covarianza di un vettore aleatorio è semi-definita positiva):

$$\begin{aligned}\sum_{i,j=1}^{n} R(t_i - t_j)\, a_i a_j &= \sum_{i,j=1}^{n} E\left[X_{t_i} X_{t_j}\right] a_i a_j = \sum_{i,j=1}^{n} E\left[a_i X_{t_i} a_j X_{t_j}\right] \\ &= E\left[\sum_{i,j=1}^{n} a_i X_{t_i} a_j X_{t_j}\right] = E\left[\sum_{i=1}^{n} a_i X_{t_i} \sum_{j=1}^{n} a_j X_{t_j}\right] \\ &= E\left[\left(\sum_{i=1}^{n} a_i X_{t_i}\right)^2\right] \geq 0.\end{aligned}$$

Ora, un teorema sulle trasformate di Fourier dice che la trasformata di una funzione *definita* non negativa, è una funzione a valori non negativi.

3.6.1 Esempio: il white noise

Per esso abbiamo dimostrato in passato che

$$R(n) = \sigma^2 \cdot \delta(n)$$

quindi

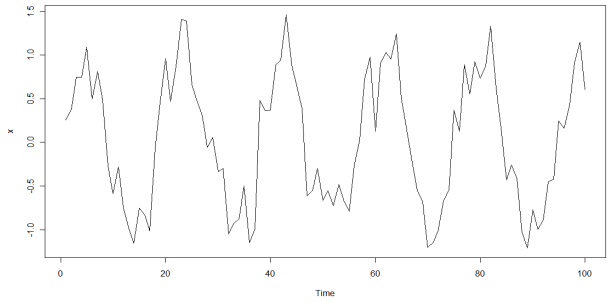
$$S(\omega) = \frac{\sigma^2}{\sqrt{2\pi}}, \qquad \omega \in \mathbb{R}.$$

La PSD è costante. Da questo deriva il nome *white noise* (un rumore con tutte le componenti di Fourier egualmente attive, come lo è lo spettro della luce bianca, approssimativamente si intende).

3.6.2 Esempio: serie periodica perturbata.

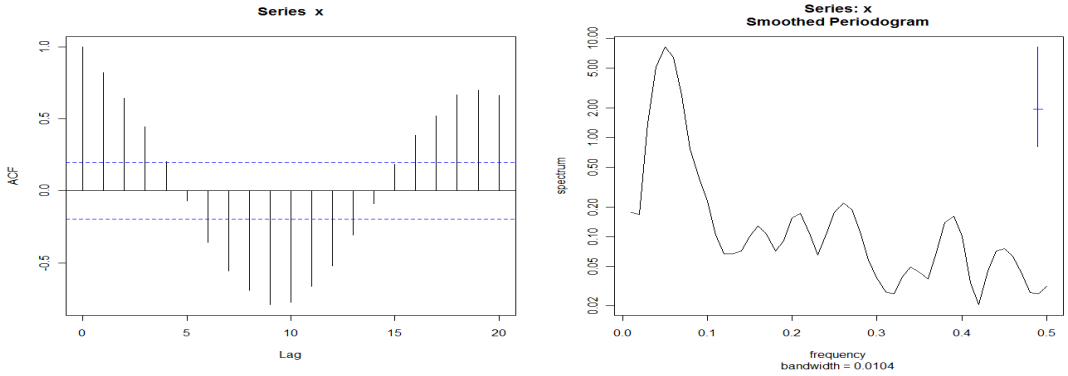
Descriviamo questo esempio solo numericamente, ma si riveda il paragrafo sulle trasformate generalizzate, per un confronto di idee. Tramite il software R produciamo la seguente serie temporale

```
t <- 1:100
x<- sin(t/3)+0.3*rnorm(100)
ts.plot(x)
```



La funzione di autocorrelazione empirica, mostra già una notevole periodicità, confermata dalla PSD numerica, opportunamente smussata dallo specifico algoritmo usato da R:

```
par(mfrow=c(1,2)); acf(x), spectrum(y,span=c(2,3))
```



3.6.3 Noise di tipo pink, brown, blue, violet

In certe applicazioni si incontrano PSD di tipo speciale a cui sono stati dati nomi analoghi a white noise. Usiamo la dizione inglese anche per essi. Si rammenti che il white noise ha PSD costante. Il *pink noise* ha PSD

$$S(f) \sim \frac{1}{f}.$$

Il brown noise:

$$S(f) \sim \frac{1}{f^2}.$$

Il blue noise:

$$S(f) \sim f \cdot 1_\Lambda$$

ed il violet noise

$$S(f) \sim f^2 \cdot 1_\Lambda$$

dove Λ è tale che $0 < \Lambda < 1$, e come sempre

$$1_\Lambda(f) = \begin{cases} 1 & \text{se} \quad 0 \le f \le \Lambda \\ 0 & \text{altrimenti} \end{cases} .$$

3.6.4 Il teorema di Wiener-Khinchin

Il seguente teorema è è spesso enunciato senza ipotesi precise. Una delle ragioni è che si può dimostrare a diversi livelli di generalità, con diversi significati dell'operazione di limite (si tratta di un limite di fuzioni). Daremo ora un enunciato rigoroso sotto ipotesi molto precise sulla fuzione di autocorrelazione $R(n)$, dimostrando una convergenza piuttosto forte. L'ipotesi (un po' strana, ma soddisfatta in tutti i nostri esempi) è:

$$\sum_{n\in\mathbb{N}} R(n)^p < \infty \text{ per qualche } p \in (0,1) . \tag{3.7}$$

Osservazione 62 *L'ipotesi precedente implica*

$$\sum_{n\in\mathbb{N}} |R(n)| < \infty$$

in quanto

$$\begin{aligned} \sum_{n\in\mathbb{N}} |R(n)| &= \sum_{n\in\mathbb{N}} |R(n)|^p \, |R(n)|^{1-p} \\ &\le \sup_{n\in\mathbb{N}} |R(n)|^{1-p} \sum_{n\in\mathbb{N}} |R(n)|^p < \infty \end{aligned}$$

(la successione è limitata essendo infinitesima, come conseguenza del fatto che $\sum_{n\in\mathbb{N}} |R(n)|^p < \infty$*). Quindi, sotto tale ipotesi, sappiamo che* $\frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} e^{-i\omega n} R(n)$ *converge uniformemente a* $S(\omega)$ *per* $\omega \in [0, 2\pi]$.

Teorema 29 (Wiener-Khinchin) *Se* $(X(n))_{n\in\mathbb{Z}}$ *è un processo stazionario in senso lato che soddisfa l'ipotesi (3.7), allora*

$$S(\omega) = \lim_{N\to\infty} \frac{1}{2N+1} E\left[\left| \widehat{X}_{2N}(\omega) \right|^2 \right] .$$

Il limite è uniforme in $\omega \in [0, 2\pi]$. *Qui* X_{2N} *è il processo troncato* $X \cdot 1_{[-N,N]}$. *In particolare, se ne deduce che* $S(\omega)$ *è reale non-negativa.*

Proof. Diamo prima l'idea euristica su cui è basata la dimostrazione. I dettagli verranno sviluppati nel seguito.

Per definizione di $R(t)$ e per stazionarietà del processo, $R(t) = E[X(t+n)X(n)]$ per ogni n, quindi se sommiamo questa uguaglianza $2N+1$ volte, per $|n| \leq N$, troviamo

$$R(t) = \frac{1}{2N+1} \sum_{|n|\leq N} E[X(t+n)X(n)]$$

per ogni valore di $t \in \mathbb{Z}$. Quindi anche

$$R(t) = \frac{1}{2N+1} E\left[\sum_{|n|\leq N} X(t+n)X(n)\right].$$

Eseguiamo ora la trasformata rispetto a t di ambo i membri: siccome $S(\omega) = \widehat{R}(\omega)$, usando poi la linearità della trasformata, troviamo

$$S(\omega) = \frac{1}{2N+1} E\left[\mathcal{F}\left(\sum_{|n|\leq N} X(t+n)X(n)\right)(\omega)\right]. \tag{3.8}$$

Abbiamo cambiato valore atteso e trasformata, cosa che si può mostrare essere lecita. Fino a qui (accettando questo scambio, che è basato su teoremi esterni al corso) è tutto rigoroso.

L'espressione $\sum_{|n|\leq N} X(t+n)X(n)$ è simila alla convoluzione. Se fosse $\sum_{n\in\mathbb{Z}} X(t+n)X(n)$ sarebbe esattamente $(X * X)(t)$. Ma attenzione a non pensare che sia approssimativamente corretto sostituire $\sum_{|n|\leq N} X(t+n)X(n)$ con $\sum_{n\in\mathbb{Z}} X(t+n)X(n)$. Se lo facessimo, avremmo che il secondo membro della (3.8) sarebbe

$$\frac{1}{2N+1} E[\mathcal{F}(X*X)(\omega)] = \frac{1}{2N+1} E\left[\left|\widehat{X}(\omega)\right|^2\right]$$

che è assurda, in quanto valendo per ogni N implicherebbe al limite $S(\omega) = 0$. L'errore in quest'approssimzione sta nel fatto che $(X*X)(t)$ non ha senso tradizionale, essendo convoluzione di serie storiche (realizzazioni del processo X) che non stanno in l_2 (perché si tratta di un processo stazionario, si vedano i commenti fatti in precedenza).

Più ragionevole è approssimare $\sum_{|n|\leq N} X(t+n)X(n)$ con

$$\sum_{n\in\mathbb{Z}} X_{2N}(t+n)X_{2N}(n) = (X_{2N} * X_{2N})(t).$$

Con tale approssimazione troviamo a secondo membro della (3.8) sarebbe

$$\frac{1}{2N+1} E[\mathcal{F}(X_{2N}*X_{2N})(\omega)] = \frac{1}{2N+1} E\left[\left|\widehat{X}_{2N}(\omega)\right|^2\right].$$

Questo è il risultato voluto, tenendo presente che, siccome abbiamo svolto unapprossimazione nel sostituire $\sum_{|n|\leq N} X(t+n)X(n)$ con $(X_{2N}*X_{2N})(t)$, invece di un'identità esatta troviamo un'identità valida solo al limite, appunto come è formulata nell'enunciato del teorema. Questa è l'idea; ora si tratta di renderla rigorosa esaminando il resto nell'approssimazione precedente.

Passo 1. Ripartiamo quindi dall'identità (3.8). Introduciamo il resto $\rho_{N,t}$ definito da

$$\sum_{|n|\leq N} X(t+n)\,X(n) = \sum_{n\in\mathbb{Z}} X_{2N}(t+n)\,X_{2N}(n) + \rho_{N,t}$$

ed otteniamo per linearità

$$\begin{aligned} S(\omega) &= \frac{1}{2N+1} E\left[\mathcal{F}\left(X_{2N} * X_{2N}\right)(\omega)\right] + r_N(\omega) \\ &= \frac{1}{2N+1} E\left[\left|\widehat{X}_{2N}(\omega)\right|^2\right] + r_N(\omega) \end{aligned}$$

dove

$$r_N(\omega) = \frac{1}{2N+1} E\left[\mathcal{F}\left(\rho_{N,t}\right)(\omega)\right].$$

Il teorema sarà dimostrato se mostriamo che $r_N(\omega)$ converge a zero uniformemente in $\omega \in [0, 2\pi]$. A questo scopo dobbiamo esplicitare $\rho_{N,t}$ e quindi $r_N(\omega)$.

Passo 2. Vale

$$\begin{aligned} \rho_{N,t} &= \sum_{|n|\leq N} X(t+n)\,X(n) - \sum_{n\in\mathbb{Z}} X_{2N}(t+n)\,X_{2N}(n) \\ &= \sum_{n\in\Delta(N,t)} X(t+n)\,X(n) \end{aligned}$$

dove ora tenteremo di descrivere l'insieme di indici $\Delta(N,t)$.

Per $0 \leq t \leq 2N$ vale

$$\sum_{n\in\mathbb{Z}} X_{2N}(t+n)\,X_{2N}(n) = \sum_{n=-N}^{N-t} X(t+n)\,X(n)\ .$$

Per $-2N \leq t < 0$ vale

$$\sum_{n\in\mathbb{Z}} X_{2N}(t+n)\,X_{2N}(n) = \sum_{n=-N-t}^{N} X(t+n)\,X(n)\ .$$

Infine, per $t > 2N$ o $t < -2N$, vale $X_{2N}(t+n)\,X_{2N}(n) = 0$ per ogni n. In generale,

$$\sum_{n\in\mathbb{Z}} X_{2N}(t+n)\,X_{2N}(n) = \sum_{n\in[N_t^-, N_t^+]} X(t+n)\,X(n).$$

dove

$$[N_t^-, N_t^+] = \begin{cases} \emptyset & \text{se} \quad t < -2N \\ [-N-t, N] & \text{se} \quad -2N \leq t < 0 \\ [-N, N-t] & \text{se} \quad 0 \leq t \leq 2N \\ \emptyset & \text{se} \quad t > 2N \end{cases}$$

Quindi

$$\begin{aligned}\rho_{N,t} &= \sum_{|n|\leq N} X(t+n)X(n) - \sum_{n\in[N_t^-,N_t^+]} X(t+n)X(n)\\ &= \sum_{n\in\Delta(N,t)} X(t+n)X(n)\end{aligned}$$

dove

$$\Delta(N,t) = [-N,N]\setminus[N_t^-,N_t^+]$$

o esplicitamente

$$\Delta(N,t) = \begin{cases} [-N,N] & \text{se} \quad t<-2N\\ [-N,-N-t-1] & \text{se} \quad -2N\leq t<0\\ \emptyset & \text{se} \quad t=0\\ [N-t+1,N] & \text{se} \quad 0<t\leq 2N\\ [-N,N] & \text{se} \quad t>2N\end{cases}$$

Passo 3. Resta ora da dimostrare che

$$\begin{aligned}r_N(\omega) &= \frac{1}{2N+1}E\left[\mathcal{F}\left(\rho_{N,t}\right)(\omega)\right]\\ &= \frac{1}{2N+1}E\left[\mathcal{F}\left(\sum_{n\in\Delta(N,t)} X(t+n)X(n)\right)(\omega)\right]\\ &= \mathcal{F}\left(\frac{1}{2N+1}\sum_{n\in\Delta(N,t)} E\left[X(t+n)X(n)\right]\right)(\omega)\end{aligned}$$

converge a zero uniformemente in $\omega\in[0,2\pi]$ (come sopra, affermiamo senza dimostrazione che è lecito scambiare valore atteso e trasformata).

L'ipotesi $\sum_{n\in\mathbb{N}} R(n)^p < \infty$ permette di dire che esiste una successione $\varepsilon_n > 0$, con $\varepsilon_n \to 0$, tale che

$$\sum_{n\in\mathbb{N}} \frac{R(n)}{\varepsilon_n} < \infty.$$

Ad esempio basta prendere $\varepsilon_n = R(n)^{1-p}$ se $R(n) > 0$, $\varepsilon_n = 1/n$ se $R(n) = 0$. Useremo l'esistenza di ε_n tra un momento.

Scriviamo

$$\sum_{n\in\Delta(N,t)} E\left[X(t+n)X(n)\right] = \sum_{n\in\Delta(N,t)} R(t) = \varepsilon_{|t|}\frac{R(t)}{\varepsilon_{|t|}}\left|\Delta(N,t)\right|$$

dove $|\Delta(N,t)|$ è la cardinalità di $\Delta(N,t)$. Se $(2N+1)\wedge|t|$ indica il più piccolo tra $(2N+1)$ e $|t|$, vale

$$\left|\Delta(N,t)\right| = (2N+1)\wedge|t|$$

quindi

$$\frac{1}{2N+1}\left|\sum_{n\in\Delta(N,t)} E\left[X\left(t+n\right)X\left(n\right)\right]\right| = \frac{\left|R\left(t\right)\right|}{\varepsilon_{|t|}}\frac{\left(\left(2N+1\right)\wedge\left|t\right|\right)\varepsilon_{|t|}}{2N+1}.$$

Dato $\delta>0$, sia t_0 tale che $\varepsilon_{|t|}\leq\delta$ per ogni $t\geq t_0$. Prendiamo $N_0\geq t_0$ tale che $\frac{t_0}{2N+1}\leq\delta$ per ogni $N\geq N_0$. Non è restrittivo assumere $\varepsilon_{|t|}\leq 1$ per ogni t. Allora, per $N\geq N_0$, se $t\leq t_0$ vale

$$\frac{\left(\left(2N+1\right)\wedge\left|t\right|\right)\varepsilon_{|t|}}{2N+1}\leq\frac{t_0\varepsilon_{|t|}}{2N+1}\leq\frac{t_0}{2N+1}\leq\delta$$

e se $t\geq t_0$ vale

$$\frac{\left(\left(2N+1\right)\wedge\left|t\right|\right)\varepsilon_{|t|}}{2N+1}\leq\frac{\left(\left(2N+1\right)\wedge\left|t\right|\right)}{2N+1}\delta\leq\delta.$$

Abbiamo dimostrato la seguente affermazione: per ogni $\delta>0$ esiste N_0 tale che

$$\frac{\left(\left(2N+1\right)\wedge\left|t\right|\right)\varepsilon_{|t|}}{2N+1}\leq\delta$$

per ogni $N\geq N_0$, uniformemente in t. Quindi anche

$$\frac{1}{2N+1}\left|\sum_{n\in\Delta(N,t)} E\left[X\left(t+n\right)X\left(n\right)\right]\right|\leq\frac{\left|R\left(t\right)\right|}{\varepsilon_{|t|}}\delta$$

per ogni $N\geq N_0$, uniformemente in t. Quindi

$$\left|r_N\left(\omega\right)\right| = \left|\frac{1}{2N+1}\frac{1}{\sqrt{2\pi}}\sum_{t\in\mathbb{Z}} e^{-i\omega t}\left[\sum_{n\in\Delta(N,t)} E\left[X\left(t+n\right)X\left(n\right)\right]\right]\right|$$

$$\leq\frac{1}{2N+1}\frac{1}{\sqrt{2\pi}}\sum_{t\in\mathbb{Z}}\left|\sum_{n\in\Delta(N,t)} E\left[X\left(t+n\right)X\left(n\right)\right]\right|\leq\frac{1}{\sqrt{2\pi}}\sum_{t\in\mathbb{Z}}\frac{\left|R\left(t\right)\right|}{\varepsilon_{|t|}}\delta=\frac{C}{\sqrt{2\pi}}\delta$$

dove $C=\sum_{t\in\mathbb{Z}}\frac{|R(t)|}{\varepsilon_{|t|}}<\infty$. Questa è la definizione di $\lim_{N\to\infty} r_N\left(\omega\right)=0$ uniformemente in $\omega\in[0,2\pi]$. La dimostrazione è completa. ■

Questo teorema fornisce un'interpretazione della PSD. La trasformata di Fourier $\widehat{X}_T\left(\omega\right)$ identifica la struttura frequenziale del segnale. Il quadrato $\left|\widehat{X}_T\left(\omega\right)\right|^2$ elimina l'informazione riguardante la fase e mantiene quella riguardante l'ampiezza, ma nel senso dell'energia (il quadrato). E' il cosidetto spettro dell'energia. Così, in base al teorema, la PSD relativa al valore ω è l'ampiezza quadratica media della componente a frequenza $f=\frac{\omega}{2\pi}$.

Per questo la PSD è un buon strumento per identificare componenti oscillatorie in una serie temporale ed osservare la loro ampiezza. Dalla PSD, si può avere una prima impressione nelle fasi preliminari dell'analisi di una serie storica.

Osservazione 63 *Sotto ipotesi più forti che includono anche l'ergodicità, si può dimostrare che*

$$S(\omega) = \lim_{N\to\infty} \frac{1}{2N+1} \left| \widehat{X}_{2N}(\omega) \right|^2$$

senza valore atteso, ad esempio in probabilità. Si noti che $\frac{1}{2N+1} \left| \widehat{X}_{2N}(\omega) \right|^2$ *è una quantità aletoria mentre il suo limite è deterministico.*

Capitolo 4

Analisi e Previsione di Serie Storiche

4.1 Introduzione

La teoria dei processi stocastici, tramite concetti quali autocorrelazione, stazionarietà, gaussianità, e così via, fornisce schemi e idee con cui guardare agli esempi in modo matematico. In questo capitolo tenteremo di applicare queste idee ad esempi concreti. Prenderemo in esame esempi reali di serie storiche, di ambito economico/sociale/industriale (generalmente reperibili su siti quali Istat ed Eurostat), ponendoci il problema di capirne le caratteristiche e prevederne i valori futuri. In parte le nostre analisi poggeranno sui fondamenti di teoria dei processi stocastici, ma senza dubbio dovremo assumere un atteggiamento pragmatico ed accettare l'introduzione di idee intuitive, di metodi a volte basati forse più sul buon senso che sulla teoria. In sintesi:

- *la previsione di serie storiche è un'arte in cui si devono usare tutte le idee utili senza preconcetti.*

In questo primo paragrafo cerchiamo di enucleare alcune idee generali, prima di addentrarci nei metodi più elaborati e specifici che ci offre la teoria delle serie storiche. Per inciso, visto che nel capitolo teorico sui processi stocastici il termine serie storica ha avuto vari significati, specifichiamo che qui, con questo termine, si intende una sequenza finita

$$x_1, ..., x_n$$

di numeri; di solito saranno numeri ottenuti tramite osservazioni di un fenomeno reale, valori relativi ad una singola grandezza (fisica, economica ecc.), quindi si immagina che siano collegati tra loro, abbiano una qualche logica oltre ad elementi di casualità. Essi rappresentano il passato (o se vogliamo, l'ultimo x_n può essere il presente) ed il nostro scopo principale è quello di prevedere i valori futuri di questa grandezza, cioè capire come prosegue nel futuro questa serie storica. Nei problemi reali, spesso si hanno a disposizione più serie storiche, relative a grandezze collegate, ed altre informazioni quantitative e numeriche su fenomeni e grandezze collegate; saper usare questa ricchezza sarebbe fondamentale per arricchire la

previsione. Faremo dei cenni a questa possibilità che però è difficile da implementare. In linea di massima lo scopo che ci poniamo primariamente è quello di:

- *prevedere i valori futuri della serie storica* $x_1, ..., x_n$ *basandoci su tali valori passati*

e non su una più ampia mole di informazioni. E' una visione chiaramente restrittiva ma accessibile. Devieremo da questa visione restrittiva solo nella Sezione 4.4.

Una prima osservazione, sul problema di previsione appena enunciato, osservazione che purtroppo pone un'enorme restrizione, è la seguente:

- *un metodo puramente matematico non può basare le previsioni che sulla ripetizione di ciò che è avvenuto in passato, cioè sull'analisi del passato e la sua riproduzione nel futuro.*

Un mago può azzardare previsioni innovative e inattese, ma non un algoritmo matematico, salvo introdurre in esso degli elementi di aleatorietà che aggiungano variazioni casuali alla previsione, ma così facendo ci staremmo affidando al caso, non alle opportunità che offre la matematica. Un algoritmo matematico può solo analizzare i dati passati e riprodurli nel futuro. Entrambe queste fasi però, *analisi e riproduzione*, possono essere fatte in vario modo e qui entra l'arte dell'analista che conoscendo i vari metodi, giudica quale può essere più conveniente, prova, analizza per quanto può i risultati dei vari metodi ecc. (il futuro è ignoto, quindi come possiamo giudicare se un metodo sta facendo una buona previsione? vedremo dei surrogati della possibilità di giudicare le previsioni di un metodo).

Solo per dare l'idea dei gradi di libertà nelle mani dell'analista, citiamo il seguente punto fondamentale:

- *conviene usare tutta la serie storica* $x_1, ..., x_n$ *oppure solo una sua parte più recente?*

Se si osservano i valori mensili di certe grandezze economiche relative a periodi di tempo molto lunghi, es. dal 1980 ad oggi, si vede chiaramente che sono accadute varie fasi molto diverse tra loro. Allora, per prevedere il prossimo mese, è sensato utilizzare anche i dati di venti anni fa? Può essere facile (anche se non necessariamente giusto) rispondere di no, che non è sensato; ma allora, dove si taglia? Quale sotto-stringa

$$x_k, ..., x_n$$

si prende? C'è completa arbitrarietà, a scelta dell'analista. Dev'essere chiaro che la matematica non c'entra in queste scelte. Non si deve attribuire alla matematica un valore che non ha. La matematica offrirà degli algoritmi; a quale serie applicarli lo dobbiamo decidere noi (così come dovremo fare altre scelte, tra i numerosi algoritmi ad esempio). Naturalmente la matematica potrebbe venire in aiuto, nel fare queste scelte, se abbiamo la pazienza di affrontare nel dettaglio ogni segmento di questa attività previsiva, ma questo è dispendioso in termini di tempo e fatica concettuale. Ad esempio, per decidere quale sotto-sequenza $x_k, ..., x_n$ usare, una scelta ad occhio e col buon senso magari è la scelta migliore e richiede pochi secondi, ma volendo si potrebbero fare delle analisi matematiche su tutte le possibili serie $x_k, ..., x_n$, al variare di k, per capire quali sono *più omogenee*, quindi più rappresentative

di una fase attuale univoca del processo fisico, economico ecc. in questione. Non c'è però una bacchetta magica matematica per giudicare l'omogeneità di una serie; ci possono essere varie idee buone e parziali; per cui alla fine, visto il tempo che costano e la parzialità dei risultati, magari si decide di fare il taglio della serie ad occhio.

Va sempre tenuto presente che

- *il processore più potente resta la nostra mente.*

In altre parole, i migliori previsori siamo noi, non i metodi matematici. Questa frase però è vera solo in parte o sotto opportune condizioni. Ad esempio, non è detto che noi, come strumento di previsione, siamo abbastanza istruiti, abbastanza allenati. Potrebbero sfuggirci, ad esemio nella fase di analisi dei dati passati, degli aspetti che non sfuggono ad un metodo matematico. Questo non significa che esso sia superiore, solo che noi non ci siamo allenati abbastanza, non abbiamo ragionato su un numero sufficiente di esempi, comparando varie situazioni. Un enorme vantaggio che abbiamo sui metodi matematici è la capacità innata di mettere in gioco tantissime variabili contemporaneamente. Prima si diceva che studieremo la previsione di una serie basata solo sui valori di quella serie. Se ci affidiamo al nostro intuito invece che alla matematica e conosciamo il problema concreto da cui è stata estratta quella serie, non possiamo fare a meno di utilizzare un sacco di altre informazioni, per fare la previsione, oltre a quelle che provengono dai dati passati della serie stessa. Questa potenza però pone anche dei limiti: da un lato potrebbe portarci a dare previsioni meno obiettive; dall'altro potrebbe deviare la nostra attenzione da un'attenta analisi dei dati passati della serie, troppo confindenti nell'uso intuitivo di varie *informazioni esogene*. Torniamo al punto, quindi, che potremmo non esserci allenati abbastanza, potremmo fare peggio di un algoritmo matematico non perché non abbiamo la capacità di fare altrettanto bene ma perché siamo distratti da troppe informazioni, non sappiamo di dover guardare ad alcune di esse ecc. In conclusione:

- *fermo restando che il nostro metro di giudizio può essere il migliore, completiamo la nostra capacità previsiva tramite le informazioni offerete da metodi matematici.*

Va aggiunto che, se proviamo in pratica a fare previsioni ad occhio, ci scontriamo con un banale problema un po' psicologico: magari abbiamo chiarissimo in che direzione deve andare la previsione, nella sostanza, ma stabilire il valore esatto del numero x_{n+1} ci mette in imbarazzo. Intuiamo con chirezza, magari, che esso sarà un po' maggiore di x_n (per fare un esempio), ma ci blocchiamo di fronte al problema di dichiarare il suo valore numerico. Siamo degli ottimi *previsori vaghi*. Abbiamo capacità impareggibili di previsione vaga, dei più svariati fenomeni, ma se dovessimo tradurle in singoli numeri precisi ci bloccheremmo, cominceremmo a dire che però non siamo sicuri, che anche un po' di più o un po' di meno va bene lo stesso. Facciamo continuamente previsioni vaghe, senza accorgercene; si pensi a quando attraversiamo una strada; ma se dovessimo quantificare velocità o distanza dei veicoli, saremmo in grande difficoltà. Allora, nella sua crudezza, l'algoritmo matematico ci toglie da questo impaccio psicologico. Dev'essere chiaro che il valore preciso offerto dall'algoritmo non ha nessuna proprietà di assolutezza, non c'è nessun motivo per credere che sia più giusto di uno molto vicino a lui, solo che l'algoritmo lo pronucia e noi non riusciremmo a farlo.

Se, giustamente, questi ultimi commenti possono lasciare insoddisfatti perché sforano nello psicologico, oppure perché sollevano (correttamente) il dubbio che il risultato numerico di un algoritmo matematico di previsione non vada preso alla lettera, li si prenda allora come il punto di partenza di un approfondimento della teoria a cui accenneremo:

- *quando si fa una previsione, andrebbe dichiarato un intervallo di confidenza;*

ed anche senza far riferimento a questo concetto prettamente statistico,

- *andrebbe dichiarato un ventaglio di valori possibili, non un singolo valore, magari corredando di informazioni circa la zona più probabile del ventaglio.*

E' un discorso non banale da quantificare, ma corrisponde proprio al fatto che con la nostra intuizione facciamo previsioni a ventaglio, non puntuali, di solito, corredate da diverse plausibilità dei valori del ventaglio; tutto però a livello intuitivo, informale, lievemente impreciso. Gli algoritmi matematici possono permetterci di quantificare queste opinioni.

Riprendiamo il discorso fatto sopra circa le due fasi di *analisi e riproduzione*, che i diversi algoritmi eseguono in modo diverso. Che significa analizzare una serie data? Innanzi tutto, vale sempre la pena di iniziare con un'analisi non strettamente matematica, ma fatta col buon senso, fatta quindi dall'analista, ad occhio. L'analista deve

- *raffigurare la serie storica*

nel modo più chiaro possibile e soffermarsi a

- *guardarla, valore per valore, a gruppi di valori, cercando somiglianze, ripetizioni, anomalie, riconoscendo anche significati se possibile.*

Ad esempio, se si tratta dei valori mensili delle vendite di un prodotto, e si vede che a luglio c'è un picco (un valore più alto del solito), tutti gli anni, questa è un'informazione da registrare, da tener presente nel seguito, forse interpretabile in modo ovvio se conosciamo la natura del prodotto in questione. Molto importante, ad esempio, può essere accorgersi che certi picchi, pur ripetendosi di anno in anno, però cambiano lievemente di mese (un anno a luglio, l'altro a giugno). Si deve ammettere che tra le ingenuità maggiori che si riscontrano in chi esegue analisi matematiche di serie storiche c'è quella di non essersi soffermato a sufficienza a guardare col buon senso la serie, cercando di cogliere il maggior numero di infomazioni possibili.

Certamente però bisogna anche essere istruiti circa quali informazioni è bene cercare. Qui si apre il discorso su cosa sia la fase di *analisi di una serie storica*. L'analisi intuitiva ad occhio, a cui abbiamo appena accennato può voler dire molte cose, come abbiamo detto, tutte forse molto importanti. L'analisi più propriamente matematica, invece, ha alcune linee guida di carattere generale. La prima, indubbiamente, è

- *capire se c'è un trend, e cercare di isolarlo, quantificarlo.*

Il trend è ciò che non hanno le realizzazioni dei processi stazionari. E' un concetto un po' vago da definire ma molto intuitivo. Se una serie, pur fluttuando, ha una tendenza (=trend) a crescere, o a decrescere, si dice che ha un trend, appunto. Magari il trend cambia nel tempo: per un po' c'è una tendenza a crescere, poi diventa a decrescere. Qui si annida la vaghezza del concetto: quanto lungo dev'essere il periodo di tendenza alla crescita, per parlare di trend? Cinque valori consecutivi in crescita sono un trend o solo una fluttuazione? Ovviamente un processo stazionario può avere, in una sua realizzazione, cinque valori in crescita. Anche dieci, ma la cosa diventa sempre più improbabile. Isomma, col buon senso, di fronte ad una serie specifica, cercheremo di capire se certe manifestazioni di crescita o decrescita di sottosequenze di valori vanno interpretate come trend oppure come fluttuazioni. Regole vincolanti non ce ne possono essere (verrebbero contraddette da certe realizzazioni di certi processi stazionari, pur poco probabili). Ovviamente, in vari casi di serie concrete del mondo reale, tutta la serie ha un trend evidente, quindi poco fa stavamo discutendo delle situazioni più intricate. Molto spesso, cioè, il trend è chiarissimo.

Il trend è un'informazione spesso importante di per sé, la prima informazione da evidenziare e comunicare ai nostri interlocutori, se stiamo analizzando una serie storica per qualcuno. Ogni giorno, se apriamo i giornali, ci viene parlato del trend di certe grandezze economiche o finanziarie. E' l'informazione per eccellenza.

Oltre che informazione di *analisi* della serie, essa è forndamentale per fare *previsioni*. O meglio, questo è il primo esempio concettuale in cui vediamo che per fare previsioni serve aver fatto analisi. Se conosciamo il trend e sappiamo *estrapolarlo*,

- *l'estrapolazione del trend costituisce già una prima previsione.*

Per estrapolazione si intende un qualsiasi procedmento che prolunga una curva, nota su un intervallo $[0, T]$, oltre il valore T. Se la curva, su $[0, T]$, è un segmento di retta, ovviamente come estrapolazione si prenderà il proseguimento della retta. Lo stesso di fa se la curva è un polinomio, almeno se di grado basso, come una parabola. Meno univoco è come estrapolare una curva più complessa o solamente un numero finito di punti, che non cadano in modo banale su una retta o un parabola. Comunque ci sono vari metodi, e ne studieremo alcuni.

Anche la fase di analisi del trend sarà fortemente automatizzata: oltre a riconoscerne l'esistenza ad occhio, avremo vari strumenti che lo mettono in evidenza e sopratutto lo quantificano, cioè forniscono a partire da $x_1, ..., x_n$ una nuova sequenza $\overline{x}_1, ..., \overline{x}_n$ che sia il trend della precedente. Essa non è univoca, ogni algoritmo trova la sua univocamente secondo una certa logica, ma ci sono varie logiche, vari algoritmi. Come sempre, quindi, si tratta di ausilii all'analisi, non di verità assolute. All'analista è sempre rimandato il compito di giudicare e scegliere.

A parte il trend, la cosa poi più importante da cercare in una serie storica sono

- *le ripetizioni, le periodicità, la stagionalità, le ricorrenze cicliche, le somiglianze di un periodo con un altro.*

Di nuovo, alcune di queste si vedono ad occhio, altre possono essere evidenziate da algoritmi. Come per il trend, esse vanno identificate, quantificate e poi riprodotte nel futuro. Mentre per il trend si tratta di estrapolare una tendenza, qui si tratta di ripetere un comportamento, di copiarlo dal passato al futuro.

Queste sono alcune delle linee guida. Che fare ora, di fronte ad una serie storica: visualizzarla, meditarla, eventualmente tagliarla, identificare e quantificare trend e ripetizioni, estrapolare il trend e ricopiare le ripetizioni? Questa è una strada, tutt'altro che trascurabile. Però ce ne sono altre, che in sostanza sono l'automatizzazione di tutto questo in un singolo algoritmo (esclusa la fase di visualizzazione e meditazione ad occhio). Due grandi classi di algoritmi si propongono questo scopo: i modelli ARIMA e i metodi riassunti sotto il nome Holt-Winters. I modelli regressivi sono poi una variante degli ARIMA che permette di inglobare fattori esogeni. Questi algoritmi sono basati sull'idea di *modello*:

- *si cerca un modello ricorsivo aderente ai dati, che ne cattura la struttura, e lo si usa per la previsione.*

I modelli ricorsivi hanno caratteristiche specifiche adatte a catturare trend e ripetizioni (periodicità, stagionalità), ma, ameno nel caso degli ARIMA, ache altri aspetti strutturali magari meno evidenti (forse però anche meno comuni nella realtà).

Volendo poi ci sono altri metodi oltre ad ARIMA ecc., come quelli markoviani (che tratteremo in un capitolo successivo), quelli legati alle reti neurali ed altri ancora. Non li tratteremo qui. Iniziamo quindi questo capitolo studiando un po' di teoria degli ARIMA, di Holt-Winters dei metodi regressivi. Nello studio della teoria si vedrà che essa è ispirata alle idee esposte in questa introduzione. Poi nella sezione di esercizi sulle serie storiche metteremo in pratica sia la versione diretta della ricerca di trend e ripetizioni e loro uso per la predizione, sia i metodi automatici che l'implementano tramite equazioni ricorsive.

4.1.1 Metodi elementari

Concludiamo questa sezione introduttiva menzionando alcuni metodi davvero elementari per la previsione, sottolineando anche i loro limiti.

Un metodo consiste semplicemente nel ripetere l'ultimo valore. Se la serie nota è $x_1, ..., x_n$, si prevede il prossimo valore ponendo (chiamiamo p_{n+1} la previsione)

$$p_{n+1} = x_n. \tag{4.1}$$

Un passo più elaborato è il metodo detto di *media mobile*: a due passi è

$$p_{n+1} = \frac{x_n + x_{n-1}}{2}$$

e si generalizza in modo ovvio a più passi ($\frac{x_n+x_{n-1}+x_{n-2}}{3}$ ecc.). Al limite ha anche senso la media complessiva:

$$p_{n+1} = \frac{x_n + x_{n-1} + ... + x_1}{n} \tag{4.2}$$

Su questa si potrebbe fare un ragionamento a parte. Quando si hanno dei dati storici di una grandezza come il volume mensile di vendite di un prodotto, rappresentati dalla serie storica $x_1, ..., x_n$, si può ignorare la struttura temporale e considerare $x_1, ..., x_n$ come un campione sperimentale estratto dalla v.a. $X =$ "volume mensile di vendite di quel prodotto". In senso stretto un campione dovrebbe avere proprietà di indipendenza delle componenti, in genere

violate per serie temporali, ma in mancanza di meglio accettiamo questa imprecisione e pensiamo a $x_1, ..., x_n$ come ad un campione sperimentale. Che possiamo prevedere allora per il prossimo valore p_{n+1}? La media del campione è la previsione più immediata. Ogni previsione che si discosti dalla media, ad esempio $\mu + 2\sigma$, non si capisce perché dovrebbe essere migliore. Casomai, si può discutere se invece della media non sia meglio la mediana. Oppure si può prendere la media teorica di una densità modellata sui dati, anche se nella maggior parte dei asi questo equivale a prendere la media aritmetica. Varianti a parte, usare la media aritmetica per prevedere p_{n+1} corrisponde al considerare $x_1, ..., x_n$ come un campione sperimentale piuttosto che una serie storica, ignrando la struttura temporale dei dati. In situazioni veramente molto aleatorie e poco strutturate (cioè senza trend e periodicità evidenti), questa strategia è quasi l'unica che ha senso (anche se ad esempio gli AR ce la mettono tutta per scoprire strutture nascoste). Aggiungiamo che, se si adotta questa strategia, cioè quella del calcolo della media dei dati, è più che mai doveroso corredare la previsione di un intervallo di confidenza.

Questi metodi sono tutti sottocasi degli AR che studieremo nella prossima sezione. Quando chiediamo al software di calcolare i coefficienti di un modello AR, i valori dei coefficienti presenti negli esempi appena descritti sono contemplati, quindi se non vengono scelti vuol dire che ci sono valori migliori dei coefficienti. Ad esempio, il software decide che invece di $p_{n+1} = \frac{x_n + x_{n-1}}{2}$ è meglio usare $p_{n+1} = 0.7 \cdot x_n + 0.3 \cdot x_{n-1}$, perché più aderente ai dati. Quindi, in linea teorica, usando gli AR si comprende anche l'uso di questi modelli elementari.

Tuttavia, va tenuto presente che il metodo della media mobile è di fatto un modello ben preciso. Sceglierlo significa credere che la cosa migliore per riassumere i dati passati sia fare la media aritmetica. Nelle serie finanziarie questo è ciò che spesso si pensa. La ragione è la loro enorme casualità, tale che ad ogni istante in un certo senso la serie si scorda del passato (non è proprio così) e può crescere o decrescere allo stesso modo. Allora, predire ad esempio il valore attuale (4.1) ha una sua logica, che può essere più veritiera della apparente struttura identificata dai modelli AR (struttura identificata cercando di fittare i dati meglio possibile, ma può accadere che la struttura riscontrata in quei dati sia finta, casuale essa stessa).

La scelta della media complessiva (4.2) ha una sua logica. Se si immagina che i dati non abbiano alcuna struttura temporale, siano cioè puri campioni casuali di una grandezza (come le misurazioni sperimentali di certe caratteristiche degli oggetti di un lotto), allora è inutile o addirittura fuorviante cercare forzatamente una struttura con un metodo come AR o Holt-Winters. Megli considerare i numeri della serie come un campione sperimentale di una singola grandezza aleatoria X e stimare dai dati la sua densità di probabilità o almeno le sue princpali grandezze medie (media e deviazione standard in primis) In una tale situazione, la miglior previsione del valore successivo, p_{n+1}, è la media di X, che abbiamo stimato con la media aritmetica (4.2). Anche la mediana può avere una sua logica.

Infine, se (come detto in precedenza) s ritiene che non tutta la serie storica sia rappresentativa della situazione presente ma solo na finestra recente di valori $x_k, ..., x_n$, e come poco fa si ritiene che la struttura temporale non sia interessante ma solo la distribuzione statistica dei valori, allora si cercherà di fittare una densità di probabilità ai soli valori $x_k, ..., x_n$, e la predizione del valore sucessivo p_{n+1} si farà con la media artimetica di questi numeri. Questa è una motivazione alla base del metodo di media mobile che lo può far preferire a metodi basati su modelli più elaborati ma magari meno veritieri, perché adattati ai valori specifici

della serie come se essi avessero una struttura temporale che non c'è (se pesiamo che non ci sia; qui entra il giudizio dell'analista).

4.1.2 Decomposizione di una serie storica

Poco sopra abbiamo sottolineato come trend e stagionalità siano le due caratteristiche principali da cerare di identificare, mettere in evidenza, per poter fare previsioni. Immaginiamo allora che valga una decomposizione di un processo X_n (o una serie storica) in tre componenti, secondo la formula

$$X_n = T_n + S_n + \varepsilon_n.$$

Le tre componenti non sono definibili in modo rigoroso ed univoco, quindi il discorso che stiamo facendo qui deve essere inteso come un discorso di sostanza, non una vera teoria.

L'idea è che la componente T_n racchiuda il trend, S_n la stagionalità, ed ε_n sia una *componente stazionaria*, magari un white noise, oppure un processo stazionario più complesso.

Idealmente, la strategia di analisi sarebbe: identificare trend e stagionalità in X_n, sottrarli, cioè calcolare

$$\varepsilon_n = X_n - T_n - S_n$$

ed analizzare poi il processo stazionario (o la serie storica) ε_n con strumenti propri dei processi stazionari, ad esempio i modelli ARMA.

Alcuni dei metodi che vedremo saltano però il primo passaggio e creano modelli validi direttamente per X_n, modelli che inglobano trend e stagionalità. Questo vale sia per certi ARIMA, sia per il metodo di Holt-Winters.

Se si volesse seguire la strada della decomposizione, bisogna trovare T_n e S_n, o meglio trovare dei possibili T_n e S_n, visto che non sono univocamente definibili.

Un modo per trovare T_n è quello della regressione, lineare o non lineare; non ci dilunghiamo su di esso ora perché usa strumenti che vanno appresi in altre parti del corso. Un'altro metodo semplice è quello della media mobile. Usando una finestra un po' ampia, il metodo della media mobile crea un profilo medio tra i dati. Si intende che lo dobbiamo usare sui dati noti a partire dall'inizio, non sugli ultimi per prevedere il futuro. Se $x_1, ..., x_n$ è la serie storica, e n_0 è la finestra della media mobile, usando i primi n_0 valori $x_1, ..., x_{n_0}$ si calcola la media $y_0 = \frac{x_1+...+x_{n_0}}{n_0}$. Poi, usando $x_{1+1}, ..., x_{n_0+1}$ si calcola la media $y_1 = \frac{x_{1+1}+...+x_{n_0+1}}{n_0}$, e così via, in generale

$$y_k = \frac{x_{1+k} + ... + x_{n_0+k}}{n_0}.$$

I numeri $y_0, y_1, ...$ hnno un grafico molto più regolare di $x_1, ..., x_n$, meno fluttuante, che può essere preso come trend. La regolarità del grafico aumenta con n_0 (per $n_0 = 1$ fluttua come la serie originaria, poi via via meno aumentando n_0).

C'è arbitrarietà di scelta anche circa il posizionamento temporale di questi valori. Ad esempio, se $x_1, ..., x_n$ sono i valori mensili di una grndezza a partire da gennaio 2000, e se abbiamo scelto $n_0 = 12$, il valore y_0 rappresenterà il trend di gennaio 2000, di dicembre 2000 (12 mesi dopo), di luglio 2000? A noi la scelta.

Trovato il trend T_n con un metodo che ci convinca, lo sottraiamo e consideriamo la nuova serie o processo

$$Z_n = X_n - T_n.$$

Ora bisogna identificare la stagionalità. Intanto va identificato il periodo P, cioè ogni quanto secondo noi le coe si ripetono (approssimativamente). A volte ci sono banali ragioni stagionali o economiche per decidere P (es. $P = 12$ in molte ovvie situazioni), a volte potremmo trovare P a partire dai dati osservando i picchi dell'autocorrelazione empirica, `acf`. Deciso P, un modo banale per calcolare la componente S_n è quello di fare la media dei valori sui periodi. Ad esempio, per una serie mensile relativa agli anni 2000, ... , 2008, il valore S_1 di gennaio si calcola facendo la media aritmetica dei valori di tutti i mesi di gennaio, e così via. Per S_{12+1}, S_{24+1}, si prendono gli stessi valori. La serie S_n così ottenuta sarà esattamente periodica.

A questo punto si può calcolare $\varepsilon_n = X_n - T_n - S_n$, raffigurarlo (se è una serie storica) e capire se è abbastanza stazionario. Ad esso si possono applicare i modelli AR.

Il software R mette disposizione due comandi molto pratici per eseguire una decomposizione di una serie: `decompose` e `stl`. A differenza del caso dei metodi ARIMA e HW, prenderemo questi metodi un po' empiricamente, senza svilupparne una vera teoria. Comunque, spendiamo due parole su di essi. Il comando `decompose` calcola il trend T_n col metodo della media mobile, facendo una scelta simmetrica per la collocazione temporale delle medie calcolate: usa per il calcolo di T_n una finestra centrata in n, aperta un po' a destra ed un po' a sinistra. Ottenuto il trend con la media mobile, lo sottrae e calcola S_n mediando i periodi, come abbiamo detto sopra. Concettualmente, `decompose` calcola il trend in modo *locale* (usando cioè una finestra), mentre calcola la stagionalità in modo *gobale* (usando cioè tutta la serie ed ottenendo un risultato sempre uguale sui diversi periodi).

Il comando `stl` invece effettua un calcolo locale sia del trend sia della stagionalità, con un complesso sistema iterativo che non decriviamo. Si vedrà l'effetto negli esempi.

4.1.3 La media di più metodi

L'esperienza mostra a volte che facendo la media tra varie previsioni, si ottiene una previsione migliore. Non c'è nessuna base teorica di questo fatto se non la seguente, che però è basata su ipotesi. Supponiamo che le diverse previsioni siano come le realizzazioni casuali di una grandezza aleatoria P, che ha come valor medio μ_P il valore giusto che vorremmo prevedere, ma che per una serie di accidenti casuali produce valori diversi da μ_P. In questa ipotesi, se

$$p_1, ..., p_n$$

sono varie previsioni, nel senso che sono un campione estratto da P, allora la loro media aritmetica $\overline{p} = \frac{p_1+...+p_n}{n}$ sarà una stima di μ_P migliore dei singoli valori $p_1, ..., p_n$ (pur essendoci magari tra i vari valori $p_1, ..., p_n$ alcuni più vicini a μ_P di quanto non sia $\overline{p}$, però non sappiamo quali siano e quindi non li possiamo scegliere).

Nella pratica, $p_1, ..., p_n$ sono ottenuti tramite algoritmi diversi, scelte diverse di opzioni e parametri di certi algoritmi. Si può immaginare che la variabilità di queste previsioni sia simile in natura alla variabilità di un campione estratto da una variabile aleatoria? Non è ovvio credere in questa ipotesi. Tuttavia, a volte il metodo della media tra previsioni funziona. decentemente.

In una cosa però cade il paragone col campione casuale. Mentre per un vero campione casuale $p_1, ..., p_n$ estratto da P non abbiamo alcun criterio di scelta di un singolo valore k che sia più vicino a μ_P di quanto non sia $\overline{p}$, nel caso dei valori $p_1, ..., p_n$ ottenuti da n algoritmi

diversi potremmo avere dei buoni motivi per ritenere che certe previsioni sono più attendibili di altre (ad esempio, vedremo dei metodi per giudicare gli algoritmi). Quindi non è ovvio preferire $\overline{p}$ ai più accreditati dei valori $p_1, ..., p_n$.

In quest'ultimo caso, idealmente potrebbe scattare un'idea ancora migliore. Se potessimo attribuire dei pesi

$$w_1, ..., w_n$$

(cioè dei numeri $w_i \in [0, 1]$ tali che $\sum_{i=1}^{n} w_i = 1$) ai vari algoritmi di previsione, che riflettano il nostro grado di giudizio sulla loro attendibilità, allora la media pesata

$$w_1 p_1 + ... + w_n p_n$$

sarebbe una buona previsione, che mescolerebbe la filosofia della media con la maggior imortanza data ad alcuni degli algoritmi.

Il problema pratico è quello di scegliere i pesi. Un modo può essere quello puramente soggettivo, magari ottenuto con un lavoro di gruppo: si attribuiscono i pesi soggettivamente. Può sembrare senza fondamento, ma si ricordi che comunque dovremmo fare delle scelte tra le varie previsioni, quindi è impossibile sfuggire alle scelte soggettive, e quella dei pesi può essere meno drastica della banale scelta di una previsione su tutte.

Altrimenti, in situazioni in cui le previsioni si eseguano più volte successivamente nel tempo, si può acquisire un giudizio circa la bontà dei vari algoritmi confrontando le loro previsioni coi dati reali accaduti (di fatto questo è simile ad uno dei metodi di giudizio che studieremo, solo che quello lo faremo sui dati noti, mentre ciò di cui stiamo parlando ora è un confronto coi dati futuri incogniti, che nel tempo diventano progressivamente noti). Si potrebbero allora raffinare via via dei pesi, magari riaggiornandoli ad ogni nuova previsione che diventa realtà, usando poi i pesi per le previsioni successive.

4.2 Modelli ARIMA

4.2.1 Modelli AR

Definizione 41 *Si chiama modello autoregressivo di ordine p, o modello AR(p), l'equazione lineare*

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t$$

e si chiama processo AR(p) la sua soluzione (più precisamente il termine viene di solito riferito alla soluzione stazionaria). Qui p è l'ordine, $\alpha_1, ..., \alpha_p$ sono i parametri o coefficienti (numeri reali), ε_t è il termine di errore, usualmente ipotizzato essere un white noise di intensità σ^2. Il modello viene considerato o sugli interi $t \in \mathbb{Z}$, quindi senza condizioni iniziali, o sugli interi non-negativi $t \in \mathbb{N}$. In questo caso, la relazione precedente inizia da $t = p$ e devono essere specificate delle condizioni iniziali per i valori di $X_0, ..., X_{p-1}$.

Esempio 83 *Nel capitolo sui processi stocastici abbiamo esaminato il caso più semplice, il modello AR(1)*

$$X_t = \alpha X_{t-1} + \varepsilon_t.$$

Quando $|\alpha| < 1$, $E[X_0] = 0$, $Var[X_0] = \frac{\sigma^2}{1-\alpha^2}$, *la soluzione risulta un processo stazionario in senso lato, ed anche in senso stretto se* X_0 *è gaussiana. Il coefficiente di autocorrelazione decade esponenzialmente:*

$$\rho(n) = \alpha^n.$$

Osservazione 64 *Anche se una formula generale per* $\rho(n)$ *non è così semplice per un generico AR(p), il decadimento esponenziale continua ad essere vero.*

Il modello precedente non contiene intercetta ed è adatto a descrivere situazioni a media nulla. Per descrivere casi a media non nulla si può considerare la seguente generalizzazione. Si può pensare che, se X_t è il processo che ci interessa, e μ è la sua media (ipotizzata per semplicità costante), allora $(X_t - \mu)$ è un processo a media nulla e per esso consideriamo il modello AR(p) introdotto sopra

$$(X_t - \mu) = \alpha_1 (X_{t-1} - \mu) + ... + \alpha_p (X_{t-p} - \mu) + \varepsilon_t.$$

Allora X_t soddisfa

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t + (\mu - \alpha_1\mu - ... - \alpha_p\mu) \\ &= \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + b + \varepsilon_t \end{aligned}$$

cioè un modello di tipo AR(p) ma con intercetta

$$b = (\mu - \alpha_1\mu - ... - \alpha_p\mu).$$

4.2.2 Esempi particolari

Nella sezione di esercizi sulle serie storiche, esamineremo situazioni concrete di carattere economico-sociale-gestionale. Le serie storiche considerate in tale ambito sono spesso le serie dei valori mensili di una grandezza economica, ed hanno a volte un carattere stagionale, cioè risultano maggiori sempre nelle stesse stagioni, di anno in anno. I modelli AR catturano alcune caratteristiche di queste serie e forniscono, come vedremo in quella sezione, un buon strumento per fare previsioni.

In vista di tali applicazioni, segnaliamo tre casi che spesso danno buoni risultati e ci fanno capire perché viene in mente di usare i modelli ricorsivi AR per descrivere serie storiche. Il primo è semplicemente il caso AR(1), già illustrato precedentemente, magari però con intercetta per avere maggior flessibilità:

$$X_n = \alpha X_{n-1} + b + \varepsilon_n.$$

Abbiamo già visto che per $b = 0$ ed $|\alpha| < 1$ c'è una soluzione stazionaria. Per altri valori di α e b si possono però avere comportamenti diversi, quindi il modello può essere usato per descrivere altre situazioni.

Esempio 84 *Ad esempio, si pensi al caso*

$$X_n = X_{n-1} + b + \varepsilon_n.$$

Vale, iterativamente,

$$\begin{aligned} X_n &= X_{n-2} + b + \varepsilon_{n-1} + b + \varepsilon_n \\ &= X_{n-2} + 2b + \varepsilon_{n-1} + \varepsilon_n \end{aligned}$$

e così via

$$X_n = X_0 + b \cdot n + \varepsilon_1 + ... + \varepsilon_n.$$

Questo mostra che X_n ha un trend lineare di coefficiente angolare b (infatti la somma $\varepsilon_1 + ... + \varepsilon_n$ è una random walk che oscilla tra positivi e negativi con valori assoluti vagamente attorno a $\sqrt{n}$, come già osservato nel capitolo sui processi stocastici; quindi $\varepsilon_1 + ... + \varepsilon_n$ non riesce a contrastare la tendenza lineare del termine $b \cdot n$).

Osservazione 65 *Nel pragrafo precedente avevamo visto che processi a media non nulla possono essere descritti da modelli AR con un'intercetta b. Nell'esempio appena visto l'intercetta b è responsabile del trend. Queste due cose sono in contraddizione? No: un'intercetta può avere vari effetti diversi, come produrre una media non nulla oppure un trend. La cosa dipende dal legame tra tutti i coefficienti, come mostrano l'esempio precedente ed il successivo.*

Esempio 85 *In contrasto all'esempio precedente, consideriamo il caso*

$$X_n = \alpha X_{n-1} + b + \varepsilon_n$$

con

$$|\alpha| < 1.$$

Qui risulta, iterativamente,

$$\begin{aligned} X_n &= \alpha \left(\alpha X_{n-2} + b + \varepsilon_{n-1} \right) + b + \varepsilon_n \\ &= \alpha^2 X_{n-2} + (\alpha + 1)\, b + \alpha \varepsilon_{n-1} + \varepsilon_n \end{aligned}$$

e così di seguito

$$\begin{aligned} X_n &= \alpha^3 X_{n-3} + \left(\alpha^2 + \alpha + 1 \right) b + \alpha^2 \varepsilon_{n-2} + \alpha \varepsilon_{n-1} + \varepsilon_n \\ &... \\ X_n &= \alpha^n X_0 + \left(\alpha^{n-1} + ... + \alpha + 1 \right) b + \alpha^{n-1} \varepsilon_1 + ... + \alpha \varepsilon_{n-1} + \varepsilon_n. \end{aligned}$$

Ora, essendo $|\alpha| < 1$, vale

$$\alpha^{n-1} + ... + \alpha + 1 \to \frac{1}{1 - \alpha}$$

quindi no c'è alcun trend (il termine $\left(\alpha^{n-1} + ... + \alpha + 1 \right) b$ non cresce linearmente ma tende alla costante $\frac{b}{1-\alpha}$). Il processo X_n ha però una media non nulla.

Nel caso, però meno interessante, in cui $|\alpha| > 1$, i modelli AR(1) hanno comportamenti esponenziali.

Esempio 86 *Consideriamo di nuovo*

$$X_n = \alpha X_{n-1} + \varepsilon_n$$

(ora $b = 0$) ma nel caso

$$|\alpha| > 1.$$

Dalla formula

$$X_n = \alpha^n X_0 + \alpha^{n-1}\varepsilon_1 + ... + \alpha\varepsilon_{n-1} + \varepsilon_n$$

vediamo intuitivamente che X_n è esponenzialmente grande in n in valore assoluto. Una giustificazione precisa sarebbe un po' noiosa in quanto dipende dall'ampiezza e segno relativi dei vari termini X_0, ε_1 ecc. che, moltiplicati per potenze elevate di α, determinano la divergenza esponenziale. In casi molto particolari possono anche esserci delle cancellazioni tra alcuni di questi termini (es. se $\alpha X_0 = -\varepsilon_1$, vale $\alpha^n X_0 + \alpha^{n-1}\varepsilon_1 = 0$) ma è chiaro che si tratta di poche situazioni particolari.

Tra i modelli più utili per le applicazioni gestionali ci sono poi quelli con ritardo annuale, orientati a catturare la periodicità annuale delle serie storiche di grandezze a carattere stagionale.

Esempio 87 *Il modello base di questo tipo è*

$$X_n = \alpha X_{n-12} + \varepsilon_n.$$

Anch'esso, si può dimostrare che ha soluzioni stazionarie se $|\alpha| < 1$ (ad esempio usando i metodi della sezione 4.2.8). La logic dietro questo modello è semplicemente che il valore ad es. di gennaio 2007 è pari a quello di gennaio 2006, lievemente ridotto, più una perturbazione causale. La lieve riduzione, dovuta ad $|\alpha| < 1$, non è necessaria ed è anzi poco realistica se si osserva il fenomeno concreto (economico ecc.) su una scala di pochissimi anni, 3-4. La stazionarietà vale approssimativamente anche per $\alpha = 1$, su orizzonti temporali così brevi.

Esempio 88 *Più aderente a moti esempi è il modello*

$$X_n = \alpha_1 X_{n-1} + \alpha_{12} X_{n-12} + \varepsilon_n$$

eventualmente anche con intercetta. Qui si sta immaginando, ad esempio, che il valore di aprile 2007 sia in parte legato a quello di aprile 2006 ed in parte a marzo 2007. Si ammette cioè che ci sia uno strascico da un mese all'altro, senza sbalzi del tutto causali tra un mese ed il successivo; più una somiglianza con lo stesso mese dell'anno precedente. Tra i sottosempi, si può riflettere sul caso

$$X_n = \alpha X_{n-1} + (1-\alpha) X_{n-12} + \varepsilon_n$$

con $\alpha \in (0, 1)$.

Naturalmente la precisione del modello aumenta se si cosidera anche il termine $\alpha_2 X_{n-2}$, oppure $\alpha_{24} X_{n-24}$. Ma d'altra parte più termini si mettono più il modello smette di essere un vero "modello", una sorta di formula generale, mentre tenta di inseguire le particolarità

dei dati sperimentali a cui si cerca di adattarlo. Per fare previsioni, non è affatto detto che inseguire le prticolarità più minute dei dati passati sia una buona strategia: alcune particolarità saranno strutturali, quindi tenderanno a ripetersi, alte no. Non ci sono ovviamente regole circa l'ecnomia da esercitare. Anche per questo fare previsioni è un'arte e non una scienza rigorosa (e non esistono software che osino proporsi come buoni previsori automatici, alla cieca dell'arte e delle scelte che l'operatore deve decidere di fare).

4.2.3 L'operatore di traslazione temporale

Sia S l'insieme di tutte le successioni $x = (x_t)_{t\in\mathbb{Z}}$ di numeri reali.

Definizione 42 *Chiamiamo operatore di traslazione temporale (in inglese time lag operator, o backward shift) l'applicazione $L : S \to S$ definita da*

$$Lx_t = x_{t-1}, \qquad \text{for all } t \in \mathbb{Z}.$$

L'applicazione L trasforma successioni in successioni. Più proprimente dovremmo scrivere $(Lx)_t = x_{t-1}$, in quanto data una successione $x = (x_t)_{t\in\mathbb{Z}} \in S$, L calcola una nuova successione $Lx \in S$, il cui valore $(Lx)_t$ al tempo t è dato da x_{t-1}. Per brevità si omettono le parentesi e si scrive $Lx_t = x_{t-1}$.

Quest'operatore prende una sequenza e la trasla all'indietro. Di fatto, esso verrà usato più che altro come notazione: invece di scrivere x_{t-1} scriveremo Lx_t. Apparentemente c'è poco vantaggio, ma ogni tanto se ne vedono i pregi.

Se invece di S consideriamo l'insieme S^+ delle successioni $(x_t)_{t\in\mathbb{N}}$ definite solo per tempi non-negativi, non possiamo definire L in quanto, data $x = (x_t)_{t\in\mathbb{N}}$, il suo primo valore è x_0, mentre il primo valore di Lx dovrebbe essere x_{-1}, che non esiste. Ciò nonostante, a patto di trascurare il primo valore di Lx, useremo la notazione $Lx_t = x_{t-1}$ ache per le successioni $x = (x_t)_{t\in\mathbb{N}}$. In altre parole, Lx è definito solo per $t > 0$.

Osservazione 66 *L'operatore L è lineare.*

Le potenze, positive o negative, di L verranno indicate con L^k. Esse sono la composizione di L fatta k volte. Ad esempio, $L^3x = L\left(L\left(Lx\right)\right)$, enfatizzando con le parentesi che si tratta di applicazioni successive, cioè composizioni, di L. Vale

$$L^k x_t = x_{t-k}, \qquad \text{per } t \in \mathbb{Z}$$

(o, per $t \geq \max(k, 0)$, nel aso $x = (x_t)_{t\in\mathbb{N}}$).

Con queste notationi, un modello AR(p) si può scrivere nella forma

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \varepsilon_t.$$

Il vantaggio sta nel poter immaginare (per ora non rigorosamente) che valga

$$X_t = \left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)^{-1} \varepsilon_t \tag{4.3}$$

cioè che sia possibile esprimere esplicitamente la soluzione X_t tramite l'input ε_t. Se ε_t rappresenta l'errore (per noi è principalmente così), questa espressione esplicita non dice molto, al massimo può essere utile per scopi teorici: si ricordi ad esempio (capitolo sui processi) che la random walk, che è un AR(1), ammette la soluzione esplicita $X_n = X_0 + \sum_{i=1}^{n} \varepsilon_i$, che abbiamo usato per efettuare alcuni calcoli teorici. Oppure si veda il paragrafo 4.2.8 seguente. Se però utilizzassimo i modelli AR(p) come modelli input-output, in cui ε_t è un fattore, una variabile espicativa, un cotrollo, ed X_t è la variabile di output, allora è fondamentale sapere come l'output dipende dall'input. L'equazione che definisce il modello AR(p) descrive come X_t dipende dai valori precedenti dell'output stesso, più ε_t. Invece l'equazione (4.3) direbbe in modo più esplicito come X_t dipende dall'input.

Date queste motivazioni, resta il problema pratico di cosa voglia dire $\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)^{-1} \varepsilon_t$. A questo proposito si deve operare come se $\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)^{-1}$ fosse un polinomio nella variabile reale L.

Esempio 89 *Ad esempio, se*

$$1 - \sum_{k=1}^{p} \alpha_k L^k = 1 - \alpha L$$

(caso AR(1) con $\alpha_1 = \alpha$), vale

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)^{-1} = (1 - \alpha L)^{-1} = \sum_{i=1}^{\infty} \alpha^i L^i$$

quindi

$$X_t = \sum_{i=1}^{\infty} \alpha^i L^i \varepsilon_t = \sum_{i=1}^{\infty} \alpha^i \varepsilon_{t-i}.$$

Da qui possiamo calcolare ad esempio

$$\begin{aligned} Cov\left(X_t, X_{t+n}\right) &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha^i \alpha^j Cov\left(\varepsilon_{t-i}, \varepsilon_{t+n-j}\right) \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha^i \alpha^j \sigma^2 \delta\left(i - j + n\right) \\ &= \sum_{i=1}^{\infty} \alpha^i \alpha^{i-n} \sigma^2 = \alpha^{-n} \frac{\sigma^2}{1 - \alpha^2} \end{aligned}$$

da cui si vede che è un processo stazionario ($Cov\left(X_t, X_{t+n}\right)$ indipendente da t; la media si vede subito che è zero) ed abbiamo

$$\rho_n = \alpha^{-n}.$$

Si leggano però le precisazioni dell'osservazione seguente.

Osservazione 67 *Questa è in un certo senso un'altra dimostrazione di un fatto visto in un esempio del capitolo sui processi stocastici. Però ci sono dei dettagli teorici da osservare. In*

quel capitolo avevamo scelto opportunamente il dato iniziale ed avevamo ristretto l'attenzione a $|\alpha| < 1$. Qui non è stato fissato alcn dato iniziale ed apparentemente non abbiamo messo restrizioni su α. Circa il dato iniziale, va osservato che i procedmento descritto nell'esempio di questo capitolo è tale da produrre automticamente la soluzione stazionaria, se c'è; non va imposto un particolare dato iniziale. Nell'invertire $\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right)$ è insito che si troverà la soluzione stazionaria. Però ci sono ipotesi che permettono di invertire questo operatore, oppure lo vietano. Intuitivamente parlando, è come se L dovesse essere preso uguale ad 1, per cui l'identità $(1 - \alpha L)^{-1} = \sum_{i=1}^{\infty} \alpha^i L^i$ vale solo se $|\alpha| < 1$. Torneremo su questo punto nella sezione 4.2.8 .

4.2.4 Modelli MA

Definizione 43 *Si chiama modello a media mobile di ordine q, o modello $MA(q)$, l'equazione lineare*

$$X_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}$$

dove il significato dei simboli è simile a quanto descritto nella definizione di $AR(p)$. Un processo $MA(q)$ è una sua soluzione (di solito si intende quella stazionaria).

A differenza del caso AR(p), qui il processo è definito esplicitamente dal rumore, attraverso una sua media pesata. Si noti che non è una media dal tempo iniziale, ma è una media su una finestra di ampiezza q che si sposta con t (da qui il nome).

In parte l'utilità di questi modelli si riconosce quando ε_t non ha il significato di rumore ma di input, o addirittura è il processo che si sta cercando di esaminare. Allora X_t è una sua media pesata, che può servire ad esempio per effettuare predizioni future. La regola più semplice per predire il valore futuro di una serie storica è ricopiare il valore attuale, $X_t = \varepsilon_{t-1}$; a ruota (come semplicità) segue il modello predittivo usalmente chiamato *a media mobile*, dato da

$$X_t = \frac{\varepsilon_{t-1} + ... + \varepsilon_{t-q}}{q}.$$

Usando l'operatore L abbiamo

$$X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t.$$

4.2.5 Modelli ARMA

Definizione 44 *Si chiama modello $ARMA(p, q)$ (AutoRegressive Moving Average di ordini p e q) l'equazione lineare*

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t$$

o esplicitamente

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}.$$

Come sempre, chiamiamo processo ARMA(p,q) una soluzione (stazionaria) di questa equazione. Il sigificato dei simboli è simile a quello delle due definizioni precedenti.

Come per gli AR(p), il modello ora scritto si adatta bene alle situazioni a media nulla. Se vogliamo esaminare con modelli simili un processo X_t a media non nulla μ, immaginiamo che $Z_t = X_t - \mu$ soddisfi l'equazione ARMA(p,q)

$$Z_t = \alpha_1 Z_{t-1} + ... + \alpha_p Z_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}$$

per cui $X_t = Z_t + \mu$ risolverà

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + b + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q}$$

dove l'intercetta b è data da

$$b = \mu - \alpha_1 \mu - ... - \alpha_p \mu.$$

4.2.6 Operatore differenza. Integrazione

Definizione 45 *Chiamiamo operatore differenza l'operaotre $\Delta : S \to S$ definito da*

$$\Delta x_t = x_t - x_{t-1} = (1 - L)\, x_t.$$

Usiamo queste notazioni anche nel caso di successioni $x \in S_+$. Si veda la definizione di L per alcune delle notazioni usate qui.

L'operatore Δ è una specie di derivata discreta. La derivata di una fuzione lineare è una costante. Si può allora immaginare che, prendendo la derivata discreta Δ di un processo che ha un trend lineare, si ottenga un processo stazionario. Queste frasi sono vaghe (salvo mettersi in ipotesi troppo specifiche, come il seguente esempio), per cui sfuggono ad un teorema, ma resta l'utilità pratica di eseguire Δ su processi con trend per renderli "più" stazionari. Si guadagna anche sul fatto che il nuovo processo è a media nulla, se la media del processo originario era costante.

Esempio 90 *Iniziamo col verificare che l'operatore Δ non distrugge la stazionarietà. Sia X un processo stazionario e sia $Y_t = \Delta X_t$. Mostriamo che Y è ancora stazionario. Vale*

$$E[Y_t] = E[X_t] - E[X_{t-1}] = 0$$

$$\begin{aligned} E[Y_t Y_{t+n}] &= E[(X_t - X_{t-1})(X_{t+n} - X_{t+n-1})] \\ &= E[X_t X_{t+n}] - E[X_t X_{t+n-1}] - E[X_{t-1} X_{t+n}] + E[X_{t-1} X_{t+n-1}] \\ &= R(n) - R(n-1) - R(n+1) + R(n) \end{aligned}$$

quindi la media è costante, anzi nulla, e l'autocorrelzione dipende solo da n. Il processo Y è stazionario ed anzi in più di X ha che è sempre a media nulla.

Esempio 91 *Supponiamo ora che sia*

$$X_t = a \cdot t + \varepsilon_t$$

dove ε_t è stazionario. Questo è un esempio molto schematico di processo con trend lineare. Posto $Y_t = \Delta X_t$, troviamo

$$Y_t = a + (\varepsilon_t - \varepsilon_{t-1}).$$

Questo è un processo stazionario, per quanto visto sopra (però la sua media è a).

Appena si deriva, nasce il problema invers dell'integrazione. Data x possiamo calcolare $y = \Delta x$; viceversa, data y, possiamo trovare x che risolve $y = \Delta x$? Basta risolvere

$$y_t = x_t - x_{t-1}$$

trovando

$$x_t = y_t + x_{t-1} = y_t + y_{t-1} + x_{t-2} = ... = y_t + ... + y_1 + x_0.$$

Il risultato è:

Proposizione 23 *Se due successioni x ed y sono legate dalla relazione $y = \Delta x$, si può ricostruire la successione x dalla y, usando x_0, tramite la formula*

$$x_t = y_t + ... + y_1 + x_0.$$

I fatti precedenti si possono iterare. L'operatore *differenza seconda*, Δ^2, è definito da

$$\Delta^2 x_t = (1 - L)^2 x_t.$$

Per invertirlo, supponiamo che con y dato sia

$$y_t = (1 - L)^2 x_t.$$

Allora introduciamo z:

$$y_t = (1 - L) z_t$$
$$z_t = (1 - L) x_t$$

quindi prima ricostruiamo z_t da y_t:

$$z_t = y_t + ... + y_2 + z_1$$

dove

$$z_1 = (1 - L) x_1 = x_1 - x_0$$

poi ricostruiamo x_t da z_t:

$$x_t = z_t + ... + z_1 + x_0.$$

Proposizione 24 *Se due successioni x ed y sono legate dalla relazione $y = \Delta x$, si può ricostruire la successione x dalla y, usando x_0 ed x_1, tramite le formule*

$$z_1 = x_1 - x_0$$
$$z_t = y_t + ... + y_2 + z_1$$
$$x_t = z_t + ... + z_1 + x_0.$$

Tutto questo si può generalizzare a Δ^d, per ogni intero positivo d.

4.2.7 Modelli ARIMA

Definizione 46 *Si chiama modello ARIMA(p,d,q) (AutoRegressive Integrated Moving Average di ordini p, d e q) l'equazione lineare*

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d X_t = \left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t.$$

Il sigificato dei simboli è simile a quello delle definizioni precedenti.

Osservazione 68 *L'operatore $\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d$ può essere riscritto nella forma $\left(1-\sum_{k=1}^{p+d}\alpha'_k L^k\right)$ per opportuni nuovi coefficienti α'_k, quindi un modello ARIMA(p,d,q) è di tipo ARMA$(p,q+d)$. Questo punto di vista però è fuorviante, ad esempio perché negli ARIMA(p,d,q) non ci si interessa alle soluzioni stazionarie.*

Se X risolve un modello ARIMA(p,d,q), allora $Y_t := (1-L)^d X_t$ risolve il seguente ARMA(p,q):

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)Y_t = \left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t$$

e X_t si può ottenere da Y_t attraverso d integrazioni successive. Il numero d è quindi l'ordine di integrazione.

Il modo giusto di pensare a questi modelli è il seguente. Il processo Y_t, risolvendo un ARMA(p,q), è naturale che sia stazionario (in altre parole, ci interessa la soluzione stazionaria di questo ARMA(p,q)). Integrando poi d volte si trova X_t, che però non è più stazionario (si veda l'osservazione al termine del paragrafo). Le soluzioni a cui siamo interessati dei modelli ARIMA(p,d,q) non sono stazionarie, ma sono quelle tali che $Y_t = (1-L)^d X_t$ è stazionario. Una tale soluzione X_t viene d solito detta *processo* ARIMA(p,d,q).

Esempio 92 *La random walk è un ARIMA$(0,1,0)$.*

Possiamo incorporare una media non nulla (per la $Y_t = (1-L)^d X_t$) in un modello ARIMA(p,d,q) considerando il modello

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d X_t = \left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t + b \tag{4.4}$$

sempre con

$$b = \mu - \alpha_1\mu - ... - \alpha_p\mu.$$

Osservazione 69 *Se $Y_t = (1-L)X_t$ e Y_t è stazionario, non lo è X_t. L'integrazione rompe la stazionarietà, per cui le soluzioni interessanti dei modelli ARIMA non sono stazionarie. Cerchiamo di capire come mai l'integrazione produce non-stazionarietà. A titolo di esempio, abbiamo già verificato nel capitolo sui processi stocastici che la random walk non è stazionaria. La non stazionarietà della RW così come di vari altri ARIMA è di tipo complesso, cioè non*

appare semplicemente come un trend lineare ma può apparire come una crescita del tipo $\sqrt{t}$ della deviazione standard, con i valori del processo che oscillano tra positivi e negativi.

A volte invce la non stazionarietà si manfesta più semplicemente con un trend. Il caso pù semplice da capire è il caso con media non nulla (4.4). Se $d = 1$, X_t ha un trend lineare; se $d = 2$, ha un trend quadratico, e così via. Infatti, supponiamo che Y_t sia stazionario e a media $\mu > 0$ (il caso $\mu < 0$ è identico) Allora (per $d = 1$)

$$X_t = Y_t + ... + Y_1 + X_0 = \mu \cdot t + Z_t$$
$$Z_t := \widetilde{Y}_t + ... + \widetilde{Y}_1 + X_0$$

dove le v.a. $\widetilde{Y}_t = Y_t - \mu$ hanno media nulla. Il processo Z_t può crescere (in modulo) ma non abbastanza da contrastare la crescita lineare del termine $\mu \cdot t$. Si può infatti dimostrare (si deve usare il teorema ergodico del capitolo sui processi, verificando che le ipotesi per gli ARMA sono vere) che

$$\frac{\widetilde{Y}_t + ... + \widetilde{Y}_1}{t} \overset{t\to\infty}{\to} E[Y_1] = 0$$

cioè Z_t ha crescita sub-lineare. Si può anche intuire come vanno le cose pensando al caso (molto particolare) in cui le v.a. $\widetilde{Y}_t$ sono indipendenti. Vale

$$Var\left[\widetilde{Y}_t + ... + \widetilde{Y}_1\right] = t \cdot Var\left[\widetilde{Y}_1\right]$$

quindi la deviazione standard di $\widetilde{Y}_t + ... + \widetilde{Y}_1$ cresce come $\sqrt{t}$, da cui si intuisce che $\widetilde{Y}_t + ... + \widetilde{Y}_1$ non può crescere linearmente. Per $d > 1$ valgono ragionamenti simili.

4.2.8 Stazionarietà, legame tra modelli ARMA e modelli MA di ordine infinito, ipotesi generali della teoria

Abbiamo già detto che, sotto opportune condizioni, esistono soluzioni stazionarie dei modelli AR, MA ed in generale ARMA, chiamate *processi* AR, MA o ARMA seconda dei casi. Nel caso più sempice degli AR(1) abbiamo trovato che la condizione è $|\alpha| < 1$.

Aggiungiamo alcune precisazioni. La parte MA di un modello non pone restrizioni alla possibilità di avere soluzioni stazionarie. Limitatamente ai modelli MA, un modo di costruire soluzioni stazionarie è quello di far partire l'iterazione da un tempo negativo molto grande, $-N$, con valori iniziali nulli. Ciò che si osserva per tempi positivi è approssimativamente stazionario, e migliora al tende di $N \to \infty$.

La parte AR invece pone restrizioni. Bisogna assumere che le radici complesse z del polinomio

$$p(z) = 1 - \sum_{k=1}^{p} \alpha_k z^k$$

siano tutte fuori dalla palla unitaria chiusa del piano complesso, cioè soddisfino tutte la condizione

$$|z| > 1. \tag{4.5}$$

Esempio 93 *Nel caso $AR(1)$ il polinomio è $p(z) = 1 - \alpha z$ che ha, per $\alpha \neq 0$, l'unica radice $z = \frac{1}{\alpha}$ (mentre per $\alpha = 0$ non ha radici per cui la condizione precedente è soddisfatta). Vale $\left|\frac{1}{\alpha}\right| > 1$ se e solo se $|\alpha| < 1$. Così si ritrova la condizione già scoperta per la stazionarietà.*

Se vale la condizione precedente, la funzione $\frac{1}{p(z)}$ è analitica per $|z| \leq 1 + \varepsilon$ per qualche $\varepsilon > 0$. Si consideri la funzione

$$g(z) = \frac{1 + \sum_{k=1}^{q} \beta_k z^k}{1 - \sum_{k=1}^{p} \alpha_k z^k} = \frac{1 + \sum_{k=1}^{q} \beta_k z^k}{p(z)}.$$

Anch'essa è analitica per $|z| \leq 1 + \varepsilon$ per qualche $\varepsilon > 0$ ed il suo svilppo di Taylor

$$g(z) = \sum_{j=0}^{\infty} \varphi_j z^j$$

converge quindi uniformemente in $|z| \leq 1$. In particolare vale

$$\sum_{j=0}^{\infty} |\varphi_j| < \infty$$

e quindi anche

$$\sum_{j=0}^{\infty} |\varphi_j|^2 < \infty.$$

Imponiamo, per la validità di tutti fatti descritti in questa sezione, la condizione (4.5) sulle radici di $p(z)$.

Osservazione 70 *Iniziamo con un'osservazione che lasciamo a livello completamente intuitivo. Sia X un processo stazionario a media nulla, di tipo $ARMA(p, q)$, definito anche per tempi negtivi. Dalla relazione*

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t$$

fingendo di poter operare su $p(L) := 1 - \sum_{k=1}^{p} \alpha_k L^k$ come fosse un polinomio, otteniamo, usando le notazioni precedenti

$$X_t = \frac{1 + \sum_{k=1}^{q} \beta_k L^k}{1 - \sum_{k=1}^{p} \alpha_k L^k} \varepsilon_t = g(L)\, \varepsilon_t.$$

Sempre euristicamente, eseguiamo lo sviluppo di Taylor di g ottenendo

$$X_t = \sum_{j=0}^{\infty} \varphi_j L^j \varepsilon_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j}.$$

Abbiamo riscritto un processo $ARMA(p, q)$ come un MA di ordine infinito.

Vediamo più rigorosamente le cose al viceversa: dato un white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ definiamo X_t tramite questa uguaglianza. E' una definizione ammissibile e fornisce un processo stazionario, soluzione del modello ARMA(p,q) di partenza?

Teorema 30 *Sotto l'ipotesi (4.5) per le radici del polinomio p, la serie* $\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}$ *converge in media quadratica e definisce un processo* X_t *che è stazionario e soluzione del modello ARMA*(p,q) *a cui sono associati i coefficienti* φ_j*. In particolare, esiste una soluzione stazionaria di tale modello ARMA*(p,q)*.*

Proof. Non diamo tutti i dettagli della dimostrazione ma solo l'idea.

Passo 1. Intanto, g è analitica in un intorno aperto di $|z|\leq 1$ (come osservato nella sezione 4.2.8) e quindi il suo sviluppo $g(x)=\sum_{j=0}^{\infty}\varphi_j x^j$ converge uniformemente in $|z|\leq 1$. In particolare i coefficienti φ_j esistono. Consideriamo la serie $\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}$. Vale, per l'indipendenza del WN (usiamo regole valide per somme finite anche nel caso infinito; è ad esempio qui che tralasciamo alcuni dettagli della diostrazione rigorosa completa),

$$Var\left[\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}\right]=\sum_{j=0}^{\infty}\left|\varphi_j\right|^2 Var\left[\varepsilon_{t-j}\right]=\sigma^2\sum_{j=0}^{\infty}\left|\varphi_j\right|^2$$

e questa serie è finita, come osservato nella sezione 4.2.8. Da questo è possibile dimostrare che $\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}$ converge in media quadratica (bisogna usare il fatto che una successione di Cauchy in media quadratica converge).

Passo 2. Chiamiamo X_t il suo limite. E' un processo stazionario. Che la media sia costante è facile, usando di nuovo regole sui valori medi delle serie che non abbiamo spiegato nel corso:

$$E\left[\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}\right]=\sum_{j=0}^{\infty}\varphi_j E\left[\varepsilon_{t-j}\right]=0.$$

Poi vale, sempre per regole simili

$$\begin{aligned}R(s,t)&=E\left[\sum_{j=0}^{\infty}\varphi_j\varepsilon_{t-j}\cdot\sum_{k=0}^{\infty}\varphi_k\varepsilon_{s-k}\right]=\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\varphi_j\varphi_k E\left[\varepsilon_{t-j}\varepsilon_{s-k}\right]\\&=\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\varphi_j\varphi_k\delta\left(t-s-j+k\right)\end{aligned}$$

che dipende quindi solo da $t-s$. Quindi la stazionarietà è verificata.

Passo 3. Infine, dobbiamo verificare che

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)\sum_{j=0}^{\infty}\varphi_j L^j\varepsilon_t=\left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t.$$

Euristicamente è ovvio:

$$\left(1-\sum_{k=1}^{p}\alpha_k x^k\right)g(x)=\left(1+\sum_{k=1}^{q}\beta_k x^k\right)$$

per definizione di g. Il problema quindi è il passaggio da un'identità tra polinomi, eventualmente infiniti (nel senso dello sviluppo di Taylor di funzioni analitiche) ad una tra polinomi di operatori. Descriviamo gli ingredienti nel prossimo passo.

Passo 4. Siano $a(z)$ e $b(z)$ due polinomi, della forma

$$a(z) = \sum_{k=0}^{n} a_k z^k, \qquad b(z) = \sum_{k=0}^{m} b_k z^k.$$

Il loro prodotto è un polinomio, che riscriviamo nella forma canonica tramite opportuni coefficienti c_k:

$$a(z)\, b(z) = \sum_{k=0}^{n+m} c_k z^k.$$

Allora vale anche

$$\left(\sum_{k=0}^{n} a_k L^k \left(\sum_{h=0}^{m} b_h L^h x\right)\right)_t = \sum_{k=0}^{n+m} c_k L^k x_t$$

per ogni successione $x \in S$. E' sufficiente verificare l'identità per i monomi:

$$\left(a_k L^k \left(b_h L^h x\right)\right)_t = a_k b_h L^{k+h} x_t.$$

Questo è vero:

$$\left(a_k L^k \left(b_h L^h x\right)\right)_t = \left(a_k L^k \left(b_h x_{\cdot - h}\right)\right)_t = a_k b_h x_{t-h-k}$$
$$a_k b_h L^{k+h} x_t = a_k b_h x_{t-h-k}.$$

Fatta la verifica per i polinomi finiti, bisogna estenderla ai caso degli sviluppi di Taylor di funzioni analitiche. La verifica è un po' tecnica e la omettiamo. ■

Può essere istruttivo ora rileggere l'esempio 89 sul modello AR(1).

4.2.9 Funzione di autocorrelazione, primi fatti

Assumimo che X sia un processo ARMA(p, q) a media nulla, soluzione stazionaria di

$$\left(1 - \sum_{k=1}^{p} \alpha_k L^k\right) X_t = \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t.$$

Proposizione 25 (equazioni di Yule-Walker) *Per ogni $j > q$,*

$$R(j) = \sum_{k=1}^{p} \alpha_k R(j-k).$$

Proof. Ricordiamo che $R(-n) = R(n)$. Osserviamo che per ogni n ed m vale

$$E\left[X_{t-n} L^m X_t\right] = E\left[X_{t-n} X_{t-m}\right] = R(m-n).$$

Allora dalla

$$E\left[\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)X_t\cdot X_{t-j}\right]=E\left[\left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t\cdot X_{t-j}\right]$$

discende

$$R(j)-\sum_{k=1}^{p}\alpha_k R(j-k)=E\left[X_{t-j}\left(1+\sum_{k=1}^{q}\beta_k L^k\right)\varepsilon_t\right].$$

Nel caso $j>q$ la v.a. $L^k\varepsilon_t=\varepsilon_{t-k}$ è indipendente da X_{t-j}, in quanto $k\leq q<j$, quindi

$$R(j)-\sum_{k=1}^{p}\alpha_k R(j-k)=0.$$

La dimostrazione è completa. ■

Corollario 7 *Se X è un processo $AR(p)$ a media nulla, soluzione stazionaria di*

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)X_t=\varepsilon_t$$

allora, per ogni $j>0$ vale

$$R(j)=\sum_{k=1}^{p}\alpha_k R(j-k).$$

Vediamo come le equazioni di Yule-Walker, per gli AR(p), permettano di calcolare la funzione R.

Esempio 94 *Si consideri il processo $AR(1)$:*

$$X_t=\alpha X_{t-1}+\varepsilon_t.$$

Vale, per ogni $j>0$,

$$R(j)-\alpha R(j-1)=0$$

ovvero

$$R(1)=\alpha R(0)$$
$$R(2)=\alpha R(1)$$
$$\ldots$$

dove $R(0)=E\left[X_0^2\right]$. Quindi

$$R(j)=\alpha^j R(0).$$

Resta da calcolare $R(0)$. Vale

$$Var[X_t]=a^2 Var[X_{t-1}]+Var[\varepsilon_t]$$

quindi

$$R(0)=a^2R(0)+\sigma^2$$

che implica $R(0)=\frac{\sigma^2}{1-a^2}$. Questo è lo stesso risultato trovato più volte in precedenza.

Esempio 95 *Consideriamo ora un processo AR(2):*

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t.$$

Abbiamo

$$R(j) = \alpha_1 R(j-1) + \alpha_2 R(j-2)$$

pr ogni $j > 0$, ovvero

$$\begin{aligned} R(1) &= \alpha_1 R(0) + \alpha_2 R(-1) \\ R(2) &= \alpha_1 R(1) + \alpha_2 R(0) \\ &... \end{aligned}$$

Essendo $R(-1) = R(1)$, troviamo

$$R(1) = \frac{\alpha_1}{1-\alpha_2} R(0)$$

dalla prima equazione. Le altre poi permettono il calcolo di $R(2)$ e così via. Dobbiamo solo torvare $R(0)$, come nell'esempio precedente. Abbiamo

$$Var[X_t] = a_1 Var[X_{t-1}] + a_2 Var[X_{t-2}] + \sigma^2 + 2Cov(X_{t-1}, X_{t-2})$$

quindi

$$R(0) = a_1 R(0) + a_2 R(0) + \sigma^2 + 2R(1).$$

Questa è una seconda equazione tra $R(0)$ ed $R(1)$, che messa insieme all'equazione $R(1) = \frac{\alpha_1}{1-\alpha_2} R(0)$ permette il calcolo di entrambe le quantità. E' poco istruttivo arrivare alle formule esplicite finali, quindi ci accontentiamo di aver verificato che le equazioni di Yule-Walker permettono di calcolare l'autocorrelazione (modulo il calcolo dei valori iniziali di R da equazioni elementari).

Le equazioni di Yule-Walker hanno due usi principali. Il primo è quello ovvio di conoscere R a partire dal modello. In quest'ottica si suppone di aver formulto un modello ARMA e di volerne capire le proprietà.

La seconda applicazione, forse più importante, è quella della costruzione di un modello data una serie storica. La serie storica viene usata per calcolare l'acf (l'autocorrelazione empirica). Chiamiamola $\widehat{R}(j)$. Imaginiamo che essa soddisfi delle equazioni tipo Yule-Walker, per ogni $j > 0$:

$$\widehat{R}(j) = \sum_{k=1}^{p} \widehat{\alpha}_k \widehat{R}(j-k).$$

Consideriamo le prime p equazioni:

$$\widehat{R}(j) = \sum_{k=1}^{p} \widehat{\alpha}_k \widehat{R}(j-k), \qquad j = 1, 2, ..., p$$

e vediamole come equazioni, lineari, nelle incognite $\widehat{\alpha}_1, \widehat{\alpha}_2, ..., \widehat{\alpha}_p$. Risolvendole si trovano questi coefficienti, cioè un modello AR empirico

$$\left(1 - \sum_{k=1}^{p} \widehat{\alpha}_k L^k\right) X_t = \varepsilon_t.$$

Questo metodo funziona in ipotesi di stazionarietà del processo. Vedremo nei paragrafi pratici sull'uso del software che viene proposto un metodo alternativo (comando `ar.ols`, ols = ordinary least squares) basato sull'idea universale di fittare un modello ai dati tramite minimi quadrati (si minimizza, al variare dei parametri incogniti del modello, la somma dei quadrati dei residui, cioè degli scarti tra i valori sperimentali veri e quelli forniti dal modello). Il metodo dei minimi quadrati non richiede alcuna ipotesi di stazionarietà, quindi è più flessibile. Però è meno improntato all'idea di *struttura*. Se vale la stazionarietà, la ricerca dei coefficienti tramite le equazioni di Yule-Walker appare più strutturale, meno soggetto alle fluttuazioni dovute ai valori particolari dell'esperimento (anche se una quantificazione precisa di queste frasi è molto difficile, per cui vanno prese come indicazione di principio, che può essere disattesa in moti esempi). Certamente, se ad esempio si usano entrambi i metodi (nel caso stazionario) ed i valori trovati per i coefficieti sono simili, questa è una bella indicazione di bontà del modello.

4.2.10 Funzione di autocorrelazione, complementi

Continuiamo ad assumere di esaminare un processo ARMA stazionario a media nulla, sotto le ipotesi (4.5), per cui in particolare vale

$$X_t = \sum_{i=0}^{\infty} \varphi_i L^i \varepsilon_t.$$

Quindi possiamo calcolare $E\left[X_{t-j}\left(1 + \sum_{k=1}^{q} \beta_k L^k\right)\varepsilon_t\right]$ anche per $j \leq q$, il caso non trattato nella Proposizione 25.

Proposizione 26 *Sotto le ipotesiprecedenti, per ogni $j = 0, ..., q$ vale*

$$R(j) - \sum_{k=1}^{p} \alpha_k R(j-k) = \sum_{i=0}^{q-j} \varphi_i \beta_{i+j} \sigma^2.$$

Quindi, per ogni $j \geq 0$ possiamo scrivere

$$R(j) - \sum_{k=1}^{p} \alpha_k R(j-k) = \sum_{i=0}^{\infty} \varphi_i \beta_{i+j} \sigma^2 1_{i+j \in \{0,...,q\}}.$$

Proof. Posto $\beta_0 = 1$, da una formula della dimostrazione della Proposizione 25 e dall'identità $X_{t-j} = \sum_{i=0}^{\infty} \varphi_i L^i \varepsilon_{t-j}$, abbiamo

$$\begin{aligned} R(j) - \sum_{k=1}^{p} \alpha_k R(j-k) &= E\left[X_{t-j} \sum_{k=0}^{q} \beta_k L^k \varepsilon_t\right] \\ &= \sum_{i=0}^{\infty} \sum_{k=0}^{q} \varphi_i \beta_k E\left[L^i \varepsilon_{t-j} L^k \varepsilon_t\right] \\ &= \sum_{i=0}^{\infty} \sum_{k=0}^{q} \varphi_i \beta_k \delta_{i+j,k} \sigma^2 \\ &= \sum_{i=0}^{\infty} \varphi_i \beta_{i+j} \sigma^2 1_{i+j \in \{0,\ldots,q\}}. \end{aligned}$$

La dimostrazione è completa. ■

Questa è una formula generale. Un approccio più diretto al calcolo di $E\left[X_{t-j}\left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t\right]$ anche per $j \leq q$ consiste nella sostituzione del modello ARMA soddisfatto da X_{t-j}

$$E\left[X_{t-j}\left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t\right] = E\left[\left(\sum_{k=1}^{p} \alpha_k L^k X_{t-j} + \left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_{t-j}\right)\left(1 + \sum_{k=1}^{q} \beta_k L^k\right) \varepsilon_t\right].$$

I prodotti con fattori $L^k \varepsilon_{t-j}$ e $L^{k'} \varepsilon_t$ si calcolano facilmente. Il problema sono i prodotti del tipo

$$E\left[L^k X_{t-j} L^{k'} \varepsilon_t\right]$$

il più difficile dei quali è

$$E\left[L^1 X_{t-j} L^q \varepsilon_t\right].$$

Se $j \geq q$, è zero, altrimenti no, ma possiamo ripetere il trucco e procedere a ritroso passo a passo. In esempi semplici possiamo calcolare $R(j)$ in questo modo.

Esempio 96 *Si consideri*

$$X_t = \alpha X_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}.$$

Abbiamo

$$R(j) - \alpha R(j-1) = 0$$

per ogni $j > 1$, ovvero

$$\begin{aligned} R(2) &= \alpha R(1) \\ R(3) &= \alpha R(2) \\ &\ldots \end{aligned}$$

ma queste relazioni non permettono il calcolo di $R(1)$ ed $R(0)$ (mentre poi tutti gli altri si calcolano iterativaemente). Per quanto riguarda $R(1)$, vale (usando il metodo illustrato prima di sviluppare l'esempio)

$$\begin{aligned} R(1) - \alpha R(0) &= E\left[X_{t-1}(1 + \beta L) \varepsilon_t\right] = \beta E\left[X_{t-1} \varepsilon_{t-1}\right] \\ &= \beta E\left[\left(\alpha X_{t-2} + \varepsilon_{t-1} + \beta \varepsilon_{t-2}\right) \varepsilon_{t-1}\right] = \beta \sigma^2. \end{aligned}$$

Quindi $R(1)$ *è espresso in termini di* $R(0)$*. Infine*

$$Var[X_t] = a^2 Var[X_{t-1}] + \sigma^2 + \beta^2\sigma^2 + 2\alpha\beta Cov(X_{t-1}, \varepsilon_{t-1})$$

quindi

$$R(0) = a^2 R(0) + \sigma^2 + \beta^2\sigma^2 + 2\alpha\beta Cov(X_{t-1}, \varepsilon_{t-1}).$$

Inoltre,

$$Cov(X_{t-1}, \varepsilon_{t-1}) = Cov(\alpha X_{t-2} + \varepsilon_{t-1} + \beta\varepsilon_{t-2}, \varepsilon_{t-1}) = \sigma^2$$

quindi

$$R(0) = a^2 R(0) + \sigma^2 + \beta^2\sigma^2 + 2\alpha\beta\sigma^2$$

da cui calcoliamo $R(0)$*.*

4.2.11 Densità spettrale di potenza dei processi ARMA

Teorema 31 *Sotto le ipotesi (4.5), se* X *è un processo ARMA stazionario a media nulla, che soddisfi le ipotesi del teorema di Wiener-Khinchin (Capitolo sui processi), allora*

$$S(\omega) = \frac{\sigma^2}{2\pi}\left|\frac{1+\sum_{k=1}^{q}\beta_k e^{-ik\omega}}{1-\sum_{k=1}^{p}\alpha_k e^{-ik\omega}}\right|^2.$$

Proof. Ricordiamo che sotto l'ipotesi (4.5) la funzione $g(x) = \frac{1+\sum_{k=1}^{q}\beta_k x^k}{1-\sum_{k=1}^{p}\alpha_k x^k}$ ha lo sviluppo di Taylor $g(x) = \sum_{j=0}^{\infty}\varphi_j x^j$ in un intorno complesso U dell'origine che include il disco unitario chiuso.

Abbiamo (indicando con $\mathbb{Z}_T$ l'insieme degli $n \in \mathbb{Z}$ tali che $|n| \leq T/2$)

$$\widehat{X}_T(\omega) = \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}_T} e^{-i\omega n}X_n = \frac{1}{\sqrt{2\pi}}\sum_{n\in\mathbb{Z}_T}\sum_{j=0}^{\infty}\varphi_j e^{-i\omega n}\varepsilon_{n-j}$$

$$\widehat{X}_T^*(\omega) = \frac{1}{\sqrt{2\pi}}\sum_{n'\in\mathbb{Z}_T}\sum_{j'=0}^{\infty}\varphi_{j'} e^{i\omega n'}\varepsilon_{n'-j'}$$

$$E\left[\widehat{X}_T(\omega)\,\widehat{X}_T^*(\omega)\right] = \frac{1}{2\pi}E\left[\sum_{n\in\mathbb{Z}_T}\sum_{n'\in\mathbb{Z}_T}\sum_{j=0}^{\infty}\sum_{j'=0}^{\infty}\varphi_j\varphi_{j'}e^{-i\omega n}e^{i\omega n'}E\left[\varepsilon_{n-j}\varepsilon_{n'-j'}\right]\right]$$

$$= \frac{\sigma^2}{2\pi}\sum_{n\in\mathbb{Z}_T}\sum_{j=0}^{\infty}\sum_{j'=0}^{\infty}\varphi_j\varphi_{j'}e^{-i\omega n}e^{i\omega(n-j+j')} = |\mathbb{Z}_T|\frac{\sigma^2}{2\pi}\sum_{j=0}^{\infty}\sum_{j'=0}^{\infty}\varphi_j e^{-i\omega j}\varphi_{j'}e^{i\omega j'} = |\mathbb{Z}_T|\frac{\sigma^2}{2\pi}\left|\sum_{n=0}^{\infty}\varphi_n e^{-i\omega n}\right|^2.$$

La cardinalità $|\mathbb{Z}_T|$ di $\mathbb{Z}_T$ ha la proprietà $\lim_{T\to\infty}|\mathbb{Z}_T|/T = 1$, quindi

$$S(\omega) = \frac{\sigma^2}{2\pi}\left|\sum_{n=0}^{\infty}\varphi_n e^{-i\omega n}\right|^2.$$

Ora è sufficiente usare la relazione $\frac{1+\sum_{k=1}^{q}\beta_k x^k}{1-\sum_{k=1}^{p}\alpha_k x^k} = \sum_{j=0}^{\infty}\varphi_j x^j$ per $x = e^{-i\omega}$. La dimostrazione è completa.

Osservazione 71 *Si consideri il caso $q = 0$. Scriviamo la formula con $\omega = 2\pi f$*

$$S(f) = \frac{\sigma^2}{2\pi} \frac{1}{\left|1 - \sum_{k=1}^{p} \alpha_k e^{-2\pi i k f}\right|^2}.$$

Consideriamo il caso particolare in cui c'è solo il termine con $k = p$:

$$S(f) = \frac{\sigma^2}{2\pi} \frac{1}{\left|1 - \alpha_p e^{-2\pi i p f}\right|^2}.$$

In questo caso i massimi si trovano per $pf \in \mathbb{Z}$, cioè $f = \frac{1}{p}$ e suoi multipli interi. la funzione $S(f)$ mostra con chiarezza la periodicità del modello

$$X_t = a_p X_{t-p} + \varepsilon_t.$$

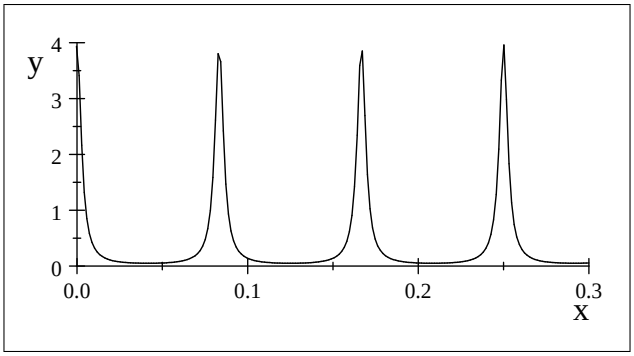
■

Esempio 97 *Per esempio, per il modello*

$$X_t = 0.8 \cdot X_{t-12} + \varepsilon_t$$

troviamo

$$S(f) = \frac{\sigma^2}{2\pi} \left| \frac{1}{1 - 0.8 \cdot e^{-2\pi i \cdot 12 \cdot f}} \right|^2$$



PSD del modello $X_t = 0.8 \cdot X_{t-12} + \varepsilon_t$

4.3 Il metodo di Holt-Winters

Lo scopo principale di questa sezione è quello di illustrare il metodo di Holt-Winters (HW), partendo dalle versioni particolari dette dello smorzamento esponenziale (SE) e dello smorzamento esponenziale con trend (SET). Vista la grande varietà di metodi di a quel punto disporremo (includendo ovvimente gli ARIMA), si pone sempre più urgentemente il problema del confronto tra essi, che perciò verrà trattato al termine di questa sezione.

4.3.1 Metodo di Smorzamento Esponenziale (SE)

Indichiamo con $x(t)$ il dato storico al tempo t, e con $p(t)$ la previsione relativa al tempo t (effettuata in un istante precedente; non è la previsione che viene effettuata al momento t, ma la previsione di ciò che sarebbe dovuto accadere al tempo t effettuta prima del tempo t). Così $p(t+1)$ sarà la previsione relativa al tempo $t+1$, e così via.

E' utile pensare che t sia il tempo presente, $t+1$ il futuro (primo tempo futuro), $t-1$ il passato (ultimo tempo prima del presente).

Il Metodo di Smorzamento Esponenziale sceglie come previsione $p(t+1)$ del futuro una media pesata tra la previsione $p(t)$ del presente, fatta in precedenza, ed il valore attuale $x(t)$ della serie storica:

$$p(t+1) = \alpha x(t) + (1-\alpha)\, p(t).$$

Il parametro α è (di solito) compreso tra 0 ed 1.

Se $\alpha = 1$, significa che decidiamo come previsione futura il valore odierno: $p(t+1) = x(t)$, cioè la pura ripetizione del presente. Se ad esempio la serie x ha delle oscillazioni, queste vengono riprodotte esattamente, in ritardo di un'unità temporale.

Se $\alpha = 0$, vale $p(t+1) = p(t)$, quindi $p(t+1) = p(t) = ... = p(1)$, cioè la previsione è costante. Scegliamo come previsione una costante (nella migliore delle ipotesi la media dei dati). Ignoriamo ogni struttura ulteriore, ogni variazione.

Se invece prendiamo $\alpha \in (0,1)$, mediamo tra questi due estremi: otterremo una previsione meno oscillante dei dati reali, ma non del tutto costante, lievemente concorde con l'ultimo dato.

La prvisione $p(t+1)$ è una media pesata tra un valore conservativo, $p(t)$, ed uno innovativo, $x(t)$.

Un'ulteriore interpretazione viene dalla formula che si ottiene applicando la ricorsione:

$$\begin{aligned} p(t+1) &= \alpha x(t) + (1-\alpha)\, p(t) \\ &= \alpha x(t) + (1-\alpha)\left(\alpha x(t-1) + (1-\alpha)\, p(t-1)\right) \\ &= \alpha x(t) + (1-\alpha)\left(\alpha x(t-1) + (1-\alpha)\left(\alpha x(t-2) + (1-\alpha)\, p(t-2)\right)\right) \end{aligned}$$

ecc. che si riscrive

$$p(t+1) = \alpha x(t) + \alpha(1-\alpha)\, x(t-1) + \alpha(1-\alpha)^2 x(t-2) + ...$$

Vediamo che la previsione futura è una media pesata di tutti i valori passati, con pesi che descrescono esponenzialmente (da qui il nome SE). Si sceglie come previsione una media pesata che dà più importanza agli ultimi valori rispetto ai precedenti. Quanta più importanza, lo decide α (α vicino ad 1 vuol dire molta più importanza ai valori più recenti).

Per certi versi somiglia ad un AR, ma l'ordine p è teoricamente infinito (anche se i pesi esponenziali diventano insignificanti dopo un certo punto in poi), e la struttura dei pesi è fissata, dipendente solo da un parametro, α, invece che da p parametri indipendenti.

Difetto certo: se la serie storica $x(t)$ ha un trend, per $\alpha = 0$ il metodo darebbe una costante, pessima previsione; per $\alpha = 1$, viceversa fa il meglio che può, ma può solo inseguire il trend in ritardo di un'unità temporale (si pensi a due rette parallele).

4.3.2 Metodo di Smorzamento Esponenziale con Trend (SET)

Indichiamo sempre con $x(t)$ il dato storico al tempo t e con $p(t)$ la previsione relativa al tempo t. L'idea ora è di avere un comportamento rettilineo (trend lineare), almeno localmente. Se siamo al tempo t, il presente, la previsione dei valori futuri $p(t+1)$, $p(t+2)$, ecc., in generale $p(t+i)$, con $i = 1, 2$, ecc. decisiamo che sia data dalla formula

$$p(t+i) = m(t) \cdot i + s(t)$$

(equazione della retta di coefficiente angolare $m(t)$ ed intercetta $s(t)$. E' utile pensare che l'asse delle ordinate sia collocato al tempo t, per farsi un'idea grafica.

L'idea di far dipendere m e s dal tempo è basilare: vogliamo sì una previsione con trend lineare, ma dobbiamo poterla modificare nel tempo, se il trend cambia.

Mentre nel metodo precedente legavamo i valori futuri di p a quelli passati, ora leghiamo i valori futuri delle grandezze ausiliarie m ed s a quelli passati. Continuiamo ad usare una logica di media pesata tra un valore conservativo ed uno innovativo. Per il oefficiente angolare m la prima idea che viene in mente è

$$m(t) = \beta(x(t) - x(t-1)) + (1-\beta)\, m(t-1)$$

media pesata tra il valore conservativo $m(t-1)$ e quello innovativo $x(t) - x(t-1)$, che è la pendenza osservata sui dati dell'ultimo periodo. Ma così ci si espone troppo alle fluttuazioni casuali dei dati: la pendenza $x(t) - x(t-1)$ può deviare marcatamente dall "pendenza media" dell'ultimo periodo. Serve una grandezza simile a $x(t) - x(t-1)$ ma più stabile, meno esposta alle fluttuazioni causali. Essa è $s(t) - s(t-1)$, come capiremo tra un momento.

Veniamo alla ricorsione per $s(t)$. Se disegnamo un grafico con due assi verticali delle ordinate, uno al tempo $t-1$ ed uno al tempo t, vediamo che l'intercetta al tempo t non deve essere simile all'intercetta al tempo $t-1$ ma al valore sulla retta $m(t-1) \cdot i + s(t-1)$. I due istanti $t-1$ e t distano di un'unità, quindi $s(t)$ è giusto che sia legata a $m(t-1) + s(t-1)$ (cioè $i = 1$). Questa è la parte conservativa della ricorsione. La parte innovativa è dire che $s(t)$ deve essere legato al valore vero $x(t)$. Quindi $s(t)$ sarà una media pesata tra $x(t)$ e $m(t-1) + s(t-1)$. In conclusione, le due equazioni ricorsive sono:

$$\begin{aligned} s(t) &= \alpha x(t) + (1-\alpha)(m(t-1) + s(t-1)) \\ m(t) &= \beta(s(t) - s(t-1)) + (1-\beta)\, m(t-1). \end{aligned}$$

In pratica è necessario calcolare prima $s(t)$ dalla prima equazione, per poterlo sostituire nella seconda.

Vediamo che il metodo è proprio innovativo rispetto a SE ed anche agli AR.

Il metodo è ottimo per catturare i trend. Ma se c'è anche un'evidente periodicità, il metodo non riesce a riconoscerla come tale, quindi la insegue come se fosse una modifica continua del trend, e commette troppi errori.

Inizializzazione di SE ed SET

Le equazioni per ricorrenza vanno inizializzate. Supponiamo che la serie temporale $x(t)$ parta da $t = 1$. Per SE, si tratta di stabilire $p(1)$, la previsione al primo istante temporale.

In questo modo, se ci troviamo al tempo $t = 1$ (nostro presente) e conosciao quindi $x(1)$, potremo poi calcolare la previsione futura $p(2)$:

$$p(2) = \alpha x(1) + (1-\alpha) p(1).$$

Quando poi ci troveremo al tempo $t = 2$, che sarà diventato il nostro presente, conosceremo $x(2)$ e potremo calcolare la previsione futura $p(3)$:

$$p(3) = \alpha x(2) + (1-\alpha) p(2).$$

E così via. Il valore di inizializzazione $p(1)$ è abbastanza casuale da scegliere. Per semplicità si può prendere ad esempio $p(1) = x(1)$.

Per SET, se ci troviamo al tempo $t = 1$ e vogliamo calcolare le previsioni $p(1+i)$, $i = 1, 2, ...$ servono i valori $m(1)$ e $s(1)$. Vista la natura di s descritta sopra (si pensi al grafico con l'asse verticale per $t = 1$), è naturale prendere $s(1) = x(1)$. La pendenza iniziale però è indecidibile senza vedere il futuro, quindi scegliamo $m(1) = 0$, salvo abbiamo informazioni diverse di tipo previsivo. Fatte queste scelte, possiamo calcolare le previsioni

$$p(1+i) = m(1) \cdot i + s(1), \qquad i = 1, 2, ...$$

Poi il tempo $t = 2$ diventerà il nostro presente. Calcoleremo $m(2)$ e $s(2)$ con le formule (ora $x(2)$ è noto)

$$\begin{aligned} s(2) &= \alpha x(2) + (1-\alpha)(m(1) + s(1)) \\ m(2) &= \beta(s(2) - s(1)) + (1-\beta) m(1) \end{aligned}$$

da usarsi nell'ordine scritto. Avendo calcolato $m(2)$ e $s(2)$, la previsione del futuro è

$$p(2+i) = m(2) \cdot i + s(2), \qquad i = 1, 2, ...$$

Si noti che il valore

$$p(1+2) = m(1) \cdot 2 + s(1)$$

calcolato al tempo $t = 1$ ed il valore

$$p(2+1) = m(2) \cdot 1 + s(2)$$

sono entrambe delle previsioni relative al tempo $t = 3$. Ci fideremo ovviamente più della seconda, in quanto basata su più dati. per così dire, $m(2) \cdot 1 + s(2)$ è una sorta di aggiornamento di $m(1) \cdot 2 + s(1)$. Lo stesso vale per i valori successivi.

Si può inizializzare SET in un secondo modo: attendere alcuni istanti in più prima di iniziare la previsione, ed utilizzarli per calcolare una retta di regressione, da usarsi come stima iniziale della pendenza. Chiaramente, più è lungo il periodo iniziale che attendiamo, più è precisa la stima della pendenza, quindi le previsioni inizieranno molto megli che con la semplice posizione $m(1) = 0$. Ma è anche vero che, se iniziamo all'istante iniziale con $m(1) = 0$, dopo alcuni istanti questa anomalia sarà stata automaticamente aggiustata dal metodo, tramite le iterazioni che correggono di volta in volta m tramite i valori degli incrementi $s(t) - s(t-1)$. Quindi alla fine le cose si equivalgono abbastanza: i primi istanti o non vengono proprio previsti oppure sono previsti un po' male; i successivi vengono previsti piuttosto bene.

4.3.3 Smorzamento esponenziale con trend e stagionalità (Holt-Winters)

Di questo metodo esiste una versione per stagionalità additiva ed una per quella moltiplicativa; descriviamo solo quest'ultima, essendo l'altra del tutto simile e forse più elementare. Si ipotizza il modello

$$x(t) = (at + b) F(t) + \varepsilon(t)$$

con $F(t)$ funzione periodica di periodo P. Per capire nel modo più semplice possibile come sono state idete le equazioni ricorsive del modello, fingiamo di non avere il rumore $\varepsilon(t)$, quindi di lavorare sull'equazione

$$x(t) = (at + b) F(t).$$

Idealmente, si introduce la grandezza ausiliaria $y(t) = \frac{x(t)}{F(t)}$ che soddisfa

$$y(t) = at + b.$$

A questa possiamo applicare quindi lo smorzamento con trend. Detta p_y la previsione di y e detti s_y, m_y i valori calcolati dal metodo SET relativamente ad y, si trova

$$p_y(t + i) = m_y(t) \cdot i + s_y(t)$$

dove (si noti che c'è $y(t)$ e non $x(t)$)

$$\begin{aligned} s_y(t) &= \alpha y(t) + (1 - \alpha)(m_y(t - 1) + s_y(t - 1)) \\ m_y(t) &= \beta(s_y(t) - s_y(t - 1)) + (1 - \beta) m_y(t - 1). \end{aligned}$$

Il problema è che per innescare questo sistema bisogna conoscere $y(t)$ e per questo bisognerebbe conoscere $\frac{x(t)}{F(t)}$, mentre $F(t)$ per ora è incognita. L'idea è di stimare anche la funzione periodica F in modo iterativo, così da aggiustarla se è il caso. Allora al posto di $y(t)$ si mette $\frac{d(t)}{F(t-P)}$, immaginando che nella struttura iterativa che troveremo alla fine il valore $F(t - P)$ sia noto (e riteniamo sia una buona approssimazione di $F(t)$ in quanto cerchiamo una F periodica).

Poi bisogna creare un'equazione iterativa per F. Un'idea ispirata alla filosofia dello smorzamento esponenziale è

$$F(t) = \gamma \frac{x(t)}{y(t)} + (1 - \gamma) F(t - P)$$

(si ricordi la definizione $y(t) = \frac{x(t)}{F(t)}$; se non si mettesse alcun termine legato al passato useremmo l'equazione $F(t) = \frac{x(t)}{y(t)}$). Però non conosciamo $y(t)$. Noti però $s_y(t)$ ed $m_y(t)$, $s_y(t)$ è una stima di $y(t)$. In definitiva, si arriva al sistema:

$$\begin{aligned} s(t) &= \alpha \frac{x(t)}{F(t - P)} + (1 - \alpha)(m(t - 1) + s(t - 1)) \\ m(t) &= \beta(s(t) - s(t - 1)) + (1 - \beta) m(t - 1) \\ F(t) &= \gamma \frac{x(t)}{s(t)} + (1 - \gamma) F(t - P) \end{aligned}$$

dove abbiamo smesso di indicare y a pedice in quanto ormai usiamo queste equazioni come equazioni finali per stimare x.

Inizializzazione di HW

L'inizializzazione qui è più complessa. Serve F su un intero periodo per innescare l'iterazione. Allora si sacrifica il primo periodo (a volte più di uno), su quello si trova una retta di regressione

$$z(t) = at + b$$

e la si usa come se fosse una stima di $y(t)$. Quindi dalla definizione $y(t) = \frac{x(t)}{F(t)}$ si stima

$$F(t) = \frac{x(t)}{at + b}.$$

In definitiva, per $t = 1, 2, ..., P$ si prendono questi valori di F e poi si pone $s(P) = aP + b$, $m(P) = a$. Si comincia quindi a prevedere il valore al tempo $P + 1$.

4.3.4 Confronto tra modelli previsionali: i) cross-validation

In generale, con questo nome si intende l'idea di suddividere il set di dati in due parti, la prima chiamata *training set*, la seconda chiamata *test set*. Si usa poi il training set per contruire il modello, con esso si effettuano le previsioni, e le si confrontano col test set. E' un'idea utilizzabile in vari ambiti, non solo per le serie storiche.

Nel caso di serie storiche, bisogna scegliere due finestre temporali, successive una all'altra, prendendo come training set la prima, quella dei dati più vecchi. Si veda l'esempio dell'esercizio della lezione 13.

Il confronto tra due modelli si può fare visivamente, oppure sulla base della deviazione standard (o della varianza) degli errori o residui, quelli calcolati sul test set; si veda di nuovo l'esercizio della lezione 13.

Non è necessario che le due serie (training e test sets) coprano tutta la serie nota. Può aver senso escludere una parte iniziale della serie nota, in quanto troppo vecchia e non più rappresentativa. Oppure escludere una parte finale (nel senso di non considerarla proprio, non nel senso di prenderla come test set), se di nuovo è anomala, non rappresentativa di situazioni tradizionali; è quanto accade nell'esercizio della lezione 13: usando come test set gli ultimi due anni, che coincidono proprio col periodo di crisi economica, tutti i metodi sbagliano parecchio le previsioni e quindi risulta abbastanza vanificato il contronto tra essi.

(Per inciso, si noti come questa analisi sia una conferma della presenza della crisi economica, nonche' un modo per quantificarla: si vede ad esempio che i mesi coi valori più alti vengono sovrastimati, nella previsione, di un 30%, cioè ciò che si è perso in quei mesi a causa della crisi; invece i mesi medio-bassi sono rimasti pressoche' inalterati.)

Quando parliamo di modello, in genere c'è l'equivoco tra modello in quanto classe dipendente da parametri, oppure specifico esempio con certi valori dei parametri. Nel senso: modello di Holt-Winters (come classe) oppure il modello HW con $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 0.1$ (singolo modello specifico). Bene, la cross-validation confronta modelli nel primo senso, classi. Infatti, quando andiamo a determinare i parametri del modello tramite training set, troveremo parametri diversi da quelli del modello basato su tutta la serie storica (modello che usiamo per le vere predizioni future). Quindi, alla fine della cross-validation, non avremo confrontato quello specifico modello HW con parametri calcolati su tutti i dati, con lo specifico modello

AR calcolato sugli stessi. Avremo confrontato la classe HW con la classe AR, relativamente alla serie storica in oggetto.

Quindi, forse, la cross-validation non va usata troppo di fino. Nell'esercizio n.9, i valori di σ sono quasi uguali (rispetto all'unità di misura del problema, dell'ordine di 100) ed anche graficamente le previsioni sembrano molto simili, pur essendo abbastanza sbagliate tutte e due. Quindi non ha molto senso concludee che HW è migliore per via del piccolissimo vantaggio in σ. Invece, una grossa differenza in cross-validation può essere una chiarissima indicazione che un certo metodo è inadatto per un certo tipo di serie storica. Ad esempio, SET non è certo adatto alla serie assai periodica dei motorcycles, lo capiamo a priori. Ma se non avessimo questa (ovvia) intuizione, applicando la cross-validation avremmo una conferma schiacciante.

4.3.5 Confronto tra modelli previsionali: ii) metodo del "conflitto di interessi"

Con questo nome, non tradizionale ed usato qui in senso scherzoso, indichiamo una variante delle idee precedenti di uso estremamente comune. In essa non si suddivide il set di dati in due parti, ma lo si usa sia come training set sia come test set! Sarebbe giusto replicare subito: ma è ovvio che un modello costruito su certi dati farà delle buone previsioni di essi stessi (da qui il nome del paragrafo). Ancor peggio: se ho una serie storica di 125 dati, posso costruire un modello a 125 parametri che descrive i dati nel modo migliore possibile: la serie storica stessa! Il suo errore di previsione relativo ai dati stessi è nullo.

Ma ovviamente, tale modello (la serie stessa) è del tutto inutile per previsioni future; non ha catturato alcun elemento strutturale da poter replicare. E' il problema dell'*overfitting*. Da qui nasce la necessità di cercare un trade-off tra precisione del modello sui dati stessi ed economia del modello in termini di numero di parametri. Il coefficiente *AIC* risponde a questa esigenza. Comunque non è di questo che vogliamo parlare.

Nonostante la critica detta sopra, scelti a priori un certo numero di modelli base, tipo HW, AR con AIC ottimale, decomposizione con stl per $k = 6$ (per esempio), si sta implicitamente ponendo una limitazione al grado di precisione raggiungibile, al numero di parametri, per cui ha senso valutare le performances relative di questi metodi, gli uni rispetto agli altri.

Ciò non esclude che, ad esempio, variando k in stl, si migliorino le performances, ma in tal caso bisogna stare attenti all'overfitting. Ad esempio, per k molto piccoli, la componente periodica di stl è quasi uguale alla serie stessa, e quindi stiamo cadendo nel paradosso descritto sopra: essa ha praticamente gli stessi gradi di libertà della serie originaria. I residui saranno piccolissimi. Ma il suo potere previsivo per il futuro è quasi nullo: possiamo al massimo replicare l'ultimo anno tale e quale, cosa che potevamo fare anche senza stl.

Fatte tutte queste premesse, che mettono in guardia dal trarre conclusioni troppo sicure dalle risposte offerte da questo metodo, vediamone l'implementazione.

Passo 1: si applica la classe di modelli scelta (HW, AR ecc.) sulla serie training, trovando i parametri del modello (es. α, β, γ per HW, oppure l'ordine p ed i coefficienti $a_1, ..., a_p$ per AR con AIC).

Passo 2: si calcolano i residui, *relativi al periodo noto della serie* (è come usare la serie stessa come test set), esclusa la fase iniziale di inizializzazione; a volte, per calcolare i residui,

bisogna prima calcolare esplicitamente le previsioni, altre volte il calcolo della previsione rimane implicito e si calcolano direttamente i residui. Torniamo in dettaglio su questo punto tra un momento.

Passo 3: si calcola la deviazione standard σ_ε dei residui, che descrive lo scostamento tipico, lungo il periodo noto, tra previsioni date dal modello e dati veri. Se si fa tutto questo relativamente a diversi modelli, sarebbe meglio escludere un intervallo comune iniziale di valori (anche se non è strettamente necessario, in quanto σ_ε già tiene conto del numero di dati). Il modello migliore (secondo questa filosofia del "conflitto di interessi") è quello con σ_ε più basso.

Notiamo che è equivalente calcolare la varianza dei residui, σ_ε^2, oppure la varianza spiegata

$$1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2}$$

dove σ_X^2 è la varianza dei dati. Naturalmente vince il modello con la *maggior* varianza spiegata.

Alcuni commenti sul passo 2. In pratica, i residui dei modelli che stiamo studiando (es. quelli dell'esercizio della lezione 13) ci sono forniti dal software, usando i comandi appropriati (vedere l'help dei singoli comandi). Però, teoricamente, di cosa stiamo parlando? Vediamo i tre metodi separatamente.

Decomposizione. La serie storica nota X_t viene scomposta nella somma di trend, stagionalità e residui

$$X_t = T_t + S_t + \varepsilon_t.$$

Per residuo al tempo t, qui si intende ovviamente il numero ε_t. Potremmo anche dire che $T_t + S_t$ è la previsione al tempo t, ma è superfluo.

AR. Si ipotizza che la serie storica nota X_t soddisfi l'equazione ricorsiva

$$X_t = a_1 X_{t-1} + ... + a_p X_{t-p} + \varepsilon_t$$

(eventualmente dopo aver sottratto un valor medio μ). Per residuo al tempo t, qui si intende il numero ε_t, calcolabile dai dati come differenza tra il valore vero X_t e quello previsto $a_1 X_{t-1} + ... + a_p X_{t-p}$:

$$\varepsilon_t = X_t - (a_1 X_{t-1} + ... + a_p X_{t-p}).$$

HW. Qui i residui si vedono in modo più indiretto. L'iterazione coinvolge grandezze ausiliarie, $s(t)$, $m(t)$, $F(t)$, che usano i dati per essere aggiornate. Note queste grandezze al tempo $t-1$, si calcola la previsione $p(t)$ del valore al tempo t con la formula

$$p(t) = F(t)(m(t-1) + s(t-1))$$

e poi i residui con la formula

$$\varepsilon_t = X_t - p(t).$$

Comunque, come dicevamo, lo fa il software, basta saperlo chiedere.

4.3.6 Esercizi sul confronto tra modelli previsionali

Gli esercizi che seguono si riferiscono all'esercizio n. 9 della sezione seguente.

Esercizio 28 *Applicare la cross-validation come nell'esercizio n. 9, ma prevedendo un paio di anni attorno al 2007, a partire dai precedenti.*

Esercizio 29 *Usare il metodo SET per l'esempio dell'esercizio n. 9 e valutare le sue prestazioni rispetto a HW, tramite cross-validation.*

Esercizio 30 *Usare il metodo SET ed il metodo HW sulla serie dei cereali e valutare le prestazioni con la cross-validation.*

Esercizio 31 *Relativamente ai dati dell'esercizio dell'esercizio n. 9,*
i) estrarre i residui dai metodi HW, AR, stl
ii) ritagliare una finestra comune (escludere un tratto iniziale comune)
iii) calcolare la varianza spiegata nei tre casi, osservando quale è migliore.

4.4 Metodi regressivi

L'esposizione qui di questo argomento pone dei problemi di consequenzialità, in quanto si basa tecnicamente sulla regressione lineare multipla che verrà descritta nel capitolo di statistica multivariata. Ci limitiamo quindi ad alcune idee, che devono essere riprese.

4.4.1 AR come regressione lineare multipla

Un modello di regressione lineare multipla è un modello della forma

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

dove le v.a. $X_1, ..., X_p$ sono dette fattori, predittori, input, la v.a. Y è l'output, la grandezza da predire, da spiegare, la variabile dipendente, ε è un termine aleatorio di errore, e $a_1, ..., a_p, b$ sono i coefficiendi del modello (b è l'intercetta). La logica è quella di avere una v.a. Y di cui vorremmo capire di più; nel momento in cui valga un modello regressivo del tipo descritto, capiamo che Y è influenzata dai fattori $X_1, ..., X_p$, secondo l'ampiezza dei coefficienti $a_1, ..., a_p$. La variabilità di Y, precedentemente oscura, viene parzialmente spiegata dal modello (parzialmente, in quanto c'è sempre il termine di errore casuale ε).

La forma algebrica di queste equazioni è evidentemente molto simile a quella dei modelli AR(p):

$$X_t = \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + b + \varepsilon_t$$

solo che qui le diverse variabili compongono un unico processo stocastico. Ma la logica è la stessa appea descritta per la regressione: si immagina che i valori assunti da X_t siano influenzati, spiegati dai valori di $X_{t-1}, ..., X_{t-p}$, tramite i pesi $\alpha_1, ..., \alpha_p$, a meno dell'errore ε_t.

E' quindi chiaro che, una volta che saranno note le procedure di calcolo della regressione lineare, queste possono essere applicate ai modelli AR(p). Siccome quelle procedure sono improntate al metodo dei minimi quadrati, in sostanza è come se si stesse applicando il comando `ar.ols` ad una serie storica.

4.4.2 Implementazione con R

Anticipando l'uso del comando `lm` che esegue la regressione lineare multila, descriviamo l'implementazione con `R` di quanto ora spiegato. Iniziamo col modello AR(1)

$$x_k = a_1 x_{k-1} + b + \varepsilon_k.$$

Dato il vettore `x` di lunghezza n, si devono costruire due vettori di lunghezza $n-1$, ovvero

```
x0<-x[2:n]
x1<-x[1:(n-1)]
```

Poi si esegue la regressione `REG<-lm(x0~x1)`. Essa pone in relazione lineare il primo termine di `x0` come output col primo termine di `x1` come input, che sono x_2 e x_1, e così via, fino all'ultimo termine di `x0` con l'ultimo termine di `x1`, che sono x_n e x_{n-1}.

Per il modello AR(2)

$$x_k = a_1 x_{k-1} + a_2 x_{k-2} + b + \varepsilon_k$$

si introducono

```
x0<-x[3:n]
x1<-x[2:(n-1)]
x2<-x[1:(n-2)]
```

e si esegue `REG<-lm(x0~x1+x2)`. La scelta fatta dei vettori pone in relazione lineare il primo termine di `x0` come output, che è x_3. coi primi termini di `x1` e `x2` come input, che sono x_2 e x_1, e così via.

Per il modello

$$x_k = a_1 x_{k-1} + a_{12} x_{k-12} + b + \varepsilon_k$$

si introducono

```
x0<-x[13:n]
x1<-x[12:(n-1)]
x12<-x[1:(n-12)]
```

e si esegue `REG<-lm(x0~x1+x12)`. La scelta fatta dei vettori pone in relazione lineare il primo termine di `x0` come output, che è x_{13}. coi primi termini di `x1` e `x12` come input, che sono x_{12} e x_1, e così via.

Dati questi comandi, possiamo leggere in `summary(REG)` le caratteristiche della regressione appena eseguita, come ad esempio la varianza spiegata R^2, l'importanza dei diversi fattori (tanto maggiore quanto più piccolo è il corrispondente $Pr(>|t|)$) e possiamo leggere i valori dei coefficienti $a_1, ..., b$. Tali valori sono anche estraibili col comando `REG$coefficients[i]` dove per $i = 1$ si ottiene l'intercetta b, per $i = 2$ si ottiene il primo dei coefficenti a e così via.

Noti i coefficienti, possiamo usare il modello per fare delle previsioni.

4.4.3 Previsione col modello regressivo

Bisogna distinguere tra due scopi previsivi entrambi importanti: il più naturale è ovviamente quello di voler prevedere il futuro; ma per motivi tecnici è anche molto importante osservare come i metodi o modelli che stiamo usando si comportano nel prevedere il passato già noto. Questa bipolarità apparirà spesso nel seguito.

Per uniformare un po' le notazioni, introduciamo il seguente modo di lavorare, che però non è obbligatorio (si possono seguire notazioni e convenzioni diverse). Sia `X` il vettore di lunghezza n che rappresenta la nostra serie storica. Introduciamo un vettore `P` che rappresenta la previsione. Se vogliamo solamente vedere come metodo si comporta nel prevedere i dati noti, prenderemo `P` della stessa lunghezza di `X`. Quindi un modo ovvio di crearlo è `P=X`. Se invece vogliamo prevedere il futuro, ad esempio i prossimi 12 mesi, prenderemo come `P` un vettore di lunghezza $n+12$, creabile con `P=1:(n+12)`. Naturalmente i vettori `P` così creati non contengono ancora le previsioni: sono come dei contenitori vuoti.

Previsione dei valori futuri tramite modelli lineari

Iniziamo col problema della previsione futura. Procediamo per esempi.

Consideriamo il modello più semplice:

$$x_k = a_1 x_{k-1} + b + \varepsilon_k.$$

e supponiamo di aver eseguito la regressione `REG<-lm(x0~x1)`. Posto

```
a1<-REG$coefficients[2]
b<-REG$coefficients[1]
P[1:n]<-X
```

eseguiamo il ciclo di for

```
for (k in (n+1):(n+12)){
P[k]=a1*P[k-1]+b
}
```

Il vettore `P` conterrà nella prima parte, da 1 a n, la serie storica nota, mentre nella seconda parte, da $n+1$ a $n+12$, conterrà la previsione dei prossimi 12 mesi. Per capire che abbiamo calcolato le cose giuste, si ragioni passo a passo:

- per $k = n+1$, la previsione `P[n+1]` del modello lineare deve essere data dalla forma del modello stesso, cioè uguale ad $a_1 x_n + b$, ovvero a `a1*X[n]+b`, che coincide con `a1*P[n]+b` in quanto abbiamo posto `P[1:n]<-X`;
- per $k = n+2$, la previsione `P[n+2]` del modello lineare deve essere data dalla forma del modello stesso, cioè uguale ad $a_1 x_{n+1} + b$, ma x_{n+1} non lo conosciamo (i dati arrivano solo fino al tempo n), quindi al suo posto usiamo la previsione al tempo $n+1$, quindi `a1*P[n+1]+b`; e così via.

La generalizzazione ad altri modelli più complicati è abbastanza immediata. Vediamo ad esempio il modello

$$x_k = a_1 x_{k-1} + a_{12} x_{k-12} + b + \varepsilon_k.$$

Eseguita `REG<-lm(x0~x1+x12)`, poniamo

```
a1<-REG$coefficients[2]
a12<-REG$coefficients[3]
b<-REG$coefficients[1]
P[1:n]<-X
```

e poi eseguiamo il ciclo di for

```
for (k in (n+1):(n+12)){
P[k]=a1*P[k-1]+a12*P[k-12]+b
}
```

Tutto questo ha senso se $n \geq 12$.

Fatte le previsioni, nasce il desiderio di raffigurare la serie storica insieme alle previsioni. Con i comandi

```
ts.plot(P, col = ''red'')
lines(X, col=''blue'')
```

si ottiene in blu il grafico dei dati noti ed in rosso la previsione dei 12 mesi successivi. Se avessimo anche i dati noti dei 12 mesi successivi, detto `X2` il vettore complessivo dei dati noti, basterebbe usare `X2` al posto di `X` nel precedente comando.

Una nota necessaria: in alcune applicazioni macroeconomiche le cose vanno bene come in questo esempio, in quanto si studiano grandezze molto stabili nel tempo, ottenute mediando su tantissimi sottosistemi (qui è la grande distribuzione di alimentari a livello nazionale). Se invece si studiano problemi a scala più piccola, come le vendite di un prodotto di una media impresa, le cose cambiano e le previsioni diventano assai meno precise.

Previsione mese per mese

Sopra, in esempi del tipo 24 dati noti e 12 futuri incogniti, abbiamo eseguito la previsione di tutti i 12 mesi futuri in blocco. Stiamo immaginando di trovarci a dicembre di un certo anno e voler prevedere le vendite dell'anno successivo, per fare un esempio.

Diversamente, un modello lineare può essere usato mese per mese, man mano che si hanno nuovi dati veri. A fine dicembre 2008, eseguiamo pure la previsione di tutto il 2009, ma poi, a fine gennaio 2009, noto il valore vero di gennaio, potremo migliorare le previsioni del resto del 2009: ad esempio, per il mese di febbraio 2009, invece di usare la formula

```
P[febbr09]=a1*P[genn09]+a12*X[febb08]+b
```

useremo la formula più precisa

```
P[febbr09]=a1*X[genn09]+a12*X[febb08]+b
```

in cui si fa uso del valore vero di gennaio 2009.

L'implementazione con `R` della previsione mese per mese va fatta appunto mese per mese, non si può scrivere in blocco all'inizio: ogni mese si deve adattare la formula generale usando i nuovi dati acquisiti ovunque è possibile nella formula. Se ad esempio ci fosse un termine del tipo $a_6 x_{k-6}$, cioè una periodicità semestrale (rara ma presente in certi fenomeni più legati a scelte sistematiche), useremmo `a6*P[k-6]` fino al sesto mese 2009, ma dal settimo potremmo usare `a6*X[k-6]`.

4.4.4 Variabili esogene, cross-correlazione, modelli ARX

Tra le caratteristiche uniche di questa metodologia (quella che implementa gli AR tramite la regressione) c'è la possibilità di inserire, tra i predittori, serie storiche diverse da quella data, serie che riteniamo possano essere utili per la previsione della serie data. Possiamo quindi

creare modelli del tipo (detti ARX, caso particolare degli ARIMAX)

$$x_k = a_1 x_{k-1} + ... + c_1 z_{k-1} + ... + b + \varepsilon_k$$

dove $z_1, ..., z_n$ è un'altra serie storica (ovviamente si possono usare, come predittori, diverse serie storiche). Ad esempio, si può immaginare che il costo di certi beni di consumo siano influenzati dal prezzo del petrolio nel mese precedente. Allora x_k è il costo del bene in oggetto, z_{k-1} è il prezzo del petrolio nel mese precedente, x_{k-1} è il costo del bene considerato, relativo al mese precedente. E' chiara la flessibilità di questo metodo. Purtroppo la sua applicazione pratica richiede pazienza ed arte.

Prima di buttarsi nell'uso di tali modelli conviene assicurarsi che ci sia un legame tra le serie storiche che si vogliono mettere insieme, qui chiamate $z_1, ..., z_n$ e $x_1, ..., x_n$. Il concetto di cross-correlazione introdotto nel capitolo sui processi stocastici è molto utile a questo scopo. Esso calcola la correlazione tra le due serie (opportunamente troncate), con tutte le traslazioni possibili. In altre parole, per ogni $k = 1, ..., n-1$ calcola la correlazione tra le serie

$$z_1, ..., z_{n-k}$$
$$x_{k+1}, ..., x_n$$

Naturalmente il risultato è statisticamente sigificativo solo per k basso. Il software R, tramite il comando `ccf`, esegue questo calcolo anche per k negativi. Attenzione quando lo si usa: l'ordine con cui si danno le due serie al comando `ccf` è essenziale; per k positivi si vedrà la correlazione tra l'una e le traslazioni in avanti dell'altra, e viceversa, ma come essere sicuri dell'ordine? Rileggendo l'help del comando. Esso recita: The lag k value returned by ccf(x,y) estimates the correlation between x[t+k] and y[t]. Quindi, supponiamo che si voglia scoprire se una certa serie storica x influisce su una y qualche mese dopo (i valori di x di gennaio si ripercuotono sui valori di y di marzo, e così via, ad esempio). Allora ccf(x,y) con k=-2 ci dice come x[t-2] è collegata a y[t].

Se la correlazione tra $z_1, ..., z_{n-k}$ e $x_{k+1}, ..., x_n$ è elevata (vicina ad 1), allora queste serie hanno un legame. Se lo riteniamo sensato dal punto di vista applicativo, possiamo estrapolare che $z_1, ..., z_{n-k}$ influisca su $x_{k+1}, ..., x_n$ e che quindi abbia senso impostare un modello ARX, che spiega la serie x non solo attraverso la propria struttura ricorsiva, ma anche facendo uso del predittore z.

Molto utile, come output della `ccf`, è scoprire per quali k la correlazione è maggiore. Questi ci dicono *con che ritardo* la serie z influisce su x (se di influenza si può parlare, si vedano gli avvertimenti nel capitolo sulla regressione). Infatti, ad esempio, se è ragionevole pensare che il prezzo del petrolio influenzi il prezzo di certi beni di consumo, il ritardo temporale con cui questo avviene magari non è ovvio a priori e può essere desunto dalle serie storiche guardando i picchi della `ccf`.

Un avvertimento. Relativente ad una serie di dati, se si aumenta l'ordine p del metodo AR, oppure se si introducono predittori come ora descritto, è chiaro che un poco aumenterà la precisione con cui il modello descrive i dati noti. Ma non è detto che questo corrispoda ad una realtà stutturale. Quindi, se la precisione aumenta pochissimo dopo aver introdotto un predittore questo probabilmente significa che esso non influisce realmente, solo fittiziamente.

4.5 Fit di una densità

Uno dei problemi più comuni nelle applicazioni della probabilità è quello di trovare una densità di probabilità che descriva adeguatamente un fenomeno, di cui si conoscono alcuni dati sperimentali. Ad esempio, note le vendite giornaliere per 20 giorni lavorativi, si può tentare di trovare una densità che descriva i dati abbastanza accuratamente.

Avere tale densità è utile per vari scopi, incluso quello di calcolare intervalli in cui cadranno i valori con elevata probabilità, soglie utili per dimensionamento di servizi, magazzini, ecc.

Questo argomento di può applicare ad esempio alle serie storiche. Prendiamo ad esempio la serie storica delle esportazioni di veicoli, lezioni 3 e seguenti. Prendiamo la serie IT3 registrata nel file R della lezione 4 (si apra ad esempio il file R della lezione 12, che ci servirà anche nel seguito). Osserviamola con ts.plot: ha un andamento piuttosto complesso, non facile da prevedere, come abbiamo poi visto nelle varie lezioni. Un modo un po' sommario di fare previsioni future può essere il seguente:

1. si escludono gli anni troppo distanti nel tempo, non più rappresentativi, conservando solo gli ultimi, ad esempio si conservano solo gli ultimi 4 anni (globalmente c'è un trend a crescere, ma negli ultimi 4 anni non è così accentuato)

2. si ignora la struttura temporale, l'eventuale stagionalità, e si usano i 48 dati come fossero 48 registrazioni passate, tutte rappresentative di un generico mese futuro in cui vogliamo prevedere le esportazioni

3. si cerca una densità che descriva i 48 dati

4. una volta trovata, possiamo ad esempio calcolare un intervallo in cui cadono le esportazioni con probabilità 90%; se ad esempio ci interessa sapere che le esportazioni sono maggiori di una certa soglia minima λ al 90%, cerchiamo il quantile al 10% della densità trovata.

Oppure, più raffinatamente, se si possiede un buon modello della serie storica, come avviene ad esempio per la serie dei *motorcycles*, lezione 12 e seguenti (si trova nel file R della lezione 12), si può lavorare nel seguente modo:

1. si usa il modello prescelto per fare previsioni

2. si considerano i residui (del passato) e si cerca una densità che li descriva; anche qui eventualmente si possono considerare solo i residui recenti, se sembra opportuno da una loro analisi grafica

3. una volta trovata una densità che descrive i residui, la si usa per calcolare intervalli simili a quelli discussi sopra, ricordandosi che tali intervalli vanno centrati nel valore previsto col modello di previsione.

4.5.1 Istogrammi e cumulative empiriche

Per prima cosa, raffiguriamo qualche istogramma del campione. Un *istogramma* è una sorta di *densità empirica*. Solo che non è univocamente determinato a partire dai dati: dipende dalle classi che si usano. I seguenti comandi vanno applicati al file R della lezione 12. Come detto sopra, visualizziamo la serie delle esportazioni di veicoli, detta IT3, col comando `ts.plot(IT3)`, decidiamo di usare solo gli ultimi 48 dati, li selezioniamo con `I.rec<-IT3[(168-48):168]`, li ri-plottiamo per sicurezza con `ts.plot(I.rec)`, poi vediamo due istogrammi con `par(mfrow=c(1,2)) ; hist(I.rec); hist(I.rec,10)`:

Si può anche fare un plot della cumulativa, con `plot.ecdf(I.rec)`.

4.5.2 Metodi parametrici e metodi non parametrici

L'uso di un metodo parametrico consiste nella scelta di una classe (Weibull, normale, ecc.) caratterizzata da pochi parametri (di solito 2) e ricerca dei parametri più opportuni in quella classe; confrontando i risultati relativi a più classi.

I metodi non parametrici consistono nella ricerca della densità in classi di funzioni definite da moltissimi gradi di libertà. In realtà anche queste classi sono parametrizzate, ma con un insieme di parametri di dimensione così elevata (a volte teoricamente infinita) da raggiungere un elevatissimo grado di flessibilità ed adattamento ai dati.

Gli istogrammi fatti sopra possono servire a orientarci nella ricerca della classe, per i metodi parametrici.

Riguardo ai metodi non parametrici, i comandi sono:

```
require(KernSmooth)
density <- bkde(I.rec, kernel = "normal", bandwidth=20)
plot(density, type="l")
```

Essi caricano il package `KernSmooth` (kernel smoothing) che non è tra quelli usuali, quindi non viene caricato di default. Questo package si occupa di trovare densità di probabilità non parametriche, usando procedimenti di smoothing basati su opportuni kernel. Leggendo l'help vediamo che ci sono vari kernel. Nel grafico successivo ne abbiamo usato un altro.

Il pregio di questo metodo è di adattarsi ai dettagli dell'istogramma. Ad esempio una bi-modalità. Si possono sovrapporre con

```
hist(I.rec,10,freq=FALSE)
lines(density, type="l")
```

4.5.3 Stima dei parametri

Supponiamo di aver scelto una classe e di voler determinare i parametri ottimali.

Ci sono varie strade. Le due più classiche sono il *metodo di Massima Verosimiglianza* ed il *metodo dei momenti*. In alternativa, ad esempio, si possono ottimizzare i parametri secondo un indicatore da noi scelto, come la distanza L^1 descritta nel seguito. Qui descriviamo solo la massima verosimiglianza.

Data una densità $f(x)$, dato un possibile risultato sperimentale x', il numero $f(x')$ non è la probabilità di x' (essa è zero). Viene però detta la *verosimiglianza* di x'.

Dato poi un campione $x_1, ..., x_n$, il prodotto

$$L(x_1, ..., x_n) = f(x_1) \cdots f(x_n)$$

porta il nome di *verosimiglianza* di $x_1, ..., x_n$. Dietro la scelta del prodotto c'è un pregiudizio di indipendenza del campione, che andrà valutato caso per caso.

Quando la densità dipende da alcuni parametri, diciamo ad esempio a, s, scriveremo $f(x|a, s)$ ed anche $L(x_1, ..., x_n|a, s)$, alludendo informalmente ad una sorta di concetto di densità condizionata (in statistica bayesiana è esattamente così).

Il metodo di massima verosimiglianza (ML, Maximum Likelihood) prescrive quanto segue: dato il campione $x_1, ..., x_n$, trovare la coppia (a, s) che massimizza $L(x_1, ..., x_n|a, s)$.

Se si trattasse di probabilità, sarebbe come dire: sotto quale scelta dei parametri, il campione è più probabile?

Siccome quasi tutte le densità sono legate alla funzione esponenziale e comunque a prodotti, algebricamente conviene calcolare il logaritmo: $\log L(x_1, ..., x_n|a, s)$. Massimizzare il logaritmo è equivalente.

Se la funzione è derivabile in (a, s), e l'eventuale massimo è all'interno del suo dominio di definizione, deve valere

$$\nabla_{(a,s)} \log L(x_1, ..., x_n|a, s) = 0.$$

Queste sono le *equazioni di ML*. In vari casi, esse sono esplicitamente risolubili. In altri casi, si deve ricorrere a metodi numerici di ottimizzazione.

Il software `R` formisce un comando in grado di calcolare le stime di ML dei parametri di molte classi, sia nei casi esplicitamente risolubili sia in quelli numerici. E' il comando `fitdistr`. Applichiamolo ai nostri casi:

```
require(MASS)
fitdistr(I.rec, "weibull")
fitdistr(I.rec, "normal")
```

Quest'ultimo calcola semplicemente:

```
mean(Medi)
sd(Medi)
```

4.5.4 Confronto grafico tra densità e istogrammi e Q-Q plot

Il primo confronto grafico da fare è tra istogramma e densità ottenuta col fit. Ad esempio

```
a<-...
s<-...
```

(i valori ottenuti con `fitdistr(I.rec, "weibull")`)

```
x<-...
```

(qui bisogna scegliere un intervallo opportuno, cosa che si fa osservando l'istogramma)

```
hist(I.rec,10,freq=FALSE)
y<-dweibull(x,a,s)
lines(x,y)
```

Il secondo confronto grafico è quello del qqplot. Per spiegare questo metodo bisogna premettere il concetto di quantile, di per sé fondamentale.

E' l'inverso della cdf. In tutti gli esempi da noi trattati, la cdf $F(x)$ è una funzione continua e strettamente crescente (salvo magari essere costantemente nulla per $x < 0$, e costantemente pari a 1 per $x > 1$). Allora, dato $\alpha \in (0,1)$, esiste uno ed un solo numero reale q_α tale che

$$F(q_\alpha) = \alpha.$$

Il numero q_α verrà detto *quantile* di ordine α. Se ad es. $\alpha = 5\%$, viene anche detto *quinto percentile* (se $\alpha = 25\%$, 25° percentile, e così via). Inoltre, 25° percentile, 50° percentile, 75° percentile vengono anche detti primo, secondo e terzo *quartile*.

La seconda premessa consiste nel dichiarare con chiarezza come si costruisce la cdf empirica $\widehat{F}(x)$: dato il campione $x_1, ..., x_n$, lo si ordina in modo crescente e, detto $x'_1, ..., x'_n$ il risultato, si pone

$$\widehat{F}\left(x'_i\right) = \frac{i}{n}.$$

A volte si preferisce prendere

$$\widehat{F}\left(x'_i\right) = \frac{i - 0.5}{n}$$

per trattare più simmetricamente il primo e l'ultimo valore.

Se un campione provenisse da una cdf $F(x)$, dovremmo avere che $\widehat{F}(x'_i)$ è circa uguale a $F(x'_i)$. Applicando ad ambo i termini la funzione inversa di F, che è la funzione quantile, otteniamo che

$$q_{\widehat{F}(x'_i)}$$

dovrebbe essere circa uguale a $x'_i = q_{F\left(x'_i\right)}$. Ma allora i punti

$$\left(x'_i, q_{\widehat{F}(x'_i)}\right)$$

staranno all'incirca sulla bisettrice del primo quadrante. Basta quindi rappresentare questi punti per avere un'idea della bontà della densità scelta. Il qqplot è il grafico di questi punti $\left(x'_i, q_{\widehat{F}(x'_i)}\right)$.

Per tracciare un qqplot, quindi, bisogna disporre di due stringhe: i dati x'_i riordinati ed i quantili $q_{\frac{i}{n}}$, o numeri molto vicini (tanto l'importante è l'effetto grafico, non il dettaglio numerico). Il software ordina da solo i dati, quando si usa il comando `qqplot`. Basta quindi usare il comando (ad es. per la Weibull):

```
qqplot(qweibull((1:48)/49,a,s),I.rec)
```

4.6 Esercizi sulle serie storiche

Nei seguenti esercizi si devono scaricare dati da rete, creare file, analizzare i dati ed eseguire previsioni co software R. Dopo le sottosezioni degli esercizi c'è un breve appendice con suggerimenti pratici sull'uso di R, soprattutto per quanto riguarda la gestione delle cartelle e file di ciascun esercizio ed il caricamento dati.

4.6.1 Esercizio n. 1 (veicoli 1; fasi iniziali)

- *Percorso da eseguire*: aprire Eurostat, Statistics Database, Statistics by theme, External trade, database, External trade detailed data, EU27 Trade since 1995 by BEC; select data, FLOW: export, INDICATORS: VALUE_IN_EUROS, PARTNERS: US, PERIOD: all, PRODUCT: 510 (TRANSPORT EQUIPMENT AND PARTS AND ACCESSORIES THEREOF / PASSENGER MOTOR CARS), REPORTER: DE, FR, GB, IT; update, viewtable, period: sort ascending (se necessario); download come file Excel, copiata la pagina su proprio file *export_ veicoli.xls* foglio 1.

- *Predisposizione cartella di lavoro*: dove risulta più comodo, creare una cartella *Esercizio1*, dove salvare *export_ veicoli.xls* e creare un documento word (o simile) con comandi, commenti, ecc.

- *Caricamento in R*:

  1) aprire R, pulire usando rimuovi tutti gli oggetti sotto varie.

  2) Copiare su R il comando `IT <- scan("clipboard",dec=',')` senza dare l'invio.

  3) Copiare dal file *export_ veicoli.xls* foglio 1, i dati italiani da gennaio 1995 a dicembre 2005.

  4) Tornare su R e dare invio. I dati sono caricati. Per vederli: `ts.plot(IT)`

  (Nota: se i dati su Excel sono scritti in modo particolare, R può non riconoscerli come numeri. Ad esempio, 11.849.488 non dovrebbe andargli bene, mentre 11849488 sì. Bisogna allora modificare lo stile su Excel, nella parte relativa a Numero)

  5)Salvare questo file nella cartella *Esercizio1*, con salva area di lavoro, usando ad es. il nome *Esercizio1.RData*.

  6) Chiudere il software, rispondendo no a salva area di lavoro?.

  7) Riaprire R tramite l'icona di R che ora è visibile nella cartella suddetta (file *Esercizio1.RData*).

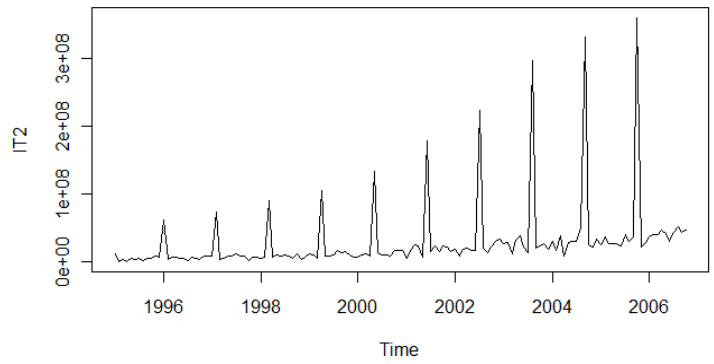
- La fase preliminare è terminata.

Eseguita questa fase preliminare, provare di nuovo:

```
ts.plot(IT)
```

Arricchire il grafico con le date:

```
IT2 <- ts(IT, frequency=12,start=c(1995,1)); ts.plot(IT2)
```



Nota: il comando `ts(IT, frequency=12,start=c(1995,1))` non ha solo attribuito le date ma ha eseguito una modifica strutturale alla stringa di numeri IT: le ha attribuito una "frequenza", a nostra scelta (siamo noi a decidere che, trattandosi di valri mensili, la frequenza naturale è 12). Questo nuovo attributo, la serie IT2 se lo porterà dietro sempre e sarà essenziale per il funzionamento di alcuni comandi R di analisi e previsione che possono lavorare solo su serie con frequenza prestabilita.

Fatte queste operazioni generali, ottenuto cioè finalmente il grafico, lo si osservi con attenzione, meditando su tutte le caratteristiche che saltano all'occhio. Evidente ad esempio è il trend crescente. E' anche evidente una periodicità annuale (ma su questa torneremo).

Qualcosa però non va bene in questo grafico: cosa?

4.6.2 Esercizio n. 2 (veicoli 2; decomposizione, stagionalità)

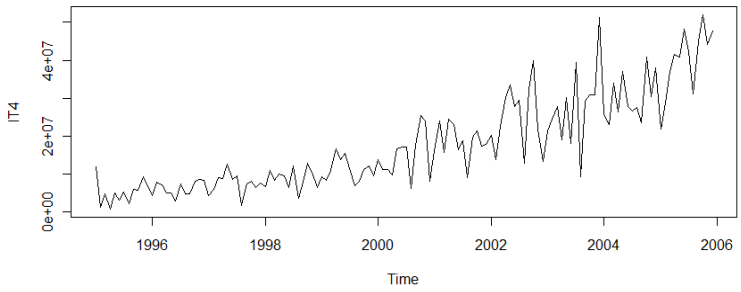
- Creare una cartella *Esercizio2*, copiarci *export_ veicoli.xls* e creare un file per comandi e commenti.
- Aprire il file *export_ veicoli.xls*, copiare l'intera tabella su foglio 2, depurare la tabella dei valori annuali cumulati (eliminare proprio le colonne).
- Aprire R dalla cartella *Esercizio1*, salvarlo col nome *Esercizio2*.RData nella cartella *Esercizio2* (magari uscire e riaprire).
- Copiare il comando `IT3 <- scan("clipboard",dec=',')` su R, senza dare l'invio.
- Copiare i dati italiani (gennaio 1995, dicembre 2005) di questa nuova tabella, tornare su R e dare invio. I dati sono caricati. Salvare di nuovo il file con lo stesso nome (va riscritto).

Per vedere i dati:

```
ts.plot(IT3)
```

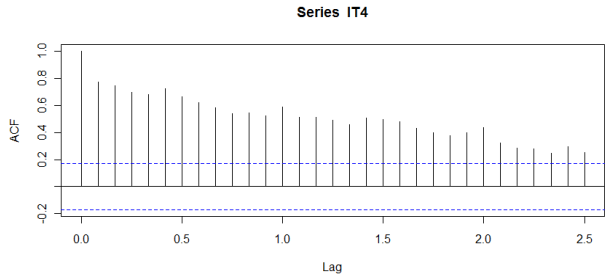
Per vederli con le date:

```
IT4 <- ts(IT3, frequency=12,start=c(1995,1)); ts.plot(IT4)
```



Iniziamo l'analisi della serie storica. Primi elementi: autocorrelazione

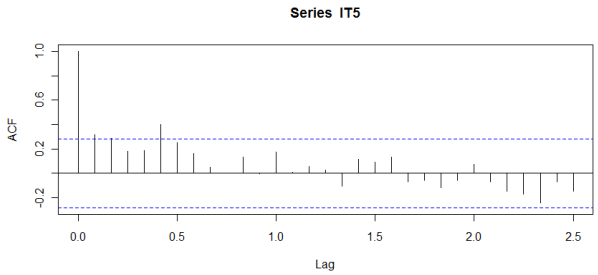
```
acf(IT4)
acf(IT4,30)
```



Non evidenzia marcate periodicità. Come mai?

Possiamo esaminare se ci sono periodicità locali. Prendiamo solo i dati degli ultimi 4 anni:

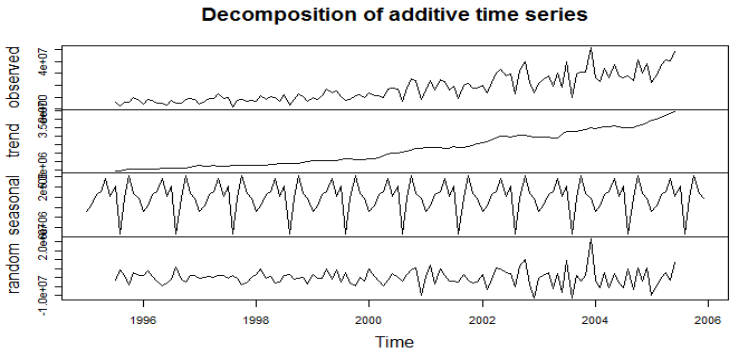
```
IT5<-window(IT4,2002)
acf(IT5,30)
```



Nemmeno così emerge nulla di rilevante.

Vediamo in un colpo solo trend, periodicità e residui:

```
plot(decompose(IT4))
```

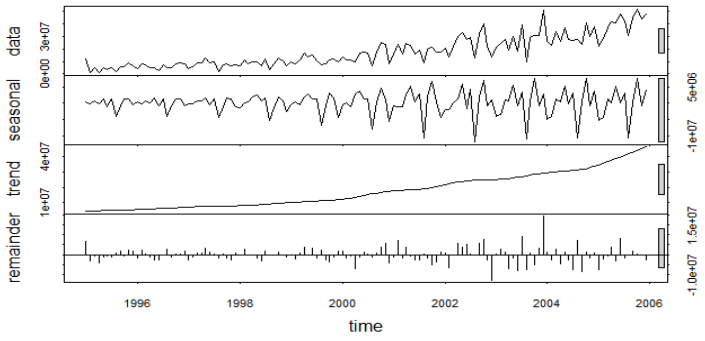


I valori delle componenti sono rilevanti per capirne l'importanza. Si possono vedere anche numericamente scrivendo

```
decompose(IT4)
```

Il comando `decompose` esegue una decomposizione globale; la periodicità è la stessa lungo tutta la serie. Invece, il seguente comando esegue una decomposizione locale:

```
plot(stl(IT4, 6))
```

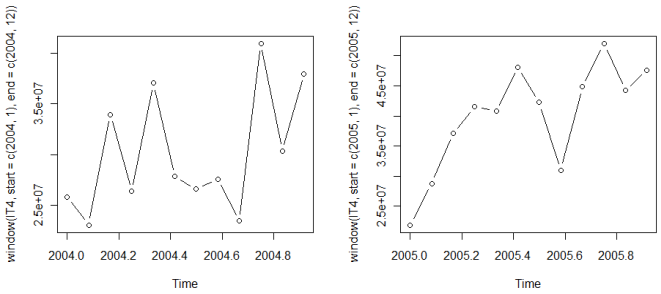


Cambiare il parametro 6 osservando le differenze.

Da queste prime indagini non si capisce bene se ci sia una struttura annuale. L'acf non la rileva. Questi due comandi di decomposizione producono comunque, di default, una componente periodica, ma si osservi quanto essa è piccola rispetto alle altre componenti. Il trend invece è rilevato in modo essenziale da questi due metodi. Si osservi anche che nulla è univoco: trend ecc. variano, se pur di poco, al variare del metodo e del parametro di `stl`.

La periodicità annuale di questa serie è davvero poco evidente. Per assicurarci di come stanno le cose, analizziao al microscopio, ad occhio, la struttura annuale:

```
par(mfrow=c(1,2))
ts.plot(window(IT4, start=c(2004,1), end=c(2004,12)),type=b)
ts.plot(window(IT4, start=c(2005,1), end=c(2005,12)),type=b)
```



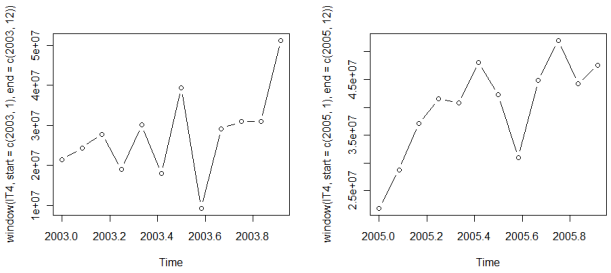
Intanto commentiamo i comandi. Il comando `par(mfrow=c(h,k))` crea finestre grafiche con h righe e k colonne di sotto-disegni. Attende che gli vengano dati $h \cdot k$ comandi di `plot`, che metterà nella griglia. Attenzione che il comando resta inserito, per così dire, anche nelle elaborazioni successive, per cui se nella finestra grafica successiva si vuole ad esempio un plot singolo, lo si deve richiedere esplicitamente con `par(mfrow=c(1,1))`.

Il comando `window(IT4,...)` ritaglia una finestra, quella relativa alle specifiche scritte nel comando. Si veda `?window` per ulteriori dettagli.

La specifica `type=b` (si veda `?plot` per tutte le alterntive possibili) disegna il grfico con *b*oth *l*ines and *p*oints (altrimenti con `type=l` si ottenevano solo *l*ines e con `type=p` si ottenevano solo *p*oints). Quando si vuole vedere un grfico al mocroscopio conviene usare entrambi perché danno entrambi informazioni: le linee danno la visione d'insieme, i punti permettono di vedere i singoli valori (si noti che a volte due linee successive di un grafico a linee sono allineate o quasi e quindi non si percepisce dove sta il punto si suddivisione, se non lo si disegna esplicitamente).

Fatti questi commenti sul software, torniamo all'analisi della serie. I due anni, 2004 e 2005, mostrano assai poco in comune. Hanno un generico trend crescente, ma questo è solo il sottoprodotto del trend generale crescente, non è una stagionalità. Forse gli ultimi 3 valori annuali hanno un pattern simile, ma è poca cosa (volendo la si potrebbe estrapolare). Si provino altri confronti, per capire di più ad esempio

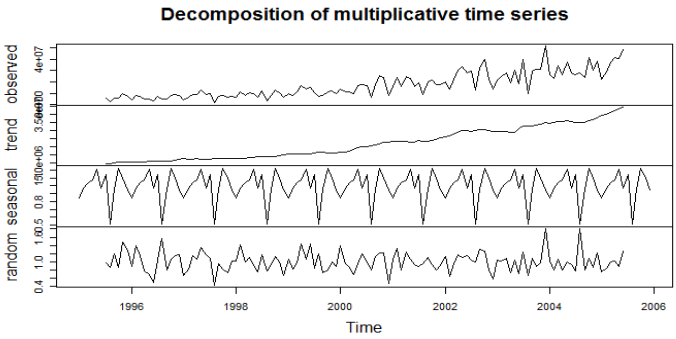
```
ts.plot(window(IT4, start=c(2003,1), end=c(2003,12)),type=b)
ts.plot(window(IT4, start=c(2005,1), end=c(2005,12)),type=b)
```



da cui emerge che agosto è il mese peggiore (così non è stato però nel 2004!), mentre ad esempio il pattern degli ultimi tre valori si è perso, quindi forse non vale la pena di evidenziarlo. Si facciano altre prove. E' essenziale conoscere la serie che si studia nei minimi dettagli.

Per scoprire empiricamente le opportunità offerte dal software R si provi anche decompose moltiplicativo:

```
plot(decompose(IT4, type = multiplicative))
```



Cercare nella guida le formule di `decompose`, ad esempio digitando `?decompose`, per capire la differenza teorica tra i due metodi.

Salvare il file nuovamente, altrimenti IT4 non resterà registrato per il futuro.

Per riassumere questo esercizio: abbiamocercato in tutti i modi di capire se ci fosse una stagionalità da sfruttare per le previsioni future. Forse c'è qualche traccia, ma non così accentuata.

4.6.3 Esercizio n. 3 (veicoli 3; previsione tramite decomposizione)

- Creare una cartella *Esercizio3* e creare un file per comandi e commenti.
- Aprire R dalla cartella *Esercizio2*, salvarlo col nome *Esercizio3.RData* nella cartella *Esercizio3* (magari uscire e riaprire).
- Sotto "varie", vedere "elenco degli oggetti": compaiono `IT`, `IT3`, `IT4` e fore altri, dipenda dai salvataggi eseguiti). Se non compaiono, sono stati sbagliati dei salvataggi negli esercizi precedenti.

In questo esercizio cerchiamo di fare la previsione dei dati del 2006, utilizzando solo gli strumenti visti fino ad ora. Decidiamo, a titolo di esercizio, di vole prevedere tutto il 2006, non solo il primo mese successivo ai valori noti.

Metodo 0: *ricopiare gli ultimi valori.* Nella sezione introduttiva abbiamo citato il metodo banale di previsione $p_{n+1} = x_n$, la copia dell'ultimo valore. Volendo una previsione di un anno, ricopiamo tutto l'anno precedente, il 2005. Sappiamo ritagliarlo col comando

```
anno2005 <- window(IT4, start=c(2005,1), end=c(2005,12))
```

oppure, essendo l'ultimo, con il comando abbreviato

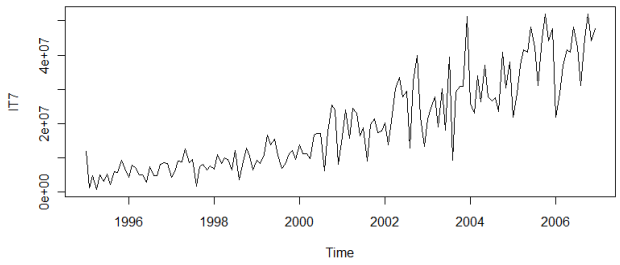
```
anno2005 <- window(IT4,2005)
```

Incolliamolo poi alla serie IT4:

```
IT6 <- c(IT4,anno2005)
ts.plot(IT6)
```

e poi arricchiamo con le date:

```
IT7 <- ts(IT6, frequency=12,start=c(1995,1)); plot(IT7)
```



Dal punto di vista grafico non è male. Cerchiamo ora di fare altre previsioni.

Metodo 1: *ad occhio*. Al posto del vettore `anno2005`, inserire un vettore

```
P<-c(...,...,...,...,...,...,...,...,...,...,...,...)
```

scelto da noi ad occhio come possibile previsione. Conviene fare questo esercizio per realizzare le nostre capacità previsive e le difficoltà che si incontrano (si rilegga l'introduzione al capitolo). Modificare i valori del vettore P in modo da ottenere un risultato che torni bene visivamente.

Metodo 2: *tramite la scomposizione ottenuta con* `stl(IT4, k)` (il parametro k è a scelta). Se indichiamo con X la serie storica, T il trend, S la componente stagionale, ε l'errore, legati da

$$X = T + S + \varepsilon$$

dobbiamo estrapolare in avanti di un anno T ed S, che chiamiamo T^* ed S^*, e definire la previsione

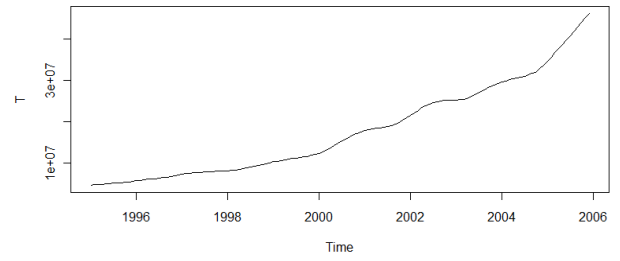
$$P = T^* + S^*.$$

Intanto, si esegua (prendiamo $k = 6$ per provare)

```
DEC <- stl(IT4, 6); plot(DEC)
```

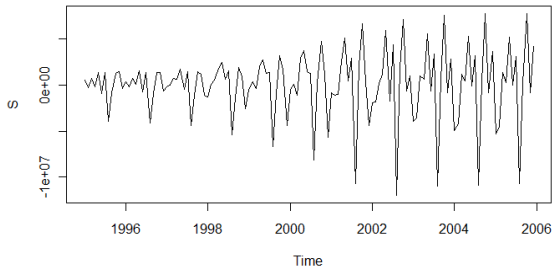
Poi si esegua

```
T <- DEC$ time.series[,2]; plot(T)
```



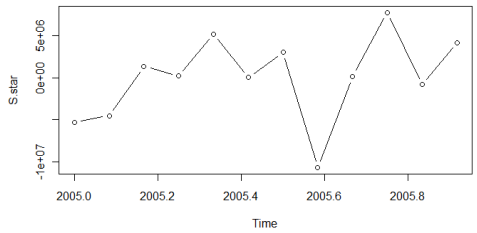
da cui riconosciamo che T è il trend; analogamente

```
S <- DEC$ time.series[,1]; plot(S)
```



è la componente stagionale. Dobbiamo ora estrapolare una previsione ragionevole T^* ed S^* e poi sommarle. Circa S^*, la cosa più semplice e naturale è ricopiare l'ultimo periodo:

```
S.star <- window(S,2005); ts.plot(S.star,type=b)
```



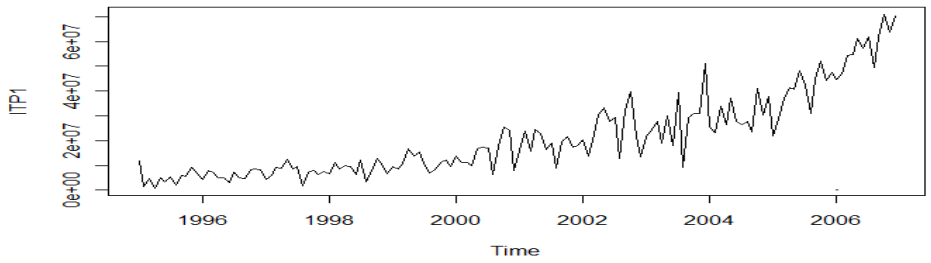
Circa T^*, in questo primo esercizio sull'argomento ci limitiamo ad usare una retta di regressione, forse lo strumento più semplice per estrapolare una curva. Il problema sta nella scelta della finestra su cui fare la regressione. Riguardando il grafico di T si vede che se consideriamo l'ultimissimo periodo, diciamo solo il 2005, la retta sarà più pendente, se usiamo invece 2004 e 2005 sarà un po' meno pendente e così via. Non c'è ovviamente una regola matematica, la matematica qui non c'entra niente. Dobbiamo decidere noi se riteniamo che il futuro, l'anno 2006, sarà più simile al 2005 o genericamente ad un passato più ampio. Forse si potrebbe propendere per l'uso del solo 2005 se si considera che la pendenza è andata sempre crescendo, di anno in anno, quindi usare per il 2006 una pendenza minore di quella del 2005 è un controsenso. Al tempo stesso, l'analista che conosca il mercato specifico di questo prodotto industriale, potrebbe essere al corrente di informazioni tendenziali che lo portino ad essere più prudente.

In assenza di idee migliori, o comunque per sviluppare un tentativo fino in fondo, estrapoliamo linearmente il solo 2005. Dobbiamo usare qui i comandi R relativi alla regressione, che verranno ripresi più avanti:

```
L <- length(IT4)
x<-(L-12):L; y <- IT4[(L-12):L]
reg <- lm(y ~x)
a <-reg$coefficients[2]; b<-reg$coefficients[1]
T.star <- a*((L+1):(L+12))+b
```

Vediamo intanto il risultato complessivo:

```
P <- S.star+ T.star
ITP <- c(IT4,P); ITP1 <- ts(ITP, frequency=12,start=c(1995,1)); plot(ITP1)
```



Forse ad occhio è un po' troppo ottimista. La scelta della regressione solo sul 2005 forse non è corretta. La ragione forse è che la stima del trend del 2005 fatta dal comando stl potrebbe essere poco attendibile, toppo locale per così dire, e quindi non andrebbe estrapolata.

Si notino le innumerevoli possibilità di scelta che abbiamo: vari metodi, e solo nell'ultimo possiamo cambiare k in stl, oppure la finestra su cui fare la regressione.

Esercizio: fate la vostra scelta. Ci si deve immergere nel senso di responsabilità che ha un operatore a cui è chiesto di fare una previsione, su cui si baseranno investimenti, politiche aziendali. Fate una previsione su cui potreste sommettere.

4.6.4 Esercizio n. 4 (veicoli 4; modelli AR)

- Creare una cartella *Esercizio4*, creare un file per comandi e commenti, aprire R dalla cartella *Esercizio3*, salvarlo col nome *Esercizio4.RData* nella cartella *Esercizio4* (magari uscire e riaprire). Se l'esercizio 3 è stato svolto correttamente, chiedendo "elenco degli oggetti" compaiono IT, IT3, IT4 (ed altri).

In questo esercizio iniziamo lo studio della serie IT4 tramite modelli ARIMA, o più precisamente modelli AR (auto regressivi). Si digiti

```
?ar
```

Dare una lettura molto sommaria al contenuto e forma di questo comando (alla lunga, bisognerà abituarsi a leggere l'help di R).

Usare il comando (aic è l'indice di Akaike):

```
ar.best <- ar(IT4, aic = T)
ar.best
```

ed osservare il risultato:

Call:

ar(x = IT4, aic = T)

Coefficients:

1	2	3	4	5	6	7	8	9	10	11	12
0.37	0.24	0.07	0.05	0.30	0.01	-0.14	-0.03	-0.09	-0.06	-0.10	0.31

Order selected 12 sigma^2 estimated as 4.715e+13

La varianza dei residui, ottenibile anche col comando `ar.best$var`, è pari a $4.715e+13$. Il comando ha scelto il modello col miglior AIC. Il modello è

$$(X_n - \mu) = a_1(X_{n-1} - \mu) + \ldots + a_{12}(X_{n-12} - \mu) + \varepsilon_n$$

dove $\mu =$`mean(IT4)` e la varianza dell'errore è appunto la varianza non spiegata.

Dobbiamo spiegare vari concetti. Cosa sia l'indice di Akaike e cosa sia la variaza spiegata, in vista del giudizio sul modello trovato che discuteremo sotto.

1) Il *criterio di Akaike* (Akaike Information Criterion, AIC) consiste nel calcolare

$$AIC = 2k + n \log(RSS)$$

dove k è il numero di parametri del modello ed n il numero di osservazioni, e scegliere il modello col minor AIC. A volte il software calcola delle grandezze legate in modo affine formula precedente (es. $2k + n \log(2\pi RSS/n) + n$ che differisce dalla precedente per una costante), che comunque assolvono lo stesso scopo di confrontare diversi modelli tra loro. La quantità AIC può anche essere negativa ($\log(RSS)$ può essere arbitrariamente negativo). Per questo, spesso il software calcola AIC rispetto ad un valore di riferimento (cioè aggiunge una costante) in modo da avere valori positivi. Se si stanno confrontando metodi, si può prendere l'AIC del modello migliore come punto di riferimento.

Se si minimizzasse solo RSS lasciando libero il numero di parametri, si troverebbe il p massimo possibile e si cadrebbe in overfitting (pessimo per la predizione). Se si minimizzasse solo k, si troverebbe sempre il modello banale $X_n = b$, b dato dalla media dei dati. Diventa come un problema di minimizzazione multiobiettivo. Si minimizza la somma per cercare una situazione intermedia. Si prende $\log(RSS)$ invece che RSS per riportare il valore di RSS ad un livello comparabile con k (questo commento è vago, dà solo l'idea del problema).

2) La *varianza spiegata* è un concetto di carattere generale, che ritroviamo nei più svariati contensi, e deve essere definito volta per volta a seconda del contesto. Per una serie storica $x_1, ..., x_n$, da un lato c'è la varianza (empirica) della serie, che indichiamo con S_X^2, dall'altro, una volta trovato un modello (es. $X_n = T_n + S_n + \varepsilon_n$), si può calcolare la varianza (empirica) dei residui $\varepsilon_1, ..., \varepsilon_n$, , che indichiamo con S_E^2. Concettualmente, la prima, S_X^2, rappresenta l'imprevedibilità dei dati originari, la seconda, S_E^2, l'imprevedibilità rispetto a ciò che può prevedere il modello, cioè l'imprevedibilità rimasta dopo aver scoperto il modello. Allora la grandezza $\frac{S_E^2}{S_X^2}$ rappresenta la percentuale di imprevedibilità rimasta dopo aver scoperto il modello, mentre la percentuale di imprevedibilità spiegata dal modello, detta *varianza spiegata*, è

$$\text{varianza spiegata} = 1 - \frac{S_E^2}{S_X^2}$$

Per convenzione, col termine varianza spiegata si intende la *percentuale* di varianza spiegata, che ha il pregio di essere un numero tra 0 ed 1, quindi a carattere universale, per così dire: possiamo cioè apprezzare la bontà di un metodo sia rispetto ad un altro (per questo bastava la varianza dei residui) sia rispetto ad una generica esperienza sviluppata su tanti esempi. Volendo, si potrebbe anche impostare un test statistico, su questo indicatore universale.

Dopo queste precisazioni, torniamo all'esercizio.

Avendo scoperto il modello, possiamo usarlo ad esempio per fare delle predizioni (più avanti).

La varianza della serie originaria è

```
> var(IT4)
```

[1] 1.546497e+14

Per cui la cosiddetta varianza spiegata, cioè $1 - \frac{var(residui)}{var(IT4)}$, è

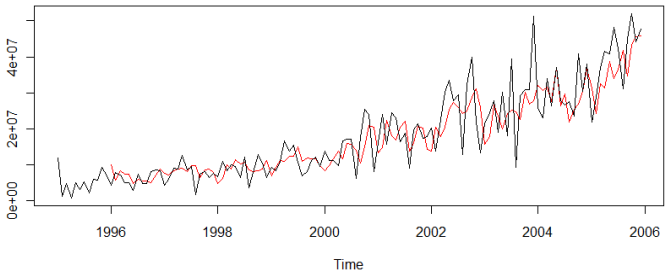
```
> varianza.spiegata <- 1- ar.best$var/var(IT4)
> varianza.spiegata
```

[1] 0.695

Abbiamo catturato circa il 70% della variabilità delle esportazioni. Resta un 20% di variabilità che per noi è del tutto casuale, imprevedibile, non compresa.

Possiamo tracciare in sovrapposizione le due serie (quella dei valori veri e quella prodotta dal modello):

```
ar.best.values <- IT4-ar.best$resid
ts.plot(ar.best.values,IT4,col=c(red,black))
```



Il risultato grafico mostra i limiti della soluzione trovata (forse non per colpa nostra ma a cusa dell'arbitrarietà delle fluttuazioni). La previsione (in rosso) tende ad essere una sorta di media locale dei valori, non cattura le fluttuazioni.

Svolgiamo un esercizio di carattere accademico:

```
x<-rnorm(10000)
xx <- ts(x, frequency=12,start=c(1995,1)); ts.plot(xx)
ar(xx, aic = T)
```

Call:

ar(x = xx, aic = T)

Order selected 0 sigma^2 estimated as 0.9973

Il comando `ar` funziona: ha riconosciuto di avere a che fare con un white noise, cioè un AR(0).

Esercizio per esperti. Con un po' di fatica di programmazione, possiamo calcolare la serie prodotta dal modello, non come differenza tra la serie vera ed i residui, ma usando il modello stesso; e poi tracciare in sovrapposizione le due serie. Mostriamo tale risultato e quello precedente in un unico disegno. Chi se la sente, svolga l'esercizio senza leggere le seguenti istruzioni.

```
ord<- ar.best$order
```

```
a <- ar.best$ar
P<-IT4
for (k in (ord+1):length(IT4)) {
P[k] <- sum(rev(a)*IT4[(k-ord):(k-1)])+mean(IT4)*(1-sum(a))
}
par(mfrow=c(2,1))
ts.plot(P,IT4,col=c(red,black))
ar.best.values <- IT4-ar.best$resid
ts.plot(ar.best.values,IT4,col=c(red,black))
```

4.6.5 Esercizio n. 5 (veicoli 5; proseguimento sugli AR)

- Creare una cartella *Esercizio5*, creare un file per comandi e commenti, aprire R dalla cartella *Esercizio4*, salvarlo col nome *Esercizio5.RData* nella cartella *Esercizio5* (magari uscire e riaprire). Se l'esercizio 5 è stato svolto correttamente, chiedendo "elenco degli oggetti" compaiono per lo meno `IT`, `IT3`, `IT4`.

Primo scopo di questa lezione è introdurre il comando `ar.ols`, spiegando cosa non va nell'uso di `ar` fatto precedentemente.

La serie che stiamo esaminando è visibilmente non stazionaria; bisognerebbe quindi analizzarla con metodi che accettano la non-stazionarietà. La teoria AR tradizionale è legata alla stazionarietà ed il metodo di fit dei coefficienti del comando `ar`, method = "yule-walker" (di default), è basato su formule teoriche che dipendono da questa ipotesi, le equazioni di Yule-Walker descritte nella sezione ARIMA. Se la serie fosse stazionaria, il metodo di Yule-Walker sarebbe plausibimente migliore. Ma il nostro esempio non è stazionario ed il metodo di Yule-Walker diventa privo di fondamento. Se però si usa un metodo di fit più banale, come i minimi quadrati dei residui, il modello AR può essere applicato anche a casi non stazionari. Questo è ciò ch fa il comando `ar.ols`.

Si può agire in due modi: o si usa `method = ols` (ordinary least squares) in `ar`, oppure si usa direttamente la funzione `ar.ols`.

```
?ar.ols
```

(da leggere sommariamente)

```
ar.ols.best <- ar.ols(IT4, aic = T)
ar.ols.best
```

Call:
ar.ols(x = IT4, aic = T)
Coefficients:
(riportiamo per brevità solo quelli con valore assoluto >0.1

5	6	9	10	11	12	15	17	18	19
0.29	0.11	-0.11	-0.11	-0.11	0.28	0.16	0.17	0.29	0.28

Intercept: 4842084 (834852)
Order selected 19 sigma^2 estimated as 2.589e+13

La varianza dei residui è solamente pari a $2.589e+13$ contro il $4:715e+13$ del metodo `ar`. Il comando ha scelto il modello col miglior AIC ed usato proprio la somma dei quadrati dei residui come criterio da minimizzare, per questo tale varianza è migliorata. Il modello è

$$(X_n - \mu) = a_1(X_{n-1} - \mu) + \ldots + a_{19}(X_{n-12} - \mu) + \varepsilon_n$$

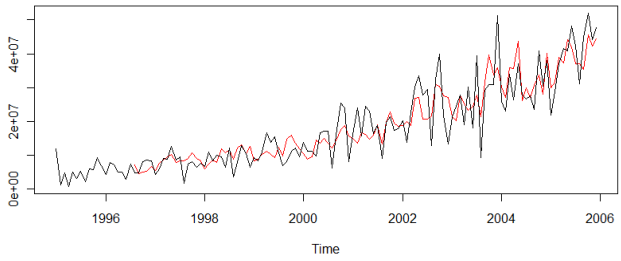
dove $\mu =$ `mean(IT4)`. La varianza spiegata è

```
varianza.spiegata <- 1- ar.ols.best$var/var(IT4)
varianza.spiegata
```

[1,] 0.83

Abbiamo catturato l'83% della variabilità delle esportazioni, molto meglio che con l'altro metodo. Vediamo come appaiono i fit:

```
ar.ols.best.values <- IT4-ar.ols.best$resid
ts.plot(ar.ols.best.values, IT4,col=c(red,black))
```



Il modello osa un po' di più, in termini di fluttuazioni, ma bisogna ammettere che a livello grafico faremmo fatica ad esprimere una preferenza (questo mostra l'utilità di certi indicatori numerici come la varianza spiegata).

Apprendiamo ora una interessante raffigurazione nuova legata ai modelli AR. Intanto calcoliamo il modello ottimale AR(19):

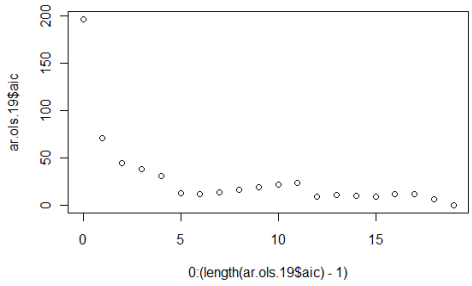
```
ar.ols.19<-ar.ols(IT4, order.max=19)
```

poi calcoliamo

```
ar.ols.19$aic
```

Esso fornisce la differenza tra l'AIC del modello con quel numero di parametri ed AIC del modello migliore, quello fittato. Siccome parte da $k = 0$, per raffigurarlo dobbiamo usare i numeri da 0:

```
plot(0:(length(ar.ols.19$aic)-1),ar.ols.19$aic)
```



Ad esempio, si vede che con 2, 3 e 4 parametri si otterrebbe un AIC che si discosta sensibilmente da quello ottimale mentre prendendo 5 parametri si ottiene un AIC decisamente più vicino a quello ottimale. Poi le cose peggiorano aumentando il numero dei parametri (perché, pur diminuendo la varianza dei residui - che diminuisce sempre per definizione di ottimo -, aumenta il numero di parametri e questo sbilancia in negativo l'AIC), fino a quando, inserendo il 12° parametro, AIC migliora decisamente, diventando quasi uguale a quello ottimale. Aumentando ulteriormente il numero di parametri si complica il modello per niente, fino a quando, con 18-19 parametri c'è di nuovo un miglioramento. Però potremmo anche accontentarci di un modello più parsimonioso.

Concludiamo che IT4 si fitta già ragionevolmente con order.max =5, altrimenti order.max =12.

Esercizio: eseguire questi comandi per l'esempio del white noise ed osservare la crescita sostanzialmente lineare di AIC (vedi definizione: il termine di RSS è circa costante visto che nessun modello migliora AR(0)).

Esercizio: eseguire il comando ar.ols con order.max =5 oppure order.max =12 e vedere i miglioramenti in termini di varianza spiegata, anche a confronto con quello ottimo. Quale sceglierete?

Soluzione.

```
ar.ols.5 <- ar.ols(IT4, order.max=5)
ar.ols.5
```

Call:
ar.ols(x = IT4, order.max = 5)
Coefficients:
1 2 3 4 5
0.1862 0.1701 0.1041 0.1652 0.3993
Intercept: 1119648 (536239)
Order selected 5 sigma^2 estimated as 3.535e+13

```
var.sp.5 <- 1- ar.ols.5$var/var(IT4)
var.sp.5
```

[1,] 0.7714451

```
ar.ols.12 <- ar.ols(IT4, order.max=12)
ar.ols.12
```

Call:

ar.ols(x = IT4, order.max = 12)

Coefficients:

1 2 3 4 5 6 7 8

0.1111 0.0989 0.0941 0.1154 0.3582 0.1207 -0.0415 -0.0006

9 10 11 12

-0.0744 -0.0395 -0.0457 0.3892

Intercept: 1799768 (601055)

Order selected 12 sigma^2 estimated as 3.082e+13

```
var.sp.12 <- 1- ar.ols.12$var/var(IT4)
var.sp.12
```

[1,] 0.8006854

Ovviamente non c'è una regola di scelta, ma visto che 12 è il numero di mesi dell'anno e quindi un modello di ordine 12 ha un'interpretazione più naturale e potrebbe essere più realistico per le predizioni rispetto ad un modello più artificioso che si è adattato eccessivamente ai dati particolari, visto inoltre che la sua varianza spiegata è già circa l'80%, lo potremmo preferire.

Riassumiamo come procedere con i modelli AR:

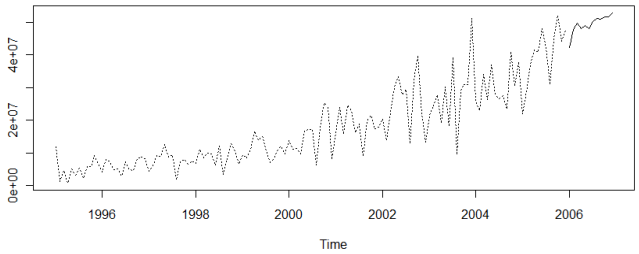
i) vedere l'ordine migliore con `ar.ols(IT4, aic = T)`

ii) vedere comunque come migliora AIC usando `plot(0:(length(ar(IT4, aic = T)$aic)-1),ar(IT4, aic = T)$aic)`

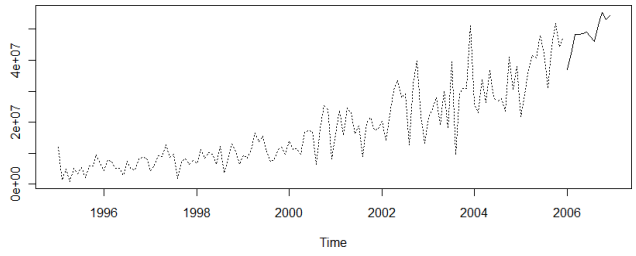
iii) eventualmente scegliere (sulla base del grafico AIC) modelli più parsimoniosi di quello ottimale ma già sufficientemente precisi.

Infine, effettuiamo la previsione con questi tre modelli: ar.ols.5, ar.ols.12, ar.ols.19. I comandi sono:

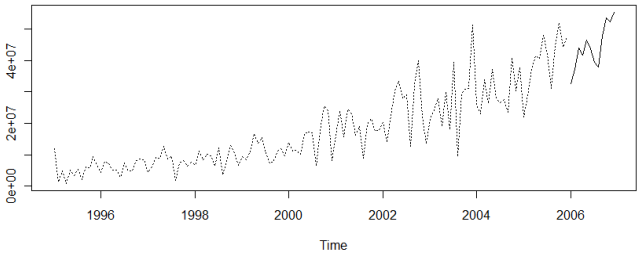
```
pred.5 <- predict(ar.ols.5, n.ahead=12) ; ts.plot(IT4, pred.5$pred, lty=c(3,1))
```




```
pred.12 <- predict(ar.ols.12, n.ahead=12) ; ts.plot(IT4, pred.12$pred, lty=c(3,1))
```




```
pred.19 <- predict(ar.ols.19, n.ahead=12) ; ts.plot(IT4, pred.19$pred, lty=c(3,1))
```



Naturalmente la scelta è difficile.

4.6.6 Esercizio n. 6 (veicoli 6; trend con SET; HW)

Creare una cartella *Esercizio7*, creare un file per comandi e commenti, aprire R dalla cartella *Esercizio6*, salvarlo col nome *Esercizio7.RData* nella cartella *Esercizio7*.

Riprendiamo l'esercizio 3 sulla previsione tramite decomposizione. Un punto debole di quell'esercizio era il metodo troppo sommario e casalingo di estrapolare il trend. Usiamo il metodo SET a questo scopo.

Tramite la scomposizione ottenuta con `stl(IT4, 6)` (k=6 tanto per fissare un valore intermedio) isoliamo trend e stagionalità:

```
DEC <- stl(IT4, 6); plot(DEC)
T <- DEC$ time.series[,2]; plot(T)
S <- DEC$ time.series[,1]; plot(S)
```

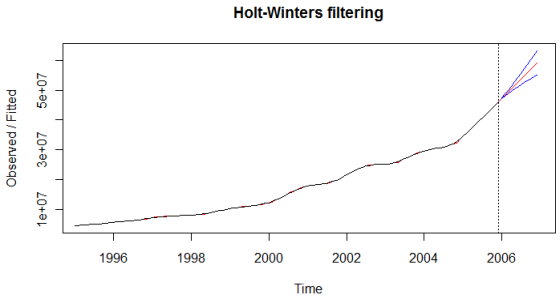
Poi calcoliamo

```
HW <- HoltWinters(T,gamma=FALSE)
HW
```

Holt-Winters exponential smoothing with trend and without seasonal component.
Call:
HoltWinters(x = T, gamma = FALSE)
Smoothing parameters:
alpha: 1
beta : 1

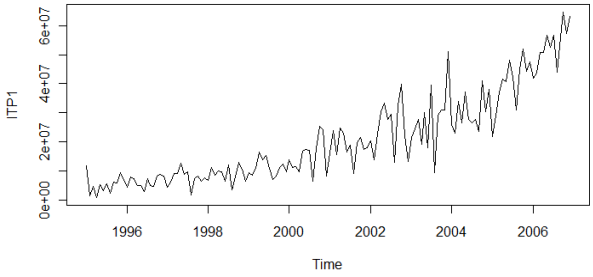
gamma: FALSE

```
p <- predict(HW, 12, prediction.interval = TRUE)
plot(HW, p)
```



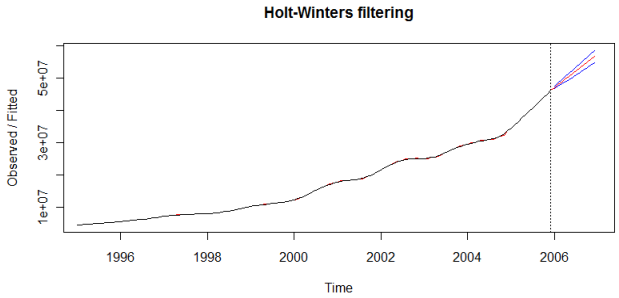
Il risultato in un certo senso non è molto diverso da una regressione lineare, ma è privo di accorgimenti ad hoc e complicazioni viste allora. La previsione complessiva si ottiene sommando la stagionalità:

```
S.star<-ts(window(S,2005), frequency=12,start=c(2006,1))
T.star <- p[,1]
P <- S.star+ T.star; ITP <- c(IT4,P); ts.plot(ITP)
```



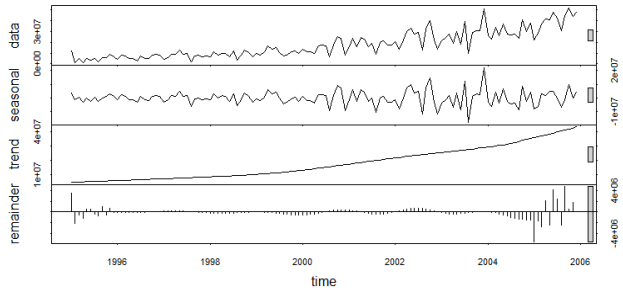
Soffre dello stesso "ottimismo" dell'estrapolazione del trend eseguito con la regressione sul 2005. Si può allentare questo ottimismo obbligando SET ad essere più conservativo. Infatti, il coefficiente beta con cui viene riaggiornata la pendenza qui è risultato uguale ad 1 (è quello che il software, con le sue minimizzazioni di errore, ha decisio fosse la scelta migliore). Un tale coefficiente in pratica ignora il passato. Obblighiamolo noi a ricordare il passato:

```
HW.beta.new <- HoltWinters(T,beta=0.1,gamma=FALSE)
HW.beta.new
p.beta.new <- predict(HW.beta.new, 12, prediction.interval = TRUE)
plot(HW, p.beta.new)
```

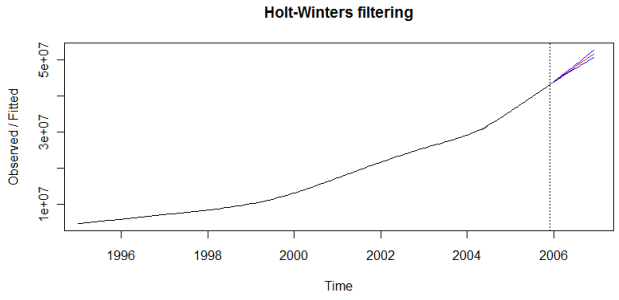


Il risultato però è solo poco diverso.
Vediamo uno scenario diverso, con $k = 3$:

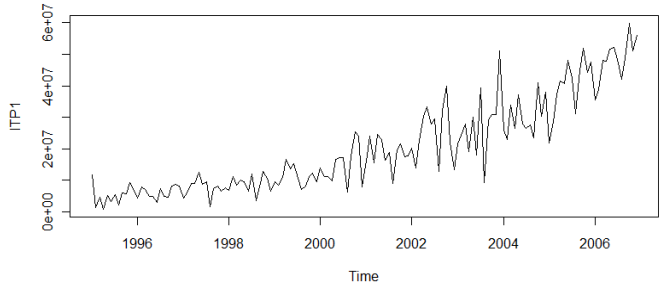
```
DEC <- stl(IT4, 3); plot(DEC)
```




```
T <- DEC$ time.series[,2]; S <- DEC$ time.series[,1]; HW <- HoltWinters(T,gamma=FALSE)
p <- predict(HW, 12, prediction.interval = TRUE); plot(HW, p)
```



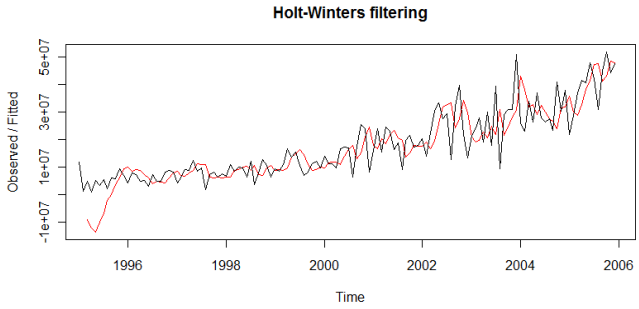

```
S.star<-ts(window(S,2005), frequency=12,start=c(2006,1))
T.star <- p[,1]
P <- S.star+ T.star; ITP <- c(IT4,P); ts.plot(ITP)
```



Forse è un po' più realistica della precedente.

Esaminiamo ora una variante. Cerchiamo il trend di IT4 direttamente col metodo SET, senza prima decomporre:

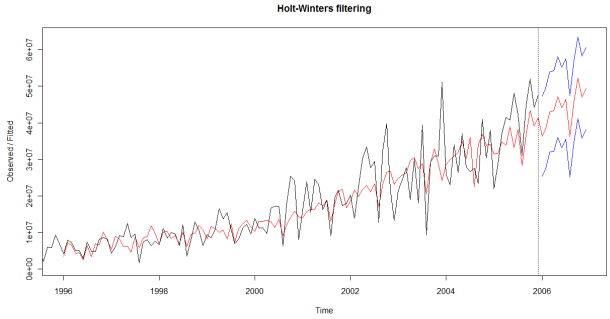
```
HW.glob <- HoltWinters(IT4,gamma=FALSE)
plot(HW.glob)
```



Non va bene, è un trend su scala temporale troppo breve.

Proviamo invece, a conclusione dell'esercizio, ad applicare il metodo completo di Holt-Winters (inclusa la periodicità) alla serie storica. Basta usare i comandi

```
HW.periodico <- HoltWinters(IT4)
HW.periodico
p.periodico <- predict(HW.periodico, 12, prediction.interval = TRUE)
plot(HW.periodico, p.periodico)
```



Non è affatto irrealistica e non differisce molto dai risultati dell'esercizio precedente.

Con questo esercizio si conclude il nostro studio della serie IT4. Naturalmente ci sarebbero altre cose da fare (es. il fid dei residui), altre varianti da provare (es. fare la media di varie previsioni), trarre le conclusioni su quale sia il metodo che ci sembra migliore, ed infine prendere i dati veri del 2006 e confrontare! Si può ad esempio calcolare la deviazione standard della differenza tra i dati veri e quelli previsti, per ciascun metodo, e vedere chi ha vinto.

4.6.7 Esercizio n. 7 (Motorcycles 1; decomposizione, AR)

- *Percorso*: Eurostat, Statistics Database, Database by themes; Industry, trade and services; Short-term business statistics (sts); Industry (NACE Rev.2) (sts_ind); Industry production index (NACE Rev.2) (sts_ind_prod); Industry production index - monthly data - (2005=100) (NACE Rev.2) (sts_inpr_m); select data: Italy; Nace R2: C3091 Manufacture of motorcycles; S_adj = GROSS; time= all (oppure dal 2000); update, view table, sort ascending. Download come file Excel, copiata la pagina su proprio file *motorcycles.xls* foglio 1.

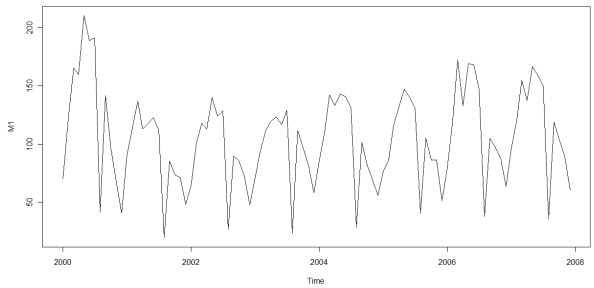
- Creare cartella *Esercizio7* con file word o simile, file *motorcycles.xls*. Caricare stringa dati da gennaio 2000 a dicembre 2007 (motorcycles relativi a Italia) in R con

  ```
  Mot <- scan("clipboard",dec=',')
  ```

  e trasformarla in M1 col comando

  ```
  M1 <- ts(Mot, frequency=12, start=c(2000,1), end=c(2007,12))
  ```

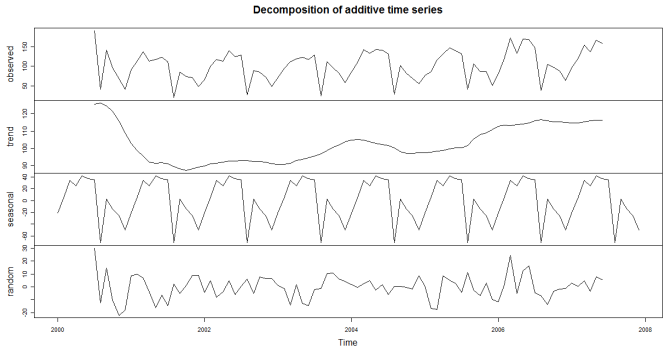
  Verificare con
  `plot(M1)`



- Salvare il file R col nome *Esercizio7.RData*, per il futuro.

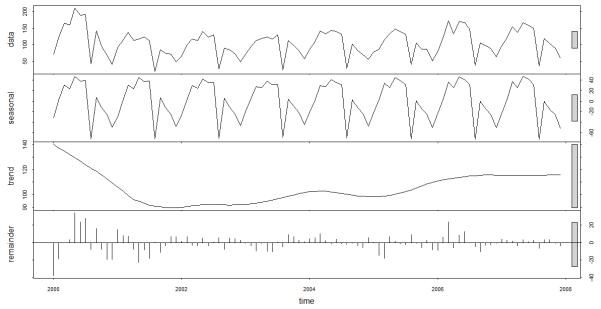
Analisi della serie. La periodicità ora è più evidente, ad occhio. Esaminiamola quantitativamente:

```
M1.dec <- decompose(M1); plot(M1.dec)
```

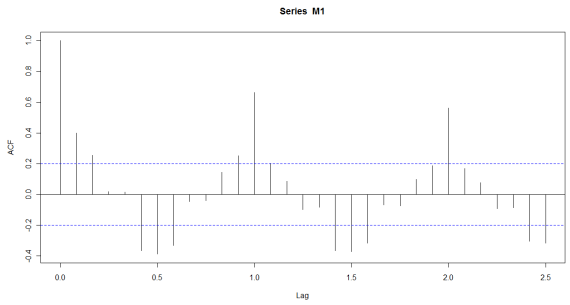


oppure:

```
M1.stl <- stl(M1, 6) ; plot(M1.stl)
```



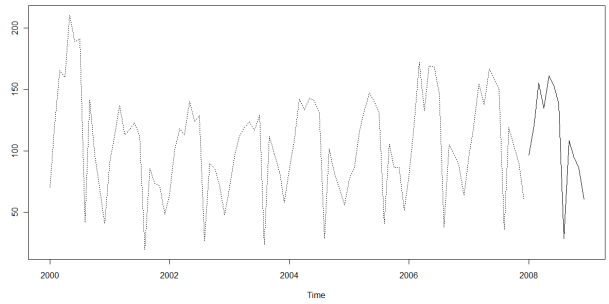

```
acf(M1,30)
```



conferma l'elevata periodicità annuale. Eseguiamo due analisi/predizioni, con metodi AR e con Holt-Winters. La più banale da eseguire è Holt-Winters:

```
HW<-HoltWinters(M1)
HW
pred<-predict(HW,12)
ts.plot(M1, pred, lty=c(3,1))
```

Smoothing parameters:
alpha: 0.4034712
beta : 0.0482223
gamma: 0.8496734



Esaminiamo gli AR:

```
ar.ols.best <- ar.ols(M1, aic = T); ar.ols.best
```

Coefficients:

1 2 3 4 5 6 7 8

0.3185 0.0216 0.1178 0.1493 -0.1315 0.0224 -0.0437 0.0281

9 10 11 12 13 14 15 16

-0.0486 0.0068 -0.0154 0.8868 -0.2547 -0.0119 -0.1464 -0.1458

17 18 19

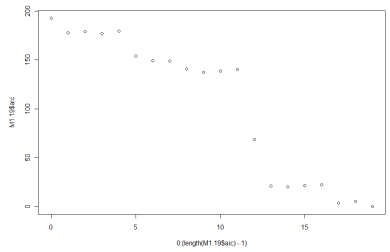
0.0898 -0.0519 0.0354

Intercept: 1.425 (1.566)

Order selected 19 sigma^2 estimated as 148.9

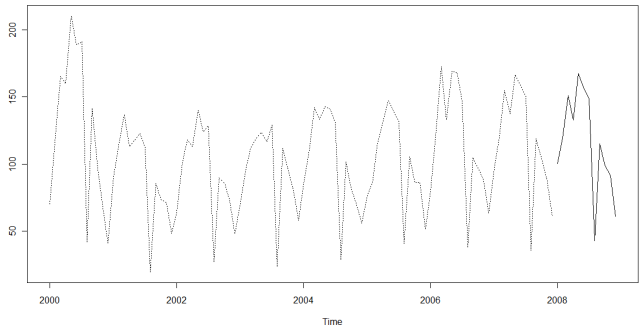
Probabilmente 19 è eccessivo. Esaminiamo l'indice AIC:

```
M1.19<-ar.ols(M1, order.max=19)
plot(0:(length(M1.19$aic)-1),M1.19$aic)
```



Chiaramente 13 contiene un miglioramento drastico, mentre I raffinamenti ulteriori possono essere trascurati a vantaggio dell'economicità del modello. Comunque per semplicità di software usiamo tutto:

```
pred <- predict(ar.ols.best, n.ahead=12) ; ts.plot(M1, pred$pred, lty=c(3,1))
```



Praticamente è lo stesso risultato di HW. Pur essendo due metodi completamente diversi, il risultato è stabile. E' un ottimo risultato.

Salvare il file.

4.6.8 Esercizio n. 8 (Motorcycles 2; HW, AR; confronti)

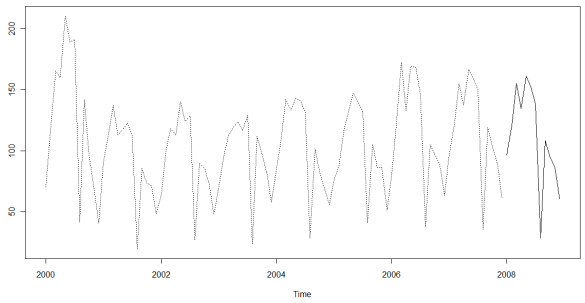
- *Preparazione cartella*: creare cartella *Esercizio8* con file word o simile, file *Esercizio8.RData* copiato dalla precedente. Controllare con list objects che ci sono Mot, M1.

Questo esercizio ha due scopi. Il primo è quello di riassumere tre routines R apprese negli esercizi passati, creando una sorta di schema di comandi che un analista può rapidamente usare in futuro su ogni serie (anche se sarebbe meglio soffermarsi su ciascuno con attenzione. Il secondo scopo è illustrare il metodo di cross-validation per un confronto.

Previsioni. Sunto di tre metodi (serve per il futuro come schema di comandi):

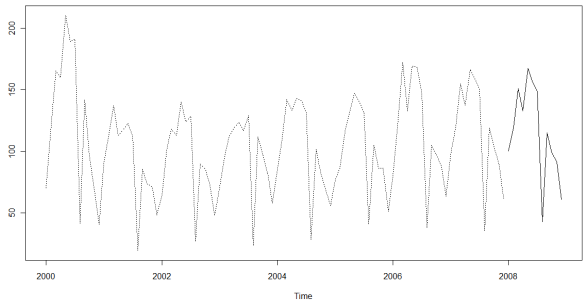
Holt-Winters:

```
HW<-HoltWinters(M1); pred<-predict(HW,12); ts.plot(M1, pred, lty=c(3,1))
```



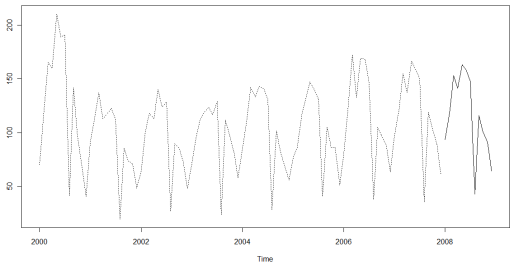
Modello AR (tipo ols) ottimale:

```
ar.ols.best <- ar.ols(M1, aic = T); pred <- predict(ar.ols.best, n.ahead=12)
; ts.plot(M1, pred$pred, lty=c(3,1))
```



Decomposione:

```
DEC <- stl(M1, 6); T <- DEC$ time.series[,2]; S <- DEC$ time.series[,1]
HW <- HoltWinters(T,gamma=FALSE); p <- predict(HW, 12); T.star <- p
S.star <- ts(window(S,2007), frequency=12, start=c(2008,1))
P <- S.star+ T.star; ts.plot(M1, P, lty=c(3,1))
```

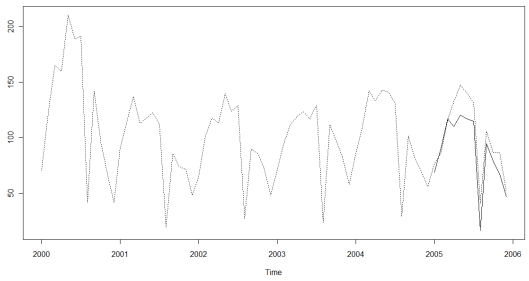


Sono tutte ragionevoli e molto simili. E' in effetti un esempio assai più schematico di quello degli esercizi 1-6.

Un motivo di questa concordanza è il fatto che i dati recenti di questa serie storica sono particolarmente regolari, in termini di trend e stagionalità. Come si sarebbero comportati i metodi ad esempio sulla base dei dati fino al 2004? Usiamo il procedimento di cross-validation, prendendo la finestra 2000-2004 come training set, e l'anno 2005 come test set.

Applichiamo i comandi alla serie ridotta, calcolando le previsioni, calcoliamo poi lo scarto quadratico medio (SQM) tra previsioni e dati veri dell'anno 2005. Vince il metodo con SQM minore. Se lo scopo di questo studio è capire quale modello tra i tre è mediamente più potente nella previsione di questa serie storica, il procedimento andrebbe ripetuto con varie finestre e magari andrebbero mediati i risultati, se non c'è univocità.

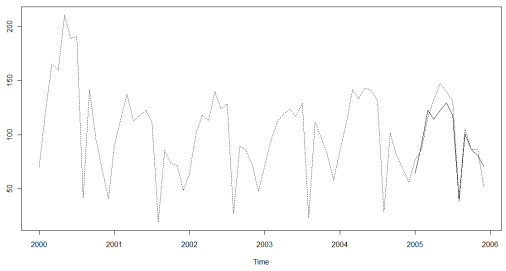
```
M.train <- window(M1, c(2000,1), c(2004,12)); M.test <- window(M1, c(2005,1),
c(2005,12))
HW.train<-HoltWinters(M.train); pred.train<-predict(HW.train,12)
ts.plot(window(M1, c(2000,1), c(2005,12)), pred.train, lty=c(3,1))
```




```
SQM.HW <- sd(M.test- pred.train); SQM.HW
```

10.45711

```
ar.ols.best.train <- ar.ols(M.train, ); pred.train <- predict(ar.ols.best.train,
n.ahead=12)
ts.plot(window(M1, c(2000,1), c(2005,12)), pred.train$pred, lty=c(3,1))
```

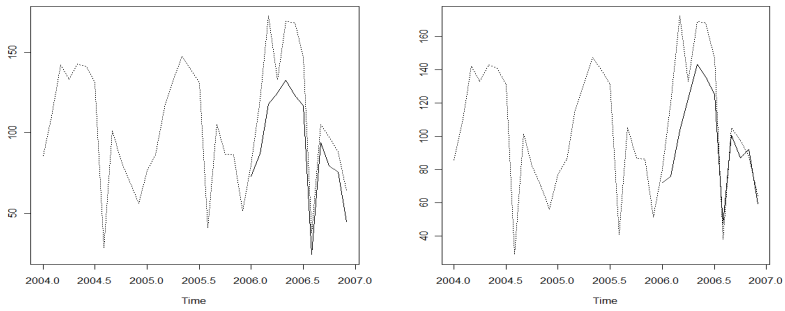



```
SQM.AR <- sd(M.test- pred.train$pred); SQM.AR
```

11.96519

Numericamente è meglio HW, anche se dal disegno non sembra. Vediamo ad esempio con la finestra 2000-2005:

```
M.train <- window(M1, c(2000,1), c(2005,12)); M.test <- window(M1, c(2006,1),
c(2006,12))
HW.train<-HoltWinters(M.train); pred.HW<-predict(HW.train,12)
ar.ols.best.train <- ar.ols(M.train, aic = T); pred.AR <- predict(ar.ols.best.train,
n.ahead=12)
par(mfrow=c(1,2))
ts.plot(window(M1, c(2004,1), c(2006,12)), pred.HW, lty=c(3,1))
ts.plot(window(M1, c(2004,1), c(2006,12)), pred.AR$pred, lty=c(3,1))
```

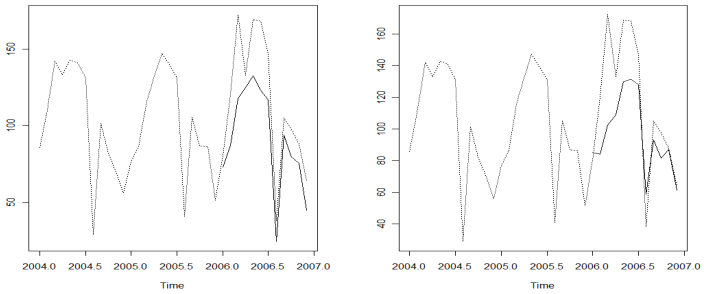



```
SQM.HW <- sd(M.test- pred.HW); SQM.AR <- sd(M.test- pred.AR$pred)
c(SQM.HW, SQM.AR)
```

15.54804 22.12240

Davvero sorprendente come HW sia meglio di AR. Può venire il dubbio che un modello AR più sintentico faccia meglio ma non è così:

```
M.train <- window(M1, c(2000,1), c(2005,12)); M.test <- window(M1, c(2006,1),
c(2006,12))
HW.train<-HoltWinters(M.train); pred.HW<-predict(HW.train,12)
ar.ols.12.train <- ar.ols(M.train, order=12); pred.AR <- predict(ar.ols.12.train,
n.ahead=12)
par(mfrow=c(1,2))
ts.plot(window(M1, c(2004,1), c(2006,12)), pred.HW, lty=c(3,1))
ts.plot(window(M1, c(2004,1), c(2006,12)), pred.AR$pred, lty=c(3,1))
```




```
SQM.HW <- sd(M.test- pred.HW); SQM.AR <- sd(M.test- pred.AR$pred)
c(SQM.HW, SQM.AR)
```

15.54804 24.21524

Su questo problema HW è davvero molto potente, come si può verificare con altre combinazioni. Va però detto che HW è un po' una scatola nera, dal punto di vista modellistico, mentre gli AR hanno una struttura più esplicita.

4.6.9 Esercizio n. 9 (Veicoli e Motorcycles, densità dei residui)

- *Preparazione cartella*: creare cartella *Esercizio9* con file word o simile, aprire nuovo file R

- Aprire poi il file *export_ veicoli.xls* dalla cartella *Esercizio2*, copiare il comando `IT3 <- scan(clipboard,dec=',')` su R, senza dare l'invio. Copiare i dati italiani (gennaio 1995, dicembre 2005), tornare su R e dare invio. Scrivere

  `IT4 <- ts(IT3, frequency=12,start=c(1995,1))`

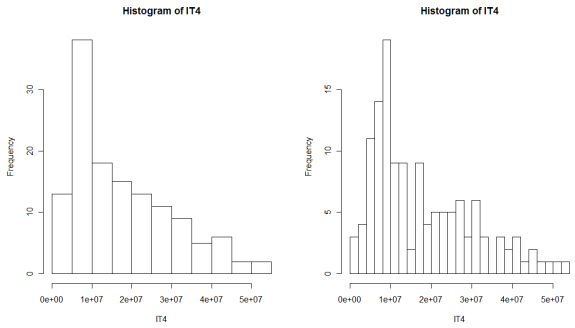
- Aprire poi il file *motorcycle.xls* dalla cartella *Esercizio7*, copiare il comando `Mot <- scan("clipboard",dec=',')` su R, senza dare l'invio. Copiare i dati italiani (gennaio 2000, dicembre 2007), tornare su R e dare invio. Scrivere

  `M1 <- ts(Mot, frequency=12, start=c(2000,1))`

- Salvare il file R come file *Esercizio9.RData*

Cominciamo l'esercizo ignorando la struttura della serie IT4, considerandola cioè come un campione speriementale delle esportazioni mensili di veicoli. Cerchiamone la distribuzione. Oltre al calcolo dei due indicatori medi più noti

mean(IT4); sd(IT4)
[1] 17723615
[1] 12435821
tracciamo un istogramma:
`par(mfrow=c(1,2)) ; hist(IT4); hist(IT4,20)`

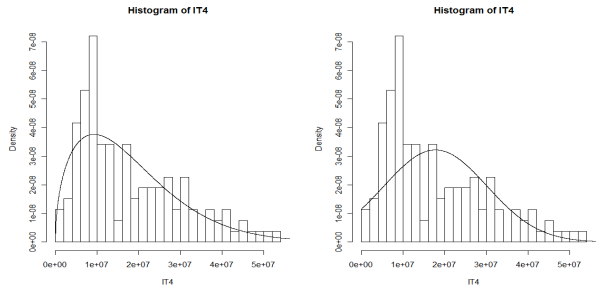


La forma è sicuramente più Weibull che gaussiana, ma per sicurezza accertiamocene col qqplot. Eseguiamo un fit Weibull ed uno gaussiano:
`require(MASS); fitdistr(IT4, weibull); fitdistr(IT4, normal)`
Loading required package: MASS

shape	scale
1.480863e+00	1.972954e+07
(9.522769e-02)	(2.965822e+03)
mean	sd
17723614.9	12388626.4
(1078291.5)	(762467.2)

Poi vediamo le due densità sovrapposte agli istogrammi:

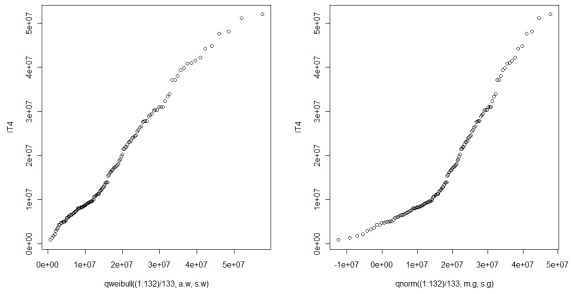
```
a.w<- 1.48; s.w<-19729540; x<-(0:8000)*10000; y.w<-dweibull(x,a.w,s.w); hist(IT4,20,fre
lines(x,y.w)
m.g<-17723614.9; s.g<-12388626.4; x<-(0:8000)*10000; y.g<-dnorm(x,m.g,s.g);
hist(IT4,20,freq=FALSE); lines(x,y.g)
```



Esaminiamo I qqplot:

`qqplot(qweibull((1:132)/133,a.w,s.w),IT4); qqplot(qnorm((1:132)/133,m.g,s.g),IT4)`

(132 è la numerosità dei dati)



Il fit Weibull è nettamente più preciso, anche se non perfetto.

Mostriamo l'utilità di questi calcoli. Calcoliamo il valore minimo delle esportazioni al 90%:

```
qweibull(0.1,a.w,s.w)
[1] 4312848
```

gen06	feb06	mar06	(*ecc.*)
4.312.848	4.312.848	4.312.848	

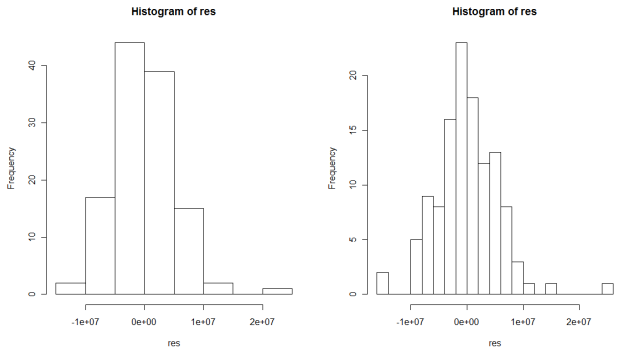
Possiamo affermare che mediamente le esportazioni saranno pari a circa 17.723.614, con valore minimo al 90% pari a circa 4.312.848. La deviazione standard di questa grandezza è molto elevata, 12435821, quindi sono stime molto imprecise.

Usiamo ora uno dei modelli, ad esempio AR(12). Dall'esercizio 5:

```
ar.ols.12 <- ar.ols(IT4, order.max=12)
pred.12 <- predict(ar.ols.12, n.ahead=12)
```

Questa è la predizione. Di quanto pensiamo possa essere sbagliata, sulla base dei residui del modello sui dati vecchi? `ar.ols.12$resid` restituisce i residui, che però iniziano da gennaio 1996. Applichiamo le analisi precedenti a questi residui:

```
res <- ar.ols.12$resid[13:132]
par(mfrow=c(1,2)) ; hist(res); hist(res,20)
```



Com'è naturale, i residui sono un po' più gaussiani. L'uso delle Weibull è reso complicato dal fatto che ci sono valori negativi; bisogna traslare tutto di `min(res)` più qualcosa:

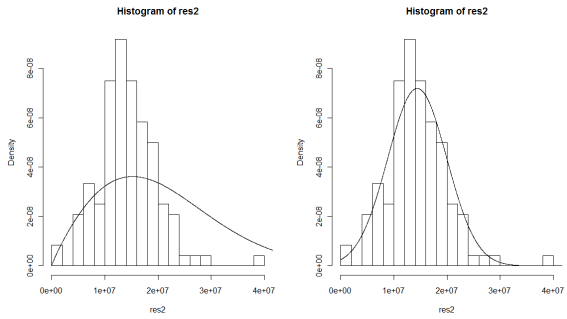
```
res2=res-min(res)+10
fitdistr(res2, weibull); fitdistr(res2, normal)
```

Loading required package: MASS

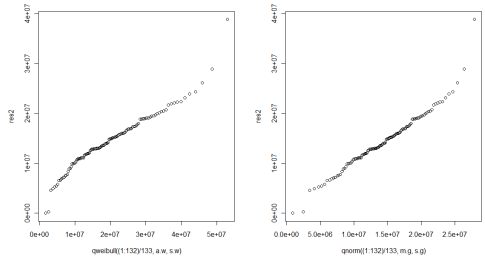
shape	scale
1.883589e+00	2.281359e+07
mean	sd
14397291.4	5551930.7

Poi vediamo le due densità sovrapposte agli istogrammi:

```
a.w<- 1.88; s.w<-22813590; x<-(0:8000)*10000; y.w<-dweibull(x,a.w,s.w); hist(res2,20,fr
lines(x,y.w)
m.g<-14397291.4; s.g<-5551930.7; x<-(0:8000)*10000; y.g<-dnorm(x,m.g,s.g);
hist(res2,20,freq=FALSE); lines(x,y.g)
```




```
qqplot(qweibull((1:132)/133,a.w,s.w),res2); qqplot(qnorm((1:132)/133,m.g,s.g),res2)
```



da cui abbandoniamo Weibull. Questi calcoli permettono ad esempio di calcolare il valore minimo delle esportazioni al 90%, per i mesi successivi ai dati noti:

```
qnorm(0.1,m.g,s.g)+min(res)-10
```

[1] -7115085

che va aggiunto alla predizione

```
pred.12
```

gen06	feb06	mar06
36833568	42080786	48266104

(*ecc.*)

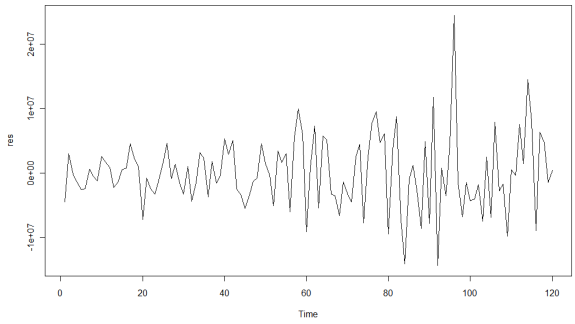
per cui il valore minimo delle esportazioni al 90% è

gen06	feb06	mar06
29.718.483	34.965.701	41.151.019

(*ecc.*)

E' molto più realistico. In verità andrebbe peggiorato un po' per la seguente ragione: abbiamo usato tutti i residui per la stima della gaussiana ma gli ultimi sono maggiori dei primi:

```
par(mfrow=c(1,1))
ts.plot(res)
```



per cui sarebbe più onesto usare solo i residui diciamoo dal 2000 in poi. Lasciamo il calcolo per esercizio.

Consideriamo ora la serie Mot.

Esercizio: Ripetere tutti i ragionamenti precedenti per la serie Mot, usando HW come modello. Nota: al posto del comando

```
res <- ar.ols.12$resid[13:132]
```

si deve usare

```
res<-HW$fitted[,1]-M1
```

4.7 Appendice

Riportiamo alcuni suggerimenti sull'uso del programma R. Ricordiamo che si può scaricare gratuitamente da rete e che, sempre in rete, si trova una grande quantità di materiale su di esso, a cui collaborano numerosissimi studiosi in tutto il mondo.

Gestione sul proprio PC delle cartelle di lavoro del corso.

1) Creare una cartella generale del corso, in una posizione facilmente raggiungibile (es. sul desktop).

2) Creare varie sotto-cartelle coi nomi degli studi principali che svolgiamo; ad esempio la cartella "Lezione 2" oppure unemployment.

3) Salvare in essa i vari file relativi, tipicamente: i) un file Excel di dati, ii) un file word (o simile) con comandi, commenti, risultati, figure, iii) una "area di lavoro R".

Si lavora così con vari file aperti: sicuramente una sessione di R ed un file word (o simile), eventualmente un file Excel di dati ed Eurostat.

Gestione aree di lavoro R

Distinguiamo la prima volta che si attiva una sessione di lavoro su un tema, dalle volte successive.

1) La prima volta si apre R, si pulisce usando rimuovi tutti gli oggetti sotto varie (infatti, salvo la primissima volta, tutte le successive l'apertura di R ricorderà l'ultima sessione, che riguarda un altro tema), si caricano i dati (nei modi illustrati a parte). Alla fine della sessione di lavoro si salva con salva area di lavoro, ad es. col nome Lezione2.RData. Chiudendo il software, conviene rispondere no a salva area di lavoro?.

2) Le volte successive basta aprire la cartella di interesse e cliccare sul file R di interesse (ha l'icona colorata). In esso restano salvati i dati immessi, coi nomi scelti precedentemente. Restano anche salvate alcune istruzioni date l'ultima volta. Se si vuole evitare che esse vengano salvate (i dati restano invece sempre salvati), fare "pulisci console" prima di chiudere (sempre rispondendo no a salva area di lavoro?). Con "elenco degli oggetti", sotto "varie", si ottiene l'elenco degli oggetti tipo data frame, vettori ecc. che sono salvati.

Caricamento veloce dei dati

Supponiamo di avere già una stringa di dati (soli numeri) su Excel, e supponiamo di volerla importare in R velocemente per analizzarla. Selezionarla e fare "copia".

Scrivere su R il comando (è come fare "incolla" su R, attribuendo il nome X):

```
X <- scan("clipboard")
```

e dare invio. X è un vettore numerico.

Se (caso più frequente) Excel usa la virgola per separare i decimali, essa va convertita in punto. Basta usare il comando:

```
X <- scan("clipboard",dec=',')
```

Capitolo 5

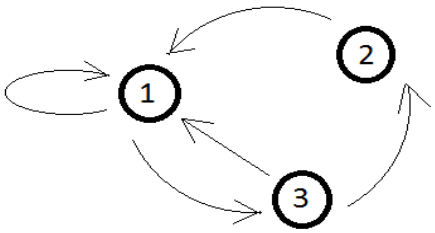
Sistemi Markoviani

5.1 Catene di Markov

5.1.1 Grafo, probabilità e matrice di transizione, probabilità di stato, proprietà di Markov

Definizione 47 *Una catena di Markov, nel senso insiemistico o algebrico del termine, è definita da un grafo orientato (con un insieme di stati al più numerabile) munito di probabilità di transizione.*

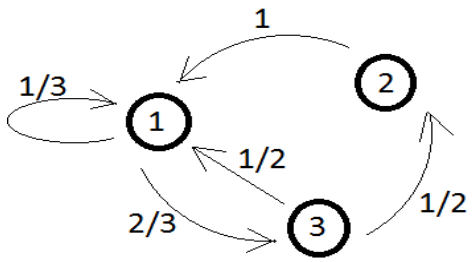
Spieghiamo i termini di questa definizione. Per *grafo orientato* intendiamo un insieme S (al più numerabile) di elementi, i vertici, detti *stati*, che usualmente disegneremo con cerchietti o anche solo punti, e da archi orientati che uniscono alcuni degli stati, che disegneremo con frecce. Sono ammesse anche le frecce che portano da uno stato in se stesso. L'unica regola da rispettare è che da ogni stato esca almeno una freccia. Queste ultime due convenzioni non fanno parte della definizione usuale di grafo orientato, quindi qui con tale termine intendiamo la struttura matematica appena descritta, anche se un po' peculiare rispetto al linguaggio tradizionale. Gli stati possono essere denominati come si vuole, a seconda dell'esempio, ma nell'esposizione teorica li numereremo tramite gli interi positivi (stato 1, stato 2 ecc.).



Le *probabilità di transizione* sono numeri $p_{ij} \in [0,1]$ associati a ciascuna coppa di stati $i, j \in S$, incluso il caso $j = i$. Devono soddisfare unicamente la regola

$$\sum_{j \in S} p_{ij} = 1.$$

Nell'interpretazione applicativa il numero p_{ij} va pensato come la probabilità di effettuare la transizione dallo stato i allo stato j. Quindi il numero p_{ij} va scritto, nel disegno del grafo, sulla freccia che porta dallo stato i a j. Quando manca una freccia, è il caso $p_{ij} = 0$.



Si possono riassumere questi elementi in una matrice

$$P = (p_{ij})_{i,j \in S}$$

quadrata, con tante righe (o colonne) quanti gli stati (anche infiniti). Viene detta *matrice di transizione*. Una catena di Markov (per così dire insiemistica o algebrica) è definita quindi o da una matrice di transizione o da un grafo orientato corredato di probabilità di transizione. Per la caten di Markov della figura precedente la matrice è

$$P = \begin{pmatrix} 1/3 & 0 & 1/3 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Arricchiamo questa visione, peraltro già esaustiva per molte applicazioni, con alcuni elementi più propriamente stocastici.

Definizione 48 *Dato un insieme S di stati, al più numerabile e data una matrice di transizione P relativa a tali stati, chiamiamo processo (o catena) di Markov ad essi associata un processo stocastico $(X_n)_{n \in \mathbb{N}}$ che assuma valori in S e tale che valga:*

$$\begin{aligned} p_{i_n, i_{n+1}} &= P\left(X_{n+1} = i_{n+1} | X_n = i_n\right) \\ &= P\left(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0\right) \end{aligned}$$

per ogni valore degli indici e degli stati..

A posteriori questa definizione sintetica risulterà chiara ma arriviamoci progressivamente attraverso una serie di ragionamenti e costruzioni.

Introduciamo, a partire dagli elementi sopra descritti, un processo stocastico $(X_n)_{n \in \mathbb{N}}$, che chiameremo anch'esso catena di Markov (ora in senso propriamente stocastico).

Operativamente, il processo stocastico è definito in questo modo. La v.a. X_t è una variabile discreta che assume come valore uno qualsiasi degli stati, con probabilità che indicheremo con

$$p_i^{(n)} := P\left(X_n = i\right), \qquad i \in S, n \in \mathbb{N}$$

(i è il generico stato, un vertice del grafo). X_n è lo "stato del sistema al tempo t" e $p_i^{(n)}$ è la probabilità che il sistema si trovi nello stato i al tempo t. I valori $p_i^{(n)}$ non vengono specificati a priori, non sono per così dire i mattoni elementari di costruzione. Invece, vengono specificate le probabilità condizionali

$$P\left(X_{n+1}=j|X_n=i\right)=p_{ij}, \qquad i,j\in S, n\in\mathbb{N}$$

interpretate appunto come probabilità di transizione. p_{ij} è la probabilità che il sistema effettui la transizione da i a j al tempo n; più precisamente, la probabilità che, trovandosi al tempo n in i, transisca in j al passo successivo, cioè si trovi in j al tempo $n+1$. Il sistema effettua una transizione da i a j, al tempo n, con probabilità p_{ij}. I numeri p_{ij} sono dati ed hanno la proprietà $\sum_j p_{ij}=1$. Si sta quindi imaginando che al tempo n il sistema occupi lo stato i, e debba effettuare una transizione, per portarsi in un qualche stato j al tempo $n+1$; le probabilità p_{ij} quantificano la probabilità di effettuare la transizione a questo o quello stato.

Implicito in questa regola è il fatto che tali probabilità non dipendano da t ma solo dagli stati i e j, fatto che si esprime dicendo che stiamo considerando catene di Markov temporalmente *omogenee*. Si potrebbe studiare anche il caso non omogeneo ma esso sfugge ad una semplice descrizione grafica e quindi risulta piuttosto astratto.

Di una catena di Markov, intesa ora come processo $(X_n)_{n\in\mathbb{N}}$, va inoltre specificata la distribuzione di probabilità al tempo zero, oppure lo stato di partenza:

$$p_i^{(0)}:=P\left(X_0=i\right), \qquad i\in S.$$

Se si specifica che la catena parte dallo stato i_0, significa che $p_{i_0}=1$, $p_i=0$ per ogni $i\neq i_0$. Il vettore

$$p^{(0)}=\left(p_i^{(0)}\right)_{i\in S}$$

non è noto a partire da P, è un'informazione indipendente.

A questo punto, le *probabilità di stato al tempo n*, cioè i numeri $p_i^{(n)}$, si possono calcolare univocamente ed esplicitamente a partire dalla distribuzione iniziale e la matrice di transizione:

$$p_i^{(n)}=\sum_{i_{n-1}}\sum_{i_{n-2}}\cdots\sum_{i_0}p_{i_0}^{(0)}p_{i_0i_1}\cdots p_{i_{n-2}i_{n-1}}p_{i_{n-1}i} \tag{5.1}$$

Infatti,

$$\begin{aligned}
P\left(X_n=i\right)&=\sum_{i_{n-1}}P\left(X_n=i|X_{n-1}=i_{n-1}\right)P\left(X_{n-1}=i_{n-1}\right)\\
&=\sum_{i_{n-1}}p_{i_{n-1}i}P\left(X_{n-1}=i_{n-1}\right)\\
&=\sum_{i_{n-1}}p_{i_{n-1}i}\sum_{i_{n-2}}P\left(X_{n-1}=i_{n-1}|X_{n-2}=i_{n-2}\right)P\left(X_{n-2}=i_{n-2}\right)\\
&=\sum_{i_{n-1}}\sum_{i_{n-2}}p_{i_{n-1}i}p_{i_{n-2}i_{n-1}}P\left(X_{n-2}=i_{n-2}\right)
\end{aligned}$$

e così via. Indicando con $p^{(n)}$ il vettore $\left(p_i^{(n)}\right)_{i\in S}$, vale in forma vettoriale

$$p^{(n)} = p^{(n-1)}P = ... = p^{(0)}P^n.$$

Le potenze della matrice di transizione, applicate a sinistra al vettore di stato iniziale, forniscono le probabilità di stato ad ogni istante successivo.

Potrebbe sembrare che il processo $(X_n)_{n\in\mathbb{N}}$ sia così univocamente determinato. Non è così. Serve imporre una regola altrimenti non è possibile calcolare univocamente probabilità del tipo $P(X_2 = k, X_1 = j, X_0 = i)$. Si assume che valga la seguente regola, detta *proprietà di Markov*:

$$\begin{aligned}&P(X_{n+1} = i_{n+1}|X_n = i_n)\\ &= P(X_{n+1} = i_{n+1}|X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0)\end{aligned}$$

per ogni valore degli indici e degli stati. Essa significa che la conoscenza dello stato "presente" i_n oppure la conoscenza del "presente" e di tutto il passato, producono le stesse previsioni sul futuro. Il futuro è idipendente dal passato, noto il presente. Se pensiamo al grafo ed al processo costruito su di esso, l'idea è che quando il processo occupa, al tempo n, lo stato i, per la determinazione futura del moto (cioè degli stati che verranno occupati) serve solo sapere che ci troviamo in i, non serve ricordare da quali stati siamo passati precedentemente. Se così non fosse, la descrizione algebrica tramite grafo crollerebbe, bisognerebbe introdurre delle complicazioni per tenere memoria del passato.

Le catene di Markov sono quindi sistemi senza memoria.

Assunta la proprietà di Markov, vale, per ogni valore di stati e tempi

$$\begin{aligned}&P(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0)\\ &= P(X_n = i_n|X_{n-1} = i_{n-1}, ..., X_0 = i_0)\, P(X_{n-1} = i_{n-1}, ..., X_0 = i_0)\\ &= P(X_n = i_n|X_{n-1} = i_{n-1})\, P(X_{n-1} = i_{n-1}, ..., X_0 = i_0)\\ &= p_{i_{n-1}i_n} P(X_{n-1} = i_{n-1}, ..., X_0 = i_0)\end{aligned}$$

da cui si può ripetere il conto ricorsivamente ed ottenere

$$P(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = p_{i_0}^{(0)} p_{i_0 i_1} \cdots p_{i_{n-1}i_n}.$$

Vediamo quindi che:

Proposizione 27 *Sotto la proprietà di Markov, la matrice di transizione P ed il vettore $p^{(0)}$ determinano tutte le marginali $P(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0)$, quindi determinano univocamente il processo stocastico $(X_n)_{n\in\mathbb{N}}$.*

Nota per gli esercizi. Per calcolare in un dato esempio la probabilità $p_i^{(n)}$ con i ed n (basso) specificati si può procedere in due modi. Il primo è puramente algebrico e consiste nel calcolo della potenza P^n. L'altro consiste nel capire graficamente la formula (5.1), identificando tutti i cammini che in n passi portano nello stato i, cammini che conviene elencare esplicitamente, calcolando poi le probabilità lungo ciascun cammino e poi la loro somma. Se

la matrice di transizione è ricca di zeri, cioè ci sono poche frecce nel grafo, e se esse sono strutturate bene rispetto ad i, forse si fa prima col secondo metodo che col primo. Se invece si deve calcolare tutto $p^{(n)}$, forse il primo metodo è più veloce. Vedremo vari esempi negli esercizi proposti (risolti).

Esempio 98 *Immaginiamo di voler costruire un automa che si comporti in modo simile ad un essere vivente "semplice", o reativamente ad una serie "semplice" di sue azioni. Supponiamo che l'essere vivente possa trovarsi in 4 situazioni possibili, ad esempio, inerte, vigile ma fermo, in fuga, in azione di attacco. Supponiamo poi che, dopo numerose osservazioni del comportamento di questo essere, si possa dire quanto segue: se inerte, così resta un tempo arbitrario, dopo di che diventa vigile ma fermo; se vigile ma fermo, può tornare inerte, oppure mettersi in fuga oppure all'attacco; e dopo molte osservazioni si vede che nel 50% dei casi torna inerte, nel 20% si mette in fuga, nel restante 30% in attacco; se in fuga, torna fermo e vigile al termine della fuga; similmente, se in attacco, torna fermo e vigile al termine dell'attacco. Possiamo allora descrivere questo sistema con 4 stati: 1 = inerte, 2 = fermo vigile, 3 = in fuga, 4 = in attacco. Le frecce presenti, con le relative probabilità, sono*

$$
\begin{aligned}
&1 \xrightarrow{1} 2\\
&2 \xrightarrow{1/2} 1, 2 \xrightarrow{1/5} 3, 2 \xrightarrow{3/10} 4\\
&3 \xrightarrow{1} 2\\
&4 \xrightarrow{1} 2.
\end{aligned}
$$

La matrice di transizione è

$$
P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/5 & 3/10 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.
$$

Ovviamente da qui a costruire un automa ce ne passa, ma è un inizio. Una catena di Markov è facile da simulare, quindi è facile generare moti causali una volta costruita la catena associata ad un problema. Nella descrizione ora data manca, rispetto alla realtà, un elemento fondamentale: il tempo trascorso in uno stato prima di effettuare una transizione. Anche questo tempo sarà aleatorio. Ci sono due modi di introdurlo. Un primo modo consiste nel mettere anche le frecce che mandano uno stato in se stesso, con opportune probabilità:

$$
P = \begin{pmatrix} p_1 & 1-p_1 & 0 & 0 \\ \frac{1}{2}(1-p_2) & p_2 & \frac{1}{5}(1-p_2) & \frac{3}{10}(1-p_2) \\ 0 & 1-p_3 & p_3 & 0 \\ 0 & 1-p_4 & 0 & p_4 \end{pmatrix}.
$$

Si sta immaginando che il tempo ora esista e la catena di Markov esegua un passo ogni intervallino Δt (di ampiezza da decidere). Ad esempio, dallo stato 1, con probabilità p_1 si sposta nello stato 2 nel tempo Δt, altrimenti resta dov'è. Se p_1 è piccolo, resterà per molto tempo nello stato 1, prima di effettuare la transizione. Così, modulando i valori p_i, si creano

temi aleatori di attesa negli stati, a piacere. Un altro modo di introdurre il tempo di attesa in uno stato è quello di usare i processi a salti della prossima sezione: si introducono degli orologi aleatori, associati a tutte le transizioni possibili, che "suonano" dopo un tempo aleatorio di caratteristiche statistiche date, provocando la relativa transizione nell'istante in cui suonano (poi vanno azzerati tutti e riavviati).

5.1.2 Misure invarianti

Definizione 49 *Relativamente ad una catena data, cioè ad una matrice di transizione P su un insieme di stati S, una misura invariante (detta anche distribuzione di probabilità invariante, o stazionaria) è un vettore*

$$\pi = (\pi_i)_{i \in S}$$

di numeri $\pi_i \in [0,1]$, con $\sum_{i \in S} \pi_i = 1$, tale che

$$\pi = \pi P.$$

Vale equivalentemente

$$\pi = \pi P^n \text{ per ogni } n \in \mathbb{N}.$$

Per questo si dicono invarianti (o stazionarie, o di equilibrio): se il sistema parte al tempo zero con la distribuzione di probabilità π, lo troviamo successivamente nei vari stati sempre con la stessa probabilità (infatti πP^n è la distribuzione di probabilità al tempo n). Non significa ovviamente che lo stato resti sempre lo stesso: il sistema transisce da uno stato all'altro ma occupa il generico stato i sempre con la stessa probabilità π_i.

L'interesse per queste distribuzioni non è però legato al fatto di partire da esse al tempo zero ma di convergere verso di loro quando il tempo tende all'infinito. Si sta pensando alla situazione comune a tanti sistemi reali in cui c'è un transitorio dopo di cui si passa, al limite, ad un regime stazionario. Bene: le distribuzioni stazionarie dovrebbero descrivere il regime stazionario, quello che si osserva dopo che è passato un po' di tempo. Ad esempio: all'apertura di una banca c'è un momento iniziale di sovraffollamento causato dalla gente che attedeva l'apertura all'esterno. Dopo un certo numero di minuti però quel traffico iniziale è stato smaltito e l'affollamento della banca diventa quello a regime, in cui entrano ed escono persone in modo casale ma statisticamente regolare. Se pensiamo che X_n descriva il numero di persone in banca al tempo n, passato il transitorio, questo numero non è costante, ma la probabilità che valga i lo è (all'incirca).

Sotto ipotesi opportune che non possiamo discutere in questo momento vale

$$\lim_{n \to \infty} p^{(n)} = \pi.$$

Qunado questo vale, si dice che c'è *convergenza all'equilibrio*. I teoremi che garantiscono questo fatto vengono detti teoremi ergodici.

Nota per gli esercizi. Le misure invarianti si possono calcolare algebricamente risolvendo l'equazione $\pi = \pi P$. Tuttavia, è utile e istruttivo arrivare ad un sistema di equazioni che permette il calcolo di π per via grafica, col metodo detto del *bilancio di flusso* (naturalmente

il sistema che si ottiene col bilancio di flusso è equivalente al sistema $\pi = \pi P$, ma può apparire lievemente diverso a prima vista). Si concentra l'attenzione su uno stato i, osservando le frecce entranti e quelle uscenti. Chiamiamo probabilità entrante il numero

$$\sum_{k \in S, k \neq i} \pi_k p_{ki}$$

ovvero la somma di tutte le probabilità entranti da ogni stato k, intendendo che passi da k a i la percentuale p_{ki} della massa π_k che si trova in k. Anaogamente chiamiamo probabilità uscente il numero

$$\sum_{j \in S, j \neq i} \pi_i p_{ij}.$$

Questi due numeri devono uguagliarsi:

$$\sum_{k \in S, k \neq i} \pi_k p_{ki} = \sum_{j \in S, j \neq i} \pi_i p_{ij}.$$

Questo è il bilancio di flusso nello stato i. Ripetendo questa operazione in tutti gli stati si trova un sistema tante equazioni quante sono le incognite π_i. Tuttavia questo sistema è ridondante: un'equazione qualsiasi può essere ottenuta combinando opportunamente le altre. In questo modo non si arriverebbe quindi a determinare univocamente π (a parte i problemi di non unicità di ci parleremo nel prossimo paragrafo). Si deve quindi aggiungere l'equazione

$$\sum_{i \in S} \pi_i = 1.$$

Esercizio 32 *Mostrare che le equazioni di bilancio di flusso sono equivalenti al sistema $\pi = \pi P$.*

Osservazione 72 *La particolarità di dover aggiungere l'equazione $\sum_{i \in S} \pi_i = 1$ è presente anche se si risolve il sistema $\pi = \pi P$ (infatti i due sistemi sono equivalenti). Anzi, è proprio dall'equazione $\pi = \pi P$ che si vede chiaramente il problema: $\pi^{(0)} = 0$ è sempre soluzione, per cui se ne troviamo un'altra non nulla $\pi^{(1)}$, già sono due e poi lo sono anche tutte le combinazioni lineari $\alpha\pi^{(0)} + \beta\pi^{(1)}$ con α e β reali qualsiasi. E' quindi ovvio che il solo sistema $\pi = \pi P$ non può identificare un'unica soluzione (a parte i problemi eventuali di non unicità che esamineremo nel paragrafo della classificazione degli stati).*

Citiamo due classi particolari che a volte permettono di semplificare il calcolo delle distribuzioni invarianti.

Definizione 50 *Si dice che π soddisfa l'equazione di equilibrio dettagliato, o che π è reversibile, se*

$$\pi_i p_{ij} = \pi_j p_{ji}$$

per ogni $i \neq j$.

Si dice che P è bistocastica se

$$\sum_i p_{ij} = 1$$

(cioè se P^T è stocastica).

Proposizione 28 *Se π soddisfa l'equazione di equilibrio dettagliato allora π è invariante.*

Se P è bistocastica, su uno spazio di stati S finito, di cardinalità n, allora la distribuzione uniforme

$$\pi_i = \frac{1}{n} \text{ per ogni } i \in S$$

è invariante.

Infine, se P è una matrice simmetrica, allora è bistocastica e la distribuzione uniforme soddisfa l'equazione di equilibrio dettagliato; quindi è invariante.

Proof. Se vale l'equazione di equilibrio dettagliato allora

$$(\pi P)_i = \sum_j \pi_j p_{ji} = \sum_j \pi_i p_{ij} = \pi_i$$

quindi π è invariante. Se P è bistocastica su uno spazio di stati S finito, detta π la distribuzione uniforme, vale

$$(\pi P)_i = \sum_j \pi_j p_{ji} = \sum_j \frac{1}{n} p_{ji} = \frac{1}{n} = \pi_i$$

quindi π è invariante. Infine, se P è una matrice simmetrica, allora P^T (essendo uguale a P) è stocastica, quindi P è bistocastica. Questo già assicura che la distribuzione uniforme sia invariante. Ma in più vale

$$\pi_i p_{ij} = \frac{1}{n} p_{ij} = \frac{1}{n} p_{ji} = \pi_j p_{ji}$$

dove l'uguaglianza in mezzo deriva dalla simmetria di P. Quindi la distribuzione uniforme soddisfa anche l'equazione di equilibrio dettagliato. ■

5.1.3 Classificazione degli stati

Limitiamo la discussione al caso di una catena finita, anche se molte cose valgono in generale.

Definizione 51 *Se esiste un percorso*

$$i \to i_1 \to i_2 \to \cdots \to i_n \to j$$

con probabilità di transizione non nulle (cioè $p_{i,i_1} p_{i_i i_2} \cdots p_{i_n j} > 0$) diciamo che i comunica con j. Scriviamo $i \rightsquigarrow j$.

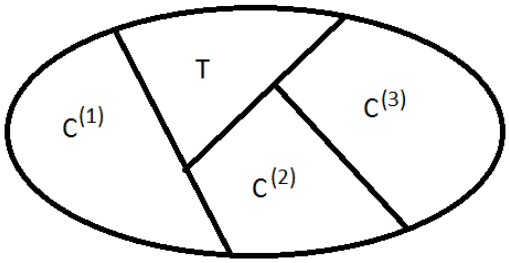
Definizione 52 *Uno stato i si dice transitorio se esiste uno stato j tale che $i \rightsquigarrow j$ ma non vale $j \rightsquigarrow i$. Gli altri stati si dicono ricorrenti (uno stato i è ricorrente se, quando vale $i \rightsquigarrow j$, allora vale anche $j \rightsquigarrow i$). Uno stato si dice assorbente se comunica solo con se stesso (altri possono portare a lui, ma non lui ad altri).*

Definizione 53 *Una classe (famiglia) di stati $S_0 \subset S$ si dice chiusa se non comunica con gli stati ad essa esterni ($i \in S_0$, $i \rightsquigarrow j$ implica $j \in S_0$) (dall'esterno si può entrare in S_0 ma da S_0 non si può uscire).*

Definizione 54 *Una classe chiusa S_0 si dice irriducibile (o chiusa irriducibile) se non esiste $S_0' \subset S_0$, S_0' strettamente più piccola di S_0, S_0' chiusa.*

Ogni classe chiusa, ed in particolare ogni classe irriducibile, può essere considerata come una catena a se stante. Parleremo della catena *ristretta* alla classe chiusa.

Una catena di Markov finita si decompon quindi in un certo numero di classi irriducibili $C^{(1)}, ..., C^{(k)}$ più un insieme T di stati transitori.



Teorema 32 *i) Ogni catena di Markov finita ha almeno una misura invariante.*

ii) Ogni misura invariante è nulla sugli stati transitori.

iii) Se la catena è irriducibile, ha una sola misura invariante. In particolare, ogni classe irriduibile di una catena di Markov, se considerata come catena a se stante (cioè se consideriamo la catena ristretta a tale classe), ha un'unica misura invariante.

iv) Se $C^{(1)}$, ... , $C^{(k)}$ sono le classi irriduibili di S e $\pi^{(1)}$, ... , $\pi^{(k)}$ sono le relative misure invarianti (uniche), le misure della forma

$$\alpha_1 \pi^{(1)} + \cdots + \alpha_k \pi^{(k)}$$

sono tutte e sole le misure invarianti di S.

Non dimostriamo questo importante teorema, ma ne spieghiamo alcuni simboli. Supponiamo che $C^{(1)}$ sia una classe irriducibile strettamente più piccola di S. Possiamo considerare la catena ristretta a $C^{(1)}$. It teorema dice che essa ha un'unica misura invariante $\pi^{(1)}$. Tuttavia, si deve osservare che $\pi^{(1)}$ è un vettore con un numero di componenti pari alla cardinalità di $C^{(1)}$. Aggiungendo zeri a tutte le altre componenti di S (quelle del complementare di $C^{(1)}$, fatto di altre classi irriducibili e dell'insieme degli stati transitori).

5.1.4 Convergenza all'equilibrio e proprietà ergodiche

Senza dimostrazioni, diamo però alcuni risultati riguardandi il concetto di convergenza all'equilibrio descritto nel paragrafo 5.1.2 ed aggiungiamo alcuni elementi di teoria ergodica per le catene di Markov.

Definizione 55 *Una matrice di transizione P si dice regolare se esiste $m > 0$ tale che*

$$(P^m)_{ij} > 0 \text{ per ogni } i, j \in S.$$

Teorema 33 *Se P è regolare su uno spazio degli stati S finito, allora esiste una ed una sola misura invariante π e per essa vale la convergenza all'equilibrio*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$$

per ogni $i, j \in S$.

Proposizione 29 *Se P è irriducibile, su uno spazio degli stati S finito, ed esiste almeno uno stato i tale che*

$$p_{ii} > 0$$

allora è regolare.

Teorema 34 *Se P è irriducibile, su uno spazio degli stati S finito, $(X_n)_{n\geq 1}$ indica un processo di Markov associato a P (con qualsiasi distribuzione iniziale al tempo zero) e π è la distribuzione invariante di P (sappiamo che è unica), allora per ogni funzione $f : S \to \mathbb{R}$ vale la convergenza in probabilità (anche quasi certa)*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \pi(f)$$

dove $\pi(f)$ è la media di f rispetto a π:

$$\pi(f) = \sum_{i\in S} f(i)\,\pi_i.$$

Si noti che $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ è una variabile aleatoria.

Se $(X_n)_{n\geq 1}$ fosse un processo stazionario e π fosse la legge di X_n (cosa vera se X_1 avesse legge π, visto che π è invariante), allora sarebbe

$$\pi(f) = E[f(X_1)]$$

cioè il terema ergodico precedente avrebbe esattamente la forma di quello visto nel capitolo sui processi stazionari.

Qui la situazione è meno generale che in quel capitolo, perché stiamo esaminando catene di Markov finite. Ma per certi versi è più generale, sia perché possiamo predere f qualsiasi (mentre nella versione base del teorema ergodico facevamo solo le medie $\frac{1}{n}\sum_{i=1}^{n} X_i$), sia perché il processo $(X_n)_{n\geq 1}$ non è necessariamente stazionario. Inoltre, si noti che non assumiamo alcuna scorrelazione asintotica. Il fatto di non aver bisogno dell'ipotesi di stazionarietà è una proprietà simile a quella del teorema 33: anche se non si parte con distribuzione π, asintoticamente è come se si avesse distribuzione π. Ciò dipende dall'ipotesi di irriducibilità. Se si vuole però la vera e propria convergenza all'equilibrio l'irriducibilità non basta (basta cioè per le medie temporali ma non per i singoli valori temporali, come si può capire facendo esempi periodici). Per quanto riguarda poi la scorrelazione asintotica, di nuovo questa è prodotta automaticamente dall'ipotesi di irriducibilità.

5.2 Esercizi

Esercizio 33 *Consideriamo la catena di Markov su* $E = \{1, 2, 3, 4, 5\}$ *associata alla seguente matrice di transizione*

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{pmatrix}.$$

i) Qual è la probabilità, partendo da 5, di essere in 4 dopo 4 passi?
ii) Classificare gli stati e trovare le classi irriducibili.
iii) Determinare tutte le probabilità invarianti della catena.

Esercizio 34 *Consideriamo due agenti finanziari. Ogni ora, ciascun agente compie un'azione, scelta tra due possibilità* A *e* B*. Ci sono pertanto quattro possibilità di azioni scelte dai due agenti:* (A, A)*,* (A, B)*,* (B, A)*,* (B, B) *(ad esempio,* (A, B) *significa che il primo agente sceglie* A *ed il secondo* B*).*
Quando si realizza (A, A)*, il primo agente guadagna 10 ed il secondo 0. Quando si realizza* (A, B)*, il primo guadagna 0 ed il secondo 10. Quando si realizza* (B, A)*, il primo guadagna 0 ed il secondo 10. Quando si realizza* (B, B)*, il primo guadagna 10 ed il secondo 0.*
I due agenti scelgono in modo indipendente l'uno dall'altro, scegliendo però in base al proprio guadagno dell'ora precedente: se hanno guadagnato 10 conservano la scelta precedente, altrimenti la modificano con probabilità 1/2.
i) Descrivere il problema con una catena a 4 stati, stabilire tutte le proprietà di tale catena e calcolare il guadagno medio di ciascun agente all'equilibrio.
ii) Rispondere alle stesse domande nella seguente variante del caso precedente: i due agenti scelgono in modo indipendente l'uno dall'altro; il primo, se guadagna 10 cambia, mentre se guadagna 0 cambia con probabilità 1/2; il secondo, se guadagna 10 conferma la scelta precedente, mentre se guadagna 0 cambia (si consiglia di rileggere più volte questo testo).
iii) Calcolare, nel caso (ii), la probabilità di trovarsi in (A, A) *partendo da* (B, B) *in* n *passi (*n *qualsiasi), traendo delle conclusioni anche in relazione a fatti scoperti al punto (ii). Calcolare poi la probabilità di trovarsi in* (A, A) *partendo da* (A, B)*, in* 8 *ed in* 9 *passi. Cercare infine di capire se vale la convergenza all'equilibrio partendo da* (A, B)*.*

5.3 Processi di Markov a salti

5.3.1 Sistemi a eventi discreti

Portano questo nome tutti quei sistemi in cui possiamo classificare le situazioni possibili secondo un numero finito o al più numerabile di possibilità, dette *stati* del sistema; il sistema si trova ad ogni istante in un certo stato, lì resta per un certo tempo, poi effettua una *transizione* ad un altro stato, e così via.

Dal punto di vista matematico gli ingredienti sono gli *stati* e le *transizioni* (a cui aggiungeremo altri enti come i tempi aleatori di attesa per una transizione). Di fronte ad ogni nuovo

problema bisogna saper elencare gli stati e le transizioni possibili. Le catene di Markov della sezione precedente sono un esempio, ma ora andremo oltre.

Esempio 99 *Osserviamo una coda comune, come quella alla cassa in una banca. Supponiamo qui solo per semplicità che ci sia un solo sportello aperto e quindi tutti gli utenti si mettano in coda a quell'unica cassa. Se osserviamo questo sistema ad un generico istante possiamo individuare le seguenti situazioni (stati) possibili:*

1. *la coda è vuota e nessuno è alla cassa a farsi servire; indichiamo con 0 questo stato (0 persone nel sistema)*
2. *la coda è vuota e c'è una persona alla cassa che viene servita; indichiamo con 1 questo stato (1 persona nel sistema)*
3. *ci sono k-1 persone in coda, più quella in fase di servizio; indichiamo con k questo stato (k persone nel sistema).*

Si rifletta riconoscendo che ogni altra descrizione è equivalente. Gli stati sono quindi i numeri interi non negativi: lo stato n sta a indicare che ci sono n persone nel sistema (n-1 in coda, una che viene servita). Poi bisogna individuare le transizioni possibili. Se ammettiamo che possa entrare una persona alla volta nella banca, può avvenire la transizione

$$k \to k+1$$

(quando arriva una persona nuova). Inoltre avviene la transizione

$$k \to k-1$$

quando il servizio di una persona viene completato e questa esce dal sistema. Non ci sono altre transizioni possibili.

E' necessario però fare un'osservazione, per eliminare un dubbio che può essere nato nell'esempio precedente. Dobbiamo considerare come stati solo le situazioni un po' durature, non quelle che si presentano per pochi istanti ed hanno una durata inessenziale rispetto al tempo complessivo che trascorre. Ad esempio, quando una persona termina di essere servita, esce; non consideriamo però tra gli stati quello in cui c'è una persona che sta uscendo: la durata di questo avvenimento è brevissima ed oltre tutto inessenziale. Analogamente non consideriamo come stato la situazione in cui una persona sta entrando in banca. Analogamente, è vero che tra l'istante in cui una persona completa il servizio e l'istante in cui il cliente successivo arriva alla cassa, passano alcuni secondi, ma anch'essi sono inessenziali rispetto al tempo di servizio di un cliente o rispetto ai tempi di attesa in coda; quindi non consideriamo come stato quella situazione intermedia.

Infine, escludiamo un'ulteriore eventualità: quella che accadano due transizioni contemporaneamente. Ad esempio, escludiamo che possa contemporanemanete finire il servizio un cliente ed entrarne un altro in banca. Dal punto di vista matematico questa contemporaneità sarà impossibile sotto le ipotesi di tempi aleatori esponenziali che imporremo tra breve. Al di là di questa motivazione rigorosa, conviene semplicemente ritenere che sia impossibile dal

punto di vista pratico che accada la contemporaneità esatta; questo semplifica le cose, ad esempio evita che si debbano introdurre complicate transizioni che tengano conto di eventi simultanei.

Vediamo un altro semplice esempio.

Esempio 100 *Una macchina svolge lavorazioni a ciclo continuo. Può però trovarsi in vari stati: quello di lavoro a massimo regime; quello di lavoro a regime ridotto a causa della necessità di raffreddamento; quello di fermo a causa di manutenzione. Usiamo quindi tre stati, che ad esempio potremmo indicare con tre lettere (rispettivamente M, R, F). Poi bisogna vedere che transizioni sono possibili. Supponiamo che un termostato (più un sistema di raffreddamento) regoli il passaggio da M ad R, nei due sensi. Inoltre, supponiamo che sia da M sia da R si possa passare ad F a causa di un guasto o deterioramento. Infine, supponiamo che da F si passi solo a R, in quanto è necessaria una prima fase di lavoro a regime ridotto di riscaldamento o rodaggio, prima di mettersi a regime massimo. Le transizioni possibili sono quindi*

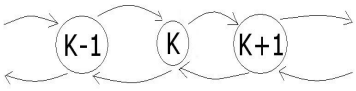
$$\begin{aligned} M &\to R, \quad R \to M \\ M &\to F, \quad R \to F \\ F &\to R. \end{aligned}$$

Se poi le fasi di fermo fossero di diverso tipo, ad esempio alcune di manutenzione programmata ed altre di riparazione guasti, dovremmo sdoppiare lo stato F.

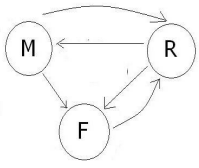
5.3.2 Stati e grafi

Il modo più semplice di raffigurare una situazione del tipo descritto sopra è di disegnare gli stati possibili e tracciare delle frecce tra uno stato e l'altro se esiste la transizione tra essi. Un insieme di stati e frecce è un grafo.

Nell'esempio 99 gli stati sono tutti gli interi non negativi, quindi non è possibile raffigurare completamente il grafo; ci si accontenta di tracciarne un pezzo rappresentativo. Si veda la prima figura. La seconda figura descrive invece il sistema del secondo esempio.



Grafo dell'esempio 1.



Grafo dell'esempio 2.

5.3.3 Tempi di permanenza aleatori

In certi esempi la durata della permanenza in un certo stato è nota e deterministica: è il caso delle manutenzioni programmate, oppure di certi servizi che richiedono un tempo ben preciso, o di certe lavorazioni industriali ben precise e programmate. In altri esempi la durata è aleatoria: è il caso più comune nelle code di servizio. Tutta la teoria che descriveremo in questa scheda è relativa al caso di permanenze aleatorie. Per variare un po' i nostri esempi, a volte includeremo nello schema delle permanenze aleatorie anche alcuni casi che lo sono assai poco; si capirà che questo ha valore puramente didattico, per esercitarsi con la matematica.

Introduciamo una variabile aleatoria per ciascuna transizione possibile. Ad esempio, se in un grafo c'è la transizione dallo stato A allo stato B, introduciamo una v.a. T_{AB} che indica il tempo che il sistema trascorre in A prima di effettuare la transizione in B.

Che tipo (gaussiano ecc.) di variabili si usano per questo scopo? Bisognerebbe analizzare ogni specifico esempio reale effettuando opportune rilevazioni statistiche e decidere sulla base dei dati. Noi qui, per semplicità ed uniformità, operiamo una scelta drastica: tutti i tempi di attesa T_{AB} saranno *esponenziali*. Un tempo di attesa esponenziale è caratterizzato dal suo parametro λ_{AB}, reciproco del tempo medio di attesa:

$$E[T_{AB}] = \frac{1}{\lambda_{AB}}.$$

Il numero λ_{AB} può essere interpetato come numero di eventi per unità di tempo, o tasso di transizione; torneremo su questo più avanti.

Siccome vale, per le esponenziali, la proprietà di assenza di memoria, accade quanto segue: supponiamo che il sistema sia fermo nello stato A già da un po' e che io, osservatore, inizi ora ad osservare il sistema. Allora il sistema rimarrà in A prima di transire a B per un tempo esponenziale di parametro λ_{AB}. In altre parole, il sistema non conserva memoria del fatto di essere stato già per un po' in A; ad ogni istante è come se ripartisse da capo. E' chiaro che questa è un'idealizzazione, quasi mai verificata negli esempi reali; rappresenta la maggior restrizione della teoria che stiamo per svolgere, ma anche la sua potenza dal punto di vista del calcolo esplicito. Se si abbandona il mondo delle v.a. esponenziali, non si riesce a calcolare quasi più nulla esplicitamente, per cui si deve ricorrere al metodo simulativo (creazione di un programma al computer che simula l'evoluzione temporale del sistema).

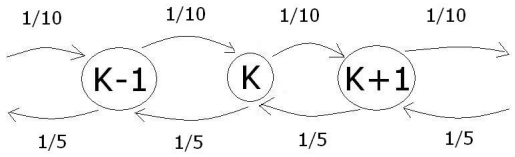
Riassumiamo: ad ogni transizione $A \to B$ associamo un tempo aleatorio esponenziale T_{AB} di parametro λ_{AB}. Quindi possiamo scrivere il numero λ_{AB} sulla freccia del grafo, per ricordarcelo. Torniamo all'esempio 99 del paragrafo precedente. Supponiamo che il tempo di servizio $T_{k,k-1}$, cioè il tempo che si deve attendere prima che una persona in fase di servizio abbia finito, sia mediamente di 5 minuti:

$$E[T_{k,k-1}] = 5 \text{ min.}$$

mentre il tempo tra un arrivo e l'altro, $T_{k,k+1}$, cioè il tempo che si deve attendere per l'arrivo di un nuovo cliente nel sistema, abbia media 10 minuti:

$$E[T_{k,k+1}] = 10 \text{ min.}$$

Tracceremo allora il grafo arricchito della figura 3. Infatti, il parametro $\lambda_{k,k-1}$, reciproco del tempo medio, è $1/5$, mentre $\lambda_{k,k+1}$ è $1/10$.



Grafo dell'esempio 1, con tassi di transizione

Stiamo scordando qualcosa? No, per via dell'assenza di memoria. Spieghiamoci meglio. Se non valesse la proprietà di assenza di memoria, i dati del grafo della figura 3 non sarebbero esaustivi, non permetterebbero cioè in ogni momento di capire cosa accadrà al sistema. Se ci troviamo nello stato k, grazie all'assenza di memoria, tutto riparte da ora; quindi possiamo sapere che tempo (aleatorio) dobbiamo attendere prima di effettuare la prossima transizione, e sappiamo dal grafo quali transizioni sono possibili e dopo quanto tempo.

5.3.4 Catene di Markov e processi di Markov a salti

I sistemi a stati finiti ora descritti, con tempi di attesa esponenziali, sono anche detti processi di Markov a salti. A volte vengono detti anche catene di Markov, per quanto questo nome sia meglio riservarlo alle catene a tempo discreto della sezione prcedente. Qui invece il tempo varia con continuità: gli stati sono discreti, a certi istanti aleatori avvengono le transizioni, ma gli istanti di transizione sono numeri reali positivi qualsiasi.

La continuità del tempo è anche all'origine della proprietà matematica secondo cui non possono avvenire due transizioni contemporaneamente. Dati due tempi aleatori esponenziali T_1 e T_2, di parametri qualsiasi, si può dimostrare che

$$P\left(T_1 = T_2\right) = 0.$$

5.3.5 Quale transizione tra varie possibili?

Fissiamo l'attenzione sul solito esempio 99. Se siamo nello stato k, cioè nel sistema ci sono k utenti ($k-1$ in coda ed uno in servizio), ci vuole un tempo $T_{k,k-1}$ per completare il servizio della persona alla cassa, e ci vuole il tempo $T_{k,k+1}$ per l'arrivo di un nuovo cliente in banca. Ma questi due eventi, a partire dallo stato k, non avverranno entrambi: se finisce prima il servizio, si passa allo stato $k-1$ e quindi non si è più nello stato k (quindi smette l'attesa della transizione $k \to k+1$). Per l'assenza di memoria, non importa quanto tempo siamo rimasti in k in attesa dell'arrivo di un nuovo cliente. Ora siamo in $k-1$: a questo nuovo stato sono associati due tempi, $T_{k-1,k}$ che rappresenta il tempo di attesa dell'arrivo di un nuovo cliente, e $T_{k-1,k-2}$ che rappresenta il tempo di attesa del servizio del cliente che ora si trova alla cassa.

Una immagine di sicuro effetto per consolidare l'osservazione precedente è quella degli orologi aleatori esponenziali che suonano allo scadere del loro tempo di attesa. Quando

siamo in uno stato k, accendiamo due orologi. Uno è relativo alla transizione $k \to k-1$: quando suona, dobbiamo effettuare quella transizione. L'altro è relativo a $k \to k+1$. Il primo che suona detta la transizione da effettuare; l'altro allora viene semplicemente spento. Fatta la transizione (cosa istantanea) attiviamo i nuovi orologi aleatori relativi al nuovo stato in cui ci troviamo.

5.3.6 Tempo di permamenza

Capita la questione precedente, ci accorgiamo che, se siamo nello stato k (continuiamo a riferirci all'esempio 99 per chiarezza), ci sono in realtà tre tempi aleatori in gioco:

$$T_{k,k-1}, \quad T_{k,k+1}, \quad T_k^{perm} = \min\left(T_{k,k-1}, T_{k,k+1}\right).$$

Il tempo T_k^{perm} è il tempo di permanenza nello stato k. Ci muoviamo da k quando suona il primo orologio, cioè all'istante $\min\left(T_{k,k-1}, T_{k,k+1}\right)$.

Il teorema sul minimo di v.a. esponenziali indipendenti dice che anche T_k^{perm} è una v.a. esponenziale, ed il suo parametro è

$$\lambda_k^{perm} = \lambda_{k,k-1} + \lambda_{k,k+1}.$$

Non scriviamo però questo numero sul grafo sia perché è sovrabbondante (è la somma dei numeri già scritti) sia perché ha un ruolo decisamente inferiore.

Come mai allora ne parliamo? Il problema è quello di decidere i numeri λ_{AB} negli esempi. Quando esaminiamo un problema concreto, prima di tutto dobbiamo individuare gli stati (come abbiamo sottolineato sin dall'inizio), poi le transizioni possibili $A \to B$, e finalmente i tassi di transizione λ_{AB}. Ma chi sono? Sono i reciproci dei tempi medi di attesa. Qui nasce il potenziale equivoco: attesa di cosa? Non l'attesa della transizione fuori da A. La transizione fuori da A avviene all'istante T_k^{perm}. Non quindi il tempo di permanenza in A.

Bisogna ragionare relativamente alla specifica transizione $A \to B$, isolandola dalle altre possibili, per così dire (se la si mescola si rischia di ricadere in T_k^{perm}). Rivediamo il solito esempio 99. Se ci sono k persone nel sistema, cioè siamo nello stato k, esaminiamo la transizione $k \to k-1$. Essa avviene quando si completa il servizio della persona alla cassa, e ciò richiede un tempo medio di 5 minuti, quindi $T_{k,k-1}$ è un tempo aleatorio esponenziale di parametro $\lambda_{k,k-1} = 1/5$. E' tutto abbastanza semplice, basta non confondersi e mettersi a pensare al fatto che contemporaneamente attendiamo l'arrivo di un nuovo cliente ecc. Ogni transizione va esaminata a se stante.

5.3.7 Prima l'una o l'altra?

Supponiamo che da uno stato A si possa transire in B o C. Sono quindi attivi i tempi $T_{A,B}$ e $T_{A,C}$, di tassi $\lambda_{A,B}$ e $\lambda_{A,C}$, rispettivamente.

Ci chiediamo: è più probabile passare a B o a C? Più quantitativamente, possiamo chiederci: con che probabilità la transizione sarà $A \to B$ (invece che $A \to C$)? Tale probabilità è pari a

$$P\left(T_{A,B} < T_{A,C}\right).$$

Omettiamo il calcolo, non semplicissimo. Il risultato per fortuna è semplice:

$$\frac{\lambda_{A,B}}{\lambda_{A,B}+\lambda_{A,C}}.$$

Ripetiamo. Una transizione prima o poi ci sarà. La probabilità che si tratti della transizione $A \to B$ è $\frac{\lambda_{A,B}}{\lambda_{A,B}+\lambda_{A,C}}$, mentre la probabilità che si tratti della transizione $A \to C$ è $\frac{\lambda_{A,C}}{\lambda_{A,B}+\lambda_{A,C}}$.

Questa formula si generalizza al caso di tante possibili transizioni in uscita da A: se da A si può andare in $B_1, B_2, B_3, ...$ allora la probabilità che la transizione sia proprio $A \to B_i$ è

$$\frac{\lambda_{A,B_i}}{\lambda_{A,B_1}+\lambda_{A,B_2}+...}.$$

5.3.8 Regime stazionario o di equilibrio

Alcune delle considerazioni che svolgeremo in questo paragrafo sono simili quelle sulle misure invarianti delle catene di Markov. Tuttavia, per rendere più indipendente la lettura delle varie sezioni, accettiamo alcune ripetizioni.

Fino ad ora abbiamo solo discusso come descrivere un sistema a eventi discreti tramite stati, transizioni, tassi. Ma cosa è possibile calcolare? Ad esempio le probabilità degli stati in regime stazionario.

Innanzi tutto bisogna capire intuitivamente di cosa stiamo parlando. Immaginiamo la banca. All'apertura, la coda è vuota. Questo non è regime stazionario. Se, come accade spesso, si sono accumulate fuori dalla banca molte persone, pochi istanti dopo l'apertura queste persone sono all'interno della banca, in coda. La coda in questo momento è molto lunga, in modo anomalo, per il fatto che non ci sono ancora stati servizi e si era accumulata la gente fuori. Anche questo non è regime stazionario.

Dopo un po', ad esempio una mezz'ora, le cose si sono rimesse a posto: la cassa ha smaltito l'eccesso iniziale di clienti, ora lavora a regime, nuova gente arriva, ma non più a gruppi come al primo istante, bensì con regolarità (pur aleatoria). Siamo in una situazione a regime, una situazione di equilibrio.

Si noti bene che stazionarietà ed equilibrio qui non significa (come in sistemi deterministici) che lo stato non cambia più. Lo stato cambia, casualmente, il sistema fluttua tra i vari stati, ma in modo statisticamente ripetitivo, senza ad es. permanere in valori anomali dello stato.

Supponiamo di esaminare un sistema a eventi discreti quando esso si trovi a regime. Per ciascuno stato A introduciamo la probabilità di trovarsi nello stato A, che indichiamo con π_A. Ad esempio, quando la banca è a regime, introduciamo la probabilità di avere la banca vuota, π_0; la probabilità di avere una persona allo sportello e nessuna in coda, π_1; e così via. Intuitivamente con π_A intendiamo la *frequenza relativa di tempo* in cui osserviamo lo stato A. Ad esempio, per quanto tempo rispetto ad un'ora di lavoro troveremo la banca vuota? Il numero π_0 è una sorta di astrazione di questa frequenza relativa.

La frequenza relativa purtroppo è legata al lasso di tempo che si considera e addirittura all'esempio particolare di osservazione; il numero p_A invece ha un ruolo più assoluto; per questo diciamo che è un'astrazione della frequenza relativa.

Se indichiamo gli stati del sistema con $A_1, A_2, ...$, il vettore

$$(\pi_{A_1}, \pi_{A_2}, ...)$$

viene chiamato in vari modi, ad es. *misura invariante*, vettore delle *probabilità invarianti*, misura d'equilibrio, misura stazionaria e così via.

Ci poniamo il problema di calcolare i numeri π_A. Esaminiamo come sempre l'esempio 99, estrapolando poi una regola generale. Vogliamo calcolare i numeri π_k per ogni intero $k \geq 0$. La possibilità di effettuare questo calcolo viene dal fatto che questi numeri sono legati tra loro da equazioni, le cosidette *equazioni di bilancio di flusso*. Prima diamo la regola, poi cerchiamo di darne una ragionevole spiegazione. La regola si ricorda così. Si deve immaginare la probabilità π_k come una porzione di massa. Dallo stato k esce una certa percentuale di questa massa, precisamente

$$\pi_k \cdot (\lambda_{k,k-1} + \lambda_{k,k+1})$$

ovvero il prodotto di π_k per la somma dei tassi uscenti da k. Nello stato k entra poi un percentuale della massa degli stati che possono transire in k. Possono transire in k gli stati $k-1$ e $k+1$. Dallo stato $k-1$ transisce la percentuale

$$\pi_{k-1} \cdot \lambda_{k-1,k}$$

ovvero il prodotto della massa dello stato $k-1$ per il tasso di transizione da $k-1$ a k. Analogamente, dallo stato $k+1$ transisce $\pi_{k+1} \cdot \lambda_{k+1,k}$. Questo flusso di massa deve essere nullo, per essere all'equilibrio, a regime. Deve valere cioè

$$\pi_k \cdot (\lambda_{k,k-1} + \lambda_{k,k+1}) = \pi_{k-1} \cdot \lambda_{k-1,k} + \pi_{k+1} \cdot \lambda_{k+1,k}.$$

Questa è l'equazione del bilancio di flusso nello stato k.

In astratto, considerando uno stato A e gli stati $B_1, B_2, B_3, ...$ ad esso collegati, deve valere

$$\pi_A \cdot (\lambda_{A,B_1} + \lambda_{A,B_2} + ...) = \pi_{B_1} \cdot \lambda_{B_1,A} + \pi_{B_2} \cdot \lambda_{B_2,A} + ... \tag{5.2}$$

5.3.9 Dimostrazione dell'equazione (5.2)

Capita la struttura della formula, cerchiamo di dimostrarla almeno parzialmente. Indichiamo con $A^{(t)}$ l'evento "il sistema si trova in A all'istante t", con B_n l'evento "il sistema si trova in B_n all'istante t", per ogni n. Vale allora, per la formula di fattorizzazione

$$P\left(A^{(t+\varepsilon)}\right) = P\left(A^{(t+\varepsilon)}|A^{(t)}\right) P\left(A^{(t)}\right) + \sum_n P\left(A^{(t+\varepsilon)}|B_n^{(t)}\right) P\left(B_n^{(t)}\right)$$

dal momento che $A^{(t-\varepsilon)}$, $B_1^{(t-\varepsilon)}$, $B_2^{(t-\varepsilon)}$, ... è una partizione (uno almeno è vero e sono disgiunti). Abbiamo fattorizzato la probabilità di trovarsi in A un istante dopo t (cioè al tempo $t+\varepsilon$), rispetto a dove si trova il sistema all'istante t. Il sistema si trova in regime stazionario, quindi $P\left(A^{(t)}\right) = P\left(A^{(t+\varepsilon)}\right) = \pi_A$ ecc, quindi

$$\pi_A = P\left(A^{(t+\varepsilon)}|A^{(t)}\right) \pi_A + \sum_n P\left(A^{(t+\varepsilon)}|B_n^{(t)}\right) \pi_{B_n}. \tag{5.3}$$

Fin qui è tutto rigoroso. Ora dobbiamo accettare che, per ε piccolo, in prima approssimazione valga

$$P\left(A^{(t+\varepsilon)}|A^{(t)}\right) \sim P\left(T_A^{perm} > \varepsilon\right)$$
$$P\left(A^{(t+\varepsilon)}|B_n^{(t)}\right) \sim P\left(T_{B_n,A} < \varepsilon\right).$$

Intuitivamente, $P\left(A^{(t+\varepsilon)}|A^{(t)}\right)$ è la probabilità di trovarsi ancora in A dopo un tempo ε, partendo da A, quindi è la probabilità che il tempo di permanenza in A sia maggiore di ε. Similmente per la seconda uguaglianza. In realtà queste non sono esattamente delle uguaglianze, per il fatto che nel pur brevissimo tempo ε può accadere (ma ciò avviene con bassissima probabilità) che il sistema percorra un più complicato cammino tra più stati, non solo la singola transizione considerata sopra. Accettiamo questa approssimazione. Ricordiamo poi che

$$P\left(T_{B_n,A} < \varepsilon\right) = 1 - e^{-\varepsilon\lambda_{B_n,A}}$$

quindi, per lo sviluppo di Taylor dell'esponenziale, in prima approssimazione per ε piccolo vale

$$P\left(T_{B_n,A} < \varepsilon\right) \sim \varepsilon\lambda_{B_n,A}.$$

Similmente, ricordando che T_A^{perm} ha parametro $\sum_n \lambda_{A,B_n}$,

$$P\left(T_A^{perm} > \varepsilon\right) = e^{-\varepsilon\sum_n \lambda_{A,B_n}} \sim 1 - \varepsilon\sum_n \lambda_{A,B_n}.$$

Sostituendo nell'equazione (5.3) troviamo

$$\pi_A = \left(1 - \varepsilon\sum_n \lambda_{A,B_n}\right)\pi_A + \sum_n \varepsilon\lambda_{B_n,A}\pi_{B_n}$$

da cui

$$\varepsilon\pi_A\sum_n \lambda_{A,B_n} = \varepsilon\sum_n \pi_{B_n}\lambda_{B_n,A}$$

da cui finalmente l'equazione del bilancio di flusso (5.2).

5.3.10 Il sistema delle equazioni di bilancio

Supponiamo di studiare un esempio con un numero finito di stati. Per esemplificare, supponiamo inizialmente di avere solo due stati, A e B. Scriviamo il bilancio di flusso sia in A sia in B:

$$\text{bilancio in } A: \quad \pi_A\lambda_{A,B} = \pi_B\lambda_{B,A}$$
$$\text{bilancio in } B: \quad \pi_B\lambda_{B,A} = \pi_A\lambda_{A,B}.$$

Vediamo subito che queste due equazioni coincidono. Quindi è una sola equazione, nelle due incognite π_A e π_B. Serve un'altra equazione. essa è

$$\pi_B + \pi_A = 1.$$

In questo semplice esempio vediamo che per trovare π_A e π_B bisogna scrivere una equazione di bilancio di flusso (non tutte e due) ed aggiungere la condizione di probabilità totale unitaria. Bisogna cioè risolvere il sistema

$$\begin{cases} \pi_A \lambda_{A,B} = \pi_B \lambda_{B,A} \\ \pi_B + \pi_A = 1. \end{cases}$$

Vediamo per sicurezza il caso con tre stati, A, B e C, per vedere se succede la stessa cosa. Il bilancio nei tre stati è:

$$\begin{aligned} \text{bilancio in } A: &\quad \pi_A (\lambda_{A,B} + \lambda_{A,C}) = \pi_B \lambda_{B,A} + \pi_C \lambda_{C,A} \\ \text{bilancio in } B: &\quad \pi_B (\lambda_{B,A} + \lambda_{B,C}) = \pi_A \lambda_{A,B} + \pi_C \lambda_{C,B} \\ \text{bilancio in } C: &\quad \pi_C (\lambda_{C,A} + \lambda_{C,B}) = \pi_A \lambda_{A,C} + \pi_B \lambda_{B,C}. \end{aligned}$$

Qui è meno ovvio capire se sono tre equazioni indipendenti. Sommiamo però le prime due:

$$\begin{aligned} &\pi_A (\lambda_{A,B} + \lambda_{A,C}) + \pi_B (\lambda_{B,A} + \lambda_{B,C}) \\ &= \pi_B \lambda_{B,A} + \pi_C \lambda_{C,A} + \pi_A \lambda_{A,B} + \pi_C \lambda_{C,B} \end{aligned}$$

ovvero

$$\pi_A \lambda_{A,C} + \pi_B \lambda_{B,C} = \pi_C \lambda_{C,A} + \pi_C \lambda_{C,B}$$

ovvero infine

$$\pi_C (\lambda_{C,A} + \lambda_{C,B}) = \pi_A \lambda_{A,C} + \pi_B \lambda_{B,C}$$

che è proprio il bilancio di flusso in C. Quindi esso è sovrabbondante, è già incluso nei precedenti. Di nuovo, quindi, per trovare la soluzione (p_A, p_B, p_C) bisogna risolvere il sistema formato dal bilancio in due stati più l'equazione della massa unitaria:

$$\begin{cases} \pi_A (\lambda_{A,B} + \lambda_{A,C}) = \pi_B \lambda_{B,A} + \pi_C \lambda_{C,A} \\ \pi_B (\lambda_{B,A} + \lambda_{B,C}) = \pi_A \lambda_{A,B} + \pi_C \lambda_{C,B} \\ \pi_A + \pi_B + \pi_C = 1. \end{cases}$$

Si dimostra che questo sistema ha sempre almeno una soluzione. La scelta degli stati in cui fare il bilancio è indifferente. Il principio si generalizza ad un numero finito qualsiasi di stati.

Più complessa è la questione dell'unicità. Ci sono esempi con più di una soluzione. A questo scopo, invece che lavorare a livello algebrico sul sistema, conviene prima esaminare gli stati del grafo tramite i concetti di stato transitorio e ricorrente, classe irriducibile. Quando ci si è ridotti ad una classe irriducibile, lì la soluzione è unica. Non studiamo qui in dettaglio questo problema: negli esempi cercheremo di trovare la soluzione del sistema; se è unica, il problema è risolto.

5.4 Esempi dalla teoria delle code

Gli esempi del nostro corso si possono raggruppare in tre classi. La prima è quella degli esempi con pochissimi stati, in cui si risolve manualmente il sistema delle equazioni di bilancio di flusso. La seconda classe è quella delle catene di nascita e morte, in cui c'è una formula

esplicita per le probabilità invarianti. Infine, la terza classe è quella delle catene che non sono di nascita e morte ma nemmeno così facili da svolgere esplicitamente i conti; per quelle possiamo solo trovare delle relazioni da mettere nel calcolatore.

La maggor parte degli esempi che seguono è presa dalla teoria delle code, da cui il titolo della sezione, ma useremo anche altri esempi per chiarire le cose.

Iniziamo con un semplice esempio del primo tipo.

Esempio 101 *Riprendiamo l'esempio 100. Supponiamo che quando la macchina viaggia a regime massimo (M), si surriscaldi progressivamente superando la soglia ammissibile dopo un tempo aleatorio di media 15 minuti. A quel punto, il termostato comanda di passare al regime ridotto (R). In quel regime la macchina continua a lavorare e si raffredda; questa operazione di raffreddamento dura in media 3 minuti, dopo i quali la macchina torna a regime massimo. In ogni regime può capitare un guasto che porta la macchina in fermo (F). A regime massimo, questo capita mediamente dopo 60 minuti di lavoro continuato (intendiamo una fase ininterrotta a regime massimo); a regime ridotto, dopo 120 minuti. Infine, quando la macchina è ferma per riparazione, la riparazione dura in media 30 minuti. Calcolare la probabilità a regime di avere la macchina ferma.*

***Soluzione**. Il grafo è quello indicato alla figura 4. Il bilancio in M e F è*

$$\pi_M\left(\frac{1}{15}+\frac{1}{60}\right)=\pi_R\frac{1}{3}$$
$$\pi_F\frac{1}{30}=\pi_M\frac{1}{60}+\pi_R\frac{1}{120}$$

a cui aggiungiamo l'equazione

$$\pi_M+\pi_R+\pi_F=1.$$

Esprimiamo π_M e π_F in funzione di π_R dalle prime due:

$$\pi_M=4\cdot\pi_R$$
$$\pi_F=\frac{1}{2}\cdot\pi_M+\frac{1}{4}\cdot\pi_R=2\cdot\pi_R+\frac{1}{4}\cdot\pi_R=\frac{9}{4}\cdot\pi_R.$$

Ora sostituiamo nella terza equazione:

$$4\cdot\pi_R+\pi_R+\frac{5}{4}\cdot\pi_R=1$$

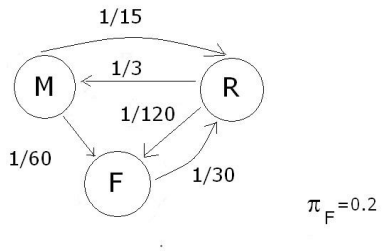
da cui

$$\pi_R=0.16$$

da cui infine

$$\pi_M=4\cdot 0.16=0.64$$
$$\pi_F=\frac{5}{4}\cdot 0.16=0.2.$$

La probabilità di fermo è 0.2.



Esempio 2 con tassi di transizione.

Vediamo ora un esempio del terzo tipo.

Esempio 102 *Una catena di montaggio è composta da due stazioni di lavoro in sequenza. Ogni pezzo che entra nel sistema viene mandato alla prima stazione, dove si mette in coda. Quando è stato lavorato viene mandato alla seconda stazione, dove pure si mette in coda. Dopo la seconda lavorazione, esce dal sistema. Tra un'entrata nel sistema e l'altra passa un tempo di media 10 minuti. La lavorazione della prima stazione richiede 5 minuti. La lavorazione della seconda ne richiede 7. Descrivere il sistema con un processo di Markov a salti.*

***Soluzione**. Se osserviamo il sistema ad un generico istante, vediamo un certo numero n di pezzi in coda alla prima stazione ed un certo numero k in coda alla seconda, includendo nelle code i pezzi in corso di lavorazione (per semplicità di linguaggio). Quindi il generico stato del sistema è una coppia (n, k) di interi non negativi. Dallo stato (n, k) si passa a $(n + 1, k)$ se entra nel sistema un nuovo pezzo, e ciò avviene con tasso $\frac{1}{10}$. Dallo stato (n, k) si passa a $(n - 1, k + 1)$ quando la prima stazione completa la lavorazione in corso; questo avviene con tasso $\frac{1}{5}$. Infine, dallo stato (n, k) si passa in $(n, k - 1)$ se la seconda stazione completa la lavorazione in corso; questo avviene con tasso $\frac{1}{7}$. Si osservi che in un problema di questa complessità non viene richiesto il calcolo della distribuzione invariante.*

Nel prossimo paragrafo affronteremo in modo sistematico gli esempi del secondo tipo, cioè i processi di nascita e morte.

5.4.1 Processi di nascita e morte

Si chiamano così tutti quelli che hanno come stati i numeri interi non negativi e come transizioni possibili quelle tra primi vicini (come nell'esempio 99). Quindi gli stati sono i numeri $k = 0, 1, 2, ...$ e le transizioni possibili sono solamente

$$\begin{aligned} & 0 \to 1, \quad 1 \to 2, \quad 2 \to 3, \quad ... \\ ..., \quad & 3 \to 2, \quad 2 \to 1, \quad 1 \to 0. \end{aligned}$$

Il bilancio di flusso nel generico stato k è

$$\pi_k \cdot (\lambda_{k,k-1} + \lambda_{k,k+1}) = \pi_{k-1} \cdot \lambda_{k-1,k} + \pi_{k+1} \cdot \lambda_{k+1,k}.$$

Questo vale per $k \geq 1$, mentre per $k = 0$ esso è semplicemente

$$\pi_0 \cdot \lambda_{0,1} = \pi_1 \cdot \lambda_{1,0}.$$

Nel seguito supponiamo che tutti i numeri λ siano strettamente positivi, altrimenti bisogna ragionare caso per caso. Ricaviamo π_1 in funzione di π_0 da questa equazione e sostituiamolo nella seconda equazione (quella di bilancio nello stato $k = 1$), ricavando anche π_2 in funzione di π_0 :

$$\pi_1 = \pi_0 \cdot \frac{\lambda_{0,1}}{\lambda_{1,0}}$$

$$\begin{aligned}
\pi_1 \cdot (\lambda_{1,0} + \lambda_{1,2}) &= \pi_0 \cdot \lambda_{0,1} + \pi_2 \cdot \lambda_{2,1} \\
\pi_2 &= \frac{\pi_1 \cdot (\lambda_{1,0} + \lambda_{1,2}) - \pi_0 \cdot \lambda_{0,1}}{\lambda_{2,1}} \\
&= \pi_0 \frac{\frac{\lambda_{0,1}}{\lambda_{1,0}} \cdot (\lambda_{1,0} + \lambda_{1,2}) - \lambda_{0,1}}{\lambda_{2,1}} \\
&= \pi_0 \frac{\lambda_{0,1}\lambda_{1,2}}{\lambda_{2,1}\lambda_{1,0}}.
\end{aligned}$$

Queste notevolissime semplificazioni si ripetono ad ogni passo: se ora prendiamo il bilancio nello stato $k = 2$ e sostituiamo tutto in funzione di π_0, troviamo

$$\pi_3 = \pi_0 \frac{\lambda_{0,1}\lambda_{1,2}\lambda_{2,3}}{\lambda_{3,2}\lambda_{2,1}\lambda_{1,0}}.$$

Per induzione si può verificare che

$$\pi_n = \pi_0 \frac{\lambda_{0,1} \cdots \lambda_{n-1,n}}{\lambda_{n,n-1} \cdots \lambda_{1,0}}.$$

Ora bisogna trovare π_0 imponendo la condizione

$$\sum_{n=0}^{\infty} \pi_n = 1.$$

Introduciamo la notazione

$$a_0 = 1, \quad a_n = \frac{\lambda_{0,1} \cdots \lambda_{n-1,n}}{\lambda_{n,n-1} \cdots \lambda_{1,0}}.$$

Poniamo inoltre

$$a = \sum_{n=0}^{\infty} a_n.$$

La condizione $\sum_{n=0}^{\infty} \pi_n = 1$ corrisponde alla condizione $\pi_0 \cdot a = 1$. Può accadere che sia $a < \infty$ oppure $a = +\infty$. Se $a = +\infty$, vediamo che è impossibile imporre la condizione $\sum_{n=0}^{\infty} \pi_n = 1$ (implicherebbe $\pi_0 = 0$, ma allora varrebbe anche $\pi_n = \pi_0 \cdot a_n = 0$ per ogni n, quindi $\sum_{n=0}^{\infty} \pi_n = 0$).

Se invece $a < \infty$, vale

$$\pi_0 = \frac{1}{a}, \quad \pi_n = \frac{a_n}{a}.$$

Avendo posto $a_0 = 1$, anche la formula $\pi_0 = \frac{1}{a}$ è un caso particolare di $\pi_n = \frac{a_n}{a}$. Abbiamo trovato:

Teorema 35 *Per un processo di nascita e morte con tutti i tassi $\lambda_{n,n-1}$ e $\lambda_{n,n+1}$ strettamente positivi, posto*

$$a_0 = 1, \quad a_n = \frac{\lambda_{0,1} \cdots \lambda_{n-1,n}}{\lambda_{n,n-1} \cdots \lambda_{1,0}},$$

se vale

$$a = \sum_{n=0}^{\infty} a_n < \infty$$

allora il sistema raggiunge il regime stazionario descritto dalla distribuzione invariante

$$\pi_n = \frac{a_n}{a}, \quad n \geq 0.$$

In generale è difficile calcolare a. Vediamo alcuni casi notevoli in cui questo è possibile.

5.4.2 Tassi costanti

Supponiamo che sia

$$\lambda_{n,n-1} = \mu, \quad \lambda_{n,n+1} = \lambda.$$

Poniamo

$$\rho = \frac{\lambda}{\mu}.$$

Vale

$$a_n = \rho^n.$$

Se $\rho < 1$, la serie geometrica $\sum_{n=0}^{\infty} \rho^n$ converge, ed ha $\frac{1}{1-\rho}$ come somma. Se invece $\rho \geq 1$, la serie diverge. Nel caso $\rho < 1$ vale allora $a = \frac{1}{1-\rho}$ e quindi

$$\pi_n = (1 - \rho)\, \rho^n.$$

La distribuzione invariante ha legge geometrica di parametro ρ.

Pensiamo ad una coda, come nell'esempio 99. La condizione $\rho < 1$ equivale a $\lambda < \mu$, quindi a

$$E[T_a] > E[T_s]$$

dove T_a indica il tempo che intercorre tra un arrivo e l'altro, T_s il tempo necessario per un servizio. Abbiamo scoperto che se il tempo medio di servizio è minore del tempo tra un arrivo e l'altro, si stabilisce il regime stazionario e siamo in grado di calcolare la distribuzione invariante. Quando invece $E[T_a] \leq E[T_s]$, cioè intercorre meno tra un arrivo e l'altro rispetto ai servizi, il servente non è in grado di far fronte agli arrivi e non si instaura un regime stazionario, bensì la coda diverge all'infinito (il numero di persone in coda cresce indefinitamente).

Le code descritte da questo esempio sono le code con un servente solo, dette $M/M/1$.

5.4.3 Tassi di crescita costanti, tassi di decrescita lineari

Supponiamo che sia

$$\lambda_{n,n-1} = n \cdot \mu, \quad \lambda_{n,n+1} = \lambda.$$

Poniamo di nuovo

$$\rho = \frac{\lambda}{\mu}.$$

Vale

$$a_n = \frac{\rho^n}{n!}.$$

La serie $\sum_{n=0}^{\infty} \frac{\rho^n}{n!}$ converge per ogni valore di ρ, ed ha e^ρ come somma. Quindi si instaura sempre il regime stazionario e vale

$$\pi_n = e^{-\rho} \frac{\rho^n}{n!}.$$

Questa, tra l'altro, è una distribuzione di Poisson di parametro ρ.

Questo esempio si incontra nelle code con infiniti serventi (ovviamente un'idealizzazione della realtà). Infatti, supponiamo che il tempo tra un arrivo e l'altro abbia media $\frac{1}{\mu}$ e che nel sistema ci siano infiniti serventi disponibili, ciascuno che serve con tempo medio $\frac{1}{\lambda}$. Ogni nuovo cliente che entra nel sistema ha subito un servente libero a disposizione, quindi inizia subito il servizio. Non c'è coda. Si passa dallo stato k allo stato $k-1$ quando un cliente tra tutti quelli in fase di servizio (cioè tutti quelli nel sistema, quindi k clienti) completa il suo servizio. Se ci sono appnto k clienti in fase di servizio, ed indichiamo con $T_s^{(1)}, ..., T_s^{(k)}$ i loro tempi di servizio, l'istante in cui il primo cliente termina il servizio è

$$T = \min\left(T_s^{(1)}, ..., T_s^{(k)}\right)$$

che, per un noto teorema, è una v.a. esponenziale di parametro pari alla somma dei parametri, quindi pari a $k \cdot \mu$. Ecco quindi che il tasso di transizione da k a $k-1$ è $k \cdot \mu$.

Le code di questo esempio vengono a volte indicate con $M/M/\infty$.

5.4.4 Coda con c serventi

Esaminiamo ora il caso intermedio, di un sistema con c serventi, $1 < c < \infty$, denotato col simbolo $M/M/c$. Il tasso di arrivo sia sempre λ e quello di servizio di un singolo servente sempre μ.

Se nel sistema ci sono c clienti o più di c, tutti i serventi sono attivi, quindi liberano una persona con tasso $c\mu$ (stesso ragionamento del caso precedente). Se però nel sistema ci sono $k < c$ persone, solo k serventi stanno lavorando, quindi liberano una persona con tasso $k\mu$. Il grafo è quello della figura 5, dove abbiamo preso $\lambda = \frac{1}{2}$, $\mu = \frac{1}{3}$. Si noti che in questo esempio numerico vale $\mu < \lambda$, cioè il singolo servitore è più lento di un singolo arrivo. Ma giocando in squadra, riescono a raggiungere il regime stazionario. Infatti, posto

$$\rho = \frac{\lambda}{c\mu}$$

vale

$$a_0 = 1, \quad a_1 = \frac{\lambda}{\mu}, \quad a_2 = \frac{\lambda^2}{(1\cdot\mu)\,(2\cdot\mu)} = \frac{\lambda^2}{2\cdot\mu^2}, \quad ...$$

fino a

$$a_{c-1} = \frac{\lambda^{c-1}}{(c-1)!\cdot\mu^{c-1}}$$

e poi, da c in avanti,

$$\begin{aligned} a_c &= \frac{\lambda^c}{c!\cdot\mu^c} \\ a_{c+1} &= \frac{\lambda^{c+1}}{c!\cdot c\cdot\mu^{c+1}} \\ a_{c+2} &= \frac{\lambda^{c+2}}{c!\cdot c^2\cdot\mu^{c+2}} \end{aligned}$$

e così via,

$$a_{c+k} = \frac{\lambda^{c+k}}{c!\cdot c^k\cdot\mu^{c+k}} = \frac{\lambda^c}{c!\mu^c}\rho^k.$$

Quindi

$$\begin{aligned} a &= \sum_{n=0}^{c-1}\frac{\lambda^n}{n!\cdot\mu^n} + \sum_{k=0}^{\infty}\frac{\lambda^c}{c!\mu^c}\rho^k \\ &= \sum_{n=0}^{c-1}\frac{\lambda^n}{n!\cdot\mu^n} + \frac{\lambda^c}{c!\mu^c}\frac{1}{1-\rho}. \end{aligned}$$

Una volta calcolato questo numero (se c è basso, lo si calcola a mano facilmente), la distribuzione invariante è

$$\begin{aligned} \pi_n &= \frac{1}{a}\frac{\lambda^n}{n!\mu^n} \quad \text{per } n = 0, 1, ..., c-1 \\ \pi_n &= \pi_{c+k} = \frac{1}{a}\frac{\lambda^c}{c!\mu^c}\rho^k \quad \text{per } n \geq c \text{ ovvero } k \geq 0 \end{aligned}$$

(attenzione: nella seconda relazione n e k sono legati dalla formula $n = c + k$).

A parte la formula finale, va notato che la condizione per la convergenza della serie, e quindi per il raggiungimento del regime stazionario, è $\rho < 1$ ovvero

$$\lambda < c\mu.$$

Il tasso di arrivo può anche superare il tasso di servizio, ma non deve superare il tasso di c serventi simultanei.

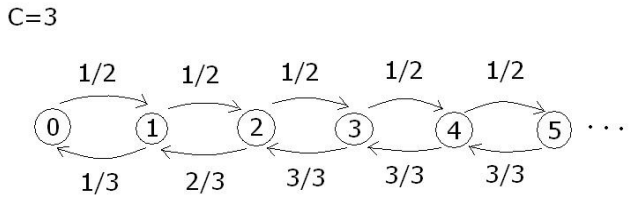


Figura 5.1: Coda con 3 serventi.

5.4.5 Nascita e morte con un numero finito di stati

Consideriamo la situazione dei processi di nascita e morte ma con stati $0, 1, ..., N$. La teoria e le formule finali sono quasi identiche: infatti abbiamo ricavato tutto partendo iterativamente da $k = 0$.

Però è più semplice il discorso legato all'esistenza del regime stazionario: esiste sempre. La distribuzione invariante esiste sempre. La grandezza a ora è data dalla somma finita

$$a = \sum_{n=0}^{N} a_n$$

che quindi è sempre finita, per cui si trova sempre

$$\pi_n = \frac{a_n}{a}, \quad n = 0, 1, ..., N.$$

Tra le piccole varianti notevoli c'è il fatto che possiamo calcolare esplicitamente la probabilità invariante nel caso

$$\lambda_{n,n-1} = \mu, \quad \lambda_{n,n+1} = \lambda.$$

Infatti, ponendo sempre $\rho = \frac{\lambda}{\mu}$, vale anche ora $a_n = \rho^n$ e si conosce il valore della seguente somma:

$$a = \sum_{n=0}^{N} \rho^n = \frac{1 - \rho^{N+1}}{1 - \rho}.$$

Quindi in questo caso

$$\pi_n = \frac{1 - \rho}{1 - \rho^{N+1}} a_n, \quad n = 0, 1, ..., N.$$

5.4.6 Valori medi notevoli

Consideriamo un processo di nascita e morte. Pensiamo ad esempio ad una coda, per avere un linguaggio più immediato. Ci chiediamo: qual'è, all'equilibrio, il numero medio di utenti nel sistema? Questa è sicuramente una delle principali grandezze che desideriamo conoscere (il gestore del servizio ragionerà sulla bontà del suo sistema di servizio in base a numeri di questo tipo).

Detto N il numero aleatorio di utenti nel sistema, all'equilibrio, vale per definizione di valor medio

$$E[N] = \sum_{n=0}^{\infty} n\pi_n$$

in quanto π_n è proprio $P(N = n)$. In generale non ci sono formule esplicite. Vediamo però alcuni esempi.

Numero medio di utenti, tassi costanti

Nel caso

$$\lambda_{n,n-1} = \mu, \quad \lambda_{n,n+1} = \lambda$$

essendo

$$\pi_n = (1-\rho)\,\rho^n$$

vale

$$E[N] = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{(1-\rho)}$$

Abbiamo usato il seguente fatto:

$$\begin{aligned}\sum_{n=0}^{\infty} n\rho^n &= \rho\sum_{n=1}^{\infty} n\rho^{n-1} = \rho\frac{d}{d\rho}\sum_{n=1}^{\infty}\rho^n = \rho\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n \\ &= \rho\frac{d}{d\rho}\frac{1}{1-\rho} = \frac{\rho}{(1-\rho)^2}.\end{aligned}$$

Per inciso, la formula $E[N] = \frac{\rho}{(1-\rho)}$ è la media di una v.a. geometrica di parametro ρ.

Numero medio di utenti, infiniti serventi

Se

$$\lambda_{n,n-1} = n \cdot \mu, \quad \lambda_{n,n+1} = \lambda$$

vale

$$\pi_n = e^{-\rho}\frac{\rho^n}{n!}$$

quindi

$$E[N] = e^{-\rho}\sum_{n=0}^{\infty} n\frac{\rho^n}{n!} = e^{-\rho}\sum_{n=1}^{\infty} n\frac{\rho^n}{n!} = \rho e^{-\rho}\sum_{n=1}^{\infty}\frac{\rho^{n-1}}{(n-1)!}$$
$$= \rho e^{-\rho}\sum_{n=0}^{\infty}\frac{\rho^n}{n!} = \rho e^{-\rho}e^{\rho} = \rho.$$

Il numero medio è semplicemente ρ. In realtà lo sapevamo: avendo osservato che la v.a. N è una Poisson di parametro ρ, già sappiamo che la sua media è ρ.

Esercizio 35 *Cercare una formula, per quanto poco esplicita, per la media $E[N]$ nel caso di una coda con c serventi.*

Numero medio di utenti in attesa

Questo titolo è solo un esempio delle diverse varianti del problema precedente. Vogliamo sottolineare il fatto che le formule precedenti davano il numero medio di utenti nel sistema, incluso quindi quello in fase di servizio. Se la richiesta è un po' diversa, come il numero medio di utenti *in attesa*, bisogna effettuare delle modifiche.

Indichiamo con N_{sist} il numero aleatorio di utenti nel sistema e con N_{att} il numero di quelli in attesa.

L'intuito potrebbe portare a dire che $N_{att} = N_{sist} - 1$ e quindi $E[N_{att}] = E[N_{sist}] - 1$. Se così fosse, non ci sarebbe nulla di nuovo da calcolare. Questa intuizione è quasi giusta ma non completamente. Infatti, la relazione $N_{att} = N_{sist} - 1$ non vale se $N_{sist} = 0$, in quanto non è $N_{att} = -1$ in tal caso. Escluso questo caso, essa è vera. Quindi non vale $E[N_{att}] = E[N_{sist}] - 1$, però l'errore è moderato.

Se vogliamo la formula esatta, si può usare il seguente lemma.

Lemma 7

$$E[N_{att}] = E[N_{sist}] - 1 + \pi_0.$$

Proof. Per definizione di N_{att} e di valor medio, vale

$$E[N_{att}] = \sum_{n=1}^{\infty}(n-1)\pi_n.$$

Questo si può scomporre e riscrivere come segue:

$$= \sum_{n=1}^{\infty} n\pi_n - \sum_{n=1}^{\infty}\pi_n$$
$$= \sum_{n=0}^{\infty} n\pi_n - \sum_{n=0}^{\infty}\pi_n + \pi_0$$
$$= E[N_{sist}] - 1 + \pi_0.$$

La dimostrazione è completa. ■

A seconda della quantità media che interessa, bisogna svolgere calcoli di questo tipo; questo era solo un esempio.

Tempo medio di permanenza di un utente nel sistema

Questo è un altro valor medio di importanza fondamentale per le applicazioni. Il suo calcolo è completamente diverso. Naturalmente stiamo sempre supponendo di essere all'equilibrio.

Bisogna evitare di confondersi tra tempo di *permanenza*, tempo di *attesa* e tempo di *servizio*. Il primo è la somma degli altri due.

Esaminiamo una coda $M/M/1$. Dobbiamo immaginare di essere un utente che arriva nel sistema: quanto tempo dovremo restare in esso, tra attesa e servizio? Un tempo aleatorio. Se capiamo la struttura di questo tempo aleatorio, possiamo calcolarne il valor medio.

Quando arriviamo nel sistema possono esserci già k utenti. Attenderemo il tempo del loro servizio più il tempo nel nostro servizio. In una coda $M/M/1$ gli utenti vengono serviti uno dopo l'altro. Indichiamo con $T^{(1)}, ..., T^{(k)}$ i tempi di servizio dei k utenti davanti a noi, e con $T^{(k+1)}$ il nostro tempo di servizio. Il nostro tempo di permanenza T_{perm} è pari a

$$T_{perm} = T^{(1)} + ... + T^{(k)} + T^{(k+1)}.$$

Insistiamo però sul fatto che questa uguaglianza è valida se, al nostro arrivo nel sistema, ci sono k persone davanti a noi. E', per così dire, un'uguaglianza *condizionata*. Comunque, in tal caso, vale

$$E\left[T_{perm}\right] = \frac{(k+1)}{\mu}$$

se μ è il tasso di servizio (abbiamo usato la linearità del valor medio).

Rimuoviamo ora la condizione che ci fossero esattamente k utenti davanti a noi. Il numero di utenti al nostro arrivo è aleatorio. Usiamo un analogo della formula di fattorizzazione, ma per i valori medi:

$$E\left[T_{perm}\right] = \sum_{k=0}^{\infty} E\left[T_{perm} | N = k\right] P\left(N = k\right)$$

dove abbiamo indicato con N il numero aleatorio di utenti davanti a noi al momento del nostro arrivo. Abbiamo calcolato sopra

$$E\left[T_{perm} | N = k\right] = \frac{(k+1)}{\mu}$$

e d'altra parte vale $P\left(N = k\right) = \pi_k$ all'equilibrio. Quindi

$$\begin{aligned} E\left[T_{perm}\right] &= \sum_{k=0}^{\infty} \frac{(k+1)}{\mu} \pi_k = \frac{1}{\mu} \sum_{k=0}^{\infty} k\pi_k + \frac{1}{\mu} \sum_{k=0}^{\infty} \pi_k \\ &= \frac{1}{\mu} \frac{\rho}{(1-\rho)} + \frac{1}{\mu}. \end{aligned}$$

Si provi a titolo di esercizio ad impostare il calcolo del tempo medio di permanenza per le code con più serventi.

5.4.7 Lancio di un dato al suono dell'orologio

Questa variante della teoria precedente si incontra alcune volte negli esempi. Supponiamo che, quando suona un orologio esponenziale che detta l'ordine di effettuare una transizione, estraiamo a sorte tra due o più possibilità ed andiamo nello stato estratto.

Ad esempio, supponiamo che il 10 per cento dell volte che una persona ha completato il servizio, si accorge di aver dimenticato di dire qualcosa e quindi si rimette in coda. Gli stati del sistema sono sempre gli interi $k \geq 0$, con transizioni tra primi vicini, ma quando starebbe per accadere la transizione $k \to k-1$ (servizio completato ed uscita dell'utente dal sistema), con probabilità $1/10$ l'utente si rimette immediatamente in coda. Quindi solo 9 volte su 10 si realizza effettivamente la transizione $k \to k-1$, mentre una volta su dieci gli utenti restano k.

In questi casi, a volte si riesce a ragionare semplicemente con buon senso, aggiustando i tassi di transizione come detta il buon senso. Altrimenti la regola è di moltiplicare il tasso per la probabilità corrisponendente. Supponiamo di essere nello stato A e che al suonare dell'orologio di tasso λ dobbiamo decidere se andare in B o C con probabilità p_B e p_C, $p_B + p_C = 1$. Allora è come se avessimo due frecce, una che porta da A a B con tasso $\lambda \cdot p_B$, l'altra che porta da A a C con tasso $\lambda \cdot p_C$. Lo schema

$$A \overset{\lambda}{\to} \begin{cases} \overset{p_B}{\to} & B \\ \overset{p_C}{\to} & C \end{cases}$$

equivale alle transizioni

$$C \overset{\lambda \cdot p_C}{\leftarrow} A \overset{\lambda \cdot p_B}{\to} B. \tag{5.4}$$

Anche se non lo dimostriamo, questo non stupisce, sia per la sua intuibilità, sia per l'analogia con un'altra regola già vista, che per così dire è il viceversa. Supponiamo infatti di partire dalle due transizioni (5.4). Introducuiamo il tempo di permanenza in A

$$T_A^{perm} = \min\left(T_{A,B}, T_{A,C}\right).$$

Esso ha tasso $\lambda_A^{perm} = \lambda_{A,B} + \lambda_{A,C}$. Possiamo interpretare la diramazione (5.4) come un singolo orologio T_A^{perm} seguito dalla scelta casuale tra B e C, operata secondo la regola spiegata in un paragrafo precedente: la probabilità di andare in B è $p_{A,B} = \frac{\lambda_{A,B}}{\lambda_{A,B}+\lambda_{A,C}}$ e così via. Ma allora il prodotto

$$\lambda_A^{perm} \cdot p_{A,B}$$

vale proprio $\lambda_{A,B}$. E' il viceversa di quanto detto sopra.

5.4.8 Il processo di Poisson

Il processo degli arrivi ad una coda di servizio, con tempo aleatorio tra un arrivo e l'altro di tipo esponenziale di parametro λ, è un processo di Poisson. Praticamente in tutti i nostri esempi di code il processo degli arrivi è di Poisson.

La definizione formale e generale di processo di Poisson è ovviamente più articolata, ma qui ci basta l'intuizione associata alle code.

La denominazione "Poisson" deriva dal fatto che, detto N_t il numero di utenti arrivati nell'intervallo $[0, t]$, la v.a. N_t è di Poisson di parametro λt. Questo fatto è stabilito da un noto e non banale teorema di legame tra v.a. esponenziali e v.a. di Poisson.

Il numero λ è detto *tasso* del processo. Esso ha varie interpretazioni. Da un lato, già sappiamo che λ è il reciproco del tempo medio tra un arrivo e l'altro. Una seconda interpretazione fondamentale è quella di *numero medio di arrivi nell'unità di tempo*: infatti

$$\lambda = \frac{E[N_t]}{t}.$$

Per questo è detto "tasso" del processo.

Due processi di Poisson indipendenti si combinano in un unico processo di Poisson di tasso pari alla somma dei tassi. Si rifletta sulla possibile giustificazione. Questo fatto si applica ad esempio alle code in cui si sa che arrivano utenti di due categorie (es. macchine a benzina e macchine diesel ad un distributore), con diversi tassi di arrivo.

5.4.9 Il processo in uscita da una coda

A volte si conosce la struttura di questo processo, a volte no. L'unica osservazione che facciamo è la seguente: se siamo all'equilibrio, il numero medio di uscite per unità di tempo è pari al numero medio di entrate. Il tasso, cioè, è lo stesso, sia in entrata che in uscita. Non dimostriamo questa importante proprietà ma invitiamo ad una riflessione intuitiva: se il numero delle uscite fosse inferiore (mediamente) alle entrate, in numero degli utenti nel sistema crescerebbe indefinitamente, quindi non sarmmo all'equilibrio; viceversa, se il numero delle uscite fosse superiore alle entrate, dopo un po' il sistema si svuoterebbe definitivamente, quindi anche in questo caso non saremmo all'equilibrio.

5.5 Esercizi

Esercizio 36 *Una stampante (del settore di produzione di una casa editrice) lavora a ciclo continuo. Ogni tanto però la qualità della stampa non è più ammissibile (alla lunga si sporcano alcune componenti), per cui si deve interrompere la stampa ed eseguire una complessa manutenzione. Si osserva che il deterioramento accade dopo un tempo di funzionamento T_f, esponenziale, di media 30 giorni, mentre la manutenzione richiede un tempo aleatorio esponenziale mediamente di un giorno.*

0) [Non è necessario risolvere questo punto, ma può aiutare] Descrivere questo sistema con un modello markoviano al fine di calcolare la probabilità a regime di trovare la stampante funzionante.

1) La casa editrice, occupandosi anche di quotidiani, non può sopportare il fermo della stampante per cui ne tiene una seconda pronta per essere utilizzata non appena la prima richiede manutenzione. Questa seconda stampante è però meno sofisticata, per cui si rompe dopo un tempo di lavoro aleatorio esponenziale di media 5 giorni e richiede un tempo esponenziale di media 1 giorno per ripartire. Appena la macchina principale viene riattivata, si interrompe l'uso della secondaria. Se la secondaria si rompe prima che la principale sia riattivata, la squadra di riparatori insiste solamente sulla principale, occupandosi della secondaria solo dopo aver fatto ripartire la principale. Descrivere il sistema con un modello markoviano.

2) Calcolare la probabilità a regime di trovarsi con entrambe le macchine ferme. Se si vuole che questa probabilità sia inferiore allo 0.001, bisogna che il tempo medio di sopravvivenza delle due macchine sia più alto: quali valori sono sufficienti? La disequazione finale nelle due variabili non deve essere risolta, ma solo scritta.

3) Supponiamo che la macchina secondaria venga sostituita da un modello nuovo di cui ancora non si conoscono bene le caratteristiche. Si osserva che, quando la macchina principale è ferma e la secondaria lavora, nel 90% dei casi viene aggiustata la macchina principale prima che si rompi anche la secondaria. Che tempo medio di funzionamento ha la nuova macchina secondaria? Non è sufficiente una risposta numerica su sola base intuitiva; casomai, si può interpretare intuitivamente il risultato ottenuto con metodi rigorosi.

4) Proseguendo il punto 1, supponiamo però che il reparto manutenzione abbia una seconda squadra, meno veloce di quella descritta sopra, che esegue una riparazione (della macchina principale o della secondaria, indifferentemente) con un tempo medio di 2 giorni. La squadra più lenta entra in gioco solo se quella più veloce è già impegnata. Inoltre, se quella veloce completa una riparazione mentre quella lenta sta lavorando, la lenta cede il lavoro alla veloce. Descrivere ora il sistema con un modello markoviano. Quando entrambe le macchine sono rotte, quanto si deve attendere mediamente prima che la stampa riparta?

Esercizio 37 *Un docente chiede ai propri studenti di realizzare un progetto come prova orale. Ogni studente può decidere in qualsiasi momento di richiedere al docente il comando del progetto da eseguire. Dopo un tempo T_{prog} lo studente consegna il progetto. Il docente impiega un tempo aleatorio esponenziale T_{corr} a correggere ciascun progetto.*

1) Semplifichiamo la situazione precedente ignorando la fase di richiesta e realizzazione del progetto. Quindi guardiamo le cose solo dal punto di vista del docente che riceve i progetti e li deve correggere. Supponiamo che il docente riceva i progetti finiti con intertempi aleatori esponenziali T_{inter} e che corregga un progetto per volta. Supponiamo che i valori medi di T_{corr} e T_{inter} siano di n gg. e 3 gg. rispettivamente. Come deve scegliere n (numero anche non intero), il docente, per avere al massimo un progetto da correggere per il 90% del tempo, a regime?

2) Continuiamo nell'ottica semplificata del punto 1. Il docente modifica la sua strategia e lavora contemporaneamente alla correzione di tutti i progetti ricevuti (ed inizia la correzione di ogni progetto appena lo riceve). Per ciascun progetto la correzione dura un tempo T_{corr} di media 5 gg. Calcolare il numero medio di compiti da correggere, a regime. Si ricorda lo sviluppo di Taylor della funzione esponenziale: $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

3) Ora prendiamo in considerazione il problema completo descritto all'inizio, con le richieste da parte degli studenti ed i loro tempi T_{prog} di esecuzione dei progetti stessi ed eliminiamo le specifiche descritte ai punti 1 e 2. Supponiamo che tra una richiesta e la successiva passi un tempo esponenziale di media 3 gg. Supponiamo che il tempo T_{prog} di esecuzione di ciascun studente sia esponenziale ed abbia media 7 gg. Supponiamo che il docente corregga i compiti secondo le modalità del punto 2. Descrivere ora il sistema, in cui si deve tener conto sia del numero di richieste attive sia del numero di compiti da correggere.

5.6 Processi nel continuo

5.6.1 Processi a tempo continuo

Si chiama *processo stocastico a tempo continuo* ogni famiglia $(X_t)_{t\geq 0}$ di variabili aleatorie indicizzata dal tempo $t \in [0, \infty)$. Un esempio: X_t = velocità del vento nella zona industriale di Livorno all'istante t. Con lo stesso nome si indicano anche i casi in cui il tempo varia su tutto l'asse reale: $(X_t)_{t\in\mathbb{R}}$; oppure su un intervallo $[0, T]$: $(X_t)_{t\in[0,T]}$ e così via per situazioni simili a queste.

Due visioni, come nel caso a tempo discreto:

- fissato t, X_t è una *variabile aleatoria*;
- se osserviamo una storia particolare, che accade in un esperimento (anche ideale), osserviamo una *realizzazione* (detta anche una *traiettoria* del processo).

Una realizzazione è una *funzione* di t, che varia nel continuo (prima, una realizzazione, cioè una serie storica, era una successione discreta). Se pensiamo alle realizzazioni come i possibili risultati degli esperimenti (esperimenti protratti nel tempo), vediamo un processo stocastico come una *funzione aleatoria*. Il caso sceglie una certa funzione tra tutte quelle possibili.

5.6.2 Più generale che tempo continuo?

E' però interessante in certe applicazioni considerare variabili aleatorie indicizzate da parametri più generali, ad esempio lo spazio, o lo spazio-tempo. Continueremo a chiamarli processi stocastici, oppure più specificamente *campi aleatori*. Ad esempio è un campo aleatorio una famiglia $\left(U_{(t,x,y,z)}\right)_{t,x,y,z\in\mathbb{R}}$ indicizzata da tempo e spazio. Un esempio concreto può essere la velocità dell'aria nel punto (x, y, z) dello spazio, all'istante t (nello studio delle predizioni atmosferiche si deve considerare questa grandezza aleatoria al variare di tempo e spazio).

Infine, per alcune applicazioni specifiche sono interessanti le famiglie di variabili aleatorie indicizzate da insiemi: $(X_A)_{A\subset\mathbb{R}^d}$. Ad esempio: X_A = quantità d'acqua piovana che cade nella regione A; oppure $N_{[a,b]}$ = numero di chiamate telefoniche che arriva ad una centrale nel periodo di tempo $[a, b]$.

In definitiva, volendo dare una definizione generale, un processo stocastico è una famiglia di variabili aleatorie indicizzata da un qualche insieme di parametri.

5.6.3 Il moto browniano

Vuole essere l'analogo della random walk, ma a tempo continuo. Nella RW si sommano incrementi indipendenti, a tempi discreti. Qui allora richiederemo che il processo sia somma di incrementi indipendenti, ma incrementi relativi a tempi qualsiasi.

Inoltre, nella RW, gli incrementi erano v.a. gaussiane. Qui si chiede lo stesso, sempre a tempi qualsiasi. Ecco le sue proprietà fondamentali, che lo *definiscono* (in modo non costruttivo come per la RW).

Definizione 56 *Un processo stocastico* $(B_t)_{t\geq 0}$ *si dice moto browniano (standard) se:*
i) $B_0 = 0$
ii) per ogni coppia di tempi $t \geq s \geq 0$*, l'incremento* $B_t - B_s$ *è una v.a.* $N(0, t-s)$
iii) gli incrementi $B_{t_n} - B_{t_{n-1}}$*, ... ,* $B_{t_1} - B_{t_0}$ *sono indipendenti, per ogni* $n \geq 1$ *e* $0 \leq t_0 < t_1 < ... < t_n$
iv) le traiettorie siano funzioni continue.

Possiamo visualizzare, simulare, delle traiettorie di un MB? Come sempre, le simulazioni impongono una discretizzazione (lo stesso vale anche se volessimo raffigurare la funzione $\sin t$). Fissiamo quindi una sequenza di tempi $t_1 < t_2 < ... < t_n$ rispetto a cui vogliamo i valori di una traiettoria. Per semplicità. prendiamo i tempi equispaziati:

$$t_k = \frac{k}{N}, \quad k = 1, ..., n.$$

Vale

$$B_{t_k} = \left(B_{\frac{1}{N}} - B_{\frac{0}{N}}\right) + \left(B_{\frac{2}{N}} - B_{\frac{1}{N}}\right) + \left(B_{\frac{3}{N}} - B_{\frac{2}{N}}\right) + ...$$

cioè il MB al generico tempo $t_k = \frac{k}{N}$ è somma di gaussiane indipendenti e con la stessa distribuzione, cioè è una RW! Basta quindi rappresentare una RW.

Si può essere quantitativamente più precisi. Supponiamo ad esempio di voler raffigurare una traiettoria browniana per $t \in [0, 5]$, usando 5000 punti. Prendiamo $N = 1000$, $k = 1, ..., 5000$. Ciascun incremento

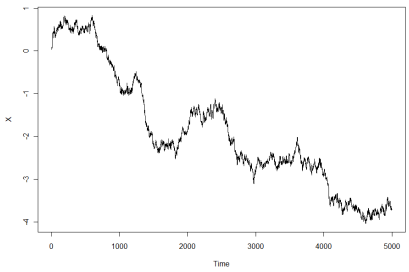
$$\left(B_{\frac{k+1}{N}} - B_{\frac{k}{N}}\right)$$

è una

$$N\left(0, \frac{k+1}{N} - \frac{k}{N}\right) = N\left(0, \frac{1}{N}\right)$$

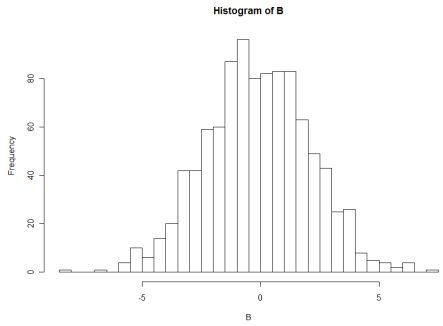
cioè una gaussiana di deviazione standard $\sqrt{\frac{1}{N}} = \sqrt{\frac{1}{1000}}$. Ecco allora i comandi:

```
L<-5000; W<-rnorm(L,0,sqrt(1/1000)); X<-1:L
X[1]<-0; X<-cumsum(W); ts.plot(X)
```



Facciamo una verifica incrociata. Il valore finale della simuolazione precedente è B_5, il MB al tempo 5. In base alle sue proprietà, dev'essere una v.a. $N(0, 5)$. Ripetiamo allora 1000 volte la simulazione precedente, vediamo un istogramma e calcoliamo media e deviazione. Ecco i risultati:

```
LL<- 1000; B <- 1:LL; L<-5000
for (i in 1:LL) {
W<-rnorm(L,0,sqrt(1/1000)); B[i]<-sum(W)}
hist(B,50)
```



White noise a tempo continuo?

Tempo discreto: RW = cumulato del WN, ovvero:

$$\text{white noise} = \text{incrementi della RW}.$$

Tempo continuo:

$$\text{white noise} = \text{derivata del MB}.$$

Però si può dimostrare che le realizzazioni del MB non sono derivabili (non esiste finito il limite del rapporto incrementale). L'idea viene dal fatto che

$$Var\left[\frac{B_t - B_s}{t-s}\right] = \frac{1}{(t-s)^2} Var\left[B_t - B_s\right] = \frac{t-s}{(t-s)^2} = \frac{1}{t-s}$$

che diverge per $t \to s$. Per dare senso al concetto di white noise, serve il concetto di derivata nel senso delle distribuzioni, che non sviluppiamo.

Moralmente, il white noise a tempo continuo è un processo W_t tale che: i) W_t è $N\left(0, \sigma^2\right)$, ii) $W_{t_1}, ..., W_{t_N}$ sono indipendenti. Ma al tempo stesso: $\sigma^2 = \infty$. La prima figura di questi appunti rende l'idea, pur essendo un'approssimazione discreta.

5.6.4 Dinamiche stocastiche

Idea comune a tanti modelli: la dinamica di un sistema è descritto da un'equazione alle differenze finite (come i processi ARIMA) o da un'equazione differenziale, però sono presenti delle incertezze:

- dati iniziali aleatori (non si conosce l'esatta configurazione iniziale)

- parametri aleatori (es. non si conosce il valore esatto di una certa concentrazione chimica)
- c'è rumore di fondo
- ci sono variabili che non sappiamo descrivere in modo deterministico.

Esempio 103 *evoluzione di una frattura in una lastra di vetro esposta al vento. L'influenza del vento viene descritta da un processo stocastico ξ_t (es. pressione all'istante t), dato a priori, con proprietà statistiche ragionevoli; l'ampiezza a_t della frattura risolve una certa equazione differenziale del tipo*

$$\frac{da_t}{dt} = f(a_t, \xi_t).$$

La classe più importante di dinamiche stocastiche nel continuo è quella delle *equazioni differenziali stocastiche*. Sono equazioni differenziali del tipo:

$$\frac{dX_t}{dt} = b(X_t, t) + \sigma(X_t, t)\frac{dB_t}{dt}$$

un analogo nel continuo delle equazioni ricorsive nel discreto, col white noise discreto W_n rimpiazzato da $\frac{dB_t}{dt}$ ("white noise" nel continuo).

La soluzione $(X_t)_{t\geq 0}$ è un processo stocastico. Possiamo simulare delle traiettorie di $(X_t)_{t\geq 0}$ discretizzando l'equazione. Nella sezione 5.7 vedremo che possiamo anche calcolare la densità di probabilità di X_t risolvendo un'equazione alle derivate parziali di tipo parabolico, detta *equazione di Fokker-Planck*.

In questa sezione introduttiva mostriamo alcuni esempi numerici.

Equilibrio stocastico

L'equazione differenziale

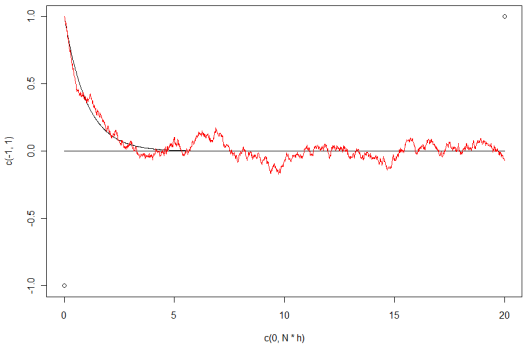
$$\frac{dX_t}{dt} = -X_t, \quad X_0 = x_0$$

è uno dei modelli più semplici di sistema con un breve transitorio seguito da una situazione di equilibrio. La soluzione è $X_t = e^{-t}x_0$, che tende esponenzialmente all'equilibrio $X = 0$.

Se aggiungiamo un white noise

$$\frac{dX_t}{dt} = -X_t + \sigma\frac{dB_t}{dt}, \quad X_0 = x_0$$

otteniamo ugualmente un sistema che rilassa all'equilibrio (la parte viscosa o dissipativa $-X_t$ continua ad agire) ma il sistema fluttua in modo casuale attorno al vecchio equilibrio deterministico. Possiamo considerare un *equilibrio statistico*, questa nuova situazione.



I due grafici sono stati ottenuti coi comandi

```
N<-2000; x0<-1; h<-0.01; s<-0.1
W<-rnorm(N,0,sqrt(h))
X.det<-1:N; X<-1:N; X.det[1]<-x0; X[1]<-x0
for (n in 1:(N-1)) {
X.det[n+1] <- X.det[n] - h*X.det[n]
X[n+1] <- X[n] - h*X[n] + s*W[n]
}
plot(c(0,N*h),c(-1,1)); lines(c(0,N*h),c(0,0), type="l")
T<-(1:N)*h; lines(T,X.det); lines(T,X,col="red")
```

Notiamo un dettaglio della discretizzazione. Abbiamo usato il metodo di Eulero esplicito (per semplicità). Indichiamo con $0 < t_1 < t_2 < ...$ gli istanti di discretizzazione, che prendiamo della forma $t_n = n \cdot h$ (h è il passo di discretizzazione), e scriviamo

$$\frac{X_{t_{n+1}} - X_{t_n}}{h} = -X_{t_n} + \sigma \frac{B_{t_{n+1}} - B_{t_n}}{h}$$

ovvero

$$X_{t_{n+1}} = X_{t_n} - h \cdot X_{t_n} + \sigma \left(B_{t_{n+1}} - B_{t_n} \right).$$

La v.a. $\left(B_{t_{n+1}} - B_{t_n} \right)$ è $N(0, h)$. Per questo abbiamo usato il comando `W<-rnorm(N,0,sqrt(h))`.

Un sistema a due stati

L'equazione differenziale

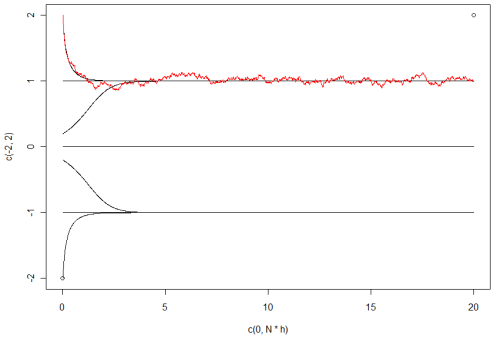
$$\frac{dX_t}{dt} = X_t - X_t^3, \quad X_0 = x_0$$

ha la proprietà che le soluzioni tendono ai due punti fissi $X = 1$ e $X = -1$ (salvo la soluzione uscente da $x_0 = 0$ che resta nulla). Ogni soluzione ha un destino preciso, $X = 1$ oppure $X = -1$.

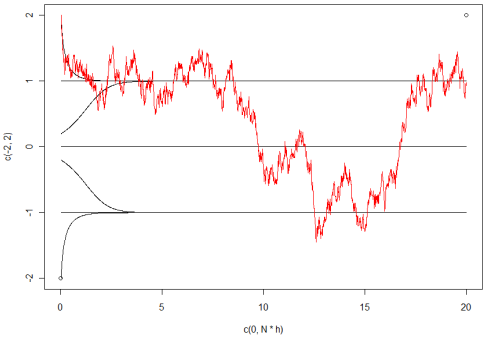
Aggiungendo un rumore

$$\frac{dX_t}{dt} = X_t - X_t^3 + \sigma \frac{dB_t}{dt}, \quad X_0 = x_0$$

si ottiene un sistema che tende provvisoriamente ad uno dei due punti fissi, però fluttua intorno ad esso, e quando una fluttuazione è sufficientemente grande, transisce all'altro punto fisso; e così via all'infinito. In fisica è chiamato l'effetto tunneling: ci sono due buche di potenziale ed il sistema, a causa delle fluttuazioni, ogni tanto cambia buca.



σ piccola



σ grande

5.6.5 Fit tramite un'equazione differenziale

Data una serie storica, si può tentare un suo fit tramite un'equazione differenziale, oltre che tramite modelli ARIMA ed altri visti in precedenza.

Se la serie storica ha proprietà gaussiane, si può tentare con un'equazione lineare del tipo

$$\frac{dX_t}{dt} = -\lambda X_t + \sigma \frac{dW_t}{dt}$$

che produce processi gaussiani. Ci aspettiamo però che il risultato sia simile a quello ottenuto con gli AR(1), se si pensa alla discretizzazione di Eulero:

$$X_{t_{n+1}} = X_{t_n} - h \cdot \lambda X_{t_n} + \sigma \left(B_{t_{n+1}} - B_{t_n} \right)$$

cioè

$$X_{t_{n+1}} = (1 - h \cdot \lambda) X_{t_n} + \sigma W_{t_{n+1}}$$

avendo posto $W_{t_{n+1}} = B_{t_{n+1}} - B_{t_n}$.

Fit nel caso non gaussiano

Assai più difficile è trovare un modello quando i dati non hanno una statistica gaussiana.

Dagli esempi precedenti sono emersi due elementi chiave di una serie storica: le sue proprietà statistiche, la sua struttura di autocorrelazione.

Se decidiamo di soprassedere su un fit preciso delle proprietà statistiche, si possono usare i metodi lineari precedenti anche nel caso non gaussiano, cercando di catturare al meglio la struttura di autocorrelazione.

Se abbiamo una serie storica stazionaria, con autocorrelazione relativamente semplice a memoria breve (come quella precedente, che va a zero dopo pochi valori), ed invece vogliamo catturare bene le proprietà statistiche non gaussiane, possiamo usare la teoria delle equazioni di Fokker-Planck, che descriveremo nella prossima sezione.

5.7 Equazioni differenziali stocastiche

Consideriamo l'equazione differenziale (o più precisamente il problema di Cauchy)

$$\frac{dx(t)}{dt} = b\left(t, x\left(t\right)\right), \quad x\left(0\right) = x_0.$$

Se tutti i termini, cioè $b\left(t,x\right)$ e x_0, sono deterministici, la soluzione sarà deterministica. Se invece o il dato iniziale x_0 oppure $b\left(t,x\right)$ è aleatoria, la soluzione sarà un processo stocastico. Il caso di un dato iniziale x_0 aleatorio è interessante ma piuttosto elementare, per cui ci concentriamo sul caso di $b\left(t,x\right)$ aleatoria. Consideriamo un caso molto particolare, in cui l'equazione ha la forma

$$\frac{dx(t)}{dt} = b\left(x\left(t\right)\right) + \sigma\left(x\left(t\right)\right)\xi\left(t\right), \quad x\left(0\right) = x_0.$$

dove b dipende solo da x e $\xi\left(t\right)$ è un processo stocastico assegnato, diciamo di media nulla e varianza unitaria, per cui σ misura la sua deviazione standard. Per essere ancora più specifici, supponiamo che $\xi\left(t\right)$ sia un white noise. Non diamo la definizione rigorosa di white noise, accontentandoci di descrivere alcuni risultati e svolgere alcune simulazioni. Notiamo solo che molto spesso un'equazione di tale tipo viene scritta nella forma

$$dx(t) = b\left(x\left(t\right)\right)dt + \sigma\left(x\left(t\right)\right)dB(t)$$

dove $B(t)$ è un moto browniano. Infatti, abbiamo già osservato altrove che il white noise è la derivata del moto browniano: $\xi\left(t\right) = \frac{dB(t)}{dt}$, uguaglianza che lega le due formulazioni dell'equazione. Il motivo che spinge a scrivere solamente $dB(t)$ e non $\frac{dB(t)}{dt}$ è che le traiettorie del moto browniano non sono derivabili, quindi in un certo senso l'equazione non è un'equazione differenziale ma solo un'equazione per incrementi $dx(t)$, $dB(t)$.

Accettando che con una certa fatica matematica si possa dar senso a tutte le espressioni ed equazioni dette sopra, vediamo i risultati. La soluzione $x(t)$ è un processo stocastico. Senza entrare nei dettagli, si può dimostrare che è un processo di Markov. Cosa molto importante, ad ogni istante $t > 0$ la v.a. $x(t)$ ha densità di probabilità $p\left(t,x\right)$ che soddisfa una certa equazione. Prima di scriverla sottolineiamo un fatto teorico non banale: anche se il dato inziale x_0 è deterministico, cioè il processo stocastico $x(t)$ al tempo $t_0 = 0$ vale identicamente x_0, con probabilità uno, quindi non ha densità, tuttavia ha densità ad ogni istante $t > 0$. Si parla infatti di processo di *diffusione*. E' come se all'istante $t_0 = 0$ ci fossero infinite particelle tutte concentrate nel punto x_0, che poi immediatamente si muovono in diverse direzioni con traiettorie erratiche (tipo moto browniano), per cui ad ogni successivo istante $t > 0$ troviamo le particelle distribuite un po' ovunque (non in modo uniforme), distribuite secondo una densità $p\left(t,x\right)$.

La funzione $p(t,x)$ soddisfa l'equazione alle derivate parziali

$$\frac{\partial p(t,x)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)\,p(t,x)\right) - \frac{\partial}{\partial x}\left(b(x)\,p(t,x)\right)$$

detta *equazione di Fokker-Planck*. Abbiamo scritto tutto nel caso di x uni-dimensionale, ma la teoria si generalizza senza difficoltà. Ci sono varianti di questa teoria per dinamiche stocastiche markoviane di vario tipo, non necessariamente descritte da equazioni differenziali stocastiche del tipo enunciato sopra. In certi casi l'equazione che corrisponde a Fokker-Planck si chiama Master Equation.

A volte interessa la soluzione $x(t)$ che esce dal dato iniziale deterministico x_0, ma altre volte può essere più interessante ragionare su soluzioni $x(t)$ non legate a dati iniziali specifici, ma aventi la proprietà di essere *stazionarie*: in particolare, aventi densità $p(t,x)$ indipendente da t. In tal caso la densità $p(x)$ risolve l'equazione

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}\left(\sigma^2(x)\,p(x)\right) - \frac{d}{dx}\left(b(x)\,p(x)\right) = 0$$

che è decisamente più semplice della precedente. Nel caso uni-dimensionale si può impostare una risoluzione generale. Infatti scriviamola nella forma

$$\frac{d}{dx}\left(\frac{1}{2}\frac{d}{dx}\left(\sigma^2(x)\,p(x)\right) - b(x)\,p(x)\right) = 0$$

da cui ricaviamo

$$\frac{1}{2}\frac{d}{dx}\left(\sigma^2(x)\,p(x)\right) - b(x)\,p(x) = C_1$$

per una opportuna costante C_1. Prendiamo il caso particolare $C_1 = 0$ e supponiamo per un momento $\sigma^2(x) > 0$.Col metodo delle variabili separate si ottiene in pochi passi

$$\begin{aligned}
\frac{1}{2}\frac{d}{dx}\left(\sigma^2(x)\,p(x)\right) &= b(x)\,p(x)\\
\frac{d\left(\sigma^2(x)\,p(x)\right)}{\sigma^2(x)\,p(x)} &= 2\frac{b(x)}{\sigma^2(x)}dx\\
\log\left(\sigma^2(x)\,p(x)\right) &= 2\int\frac{b(x)}{\sigma^2(x)}dx\\
p(x) &= \frac{1}{Z\cdot\sigma^2(x)}\exp\left(2\int_0^x\frac{b(t)}{\sigma^2(t)}dt\right)
\end{aligned}$$

per un'opportuna costante $Z > 0$. Riassumendo, data $b(x)$, se risulta

$$Z := \int_{-\infty}^{+\infty}\frac{1}{\sigma^2(x)}\exp\left(2\int_0^x\frac{b(t)}{\sigma^2(t)}dt\right)dx < \infty$$

allora

$$p(x) = \frac{1}{Z\cdot\sigma^2(x)}\exp\left(2\int_0^x\frac{b(t)}{\sigma^2(t)}dt\right)$$

è una *soluzione stazionaria dell'equazione di Fokker-Planck*. E' la densità di probabilità, ad ogni istante di tempo, di un processo stocastico stazionario che risolve l'equazione differenziale scritta sopra. Si può inoltre dimostrare che la densità $p(x)$, se esiste, è l'unica soluzione del problema precedente.

Le ultime elaborazioni dei calcoli valgono sotto l'ipotesi $\sigma^2(x) > 0$. Però ragionando caso per caso in genere si riescono ad estendere i risultati anche quando $\sigma^2(x) = 0$ per certe x, ad es. $\sigma^2(x) = 0$ per $x \leq 0$. Naturalmente si intenderà che la formula scritta sopra vale nell'intervallo delle x in cui $\sigma^2(x) > 0$.

5.7.1 Applicazione diretta

Un primo modo di applicare questa teoria è quello diretto: se sappiamo già che il processo stocastico $x(t)$ da noi esaminato soddisfa l'equazione stocastica scritta sopra, allora possiamo simularne varie cartteristiche tramite gli strumenti precedenti. Ad esempio, supponiamo di studiare una macromolecola immersa in un fluido fermo insieme a tante altre macromolecole e supponiamo di riassumere la fisica del fenomeno nell'equazione

$$dx(t) = -\lambda x(t)dt + \sigma dB(t)$$

pensando che ad ogni istante la macromolecola subisce uno spostamento $dx(t)$ dato da due componenti: lo spostamento $\sigma dB(t)$ dovuto agli urti con le macromolecole circostanti, meno il termine $\lambda x(t)dt$ che si fa carico genericamente dell'attrito o dissipazione dovuta all'interazione col fluido. Quindi l'equazione è già in nostro possesso. Possiamo scrivere l'equazione di Fokker-Planck

$$\frac{\partial p(t,x)}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 p(t,x)}{\partial x^2} + \frac{\partial}{\partial x}\left(\lambda x p(t,x)\right)$$

e tentare di simularla con programmi appositi per equazioni alle derivate parziali (in questo caso particolarissimo si può anche risolvere esplicitamente). per lo meno, possiamo affermare che

$$p(x) = Z^{-1}\exp\left(-\frac{2\lambda}{\sigma^2}x^2\right)$$

è una soluzione stazionaria di Fokker-Planck e quindi è una densità invariante. Si noti che è una densità gaussiana di media zero e varianza $\frac{\sigma^2}{2\lambda}$.

Inoltre possiamo simulare le traiettorie dell'equazione del moto, ad esempio un po' rozzamente col metodo di Eulero esplicito:

$$x(t+\triangle t) = (1-\lambda\triangle t)\,x(t) + \sigma\left[B(t+\triangle t) + B(t)\right]$$

generando gli incrementi $[B(t+\triangle t) + B(t)]$ come numeri gaussiani indipendenti di media zero e varianza $\triangle t$ (deviazione standard $\sqrt{\triangle t}$):

$$\begin{aligned}\texttt{x}[\texttt{k}+1] = (1 - \texttt{lambda} * \texttt{h}) * \texttt{x}[\texttt{k}] \\ +\texttt{sqrt}(\texttt{h}) * \texttt{sigma} * \texttt{rnorm}(1,0,1)\end{aligned}$$

dove abbiamo indicato con `h` il numero $\triangle t$.

5.7.2 Identificazione sperimentale dei parametri

Proseguiamo l'esempio precedente ma supponendo che sia noto solo il modello

$$dx(t) = -\lambda x(t)dt + \sigma dB(t)$$

nella sua struttura, non numericamente i parametri λ e σ. Supponiamo però di conoscere una realizzazione sperimentale (una serie storica)

$$x_1, x_2, ...$$

Allora possiamo da questa stimare λ e σ.

Da un lato, per certi scopi, è sufficiente conoscere il rapporto $\frac{\sigma^2}{2\lambda}$, cioè la varianza della distribuzione stazionaria. Allora, invocando un teorema ergodico, il numero

$$\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2$$

è uno stimatore di $\frac{\sigma^2}{2\lambda}$. Qui come sempre $\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$.

Se però volessimo svolgere simulazioni dell'equazione del moto, dovremmo conoscere separatamente λ e σ. Si può allora osservare che λ è legato al tempo di rilassamento (decadimento) all'equilibrio: in assenza del termine $\sigma dB(t)$ l'equazione del moto sarebbe

$$\frac{dx(t)}{dt} = -\lambda x(t)$$

la cui soluzione è $x(t) = x_0 e^{-\lambda t}$, che impiega un tempo dell'ordine di $\frac{1}{\lambda}$ a diventare molto piccola: $x(\frac{1}{\lambda}) = x_0 e^{-1} = 0.36 \cdot x_0$. Si capisce che si sta parlando dell'*ordine di grandezza* del tempo di rilassamento, altrimenti bisognerebbe stabilire a priori cosa si intende per molto piccolo. A volte si preferisce parlare del tempo di dimezzamento.

Stabilito che $\frac{1}{\lambda}$ è l'ordine di grandezza del tempo di rilassamento, si può calcololare la funzione di autocorrelazione della serie $x_1, x_2, ...$ e calcolare da essa un numero che corrisponda al tempo di rilassamento della serie storica (tempo di scorrelazione). Da qui si può stimare λ. Poi, stimato λ, si stima σ dal momento che $\frac{\sigma^2}{2\lambda}$ è stato pure stimato.

Questo procedimento è un po' vago, nel senso che non prescrive esattamente come calcolare l'analogo di $\frac{1}{\lambda}$ sulla serie storica. Però ha il pregio di far capire l'idea. Per un'identificazione più precisa di λ si può usare il seguente fatto, che non ci mettiamo a giustificare nei dettagli:

$$\begin{aligned} Cov\,(x_t, x_0) &= Cov\left(x_0 e^{-\lambda t}, x_0\right) + Cov\left(\int_0^t e^{-\lambda(t-s)} dB(s), x_0\right) \\ &= e^{-\lambda t}\frac{\sigma^2}{2\lambda} + 0 \end{aligned}$$

quindi, detta ρ l'autocorrelazione,

$$\rho\,(x_t, x_0) = e^{-\lambda t}.$$

In altre parole, la funzione $e^{-\lambda t}$ è proprio uguale all'autocorrelazione (in questo semplicissimo modello lineare). Da qui, ad esempio calcolando l'autocorrelazione sperimentale al tempo $t = 1$, si stima λ, o grossolanamente ad occhio, o con la formula

$$\lambda = -\lim_{t\to\infty} \frac{\log \rho\left(x_t, x_0\right)}{t}.$$

Si noti che bisogna fare molta attenzione alla scala temporale vera nel calcolo dell'autocorrelazione sperimentale.

Un'ultima osservazione: se l'equazione differenziale fosse stata più complessa (non lineare), non avremmo potuto calcolare $\rho\left(x_t, x_0\right)$. Allora è sufficiente simulare con R l'equazione differenziale calcolando l'acf e cercando (anche solo per tentativi) dei valori dei parametri che forniscono una acf simile a quella sperimentale.

5.7.3 Applicazione inversa

Supponiamo di avere una serie storica sperimentale $x_1, x_2, ...$, stazionaria, e di volerla descrivere tramite un modello dinamico del tipo visto sopra:

$$dx(t) = b\left(x\left(t\right)\right) dt + \sigma\left(x\left(t\right)\right) dB(t).$$

Dalle serie storiche possiamo ricavare due classi di informazioni:

- l'autocorrelazione sperimentale (acf)
- la funzione di distribuzione cumulativa empirica (ecdf).

Chiamiamo $F(x)$ una funzione che corrisponda alla ecdf: ad esempio, dopo aver esaminato la ecdf possiamo aver scelto un modello Weibull, gaussiano ecc., che chiamiamo $F(x)$. Sia $f(x)$ la densità corrispondente: $f(x) = F'(x)$.

Poniamo

$$\frac{1}{Z\sigma^2\left(x\right)} \exp\left(2\int_0^x \frac{b\left(t\right)}{\sigma^2\left(t\right)} dt\right) = f(x).$$

Nel senso: f è assegnata, b e σ sono incognite. Risparmiando i calcoli, che si possono ricostruire con un po' di pazienza, si trova la seguente equazione, nelle incognite b e σ^2:

$$2b\left(x\right) = \frac{d}{dx}\sigma^2\left(x\right) + \gamma\left(x\right)\sigma^2\left(x\right).$$

dove abbiamo posto

$$\gamma\left(x\right) = \frac{d}{dx}\log f\left(x\right).$$

Osservazione 73 *Ecco i calcoli:*

$$\begin{aligned}
\exp\left(2\int_0^x \frac{b(t)}{\sigma^2(t)}dt\right) &= Z\sigma^2(x) f(x) \\
2\int_0^x \frac{b(t)}{\sigma^2(t)}dt &= \log\left(Z\sigma^2(x) f(x)\right) \\
2\frac{b(x)}{\sigma^2(x)} &= \frac{\frac{d}{dx}\left(\sigma^2(x) f(x)\right)}{\sigma^2(x) f(x)} \\
\frac{f(x)\frac{d}{dx}\sigma^2(x) + \sigma^2(x)\frac{d}{dx}f(x)}{f(x)} &= 2b(x) \\
\frac{d}{dx}\sigma^2(x) + \sigma^2(x)\frac{d}{dx}\log f(x) &= 2b(x).
\end{aligned}$$

Si noti che possiamo giocare su molti gradi di libertà. Quindi la prima scelta che viene in mente è

$$\sigma^2 = \text{ costante}$$

da cui $\frac{d}{dx}\sigma^2(x) = 0$, $2b(x) = \gamma(x)\sigma^2$, quindi

$$b(x) = \frac{\sigma^2}{2}\frac{d}{dx}\log f(x).$$

Esempio 104 *Se ad esempio $f(x)$ è una gaussiana di media nulla e deviazione θ,*

$$f(x) = C\exp\left(-\frac{x^2}{2\theta^2}\right)$$

allora

$$\frac{d}{dx}\log f(x) = -\frac{x}{\theta^2}$$

quindi

$$b(x) = -\frac{\sigma^2}{2\theta^2}x.$$

L'equazione trovata è

$$dx(t) = -\frac{\sigma^2}{2\theta^2}x(t)dt + \sigma dB(t).$$

Chiamato λ il rapporto $\frac{\sigma^2}{2\theta^2}$, si trova il modello del paragrafo precedente. Controlliamo la compatibilità dei risultati. Nel paragrafo precedente avevamo detto che la varianza era $\frac{\sigma^2}{2\lambda}$, che per $\lambda = \frac{\sigma^2}{2\theta^2}$ diventa $\frac{\sigma^2}{2\frac{\sigma^2}{2\theta^2}} = \theta^2$. Questa è la varianza qui ipotizzata.

Osservazione 74 *Abbiamo un grado di libertà in più: possiamo scegliere la costante σ a piacere. La stimiamo in modo da avere il tempo di decadimento del modello teorico pari a quello della acf.*

Esaminiamo un caso più difficile. Supponiamo che dalle osservazioni sperimentali emerga come plausibile o opportuna una densità $f(x)$ di tipo esponenziale:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{per} \quad x \geq 0 \\ 0 & \text{per} \quad x < 0 \end{cases} .$$

Se tentiamo di seguire la strada precedente dobbiamo calcolare $\log f(x)$ che non ha senso per $x < 0$. Questo fa venire in mente un altro problema: la scelta σ costante non può funzionare in quanto l'equazione differenziale produrrebbe inevitabilmente delle traiettorie ogni tanto negative, e questo sarebbe incompatibile con una densità f concentrata sui numeri positivi. Quindi la strategia di σ costante non va più bene. Questo è il motivo per cui non abbiamo scelto σ costante sin dall'inizio, anche se questo avrebbe semplificato molti passaggi.

Prendiamo allora una funzione $\sigma(x)$ che tende a zero per $x \to 0^+$, in modo che l'effetto del moto browniano svanisca quando ci si avviciana all'origine. Per ora prendiamo

$$\sigma(x) = \sigma \cdot x^{\alpha}.$$

Allora considerando l'equazione solo per $x > 0$, dove

$$\gamma(x) = -\lambda$$

calcoliamo $2b(x) = \frac{d}{dx}\sigma^2(x) + \gamma(x)\sigma^2(x)$, ovvero

$$b(x) = \sigma^2 \alpha x^{2\alpha - 1} - \frac{\lambda}{2}\sigma^2 x^{2\alpha}.$$

Prendiamo ad esempio $\alpha = \frac{1}{2}$:

$$\begin{aligned} \sigma(x) &= \sigma\sqrt{x} \\ b(x) &= \frac{\sigma^2}{2} - \frac{\lambda\sigma^2}{2}x. \end{aligned}$$

L'equazione è

$$dx(t) = \left(\frac{\sigma^2}{2} - \frac{\lambda\sigma^2}{2}x(t)\right)dt + \sigma\sqrt{x(t)}dB(t).$$

E' interessante simularla con `R`. Lo schema di Eulero esplicito per questa equazione è

$$\begin{aligned} \texttt{x[k+1]} = \texttt{x[k]} &+ 0.5 * \texttt{sigma}\hat{}\,2 * \texttt{h} \\ &- 0.5 * \texttt{lambda} * \texttt{sigma}\hat{}\,2 * \texttt{x[k]} * \texttt{h} \\ &+ \texttt{sigma} * \texttt{sqrt(h} * \texttt{x[k])} * \texttt{rnorm(1,0,1)}. \end{aligned}$$

Purtroppo c'è un problema numerico: anche se in teoria la soluzione sta sempre nel semiasse positivo, discretizzando può capitare che un incremento del moto browniano la portino nel semiasse negativo ed in tal caso il termine $\texttt{sqrt(h} * \texttt{x[k])}$ non sarebbe ben definito. Il modo giusto di risolvere questo problema consisterebbe nello scrivere un codice più accurato di Eulero esplicito, a passo variabile, che abbrevia il passo se si supera lo zero. Per i nostri

scopi è troppo complesso. Usiamo uno stratagemma: se `x[k + 1]` diventa negativo, lo poniamo uguale a zero.

Il parametro σ è a scelta. Vediamo se lo si può usare per avere una acf simulata somigliante alla ecf empirica. Si può agire ad occhio per tentativi oppure calcolare sulla ecf empirica il numero

$$\lambda = -\lim_{t\to\infty} \frac{\log \rho_{emp}(x_t, x_0)}{t}$$

che rappresenta il tasso di decadimento a zero della ecf empirica, poi cercare σ in modo che la stessa quantità sulla serie simulata sia uguale. Attenzione sempre alla scala dei tempi.

5.8 Soluzione degli esercizi

Soluzione esercizio 33. i) In 4 passi ci sono i seguent modi di andare da 5 a 4:

$$\begin{aligned}
&5 \to 5 \to 5 \to 3 \to 4\\
&5 \to 5 \to 3 \to 4 \to 4\\
&5 \to 3 \to 4 \to 3 \to 4\\
&5 \to 3 \to 4 \to 4 \to 4
\end{aligned}$$

per cui la probabilità richiesta vale

$$\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot\frac{1}{3}\cdot 1\cdot\frac{1}{2} + \frac{1}{3}\cdot 1\cdot\frac{1}{2}\cdot 1 + \frac{1}{3}\cdot 1\cdot\frac{1}{2}\cdot\frac{1}{2} = 0.342\,59.$$

ii) Gli stati 1 e 2 comunicano tra loro e con nessun altro, quindi formano una classe chiusa irriducibile. Lo stesso vale per 3 e 4. Lo stato 5 porta in 1 (ed in 3), da cui non può tornare, quindi è transitorio.

iii) Nella classe $\{1, 2\}$ la matrice è bistocastica, quindi la misura invariante è $(\pi_1, \pi_2) = \left(\frac{1}{2}, \frac{1}{2}\right)$. Nella classe $\{3, 4\}$ il bilancio di flusso in 3 ci dà l'equazione $\frac{1}{2}\pi_4 = \pi_3$ a cui dobbiamo unire la $\pi_3 + \pi_4 = 1$. Sostituendo al prima nella seconda troviamo $\frac{1}{2}\pi_4 + \pi_4 = 1$ da cui $\pi_4 = \frac{2}{3}$, e quindi $\pi_3 = \frac{1}{3}$. Le misure invarianti del sistema complessivo hanno quindi la forma

$$\alpha\left(\frac{1}{2}, \frac{1}{2}, 0, 0, 0\right) + (1-\alpha)\left(0, 0, \frac{1}{3}, \frac{2}{3}, 0\right) = \left(\frac{\alpha}{2}, \frac{\alpha}{2}, \frac{1-\alpha}{3}, \frac{2\cdot 1-\alpha}{3}, 0\right)$$

al variare di $\alpha \in [0, 1]$.

Soluzione esercizio 34. i) Detti $1 = (A, A)$, $2 = (A, B)$, $3 = (B, B)$, $4 = (B, A)$ i quattro stati, vale ad esempio

$$\begin{aligned}
&P((A, A) \to (A, A)) = P(\text{secondo agente non cambia}) = 1/2\\
&P((A, A) \to (A, B)) = P(\text{secondo agente cambia}) = 1/2\\
&P((A, A) \to (B, B)) = 0\\
&P((A, A) \to (B, A)) = 0
\end{aligned}$$

(per il fatto che quando siamo in (A, A) il primo agente resta in A sicuramente) e così via. La matrice di transizione è

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

(si disegni anche il grafo). E' un'unica classe irriducibile, quindi c'è un'unica misura invariante. La matrice è bistocastica, quindi la misura invariante è uniforme: $\pi = (1/4, 1/4, 1/4, 1/4)$. La matrice è regolare (ad esempio perché è irriducibile e con un elemento diagonale positivo), quindi c'è convergenza all'equilibrio. Non vale il bilancio dettagliato (es. $p_{41}\frac{1}{2} \neq p_{14}\frac{1}{2}$).

Il guadagno medio all'equilibrio del primo agente è

$$\pi_{(A,A)} \cdot 10 + \pi_{(B,B)} \cdot 10 = 5.$$

Per simmetria questo è anche il guadagno medio del secondo agente.

ii) Ora vale, ad esempio,

$$\begin{aligned} P\left((A,A) \to (A,A)\right) &= 0 \\ P\left((A,A) \to (A,B)\right) &= 0 \\ P\left((A,A) \to (B,B)\right) &= 1 \\ P\left((A,A) \to (B,A)\right) &= 0 \end{aligned}$$

(per il fatto che quando siamo in (A, A) entrambi gli agenti cambiano sicuramente) e così via, facendo attenzione che ora la situazione non è più simmetrica tra i due agenti. La matrice di transizione è

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

(si disegni anche il grafo). Gli stati (A, A) e (B, B) formano una classe irriducibile con matrice ridotta $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ bistocastica e quindi misura invariante uniforme, ma non regolare, in quanto le sue potenze sono $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ stessa oppure $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Gli altri sue stati sono transitori. L'unica misura invariante è pertanto $\pi = (1/2, 0, 1/2, 0)$. Il guadagno medio del primo agente è

$$\pi_{(A,A)} \cdot 10 + \pi_{(B,B)} \cdot 10 = 10$$

mentre quello del secondo agente è nullo.

iii) Se n è dispari, la probabilità è 1, altrimenti è zero. Si vede quindi che $p_{3,1}^{(n)}$ non tende a π_1, coerentemente con la scoperta fatta sopra della non regolarità.

Se si parte da (A, B), è indispensabile connettersi a (B, B) in modo da avere poi un numero dispari di passi davanti, altrimenti il contributo è nullo. Quindi va bene andare

subito in (B,B) (poi nei restanti 9 passi si arriva in (A,B)), tragitto che ha probabilità 1/2. Oppure effettuare

$$(A,B) \to (A,B) \to (A,B)$$

e poi andare in (B,B), tragitto che ha probabilità $\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}$. Oppure effettuare

$$(A,B) \to (A,B) \to (A,B) \to (A,B) \to (A,B)$$

e poi andare in (B,B), tragitto che ha probabilità $\left(\frac{1}{2}\right)^5$. E così via, quindi la probabilità richiesta è

$$\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^7 = 0.664\,06.$$

Se si vuole la probabilità in 9 passi, con ragionamenti analoghi si trova

$$\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^8 = 0.332\,03.$$

Si intuisce che non c'è convergenza all'equilibrio. Rigorosamente, vale

$$p_{21}^{(9)} = \frac{1}{2}p_{21}^{(8)}$$

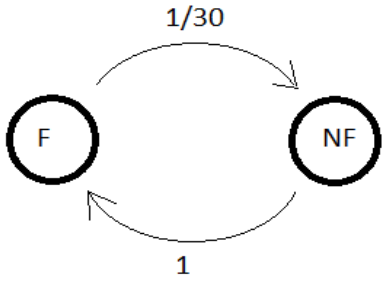
e si intuisce che in generale valga

$$p_{21}^{(2n+1)} = \frac{1}{2}p_{21}^{(2n)}$$

per cui non può accadere che $p_{21}^{(n)} \to \frac{1}{2}$.

Soluzione esercizio 36

0) Due stati, F = funziona, NF = non funziona



Il bilancio di flusso è

$$\frac{1}{30}\pi_F = \pi_{NF}$$

che inserito nella condizione $\pi_F + \pi_{NF} = 1$ produce $\pi_F + \frac{1}{30}\pi_F = 1$, $\frac{31}{30}\pi_F = 1$, $\pi_F = \frac{30}{31}$, $\pi_{NF} = \frac{1}{31}$. La probabilità a regime di trovare la stampante funzionante è $\pi_F = \frac{30}{31}$.

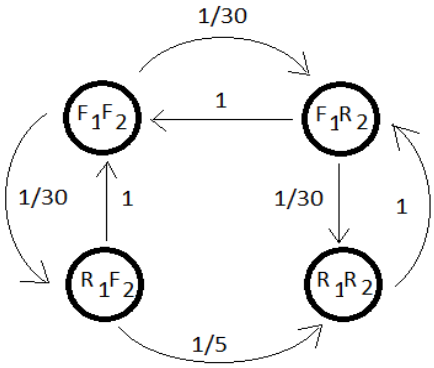
1) Stati: F_1F_2 = entrambe funzionano, F_1R_2 = la macchina principale (detta $M1$) funziona, la secondaria no, R_1F_2 = la macchina principale non funziona ma la secondaria sì, R_1R_2 = entrambe non funzionano. Transizioni e tassi:

- Nello stato F_1F_2 è attivo solo l'orologio della macchina principale (l'altra è funzionante ma ferma); quindi c'è solo la transizione $F_1F_2 \to R_1F_2$ ed il tasso è $\lambda_1 = \frac{1}{30}$ (il tempo è misurato in giorni).

- Nello stato R_1F_2 sono attivi due orologi, uno della riparazione di $M1$, l'altro di rottura di $M2$. Quindi $R_1F_2 \to F_1F_2$ con tasso $\mu = 1$, $R_1F_2 \to R_1R_2$ con tasso $\lambda_2 = \frac{1}{5}$.

- Nello stato R_1R_2 è attivo solo l'orologio di riparazione di $M1$, quindi $R_1R_2 \to F_1R_2$ con tasso $\mu = 1$.

- Nello stato F_1R_2 sono attivi due orologi, uno della riparazione di $M2$, l'altro di rottura di $M1$. Quindi $F_1R_2 \to F_1F_2$ con tasso $\mu = 1$, $F_1R_2 \to R_1R_2$ con tasso $\lambda_1 = \frac{1}{30}$.



2) Le equazioni del bilancio di flusso sono (ponendo $A = F_1F_2$, $B = R_1F_2$, $C = R_1R_2$, $D = F_1R_2$):

$$\begin{aligned} \lambda_1\pi_A &= \pi_B + \pi_D \\ \pi_B\,(1+\lambda_2) &= \lambda_1\pi_A \\ \pi_C &= \lambda_2\pi_B + \lambda_1\pi_D \\ \pi_D\,(1+\lambda_1) &= \pi_C \\ 1 &= \pi_A + \pi_B + \pi_C + \pi_D \end{aligned}$$

da cui

$$\begin{aligned} \lambda_1\pi_A &= \pi_B + \pi_D \\ \pi_B\,(1+\lambda_2) &= \lambda_1\pi_A \\ \pi_D\,(1+\lambda_1) &= \lambda_2\pi_B + \lambda_1\pi_D \text{ ovvero } \pi_D = \lambda_2\pi_B \\ 1 &= \pi_A + \pi_B + \pi_D\,(1+\lambda_1) + \pi_D \end{aligned}$$

da cui

$$1 = \frac{\pi_B\,(1+\lambda_2)}{\lambda_1} + \pi_B + \lambda_2\pi_B\,(2+\lambda_1)$$

da cui

$$\pi_B = \frac{1}{\frac{1+\lambda_2}{\lambda_1} + 1 + \lambda_2 (2 + \lambda_1)} = \frac{\lambda_1}{1 + \lambda_1 + \lambda_2 + \lambda_1 \lambda_2 (2 + \lambda_1)}$$

da cui

$$\pi_D = \frac{\lambda_1 \lambda_2}{1 + \lambda_1 + \lambda_2 + \lambda_1 \lambda_2 (2 + \lambda_1)}$$

quindi infine

$$\pi_C = \frac{\lambda_1 \lambda_2 (1 + \lambda_1)}{1 + \lambda_1 + \lambda_2 + \lambda_1 \lambda_2 (2 + \lambda_1)}.$$

La condizione è

$$\frac{\lambda_1 \lambda_2 (1 + \lambda_1)}{1 + \lambda_1 + \lambda_2 + \lambda_1 \lambda_2 (2 + \lambda_1)} \leq 0.001.$$

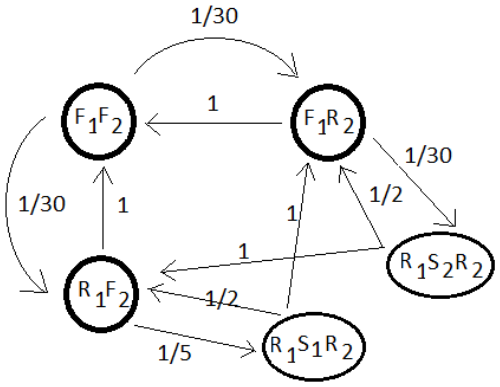
Il valore di π_C con i nostri dati è invece $5.524\,9 \times 10^{-3}$.

3) Quando ci troviamo nello stato R_1F_2 sono attivi due orologi, uno della riparazione di $M1$, l'altro di rottura di $M2$. Quindi $R_1F_2 \to F_1F_2$ con tasso $\mu = 1$, $R_1F_2 \to R_1R_2$ con tasso $\lambda_2 = \frac{1}{M}$, dove M è il tempo medio di funzionamento richiesto. Sappiamo che

$$\frac{\mu}{\mu + \lambda_2} = 0.9$$

ovvero $\frac{1}{1+\lambda_2} = 0.9$, $1 + \lambda_2 = \frac{1}{0.9}$, $\lambda_2 = \frac{1}{0.9} - 1 = \frac{0.1}{0.9}$, $M = \frac{0.9}{0.1} = 9$.

4) Modifichiamo dove necessario gli stati precedenti. Nello stato F_1F_2 le squadre di riparazione sono ferme, quindi non ci sono ambiguità. Nello stato R_1F_2 si arriva solo da F_1F_2, quindi lavora la squadra veloce: non ci sono ambiguità e continua ad essere $\mu = 1$ il tasso di ritorno a F_1F_2. Nello stato F_1R_2 lavora sicuramente la squadra veloce: ci si arriva da R_1R_2 ed in ogni caso è la squadra veloce che prende il lavoro da svolgere. Invece nello stato R_1R_2 non possiamo sapere quale squadra lavora M_1, quindi dobbiamo sdoppiare R_1R_2 in $R_1S_1R_2$ e $R_1S_2R_2$ a seconda che la macchina M_1 venga riparata dalla squadra S_1 oppure S_2. Da R_1F_2 si va per forza in $R_1S_1R_2$ (con tasso λ_2). Da F_1R_2 si va per forza in $R_1S_2R_2$ (con tasso λ_1). Da $R_1S_1R_2$ si va in F_1R_2 con tasso $\mu = 1$ ed in R_1F_2 con tasso $\mu_2 = 0.5$. Infine, da $R_1S_2R_2$ si va in F_1R_2 con tasso $\mu_2 = 0.5$ ed in R_1F_2 con tasso $\mu = 1$.



Quando ci si trova in R_1R_2 si deve attendere un tempo esponenziale di parametro 1.5 (per il teorema sul minimo di v.a. esponenziali). Quindi si deve attendere mediamente un tempo pari a $\frac{1}{1.5} = .666\,67$ giorni.

Soluzione esercizio 37

1) Consideriamo il numero di progetti pervenuti da correggere. E' una catena di nascita e morte a tempo continuo, con tasso di crescita $\lambda = \frac{1}{E[T_{inter}]} = \frac{1}{3}\ gg^{-1}$ e tasso di descrecita $\mu = \frac{1}{E[T_{corr}]} = \frac{1}{n}\ gg^{-1}$. La catena raggiunge il regime stazionario se $n < 3$. Si calcolano le probabilità invarianti

$$\pi_k = (1-\rho)\,\rho^k$$

dove $\rho = \frac{n}{3}$. Vogliamo che, a regime, con probabilità 0.9 il numero di progetti da correggere sia ≤ 2. Quindi deve essere

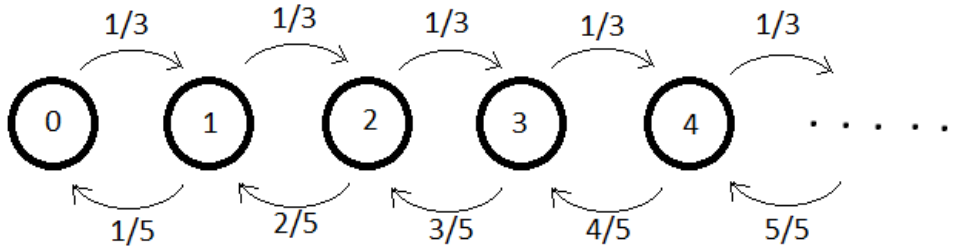
$$\pi_0 + \pi_1 = 0.9.$$

L'equazione diventa

$$(1-\rho)\,(1+\rho) = 0.9$$

ovvero $1-\rho^2 = 0.9$, $\rho^2 = 0.1$, $n^2 = 0.1 \cdot 3^2$, $n = 0.316\,23 \cdot 3 = 0.948\,69$.

2) Il numero di progetti da correggere è sempre una catena di nascita e morte ma i tassi di decrescita dipendono dallo stato. Se il numero di progetti è k, il docente sta lavorando a k correzioni, quindi completa la prima di esse con tasso $k \cdot \frac{1}{5}$.



Vale ora

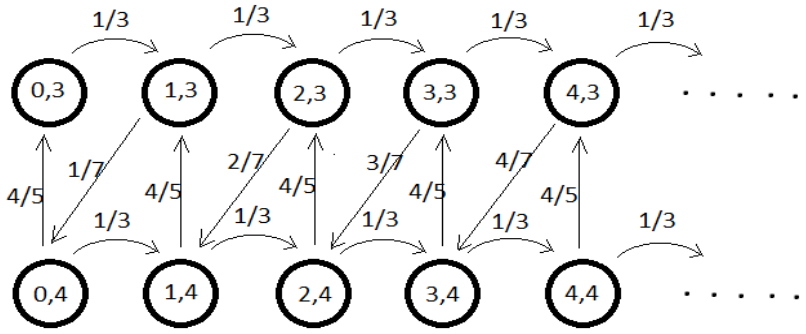
$$a_k = \frac{\frac{1}{3^k}}{k! \cdot \frac{1}{5^k}} = \frac{\left(\frac{5}{3}\right)^k}{k!}$$

$$\pi_0 = \left(\sum_{k=0}^{\infty} \frac{\left(\frac{5}{3}\right)^k}{k!}\right)^{-1} = e^{-\frac{5}{3}}, \quad \pi_k = e^{-\frac{5}{3}} \frac{\left(\frac{5}{3}\right)^k}{k!}$$

Il numero medio di compiti è

$$\begin{aligned}\sum_{k=0}^{\infty} k e^{-\frac{5}{3}} \frac{\left(\frac{5}{3}\right)^k}{k!} &= \sum_{k=1}^{\infty} k e^{-\frac{5}{3}} \frac{\left(\frac{5}{3}\right)^k}{k!} \\ &= e^{-\frac{5}{3}} \sum_{k=1}^{\infty} \frac{\left(\frac{5}{3}\right)^k}{(k-1)!} = e^{-\frac{5}{3}} \frac{5}{3} \sum_{k=1}^{\infty} \frac{\left(\frac{5}{3}\right)^{k-1}}{(k-1)!} \\ &= e^{-\frac{5}{3}} \frac{5}{3} \sum_{k=0}^{\infty} \frac{\left(\frac{5}{3}\right)^k}{k!} = \frac{5}{3}.\end{aligned}$$

3) Gli stati sono ora le coppie (n, k) dove n è il numero di richieste attive (cioè le richieste effettuate e non ancora consegnate) e k è il numero di compiti da correggere. Si passa da (n, k) a $(n + 1, k)$ con tasso $\frac{1}{3}$. Si passa da (n, k) a $(n - 1, k + 1)$ con tasso $\frac{n}{7}$. Si passa da (n, k) a $(n, k - 1)$ con tasso $\frac{k}{5}$.



Capitolo 6

Statistica Multivariata

6.1 La matrice di correlazione

La statistica multivariata in generale si pone lo scopo di esaminare i legami che intercorrono tra un numero finito di variabili. Per esempio, si vuole capire quali si esse sono più collegate.

Date le v.a.

$$X_1, ..., X_p$$

che formano un vettore aleatorio, possiamo calcolare la matrice di covarianza di questo vettore:

$$Q = \left(Cov\left(X_i, X_j\right)\right)_{i,j=1,...,p}$$

Essa fornisce una prima serie di informazioni sui legami tra le variabili, con l'unica fodamentale limitazione che si tratta di legami a due a due, non a gruppi più numerosi o globalmente tra tutte. Comunque, la matrice di covarianza è la prima informazione da mettere in gioco. Si può preferire la matrice di correlazione, in cui avendo eliminato la scala diventa più evidente ed assoluta l'iterpretazione dei numeri:

$$\left(Corr\left(X_i, X_j\right)\right)_{i,j=1,...,p} = \left(\frac{Cov\left(X_i, X_j\right)}{\sigma_{X_i}\sigma_{X_j}}\right)_{i,j=1,...,p}.$$

La statistica multivariata, essendo una parte della statistica, è ovviamente interessata all'analisi di dati sperimentali, più che di v.a. nel senso teorico. I dati relativi ad una stringa $X_1, ..., X_p$ di v.a. hanno la forma di una matrice:

	X_1	...	X_p
1	$x_{1,1}$	...	$x_{1,p}$
2	$x_{2,1}$	...	$x_{2,p}$
...	...	...	...
n	$x_{n,1}$	...	$x_{n,p}$

Ogni "unità" sperimentalmente esaminata, per esempio l'unità n.1, ha fornito una stringa di p numeri, uno per ciascuna variabile, e precisamente

$$x_{1,1} \quad ... \quad x_{1,p} \ .$$

Le righe corrispodono quindi alle unità esaminate negli esperimenti, le colonne alle variabili. Da una simile tabella, piuttosto complessa, si vorrebbero estrarre informazioni sui legami tra le variabili. In realtà c'è un altro scopo parallelo, che ora con la tabella diventa evidente: esaminare le unità, le loro somiglianze, se sono divise in gruppi, ad esempio.

Ora, data una simile matrice, si può calcolare la sua matrice di covarianza o di correlazione, in R coi comandi

`cov(A); cor(A)`

dove A è il nome dato alla matrice. Queste matrici sono la versione empirica di quelle teoriche ricordate sopra. Sono anch'esse matrici $p \times p$. Nella matrice `cov(A)`, l'elemento di posto (i, j) è la covarianza empirica tra i vettori numerici

$$\begin{array}{ccc} x_{i,1} & ... & x_{i,p} \\ x_{j,1} & ... & x_{j,p} \end{array}$$

ed è una stima della quantità teorica $Cov\,(X_i, X_j)$.

La matrice di correlazione empirica `cor(A)` fornisce immediatamente delle informazioni sui legami tra le variabili, a due a due, informazioni basate sulle osservazioni sperimentali di quelle particolari unità. Nella sezione di esercizi useremo continuamente questo comando e vedremo anche una visualizzazione del risultato.

Per andare oltre, servono nuovi elementi di statistica. Il metodo delle componenti principali è quello che più immediatamente si affianca al calcolo di `cor(A)`. Esso raggiunge il duplice scopo di mostrare visivamente i legami tra le variabili, anche un po' nel senso di gruppo (cioè non solo a due a due), ed al tempo stesso le relazioni tra le unità esaminate, gli idividui.

Le relazioni tra le unità vengono approfondite tramite altre strategie, genericamente chiamate di classificazione e clustering, che hanno lo scopo di riconoscere se le unità sono ragionevolmente suddivisibili in due o più gruppi abbastanza omogenei; ed hanno nche lo scopo di assegnare a gruppi prestabiliti delle nuove unità. Esamineremo alcuni metodi per questi scopi.

Tornando ai legami tra variabili, nasce spesso il desiderio di capire se certe variabili influiscono su altre, scoprire quali sono le variabili che provocano certi effetti e quali invece sono irrilevanti. In linea di massima la statistica non è in grado di dimostrare la presenza di relazioni causa-effetto tra variabili; è in grado di quantificare il legame che intercorre tra loro. Ipotizzando una relazione causa-effetto, la regressione lineare multipla quantifica tale relazione, scoprendo il valore dei coefficienti di un modello input-output tra le grandezze in gioco, modello che poi può essere usato per scopi di previsione, ad esempio (ed infatti torneremo sulla previsione delle serie storiche anche con questo strumento).

Infine, vedremo che il metodo delle componenti principali scopre nuove variabili, a volte interpretabili nell'ambito dell'applicazione specifica a volte no, che racchiudono la reale variabilità dei dati più delle v.a. $X_1, ..., X_p$ originarie. In un certo senso, le nuove variabili possono essere pensate come dei predittori di quelle originarie. Su questa falsariga si innesta il metodo dell'analisi fattoriale, che date delle v.a. osservate $X_1, ..., X_p$ cerca di individuarne di nuove, non osservate, che siano fattori (predittori) di quelle osservate.

6.1.1 Elevata correlazione non è sinonimo di causalità

Questo è un principio importante. Quando si riscontra un'elevata correlazione tra due variabili X ed Y, nulla indica che X *agisca* su Y, che X sia la *causa* delle variazioni di Y.

Un'ovvia ragione di tipo logico è che la correlazione è un'operazione simmetrica. Non si vede come sia possibile quindi dedurne una causalità asimmetrica, in cui una delle due variabili sia la causa (dovrebbe essere vero anche il viceversa, ma questo è assurdo nella maggior parte dei problemi causa-effetto).

Cosa può allora indicare un'elevata correlazione tra X ed Y? Almeno tre cose possibili:

- X agisce su Y
- Y agisce su X
- c'è una causa comune Z; quando Z cambia, provoca cambiamenti di X e di Y simultaneamente; noi osserviamo questi cambiamenti simultanei, quindi osserviamo una correlazione tra X e Y.

Facciamo un esempio banale di errore che si commetterebbe attribuendo ad un'elevata correlazione l'indicazione di una relazione causa-effetto: se si prende un gruppo di paesi industrializzati con un livello di sviluppo simile, si può osservare che il numero X di cittadini impegnati nell'istruzione è correlato al numero Y di cittadini impegnati nei trasporti. Ma questo è semplicemente effetto delle dimensioni delle nazioni: una nazione più grossa avrà più insegnanti e più trasporti, una più piccola ne avrà meno. Invece, è chiaro che X non è la causa di Y e viceversa.

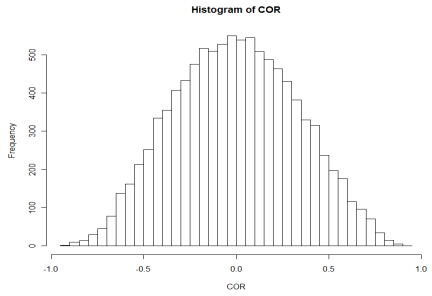
E' d'altra parte vero che un'elevata correlazione è *improbabile* tra due campioni indipendenti, come mostra il paragrafo seguente. Quindi a fronte di un'elevata correlazione non possiamo ignorare il fatto che un legame ci sia; resta però aperto il problema di quale sia (cioè se sia diretto oppure indiretto).

Analisi Monte Carlo della correlazione di campioni indipendenti

Supponiamo che $X_1, ..., X_n$ sia un campione estratto da X ed $Y_1, ..., Y_n$ un campione estratto da Y, indipendenti. Per semplicità, supponiamo che le v.a. siano tutte gaussiane e, visto che di correlazione si tratta, supponiamole standardizzate (questo non è restrittivo: la correlazione di due v.a. o due campioni è uguale a quella delle v.a. o campioni standardizzati). Che valori può assumere la correlazione? Fissato n, generiamo con R due stringhe di tal tipo e calcoliamone la correlazione, e ripetiamo questo N volte. Tracciamo un istogramma dei valori.

```
n=10; N=10000
COR <- 1:N
for (i in 1:N) {
x<- rnorm(n); y<- rnorm(n)
COR[i]<-cor(x,y)
}
```

```
hist(COR,30)
```



La forma è evidentemente simmetrica, come ci si poteva aspettare. Vediamo ad occhio che valori superiori (in valore assoluto) a 0.75 hanno una probabilità abbastanza piccola. Si potrebbe calcolare empiricamente una soglia al 95%: un numero λ che viene superato a destra solo con probabilità 0.025 (e $-\lambda$ a sinistra con probabilità 0.025, quindi globalmente 0.05). Basta ordinare il campione e prendere il valore di posto 9750:

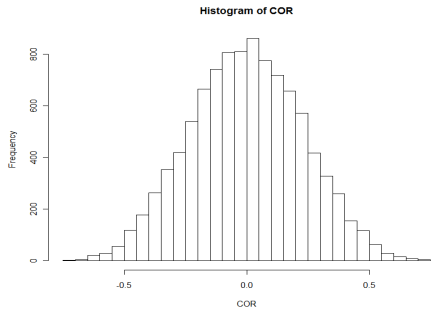
```
sort(COR)[9750]
```

[1] 0.637

La conclusione è: adottando il livello di significatività 95%, per due stringhe di lunghezza 10, un valore di correlazione superiore in modulo a 0.637 non può ritenersi casuale, cioè compatibile con l'indipendenza delle due stringhe; è indice di un legame.

Il numero 0.637 sembra piuttosto alto, ma ciò è dovuto al fatto che le stringhe sono corte. Ripetiamo lo studio per strighe lunghe 20:

```
n=20; N=10000; COR <- 1:N
for (i in 1:N) {
x<- rnorm(n); y<- rnorm(n)
COR[i]<-cor(x,y)
}
hist(COR,30)
```




```
sort(COR)[9750]
```

[1] 0.445

L'istogramma è decisamente più stretto e la soglia è molto più bassa: una correlazione superiore in modulo a 0.445 non è casuale. Per curiosità vediamo il caso 30:

```
n=30; N=10000; COR <- 1:N
for (i in 1:N) {
x<- rnorm(n); y<- rnorm(n)
COR[i]<-cor(x,y)
}
hist(COR,30)
sort(COR)[9750]
```
[1] 0.3693759

6.2 Il metodo delle componenti principali

Supponiamo di esaminare p grandezze di interesse pratico, descritte da variabili aleatorie (gaussiane, anche se questa richiesta non è necessaria) $X_1, ..., X_p$. Un esempio possono essere cinque potenziali indicatori di benessere nelle diverse regioni italiane:

X_1 = PLIC (posti letto in istituti di cura)
X_2 = SC (spese complessive per famiglia)
X_3 = SA.SC (proporzione di SC dedicata agli alimentari)
X_4 = TD (tasso di disoccupazione)
X_5 = TMI (tasso di mortalità infantile)

In tal caso, ad ogni regione italiana R possiamo associare un vettore con cinque coordinate:

$$R \leftrightarrow \begin{pmatrix} X_1(R) & X_2(R) & X_3(R) & X_4(R) & X_5(R) \end{pmatrix}$$

In generale, ad ogni dato (visto come p-upla di valori, relativi, nell'esempio, alla stessa regione italiana, uno per ogni variabile aleatoria) associeremo allo stesso modo un punto in uno spazio vettoriale di dimensione p, che ha per base proprio $X_1, ..., X_p$:

$$R \leftrightarrow \begin{pmatrix} X_1(R) & ... & X_p(R) \end{pmatrix}$$

Per non falsare l'indagine è conveniente standardizzare i dati: calcoliamo per ogni indicatore X_n la sua media μ_n e la sua deviazione standard σ_n e costruiamo una nuova tabella di dati dove sostituiamo ad ogni valore x di ogni indicatore X_n il valore standardizzato $\frac{x-\mu_n}{\sigma_n}$. In questo modo ora ogni indicatore ha la stessa media 0 e la stessa deviazione standard 1, e la matrice di covarianza Q coincide quindi con la matrice di correlazione.

Per $p = 2$, se la coppia (X_1, X_2) è un vettore gaussiano, abbiamo visto che i punti si dispongono a formare una nuvola ellissoidale, e tale rappresentazione grafica ci suggerisce alcune interpretazioni dei dati e della loro correlazione (ad esempio, se x_1 cresce allora anche x_2 tende a crescere). Per $p = 3$ otterremo invece una figura tridimensionale, che risulta di ben più difficile comprensione, mentre per valori di p superiori una qualsiasi rappresentazione grafica completa risulta impossibile (e inimagginabile). Il problema che ci poniamo è quindi quello di trovare, se esiste, un modo per avere (possibilmente in due sole dimensioni) una visualizzazione grafica d'insieme della distribuzione dei dati e della correlazione tra le variabili in esame per p maggiori di 2 o 3. Descriviamo il metodo chiamato Analisi delle Componenti Principali (abbreviato in PCA).

6.2.1 Diagonalizzazione di Q

Abbiamo visto nello studio delle gaussiane multidimensionali come le superfici di livello di una gaussiana p-dimensionale siano degli ellissoidi in dimensione p. Questo ci dice che, almeno nel caso gaussiano, i dati tenderanno a disporsi a formare un ellissoide, e saranno al solito più concentrati più ci si avvicina al centro dell'ellissoide: nel nostro esempio sugli indicatori di benessere, ci troviamo quindi con 20 punti disposti su una nuvoletta simile a un'ellissoide in 5 dimensioni. Nessuno riesce a visualizzare una tale figura, e l'idea di base del metodo PCA è quella di operare un 'cambio di variabili', cioè un cambio di base nel nostro spazio vettoriale di dimensione p, che grosso modo 'ruoti' la nuvola elissoidale in modo da poterla vedere dall'angolazione migliore, cioè in modo da averne una proiezione bidimensionale dove i dati sono il più distinti possibile tra loro.

Esempio 105 *Giusto per rendere l'idea, facciamo un esempio tridimensionale. Supponiamo di avere* 3 *variabili aleatorie e un migliaio di rilevazioni, e supponiamo che il corrispondente migliaio di punti si disponga nello spazio tridimensionale secondo un ellissoide con i tre assi lunghi rispettivamente 100, 100 e 1 (in qualche unità di misura). Se guardassimo la figura 'di taglio', vedremo solo una sottilissima striscetta di punti tutti accalcati, e non saremo in grado di vedere eventuali relazioni tra loro. Se invece guardiamo la figura 'di piatto' vediamo un cerchio pieno, che ci dà un'idea molto più realistica della disposizione tridimensionale dei dati, e ci permette di cogliere meglio eventuali relazioni tra di essi. Il fatto che uno dei tre assi della figura ellissoide sia molto piccolo (rispetto agli altri) ci dice che le variabili aleatorie 'variano di poco' in quella direzione: 'scartando' tale direzione (che è quello che facciamo proiettando l'ellissoide tridimensionale sul cerchio bidimensionale 'visto di piatto') abbiamo cioè la minima perdita di informazioni, e otteniamo quindi la 'migliore visualizzazione bidimensionale d'insieme' dei dati.*

Esempio 106 *Supponiamo ora che i dati dell'esempio precedente si dispongano invece in maniera sferica: da qualunque angolazione li guardiamo (ad esempio da sopra il polo nord) ci risulterà sempre un cerchio (delimitato in tal caso dall'equatore), ma in ogni caso vedremo sovrapposti molti dati, anche molto diversi tra loro (tutti quelli del diametro polo nord-polo sud vengono proiettati nello stesso punto!). Questo ci dice che, comunque la proiettiamo, la visualizzazione bidimensionale dei dato che ne risulterà sarà molto imprecisa, perchè in ogni caso perderemo (cioè non riusciremo a distinguere) una consistente fetta della varianza complessiva dei dati. Questo sarà, come vedremo, un esempio in cui il metodo PCA non risulta utile ed efficace.*

Torniamo alle nostre p variabili aleatorie (gaussiane). La matrice Q di covarianza (o di correlazione, dato che abbiamo standardizzato le variabili) è una matrice simmetrica, e quindi sappiamo dal teorema spettrale che è diagonalizzabile, cioè che esiste una base ortonormale di autovettori nella quale la matrice assume la forma:

$$Q' = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_p \end{pmatrix}$$

Siccome Q è una matrice definita non-negativa, sappiamo anche che gli autovalori $\lambda_1, \ldots, \lambda_p$ sono tutti non negativi. Supponiao

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

e indichiamo con $e_1, \ldots, e_p$ ei corrispondenti autovettori. Idichiamo con $u_1, \ldots, u_p$ i vettori della base canonica, es. $u_1 = (1, 0, ..., 0)$. Ci sono coefficienti e_i^j tali che

$$e_i = \sum_{j=1}^{p} e_i^j u_j$$

ed ovviamente essi sono semplicemente le coordinate dei vettori e_i nella base canonica:

$$e_i = \left(e_i^1, ..., e_i^p\right).$$

Ora, invece di eseguire la combinazione lineare di $u_1, \ldots, u_p$ combiniamo le v.a. $X_1, \ldots, X_p$:

$$V_i := \sum_{j=1}^{p} e_i^j X_j.$$

Cerchiamo di capire di che si tratta , con un esempio. Se R è una regione italiana, descritta dai 5 indicatori $X_1(R), ..., X_5(R)$, il valore $V_i(R)$ è l'ampiezza (con segno) della proiezione del punto $\begin{pmatrix} X_1(R) & ... & X_p(R) \end{pmatrix}$ sul vettore e_i. Al variare della regione R, i numeri $V_i(R)$ sono come dei punti aleatori sulla retta individuata dal vettore e_i. Quanto vale la loro varianza?

Teorema 36 $\lambda_i = Var[V_i]$.

Proof.

$$\begin{aligned} Var[V_i] &= Cov\left(\sum_{j=1}^{p} e_i^j X_j, \sum_{j'=1}^{p} e_i^{j'} X_{j'}\right) = \sum_{j,j'=1}^{p} e_i^j e_i^{j'} Cov\left(X_j, X_{j'}\right) \\ &= \sum_{j,j'=1}^{p} e_i^j e_i^{j'} Q_{jj'} \end{aligned}$$

da cui con un po' di pazienza si riconosce che $Var[V_i]$ è la componente di posto (i, i) nella matrice ottenuta trasformando Q tramite il cambio di base relativo al sistema $e_1, \ldots, e_p$. Ma allora tale componente vale λ_i, visto che in tale base la matrice Q è diagonale, con elementi λ_i sulla diagonale. ■

Gli autovalori λ_i di Q sono le varianze lungo le direzioni degli autovettori di Q. Siccome abbiamo ordinato gli autovalori λ_i in modo decrescente, e_1 è la direzione (a volte non univocamente definita) in cui abbiamo a varianza massima, e_2 quella con la varianza subito minore e così via. Potremmo essere più precisi ed enunciare e dimostrare un teorema secondo cui le direzioni scelte sono quelle che massimizzano via via la varianza, tolte quelle già trovate. Però se si pensa all'interpretazione geometrica con gli ellissoidi, è già chiaro che e_1 è la direzione dell'asse principale, e_2 quella del successivo perpendicolare ad e_1 e così via.

E' chiaro allora che la visione dei punti sperimentali secondo il piano generato da e_1, e_2 è la migliore, quella in cui i punti appaiono più sparpagliati (hanno maggior varianza). E così per lo spazio tridimensionale generato da e_1, e_2, e_3, e così via, ma solo in dimensione 2 la visualizzazione è efficace.

I vettori $e_1, \ldots, e_p$, o a volte le relative v.a. $V_1, \ldots, V_p$ definite come sopra, si chiamano *componenti principali* (nel seguito tenderemo a confondere $u_1, \ldots, u_p$ con $X_1, \ldots, X_p$ ed $e_1, \ldots, e_p$ con $V_1, \ldots, V_p$).

A titolo di esempio, riprendendo gli indicatori di benessere, elenchiamo le coordinate di V_1 rispetto alla vecchia base $X_1, \ldots, X_5$, le coordinate della variabile aleatoria X_1=PLIC nella nuova base, e le coordinate di uno stesso dato (la Toscana) rispetto alle due basi:

$$V_1 = \begin{pmatrix} -0.310 \\ -0.491 \\ 0.512 \\ 0.506 \\ 0.379 \end{pmatrix}_X \qquad X_1 = \begin{pmatrix} -0.310 \\ 0.769 \\ -0.553 \\ 0 \\ 0 \end{pmatrix}_V$$

che significa anche

$$V_1 = -0.310 \cdot PLIC - 0.491 \cdot SC + 0.512S \cdot A.SC + 0.506 \cdot TD + 0.379 \cdot TMI$$

$$X_1 = -0.310 \cdot V_1 + 0.769 \cdot V_2 - 0.553 \cdot V_3$$

$$\text{Tosc} = \begin{pmatrix} 0.126 \\ 1.093 \\ -0.796 \\ -0.645 \\ -1.356 \end{pmatrix}_X = \begin{pmatrix} -1.824 \\ -0.002 \\ 0.867 \\ -0.298 \\ 0.096 \end{pmatrix}_V .$$

Nel prossimo paragrafo si chiarirà come abbiamo trovato queste componenti.

Considerando la nuova base come i nuovi indicatori (sulla cui interpretazione torneremo in seguito), cioè le nostre nuove variabili aleatorie, la matrice diagonale che abbiamo trovato è proprio la matrice di correlazione tra queste nuove variabili (i dati sono sempre gli stessi, ma ora sono visti nella nuova base, cioè hanno altre coordinate): i valori sulla diagonale (gli autovalori) sono le varianze delle nuove variabili, mentre l'essere diagonale ci dice che queste nuove variabili sono tutte tra loro scorrelate. Essendo scorrelate, la varianza della somma delle nuove variabili aleatorie è la somma delle varianze, cioè la somma degli autovalori: possiamo quindi interpretare ogni autovalore come la parte di varianza totale spiegata dalla corrispondente nuova variabile aleatoria (torneremo in seguito su questo punto).

6.2.2 I comandi di `R`

Il programma `R` svolge tutti i conti visti fin qui in automatico con un solo comando. Una volta importata la tabella di dati (standardizzati) in una certa matrice `A`, basta la seguente linea di comando:

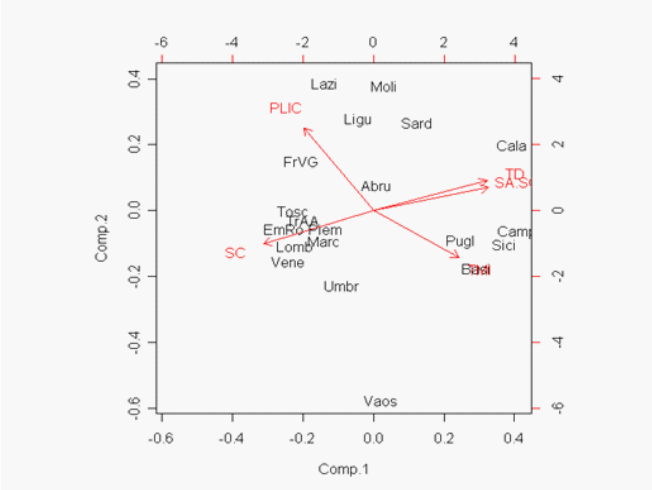
```
pca<-princomp(A)
```

`R` calcola la matrice di covarianza, la diagonalizza con una base ortonormale e ne ordina gli autovettori in ordine di autovalori decrescenti. Osserviamo che il nome `pca` vale come qualsiasi altro, nonostante l'allusione al metodo; il comando di R è `princomp(A)`.

Con il comando:

`biplot(pca)`

si ottiene un'immagine del piano principale, molto ricca di informazioni (si pensi all'elenco seguete di considerazioni, tutte immediate a partire da una sola immagine):



Essa contiene tre elementi: i due nuovi assi, la proiezione dei dati e quella dei vecchi assi. Gli assi orizzontale e verticale sono rispettivamente la prima e la seconda componente principale $Comp.1 = V_1$ e $Comp.2 = V_2$.

La proiezione di ogni dato corrisponde al punto del piano principale le cui coordinate sono le prime due coordinate del dato nella nuova base: ad esempio, tornando ai cinque potenziali indicatori di benessere del 2002, alla Toscana corrisponde il punto

$$Tosc = \begin{pmatrix} -1.824 \\ -0.002 \end{pmatrix}$$

Allo stesso modo le vecchie variabili aleatorie (cioè i vettori della vecchia base) sono rappresentate in proiezione sul piano principale, con vettori evidenziati in rosso: ad esempio, ad $X_1 = PLIC$ corrisponde il vettore

$$PLIC = -0.310 \cdot V_1 + 0.769 \cdot V_2 = \begin{pmatrix} -0.310 \\ 0.769 \end{pmatrix}.$$

Una prima analisi qualitativa può essere svolta in base ai rapporti tra i vettori che rappresentano i nostri indicatori (ortogonalità, parallelismo con versi concordi o discordi, ecc.), e ai raggruppamenti e alle posizioni dei dati. Nel nostro esempio, guardando la figura, alcune delle considerazioni che possiamo fare (per quanto naturali e più o meno note, visto che conosciamo abbastanza bene la situazione nazionale del benessere) sono:

- SC, TD e SA.SC sono tutti essenzialmente paralleli, a indicare una forte correlazione tra di loro: potremmo ad esempio leggere la loro direzione comune come un indicatore complessivo di benessere economico.

- Il verso di SC e opposto a quelli di TD e SA.SC, segno che questi indicatri sono correlati negativamente: come ci si aspetta, una maggior disoccupazione media si riflette su una minore spesa complessiva media (a TD alto corrisponde SC basso, e viceversa), mentre se la spesa complessiva media e molto bassa questa sarà, come e naturale, in gran parte dedicata agli alimentari (a SC basso corrsiponde SA.SC alto, e viceversa). Allo stesso modo, la correlazione positiva tra TD e SA.SC indica che nelle zone di più alta disoccupazione le (poche) spese sono destinate per lo più ai generi alimentari.

- PLIC e TM sono abbastanza paralleli tra loro (in analogia a quanto visto sopra potremmo leggere la loro direzione comune come un indicatore complessivo di salute), ma correlati negativamente, come è naturale.

- PLIC e TM sono abbastanza perpendicolari agli altri indicatori, segno che i due gruppi, e quindi le due direzioni indicatore complessivo di benessere economico e indicatore complessivo di salute, sono abbastanza scorrelati tra loro. Tuttavia notiamo le lievi correlazioni positive nelle direzioni che ci aspettiamo: maggiori posti letto dove ci sono maggiori spese complessive, e maggior mortalità infantile dove c'è più disoccupazione e le spese sono in prevalenza alimentari.

- L'area di maggior benessere e quella nella direzione positiva di SC, con un po' di spostamento verso PLIC. In tale zona si trovano fortemente raggruppate varie regioni (Veneto, Trentino Alto Adige, Lombardia, Piemonte, Emilia Romagna, Marche e Toscana), che pertanto risultano molto simili rispetto agli indicatori considerati.

- Le altre regioni del centro-nord (Liguria, Friuli, Lazio) non eccellono in SC ma eccellono in PLIC, a indicare una buona cura sanitaria nonostante un tenore di vita medio più modesto rispetto al gruppo precedente.

- Particolarmente negativo, sia rispetto all'asse del benessere economico che a quello della salute, risulta il raggruppamento composto da Campania, Sicilia, Basilicata e Puglia, in maniera molto più accentuata rispetto ad altre regioni meridionali o insulari (come Calabria e Sardegna) che nell'immaginario colettivo potremmo invece credere ad esse simili. Questo potrebbe indicare uno sforzo di miglioramento di alcune regioni, e potrebbe ad esempio suggerire l'analisi di altri dati più mirati per averne verica o smentita.

Osservazione 75 *Il parallelismo non è indicazione sicura di correlazione. Infatti, due vettori diversi possono avere la stessa proiezione su un piano. Bisogna sempre accertarsi dalla matrice di correlazione che ci sia davvero correlazione. La perpndicolarità invece non si guadagna per proiezione, quindi se è visibile nel piano principale, c'è davvero.*

L'orientazione delle variabili di partenza rispetto alle componenti principali può inoltre suggerire delle potenziali interpretazioni delle due componenti principali. Tornando al nostro

esempio, osserviamo la figura con davanti i dati visualizzati col comando `pca$loadings`. E ragionevole associare *Comp.*1 alle tre variabili SC, SA.SC e TD, in quanto ha componenti maggiori in tali direzioni (circa 0.5, in valore assoluto, contro i circa 0.3 nelle altre direzioni). Allo stesso modo, ha senso associare *Comp.*2 a PLIC e TMI (0.4 e 0.8 contro 0.2 e 0.3). Una possibile interpretazione delle prime due componenti principali, cioè delle nuove variabili aleatorie, potrebbe quindi essere quella dove la prima descrive il benessere di topo economico e la seconda quello relativo alla salute.

Per quantificare le osservazioni fatte sulla correlazione (positiva o negativa) o meno tra gli indicatori di partenza, è importante osservare direttamente la matrice di correlazione delle variabili di partenza, tramite il comando:

```
cor(A)
```

Nell'esempio sugli indicatori di benessere, possiamo così verificare quantoavevamo già stabilito: la forte correlazione (con il giusto segno) tra SC, SA.SC e TD, l'assenza di legame tra PLIC e SC e TD, la correlazione negativa ma non troppo marcata tra PLIC e TMI, e via dicendo. Notiamo, rispetto a quanto gi a detto basandoci sulla figura, la correlazione (anche se non forte) di TMI non solo con PLIC, ma quasi allo stesso modo anche con le tre variabili economiche, negativa o positiva nel modo che ci aspettiamo. Una rappresentazione grafica si ottiene col comando

```
plot(A)
```

che mostra, per ciascuna coppia di variabili, il grafico di dispersione dei dati, altro strumento da cui si può intuire la presenza di legame o meno tra le variabili aleatorie. Ad esempio, tornando agli indicatori di benessere, se si esegue questo comando si nota la visualizzazione grafica del legame tra TD e SA.SC da un lato, e dell'assenza di legame tra PLIC e SC dall'altro.

Con il comando:

```
pca$loadings
```

compare una tabella in cui è possibile vedere le coordinate di ogni vettore della nuova base rispetto alla vecchia base (leggendo la tabella colonna per colonna) e le coordinate dei vettori della vecchia base rispetto alla nuova base (leggendo la tabella riga per riga). Questi numeri, i loadings, contengono potenzialmente molto significato, su cui torneremo più diffusamente nell'ambito dell'Analisi Fattoriale. Se sono grandi, indicano che una variabile pesa molto sull'altra e questo può contribuire all'interpretazione delle componenti principali. Tuttavia, la figura ottenuta con `biplot(pca)` già fornisce questo tipo di informazioni in modo più immediato. Per questo i loadings sono più essenziali nel caso dell'Analisi Fattoriale, in cui non ci sono raffigurazioni così espressive.

È inoltre possibile vedere le deviazioni standard delle nuove variabili aleatorie, cioè le radici quadrate degli autovalori di `A`, semplicemente digitando

```
pca
```

6.2.3 Classifiche tramite PCA

Dato un punto sperimentale, ad esempio la regione Toscana, espresso nelle coordinate X, la sua proiezione su e_1 rappresenta la sua prima coordinata nella nuova base, e così per le altre

proiezioni. Esse si calcolano coi prodotto scalari

$$(x, e_i)$$

dove x è il vettore che definisce il punto nella base X. Questa regola vale se gli autovettori e_i sono normalizzati (lunghezza 1), altrmenti bisogna dividere per la loro lunghezza. Ad esempio, si può verificare che

$$\begin{pmatrix} -0.310 & -0.491 & 0.512 & 0.506 & 0.379 \end{pmatrix}_X \cdot \begin{pmatrix} 0.126 \\ 1.093 \\ -0.796 \\ -0.645 \\ -1.356 \end{pmatrix}_X = -1.824$$

```
v<-c(-0.310,-0.491,0.512,0.506,0.379)
x<-c(0.126,1.093,-0.796, -0.645,-1.356)
x%*%v
[1,] -1.823569
```

Questo metodo può essere usato per fare delle classiifiche tra gli individui (unità sperimentali) esaminati col metodo. Ogni individuo avrà un punteggio, dato dalla sua proiezione sulla prima componente principale, per cui i diversi individui risulteranno ordinati in una classifica e muniti di punteggio. Si veda ad esempio l'esercizio 4.

Si può naturalmente calcolare la classifica anche rispetto alla seconda componente principale, e così via. Il punto è avere un'interpretazione del risultato. Se riteniamo che la prima componente descriva una caratteristica per noi utile (a priori non misurabile) degli individui esaminati, la classifica avrà il significato corrispondente.

6.2.4 Il miglior 'punto di vista'

Convinciamoci ora del fatto che le nuove variabili $V_1, \ldots, V_n$ siano effettivamente gli assi dell'ellissoide n-dimensionale. L'asse principale dell'ellissoide è chiaramente la direzione in cui i dati sono più dispersi, cioè la direzione lungo la quale la varianza è massima. Quindi noi vogliamo trovare quel versore w per il quale $Var[w]$ sia la massima possibile. Scriviamo w nella nuova base $V_1, \ldots, V_n$:

$$w = w_1 V_1 + \ldots + w_n V_n \quad , \text{con } w_1^2 + \ldots + w_n^2 = 1$$

Siccome per variabili aleatorie scorrelate si ha $Var[X + Y] = Var[X] + Var[Y]$, possiamo facilmente calcolarne la varianza:

$$Var[w] = Var[w_1 V_1 + \ldots + w_n V_n] = w_1^2 Var[V_1] + \ldots + w_n^s Var[V_n] =$$
$$= w_1^2 \lambda_1 + \ldots + w_n^2 \lambda_n \leq \lambda_1 (w_1^2 + \ldots + w_n^2) = \lambda_1$$

cioè la varianza di un qualsiasi vettore-direzione w è minore (al più uguale) rispetto a λ_1, cioè quella di V_1! Quindi V_1 è la direzione lungo cui c'è massima varianza: l'asse principale dell'ellissoide. Tra tutte le direzioni ortogonali a V_1, cerchiamo quella con la massima varianza

(rimasta): con lo stesso procedimento, questa risulta essere V_2. E così via fino a determinare che la base $V_1, \ldots, V_n$ è esattamente composta dai versori che individuano le direzioni degli assi ortogonali dell'ellissoide n-dimensionale.

Possiamo ora proiettare il nostro ellissoide n-dimensionale sul piano (detto *piano principale*) individuato da V_1 e V_2, cioè le due nuove variabili che più evidenziano la dispersione dei dati: questo è l'angolazione sotto la quale guardare la figura n-dimensionale per avere la migliore visione (bidimensionale) d'insieme della distribuzione (n-dimensionale) dei dati.

6.2.5 Efficacia del metodo PCA

Il comando `plot(pca)` illustra la varianza lungo le diverse componenti principali, cioè le lunghezze degli assi principali della nostra figura ellissoidale, da cui e possibile farsi un'idea della dimensione dei dati, cioè di quante componenti sono necessarie o utili per analizzare i dati. Tornando al nostro solito esempio, è chiaro come Comp.4 e Comp.5 siano inutili, e quindi la dimensione dei dati sia 2 o 3. Questo significa che l'ellissoide 5-dimensionale ha in realtà solo 2 o 3 dimensioni effettive, e quindi che una rappresentazione ottimale dei dati si ottiene con una opportuna proiezione in dimensione 2 o 3. Detto altrimenti, per rappresentare le 5 variabili iniziali in realtà bastano solo 2 o 3 variabili aleatorie (cioè Comp.1, Comp.2 e, eventualmente, Comp.3).

Si possono avere i dati numerici precisi con il comando

`summary(pca)`

La prima riga riporta la deviazione standard di ogni componente principale. Essendo le componenti principali tra loro scorrelate, la varianza della somma delle nuove variabili aleatorie è la somma delle rispettive varianze: possiamo quindi calcolare per ogni componente principale la parte di varianza totale da essa spiegata, valore che viene riportato nella seconda riga. Ad esempio, per gli indicatori di benessere in esame, Comp.1 spiega circa il 67% della varianza totale, mentre Comp.2 e Comp.3 rispettivamente il 17% e l'11%. La terza riga riporta la varianza cumulativa, che e semplicemente la somma delle percentuali di varianza spiegata da quella componente principale e da tutte le precedenti (per cui e ovvio che l'ultima componente abbia varianza cumulativa 1).

La varianza cumulativa è il pricipale parametro dell'efficacia del metodo PCA, dato che quantifica quanto accurata e la visualizzazione dei dati data dal piano principale. Nel nostro esempio, le prime due componenti principali (cioè il piano principale) spiegano complessivamente l'84% della varianza totale, e quindi la rappresentazione e decisamente soddisfacente. Una rappresentazione tridimensionale, contando quindi anche Comp.3, sarebbe praticamente perfetta (95%!). In genere, si considera il metodo PCA efficace quando il piano principale rappresenta l'80 - 90% della varianza totale dei dati, cioè quando la parte di informazione persa (rappresentata dalla varianza delle altre componenti principali: Comp.3, Comp.4, eccetera) si aggira sul 10 - 20% del totale. Tuttavia, anche quando la rappresentazione bidimensionale data dal piano principale e insufficiente, il metodo PCA contribuisce comunque a comprendere meglio i dati analizzati, in particolare indicandone l'effettiva dimensione', cioè quante variabili al minimo bastano per rappresentarli efficacemente.

6.3 Modelli lineari

Questa sezione è dedicata alla descrizione sintentica di una coppia di metodi di statistica multivariata: la regressione lineare multipla e l'analisi fattoriale.

6.3.1 Introduzione: modelli lineari di legame tra variabili aleatorie

La teoria che stiamo per esporre vuole descrivere relazioni matematiche tra variabili aleatorie: in tali relazioni appariranno variabili di *input*, dette ad esempio *fattori* o *predittori*, e variabili di *output*, dette ad esempio *risposte* oppure *osservabili*. Le relazioni più semplici sono quelle lineari o affini, che ora descriveremo.

Supponiamo che certe variabili di output Y_i, $i = 1, ..., m$, siano legate a certe variabili di input X_j, $j = 1, ..., n$, dalle relazioni affini

$$Y_i = a_{i1}X_1 + ... + a_{in}X_n + b_i + \sigma_i\varepsilon_i$$
$$i = 1, ..., m$$

dove i numeri a_{ij} sono i coefficienti della relazione, i numeri b_i sono le intercette (in senso generalizzato), mentre le espressioni $\sigma_i\varepsilon_i$ sono variabili aleatorie che rappresentano gli errori presenti nella relazione, errori ascrivibili a diverse cause, o di aleatorietà intrinseca o di mancanza nostra di conoscenza e precisione nella descrizione dei legamo tra le X e le Y. Per comodità, in tali errori separiamo una parte aleatoria ε_i a media nulla (l'eventuale media non nulla dell'errore supponiamo di averla inglobata in b_i) e varianza unitaria, per cui i coefficienti σ_i rappresentano le deviazioni standard degli errori (questa convenzione è diversa da quella adottata in certe parti del capitolo sulle serie temporali, ma non dovrebbe generarsi confusione).

Una risposta, più fattori

Il caso di una sola v.a. di output Y ed uno o più fattori in input $X_1, ..., X_n$ è quello esaminato dalla regressione lineare semplice (un fattore) o multipla (più fattori).

Esempio 107 *Riprendiamo l'esempio della Sezione 6.2. Sappiamo che SA.SC, la proporzione delle spese medie familiari dedicata ai soli alimenti, è (approssimativamente, si intende) direttamente proporzionale a TD, il tasso di disoccupazione. Posto Y =SA.SC, X_1 =TD, si potrebbe studiare la regressione lineare semplice del tipo*

$$Y = a_1X_1 + b + \sigma\varepsilon$$

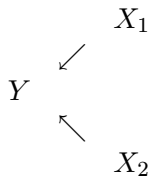
usando il comando `lm(Y~X)`*. Precisamente, se A <- read.table(file="indicatori_ benessere.txt") è il file di dati, dopo aver costruito delle variabili* `X<-A[,4]`*,* `Y<-A[,3]`*, poi si può fare* `lm(Y~X)`*. Posto ad es.* `M1<-lm(Y~X)`*, si possono poi visualizzare dei risultati scrivendo* `M1`*, oppure* `summary(M1)`*, oppure* `plot(M1)`*.*

Anticipiamo alcune cose che approfondiremo nei paragrafi successivi sulla regressione: eseguendo la regressione semplice dell'esercizio, o anche solo calcolando il coefficente R^2 (che

viene pari a 0.8198) si vede che il legame lineare c'è, buono, ma non fortissimo. Una parte della variabilità di Y resta inspiegata. In effetti, a buon senso, non è il solo tasso di disoccupazione che influenza quanto le famiglie dedicano agli alimenti rispetto al totale delle loro spese. Ci saranno altri fattori, legati ad altri aspetti generali del loro benessere economico o sviluppo sociale, come il grado di istruzione. Individuato un secondo potenziale fattore X_2, si può esaminare il modello di *regressione multipla*

$$Y = a_1X_1 + a_2X_2 + b + \sigma\varepsilon.$$

Si potrebbe rappresentare questa situazione col seguente diagramma:



Lo studio della regressione multipla con `R` è immediato: basta usare il comando `lm(Y~X1+X2)`.

Esercizio 38 *Cercare in rete un altro potenziale fattore associato ad SA.SC ed eseguire con* `R` *la regressione multipla. Quanto viene il coefficiente* R^2*? E' migliorato rispetto al caso di un fattore solo? Se non si ha voglia di cercare nuove grandezze in rete, si provi ad eseguire la regressione* `lm(SA.SC~TD + SC)`*.*

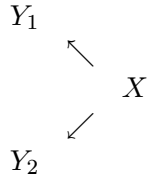
Il coefficiente R^2 (nella regressione multipla, ad esempio quella relativa all'esercizio appena enunciato) ha di nuovo il significato di varianza spiegata. E' chiaro che aumenta aggiungendo più fattori, anche se sono insignificanti; però aumenta poco o tanto a seconda della rilevanza di ciò che si è aggiunto.

Il numero $Pr(>|t|)$ è il cosidetto p-value di un test che spiegheremo. L'idea empirica è: se $Pr(>|t|)$ è piccolo, il fattore è rilevante, altrimenti no. Quanto piccolo, è soggettivo; 0.05 è una scelta condivisa da molti.

Esercizio 39 *Si eseguano le tre regressioni* `lm(SA.SC~TD + SC)`*,* `lm(SA.SC~TD)`*,* `lm(SA.SC~SC)`*. Confrontare i valori di* R^2 *e* $Pr(>|t|)$*. Trarre qualche conclusione basandosi sul buon senso.*

Un fattore, più risposte

Una situazione diametralmente opposta alla precedente è quella in cui c'è una sola X e diverse Y_i, ad esempio tre grandezze X, Y_1, Y_2 tali che Y_1 sia influenzata da X ed anche Y_2 sia influenzata dallo stesso X.



Se disponiamo di dati sperimentali per tutte queste grandezze, basta esaminare un modello lineare per la coppia (X, Y_1) e poi un'altro, separatamente, per la coppia (X, Y_2). In questo

caso non ci sarebbe nulla di nuovo rispetto al paragrafo precedente, le cui considerazioni andrebbero applicate separatamente a ciascuna Y_i. Il discorso si generalizza senza problemi al caso di più input e più output.

Completamente diverso invece è il caso in cui le due grandezze Y_1, Y_2 sono misurabili, disponiamo di loro dati sperimentali, mentre X non è misurabile, anzi forse non è nemmeno completamente ben definita. Immaginiamo ad esempio, sempre con riferimento all'esempio della Sezione 6.2, di sospettare che ci sia un fattore X che influenza Y_1 =SA.SC e Y_2 =SC. Abbiamo le misurazionii delle grandezze Y_1 e Y_2 ma non di X, di cui anzi non ci è chiaro nemmeno il significato. la domanda è: c'è un fattore, che ancora dobbiamo scoprire, mettere allo scoperto, che influenza entrambe SA.SC e SC? Un *fattore nascosto*? Che *spiega* (si suol dire) la variabilità di SA.SC e SC? Cosa c'è dietro il fatto che certe regioni hanno una minore spesa complessiva familiare ed una maggior proporzione di spesa per alimentari, rispetto ad altre regioni in cui queste grandezze sono invertite?

A modo suo, *il metodo PCA serve anche a questo scopo*: esso ha individuato una nuova grandezza aleatoria, Comp1, a cui abbiamo attribuito un significato del tipo "benessere economico", legata ad entrambe SC e SA.SC. Discuteremo sotto il metodo dell'analisi fattoriale, alternativo a PCA, tornando però anche su PCA in relazione al problema ora posto.

6.3.2 Regressione lineare semplice

Iniziamo lo studio della regressione col caso della regressione lineare *semplice*, cioè con un solo fattore, probabilmente già nota dai corsi di statistica elementare. Premettiamo un breve riassunto su covarianza e correlazione, già esposte altrove, ma che può essere utile.

Covarianza e coefficiente di correlazione

Date due v.a. X ed Y, chiamiamo *covarianza* il numero

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

La covarianza generalizza la varianza: se X ed Y sono uguali, vale

$$Cov(X,X) = Var[X].$$

Analogamente alla varianza, vale la formula (di facile dimostrazione)

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

Ricordiamo che, se X ed Y sono indipendenti allora $E[XY] = E[X]E[Y]$, ma non vale il viceversa. Da questi fatti si deduce la seguente proprietà:

Proposizione 30 *Se X ed Y sono indipendenti allora $Cov(X,Y) = 0$.*

Il viceversa però non è vero: non basta verificare la singola condizione numerica $Cov(X,Y) = 0$ per dedurre l'indipendenza. Tuttavia, nella pratica, c'è una certa (e giustificata) tendenza a ritenere che la condizione $Cov(X,Y) = 0$ sia un notevole sintomo di indipendenza.

Inoltre, si può dimostrare che, se la coppia (X, Y) è gaussiana (il concetto di coppia gaussiana verrà introdotto in seguito), allora la condizione $Cov(X, Y) = 0$ implica l'indipendenza. Anche questo fatto aiuta a confondere indipendenza e covarianza nulla.

Quando, per due v.a. aleatorie X ed Y, vale $Cov(X, Y) = 0$, diciamo che sono *incorrelate* (o scorrelate). La proposizione afferma quindi che indipendenza implica non correlazione.

La covarianza è legata alla varianza della somma: vale in generale, cioè per v.a. X ed Y qualsiasi,

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov(X, Y).$$

La dimostrazione è immediata, è semplicemente la ben nota regola del quadrato della somma. Ma da questa di capisce subito come mai abbiamo affermato, in un paragrafo della scheda n. 5, che l'indipendenza tra X ed Y implica $Var[X + Y] = Var[X] + Var[Y]$. Qui abbiamo ottenuto un risultato persino un po' più generale:

Proposizione 31 *Se X ed Y sono incorrelate (in particolare se sono indipendenti), allora*

$$Var[X + Y] = Var[X] + Var[Y].$$

Tra le regole di calcolo per la covarianza segnaliamo la linearità in ciascuno dei suoi argomenti: se X, Y e Z sono tre v.a. ed a, b, c sono tre numeri reali, allora

$$Cov(aX + bY + c, Z) = aCov(X, Z) + bCov(Y, Z)$$

e lo stesso vale per $Cov(Z, aX + bY + c)$, visto che la covarianza è simmetrica nei suoi due argomenti.

La covarianza soffre dello stesso difetto della varianza: non ha l'unità di misura e l'ordine di grandezza delle v.a. originarie. Per questo e non solo per questo, si introduce il *coefficiente di correlazione* definito da

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

L'effetto della divisione per $\sigma_X \sigma_Y$ è ancora più drastico: la grandezza $\rho(X, Y)$ è 'adimensionale', ed acquista un valore assoluto, non più relativo all'unità di misura e l'ordine di grandezza tipico dei valori di X ed Y. Si può dimostrare che vale

$$-1 \leq \rho(X, Y) \leq 1.$$

Cercheremo tra un momento, tramite lo studio della regressione lineare semplice, di sviluppare un'intuizione circa il significato di valori di $\rho(X, Y)$ vicini a +1, a -1 ed a 0.

Mentre per la covarianza vale $Cov(\lambda X, \eta Y) = \lambda\eta Cov(X, Y)$, per il coefficiente di correlazione vale $\rho(\lambda X, \eta Y) = \rho(X, Y)$. Questo è un modo matematico di apprezzare l'indipendenza dall'unità di misura e dall'ordine di grandezza del coefficiente di correlazione, in contrasto con quanto accade per la covarianza.

Regressione lineare semplice

Ipotizziamo che tre v.a. X, Y ed ε siano legate dalla relazione lineare

$$Y = aX + b + \sigma\varepsilon$$

dove a, b e σ sono numeri reali ($\sigma > 0$). Interpretiamo questa scrittura pensando che X ed Y siano legate da una relazione lineare (graficamente una retta di equazione $y = ax + b$, per cui a si dirà coefficiente angolare e b intercetta), perturbata però da un *errore* casuale $\sigma\varepsilon$. La v.a. X verrà detta *input*, o *predittore*, o fattore, la Y *output*, o quantità da predire.

Supporremo sempre che ε sia standardizzato:

$$E[\varepsilon] = 0, \quad Var[\varepsilon] = 1.$$

La deviazione standard dell'errore è inglobata in σ, la sua eventuale media in b. Supporremo inoltre che ε ed X siano indipendenti o almeno incorrelate:

$$Cov(X, \varepsilon) = 0.$$

Chiameremo *modello lineare* (semplice) la relazione precedente. Diremo anche *modello di regressione lineare* (semplice), e chiameremo *retta di regressione* la retta $y = ax + b$.

Ci poniamo due scopi:

1. trovare formule che permettano di calcolare approssimativamente a, b e σ a partire da dati sperimentali, quando si ipotizza il modello lineare ma non si conoscono i coefficienti;

2. interpretare rigorosamente il concetto di coefficiente di correlazione nell'ambito del modello lineare.

Raggiungeremo entrambi gli scopi calcolando valori medi, varianze e covarianze tra le diverse grandezze in gioco. Vale, per linearità e per la proprietà $E[\varepsilon] = 0$,

$$E[Y] = aE[X] + b.$$

Vale inotre, per le regole sulla varianza (qui usiano la scorrelazione tra X ed ε),

$$Var[Y] = a^2 Var[X] + \sigma^2.$$

Infine, per analoghe ragioni vale

$$\begin{aligned} Cov(Y, X) &= Cov(aX + b + \sigma\varepsilon, X) \\ &= aCov(X, X) + \sigma Cov(\varepsilon, X) \end{aligned}$$

da cui

$$Cov(Y, X) = aVar[X].$$

Riscriviamo queste formule in modo adatto al calcolo (iterativo) dei coefficienti a partire dai valori medi:

$$\begin{aligned} a &= \frac{Cov\left(Y,X\right)}{Var\left[X\right]} \\ b &= E\left[Y\right] - aE\left[X\right] \\ \sigma^2 &= Var\left[Y\right] - a^2 Var\left[X\right]. \end{aligned}$$

Supponiamo di avere n dati sperimentali, che in questo contesto significa avere n coppie (x_1, y_1), ... , (x_n, y_n) (n individui sono stati esaminati e per ciascuno sono stati trovati i valori di due grandezze X ed Y). Possiamo calcolare i numeri

$$\begin{aligned} &\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \\ &\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2, \quad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \\ &\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \end{aligned}$$

e considerarli come approssiamzioni (stime) rispettivamente di

$$\begin{aligned} &E\left[X\right], \quad E\left[Y\right] \\ Var\left[X\right], \quad &Var\left[Y\right] \\ &Cov\left(X,Y\right). \end{aligned}$$

Tramite queste approssimazioni possiamo stimare a, b e σ.

Interpretazione di $\rho\left(X,Y\right)$

Cerchiamo di legare il coefficiente di correlazione al coefficiente angolare: vale $\frac{Cov(Y,X)}{Var[X]} = \frac{Cov(Y,X)}{\sigma_X \sigma_Y}\frac{\sigma_Y}{\sigma_X}$ quindi

$$a = \rho\left(Y,X\right)\frac{\sigma_Y}{\sigma_X}.$$

Innanzi tutto questo chiarisce che *a non è* il coefficiente di correlazione, come invece per una sorta di gioco di parole si è spesso portati a credere. Del resto, $\rho\left(Y,X\right)$ può variare solo tra -1 e 1, mentre la pendenza di una retta può essere maggiore di quella delle bisettrici.

Vale però la regola: $a > 0$ se e solo se $\rho\left(Y,X\right) > 0$ (ed analogamente per valori negativi). Quindi $\rho\left(Y,X\right) > 0$ è indice di legame lineare diretto, cioè con coefficiente angolare positivo, mentre $\rho\left(Y,X\right) < 0$ è indice di legame lineare inverso (nel senso: una variabile cresce se l'altra cala), cioè con coefficiente angolare negativo. Almeno il segno di $\rho\left(Y,X\right)$ è facilmente interpretabile.

Supponiamo di *standardizzare sia X sia Y*. In realtà non importa che sottraiamo la media, ma è essenziale che dividiamo per la deviazione standard, in modo da ricondurci ad

avere $\sigma_X = 1$ e $\sigma_Y = 1$. In questo caso

$$a = \rho(Y, X).$$

Questo può offrire un'interpretazione più stretta. In realtà però, anche così è piuttosto faticoso capire il ruolo di condizioni tipo $\rho(Y, X) = 0.9$ rispetto a $\rho(Y, X) = 0.2$.

L'interpretazione più precisa viene invece dallo studio dell'errore. Abbiamo visto sopra che

$$\sigma^2 = Var[Y] - a^2 Var[X].$$

Sostituendo $a = \rho(Y, X)\frac{\sigma_Y}{\sigma_X}$ si trova

$$\sigma^2 = Var[Y]\left(1 - \rho^2(Y, X)\right).$$

Questo dice che la varianza dell'errore, cioè la grandezza che misura quanto preciso sia il legame lineare tra X ed Y, è tanto maggiore quanto più vicino a zero è $\rho(Y, X)$: valori vicini a zero di $\rho(Y, X)$ implicano un cattivo legame lineare (errore elevato). Viceversa, valori di $\rho(Y, X)$ vicini a ± 1 (non importa il segno!), implicano σ^2 piccolo e quindi un legame lineare stretto.

Quindi, salvo che si esegua una standardizzazione di entrambe le variabili, $\rho(Y, X)$ non è legato tanto all'inclinazione della retta di regressione quanto piuttosto alla *precisione con cui essa descrive il legame tra le variabili.*

Nel ragionamento precedente bisogna osservare che la grandezza o piccolezza di σ^2 è relativa anche alla grandezza o piccolezza di $Var[Y]$. Questa è solo una questione di unità di misura delle quantità aleatorie che stiamo esaminando. Il discorso diventa indipendente dall'unità di misura e dall'ordine di grandezza dei valori tipici di Y se introduciamo la *varianza standardizzata* dell'errore:

$$\frac{\sigma^2}{Var[Y]}.$$

Per essa vale

$$\frac{\sigma^2}{Var[Y]} = 1 - \rho^2(Y, X)$$

portando ad un ragionamento più universale circa il legame tra entità dell'errore e valore di $\rho(Y, X)$.

Infine, introduciamo alcuni nuovi nomi. Essi si ispirano all'idea che con un modello lineare stiamo cercando di dare una *spiegazione della variabilità* della grandezza Y. Abbiamo una grandezza Y, essa varia in modo imprevedibile, aleatorio, e noi vorremmo capire se queste variazioni sono almeno in parte spiegabili tramite un legame lineare con un predittore X: quando osserviamo ad es. valori di Y pià grandi della media, questo non è dovuto semplicemente al caso, ma al fatto che il predittore ha assunto valori ad es. più grandi del solito (se $a > 0$). Tutto però è pur sempre corrotto dall'errore, per cui la spiegazione della variabilità di Y offerta dalla retta di regressione non è mai una spiegazione completa.

In quest'ottica, Y ha una sua varianza, una sua variabilità. L'espressione $aX + b$ riesce a spiegarne una parte, l'altra resta non spiegata. La parte non spiegata di Y è la differenza tra Y e la parte spiegata, cioè $aX + b$. Quindi la parte non spiegata di Y è proprio l'errore $\sigma\varepsilon$ (non c'è niente di nuovo, è solo una questione di linguaggio).

Con questo nuovo linguaggio, chiamiamo *varianza spiegata* la percentuale della varianza che è stata spiegata da $aX+b$ e *varianza non spiegata* la percentuale complementare. Siccome la parte di Y non spiegata è $\sigma\varepsilon$, la varianza non spiegata è

$$\frac{\sigma^2}{Var[Y]}.$$

Quindi la varianza spiegata è

$$1-\frac{\sigma^2}{Var[Y]}.$$

Ma questa è pari a $\rho^2(Y,X)$! Siamo arrivati al seguente risultato:

Proposizione 32 *Il coefficiente di correlazione al quadrato, $\rho^2(Y,X)$, è la varianza spiegata $1-\frac{\sigma^2}{Var[Y]}$ dalla relazione lineare.*

Più $\rho^2(Y,X)$ è alto (vicino a 1) più la relazione lineare riesce a spiegare la variabilità di Y.

6.3.3 Regressione lineare multipla

Supponiamo di avere una tabella di numeri del tipo

	X_1	...	X_p	Y
1	$x_{1,1}$	...	$x_{1,p}$	y_1
2	$x_{2,1}$		$x_{2,p}$	y_2
...	...			
n	$x_{n,1}$		$x_{n,p}$	y_n

dove le colonne rappresentano diverse variabili (ad esempio X_1 = reddito, ... , X_p = numero anni istruzione, Y = spese per mostre e musei), le righe rappresentano diversi "individui" (ad esempio singole persone, oppure città o regioni di una nazione) ed i valori numerici sono noti, sono stati misurati.

Ci chiediamo se le variabili $X_1,...,X_p$ influiscono su Y. Ci chiediamo se Y dipende da $X_1,...,X_p$. Un maggior reddito induce a maggiori spese per mostre e musei? Ed un maggior gradi di istruzione (lì misurato semplicemente come numero di anni si studio)?

Immaginiamo che le variabili siano legate da una relazione funzionale, a meno di errore:

$$Y=f(X_1,...,X_p)+\varepsilon.$$

Più specificamente, per semplicità, supponiamo che la relazione sia lineare:

$$Y=a_1X_1+...+a_pX_p+b+\varepsilon$$

(b è detta intercetta, e nel caso $p=1$ il coefficiente $a:=a_1$ è detto coefficiente angolare).

A partire dalla matrice dei dati, relativamente ad una scelta dei coefficienti $a_1,...,a_p,b$, possiamo calcolare i *residui*

$$\varepsilon_i=y_i-(a_1x_{i,1}+...+a_px_{i,p}+b)$$

al variare dell'individuo $i = 1, ..., n$. Possiamo poi calcolare lo scarto quadratico medio dei residui, ancora funzione dei parametri $a_1, ..., a_p, b$,

$$SQM\left(a_1, ..., a_p, b\right) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(a_1 x_{i,1} + ... + a_p x_{i,p} + b\right)\right)^2.$$

La grandezza $SQM\left(a_1, ..., a_p, b\right)$ misura la bontà del modello lineare $Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$: se piccola, il modello è buono. Allora, innanzi tutto cerchiamo i parametri $a_1, ..., a_p, b$ che la rendono minima. Essi forniscono il migliore tra i modelli lineari. Indichiamo con $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ i parametri ottimali. Chiamiamo

$$Y = \widehat{a}_1 X_1 + ... + \widehat{a}_p X_p + \widehat{b} + \varepsilon$$

il *modello di regressione lineare multipla* (*regressione lineare semplice* nel caso $n = 1$) associato alla tabella precedente. La varianza dell'errore, o dei residui, è

$$\sigma_\varepsilon^2 = SQM\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

dove

$$\widehat{\varepsilon}_i = y_i - \left(\widehat{a}_1 x_{1,i} + ... + \widehat{a}_p x_{p,i} + \widehat{b}\right).$$

La varianza spiegata, o indice R^2, è

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$$

dove σ_Y^2 è la varianza dei dati $y_1, ..., y_n$. L'idea è che i dati $y_1, ..., y_n$ hanno una loro variabilità, descritta da σ_Y^2, in situazione di completa ignoranza; ma quando abbiamo a disposizione un modello, esso spiega i dati $y_1, ..., y_n$ in una certa misura, cioè a meno degli errori $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$. Quindi la variabilità di questi errori è la variabilità inspiegata, residua. Da qui il nome di (percentuale di) varianza spiegata per il numero $1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}$.

Calcolo dei coefficienti

Il metodo con cui abbiamo definito $\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}$ si dice metodo dei *minimi quadrati*. Si possono scrivere delle formule esplicite per il calcolo di questi coefficienti. Infatti, siccome vogliamo minimizzare la funzione $SQM\left(a_1, ..., a_p, b\right)$, che per brevità indichiamo con $f\left(a_1, ..., a_p, b\right)$, deve valere

$$\frac{\partial f}{\partial a_j}\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = 0, \qquad j = 1, ..., p$$
$$\frac{\partial f}{\partial b}\left(\widehat{a}_1, ..., \widehat{a}_p, \widehat{b}\right) = 0.$$

Vale

$$\begin{aligned}\frac{\partial f}{\partial a_j} &= -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (a_1 x_{i,1} + ... + a_p x_{i,p} + b)\right) x_{i,j} \\ &= -\frac{2}{n}\left\langle y, x_j\right\rangle + \frac{2}{n} a_1 \left\langle x_{\cdot,1}, x_{\cdot,j}\right\rangle + ... + \frac{2}{n} a_p \left\langle x_{\cdot,p}, x_{\cdot,j}\right\rangle + 2b\overline{x}_j\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial b} &= -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - (a_1 x_{1,i} + ... + a_p x_{p,i} + b)\right) \\ &= -2\overline{y} + 2a_1\overline{x}_1 + ... + 2a_p\overline{x}_p + 2b\end{aligned}$$

dove $\overline{y}$ è la media degli $y_1, ..., y_n$ e, per ciascun $j = 1, ..., p$, $\overline{x}_j$ è la media degli $x_{1,j}, ..., x_{n,j}$; ed inoltre abbiamo posto

$$\left\langle y, x_{\cdot,j}\right\rangle = \sum_{i=1}^{n} y_i x_{i,j}, \qquad \left\langle x_{\cdot,k}, x_{\cdot,j}\right\rangle = \sum_{i=1}^{n} x_{i,k} x_{i,j}.$$

Quindi deve valere

$$\begin{gathered}a_1 \left\langle x_{\cdot,1}, x_{\cdot,j}\right\rangle + ... + a_p \left\langle x_{\cdot,p}, x_{\cdot,j}\right\rangle + nb\overline{x}_j = \left\langle y, x_{\cdot,j}\right\rangle, \qquad i = 1, ..., p \\ a_1\overline{x}_1 + ... + a_p\overline{x}_p + b = \overline{y}\end{gathered}$$

Questo è un sistema di $p+1$ equazioni lineari in $p+1$ incognite, che il software risolve con facilità.

Si può anche introdurre la matrice A quadrata a $p+1$ righe e colonne ed il vettore w

$$A = \begin{pmatrix} \left\langle x_{\cdot,1}, x_{\cdot,1}\right\rangle & ... & \left\langle x_{\cdot,p}, x_{\cdot,1}\right\rangle & \left\langle x_{\cdot,1}, 1\right\rangle \\ ... & ... & ... & ... \\ \left\langle x_{\cdot,1}, x_{\cdot,p}\right\rangle & ... & \left\langle x_{\cdot,p}, x_{\cdot,p}\right\rangle & \left\langle x_{\cdot,p}, 1\right\rangle \\ \left\langle x_{\cdot,1}, 1\right\rangle & ... & \left\langle x_{\cdot,p}, 1\right\rangle & 1 \end{pmatrix}, \qquad w = \begin{pmatrix} \left\langle y, x_{\cdot,1}\right\rangle \\ ... \\ \left\langle y, x_{\cdot,p}\right\rangle \\ \left\langle y, 1\right\rangle \end{pmatrix}$$

dove 1 indica il vettore con tutti "1", ed y è il vettore delle y_i. Allora il calcolo del vettore

$$v = \begin{pmatrix} \widehat{a}_1 \\ ... \\ \widehat{a}_p \\ \widehat{b} \end{pmatrix}$$

si può vedere come la risoluzione di

$$Av = w.$$

Infine, la matrice A si ottiene col prodotto

$$A = X^T X$$

dove

$$X = \begin{pmatrix} x_{1,1} & ... & x_{1,p} & 1 \\ x_{2,1} & & x_{2,p} & 1 \\ ... & & & \\ x_{n,1} & & x_{n,p} & 1 \end{pmatrix}.$$

Si noti che questa è la matrice iniziale dei dati dove al posto delle y_i abbiamo messo 1. Inoltre,

$$w = X^T y.$$

Quindi il problema ha la forma

$$X^T X v = X^T y.$$

Il caso in cui le X_i sono aleatorie

Quando scriviamo il modello

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

abbiamo due opzioni di ragionamento, entrambe perseguibili ed utili.

La prima è che i valori assunti dalle variabili X_i siano deterministici, ad esempio fissati dallo sperimentatore che vuole esaminare l'effetto di queste variabili sulla Y. In questo caso, solo la Y è, in un certo senso, aleatoria, se pensiamo aleatorio l'errore ε.

Oppure, possiamo immaginare che le X_i siano aleatorie quanto ε (e quindi Y) e noi eseguiamo misurazioni di tutte queste grandezze aleatorie. In questa seconda ottica, ha senso eseguire il seguente calcolo.

Calcoliamo la covarianza tra Y ed X_j:

$$\begin{aligned} Cov(Y, X_j) &= Cov(a_1 X_1 + ... + a_p X_p + b + \varepsilon, X_j) \\ &= a_1 Cov(X_1, X_j) + ... + a_p Cov(X_p, X_j) + Cov(\varepsilon, X_j). \end{aligned}$$

Ricordiamo che la matrice $Q = (Cov(X_i, X_j))$ è detta matrice di covarianza del vettore aleatorio X. Supponiamo che ε sia indipendente (o almeno scorrelato) da ciascuna X_j. Troviamo

$$Cov(Y, X_j) = \sum_{i=1}^{p} Q_{ij} a_i = \sum_{i=1}^{p} Q_{ji} a_i$$

(ricordiamo che Q è simmetrica). Detto c il vettore di coordinate $Cov(Y, X_j)$ ed a il vettore di coordinate a_i abbiamo trovato

$$Qa = c.$$

Quindi

$$a = Q^{-1} c.$$

Questo risultato fornisce un modo per calcolare i coefficienti a_i a partire da una matrice di dati. Si calcola la matrice di covarianza *empirica* $\widehat{Q}$ della matrice di dati riguardante le variabili X_i, si calcola il vettore $\widehat{c}$ delle covarianze empriche tra le variabili X_j ed Y, e si calcolano i valori

$$\widehat{a} = \widehat{Q}^{-1} \widehat{c}.$$

Poi, per calcolare b, serve una nuova equazione. Essa si trova semplicemente calcolando il valor medio a destra e sinistra dell'equazione che definisce il modello:

$$E[Y] = a_1 E[X_1] + ... + a_p E[X_p] + b.$$

Allora vale

$$b = E[Y] - (a_1 E[X_1] + ... + a_p E[X_p])$$

da cui si può calcolare un valore empirico $\widehat{b}$ a partire dai dati.

Con un po' di sforzo si potrebbe riconoscere che il risultato è identico a quello ottenuto sopra con i minimi quadrati. Ci si chiede allora: cosa ha sostituito, qui, la richiesta fatta sopra che i valori ottimali fossero quelli che minimizzavano SQM? L'indipendenza tra ε e le X_i. Se per certi valori dei parametri risulta ε indipendente dalle X_i, significa che il modello è buono, nel senso che abbiamo sfruttato nel modo migliore le informazioni contenute nelle X_i, senza ritrovarci parte di quelle informazioni nel resto ε. Il resto contiene ciò che non siamo riusciti a spiegare, e non è nullo, ma l'importante è che non contenga residui legati alle variabili X_i, altrimenti significa che c'era un'altro modo di usare le X_i più efficiente.

6.3.4 Predizione con modelli regressivi

Nel Paragrafo 6.1.1 abbiamo sottolineato come l'evidenza di elevata correlazione non implichi in alcun modo l'esitenza di un legame causa-effetto. Quando impostiamo un modello di regressione, invece, stiamo ipotizzando che certe variabili giochino il ruolo di fattori, predittori, ed altre di output. Mentre nel calcolo della correlazione e nella PCA le variabili sono esaminate tutte in modo simmetrico, la simmetria è rotta a priori da noi quando impostiamo un modello regressivo.

Dobbiamo pertanto assicurarci che esista davvero una tale relazione causa-effetto, quando facciamo la regressione? Altrimenti non ha senso farla? Dipende dagli scopi. Se lo scopo del modello regressivo è solo quello di effettuare previsioni e non di sostenere l'esistenza di una relazione causa-effetto che magari non hanno alcun senso, allora va benissimo applicare i metodi di regressione anche a variabili che non hanno un legame causa-effetto, ma sono semplicemente ben correlate.

Cosa si intende per "predizione con modelli regressivi": sulla base di dati noti, si identifica quantitativamente il legame tra certi fattori $X_1, ..., X_d$ ed una variabile da predire Y (si calcolano i coefficienti del modello); in queste situazioni note, si conoscono i valori assunti da tutte le variabili $(X_1, ..., X_d, Y)$; poi, si applica il modello a situazioni nuove, dove si conoscono solo i valori assunti dalle variabili $(X_1, ..., X_d)$, usanto il modello per calcolare (predire) il valore di Y.

Bene, ha perfettamente senso applicare questa strategia anche quando le variabili in gioco non sono legate da relazioni causa-effetto; basta che sia buona la loro correlazione.

A parte una banale logica strumentale (il modello funziona bene come scatola nera per fare predizioni), ci può essere una logica dietro questo fatto? Pensiamo al caso in cui due variabili X ed Y sono molto correlate, ma non c'è relazione causa-effetto; e supponiamo invece che ci sia una variabile Z che sia causa di entrambe, ma non la possiamo misurare, magari non l'abbiamo nemmeno individuata. Troviamo la formula regressiva $Y = aX + b$ (che non ha alcun significato fisico/economico ecc.). In situazioni nuove, misurata X, prevediamo che

Y abbia il valore dato da questa formula. Che logica c'è dietro? Perché speriamo che questa equazione fornisca prodizioni sensate, se X non influenza Y? La ragione è che, se misuriamo valori di X di un certo tipo, questi sono stati causati da certi valori di Z (sconosciuti), che hanno prodotto contemporaneamente certi valori di Y, compatibili con la formula $Y = aX+b$ in quanto questa è stata determinata da valori sperimentali che seguivano la stessa logica. In un certo senso, è come se scomponessimo la freccia

$$X \longrightarrow Y$$

in due passi

$$X \longrightarrow Z \longrightarrow Y$$

dove la freccia $X \longrightarrow Z$ va intesa come l'inversione della relazione di causa-effetto $Z \longrightarrow X$ (si ricostruisce la causa che ha provocato un certo effetto). La scomposizione ovviamente è solo ideale, non conosciamo Z, quindi non svolgiamo realmente questi passaggi. Quella che abbiamo illustrato serve solo a spiegare come mai possiamo sperare che la conoscenza di X possa dire qualcosa su Y.

A livello teorico questa spiegazione ha almeno un difetto: ipotizzando che Z influisca su X, se diversi valori di Z producono lo stesso valore di X ma non di Y, dall'osservazione di quel valore di X non è possibile risalire a quale Z lo ha prodotto, e quindi quale Y debba essere poi generato. In altre parole, la freccia

$$X \longrightarrow Z$$

potrebbe essere multivoca, quindi potrebbe non essere possibile ottenere univocamente Y da X. Un po' l'errore nel modello $Y = aX + b + \sigma\varepsilon$ può farsi carico di questo, ma non oltre una certa misura.

6.3.5 Analisi fattoriale

Qualche calcolo a mano sull'Analisi Fattoriale

Consideriamo alcuni esempi semplicissimi di Analisi Fattoriale (FA, Factorial Analysis), col solo scopo di far capire alcune idee strutturali del problema.

Consideriamo il modello

$$Y_1 = a_1 X + b_1 + \varepsilon_1$$
$$Y_2 = a_2 X + b_2 + \varepsilon_2$$

cioè un fattore e due output. Ma immaginiamo di avere dati solo delle variabili (Y_1, Y_2). Anzi, X non sappiamo nemmeno a priori cosa sia, che variabile sia, se ci sia. E' possibile risalire alla X, ai coefficienti del modello?

Il problema è quello descritto sopra: misuriamo due variabili Y_1, Y_2, magari ben correlate, ma che la logica ci dice non essere in relazione causa-effetto. Ci chiediamo invece se ci sia, alle loro spalle, a monte di esse, una variabile X che sia loro causa, che le "spieghi", nel senso che spieghi come mai Y_1 ed Y_2 variano in modo coordinato (sono correlate). Si tratta

di *spiegare le variazioni (coordinate) degli output*. In termini matematici, *spiegare la matrice di covarianza* Q_Y di Y.

Abbiamo enfatizzato il problema illustrando il caso in cui X sia causa di Y_1 e Y_2, ma non è necessario che sia proprio così. Magari si tratta solo di rintracciare una variabile riassuntiva X, di cui Y_1 e Y_2 siano manifestazioni misurabili. Ad esempio X può essere il grado di benessere economico, e le Y_i essere varie misurazioni di indicatori di benessere (spese per cultura, per vacanze ecc.).

Si noterà che in tutti i nostri esempi prendiamo sempre meno fattori che output, altrimenti varrebbe la risposta banale: un fattore per ogni output. Se accettassimo di cercare un numero di fattori pari (o addirittura superiore) agli output, fattori che spieghino gli output, la risposta banale sarebbe prendere come fattori gli output stessi. Essi spiegherebbero perfettamente tutta la variabilità degli output. Solo imponendo il vincolo che i fattori sono di meno, sopravvive un problema non ovvio di spiegare delle variazioni coordinate degli output.

Se abbiamo una tabella di dati per le variabili $Y = (Y_1, Y_2)$ calcoliamo dai dati la matrice di correlazione

$$Q_Y = \begin{pmatrix} \sigma^2_{Y_1} & Cov\,(Y_1, Y_2) \\ Cov\,(Y_1, Y_2) & \sigma^2_{Y_2} \end{pmatrix}.$$

Non abbiamo altro (eventualmente i valori medi delle Y_1, Y_2) per tentare di risalire al modello.

A livello teorico, se vale un modello di questo genere, con $\varepsilon_1, \varepsilon_2, X$ indipendenti (ricordiamo che questa richiesta, nella regressione, rimpiazzava la minimizzazione dei quadrati), vale

$$\begin{aligned} Cov\,(Y_1, Y_2) &= a_1 a_2 \sigma^2_X \\ \sigma^2_{Y_1} &= a_1^2 \sigma^2_X + \sigma^2_{\varepsilon_1} \\ \sigma^2_{Y_2} &= a_2^2 \sigma^2_X + \sigma^2_{\varepsilon_2}. \end{aligned}$$

Supponiamo $\sigma^2_X = 1$ altrimenti questa grandezza la si fa rientrare nei coefficenti incogniti a_1 e a_2. Quindi

$$\begin{aligned} a_1 a_2 &= Cov\,(Y_1, Y_2) \\ a_1^2 + \sigma^2_{\varepsilon_1} &= \sigma^2_{Y_1} \\ a_2^2 + \sigma^2_{\varepsilon_2} &= \sigma^2_{Y_2}. \end{aligned}$$

Sono tre equazioni nelle quattro incognite $(a_1, a_2, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2})$. Ci sono quindi (almeno in linea di principio, visto che è un problema nonlineare, quindi non del tutto banale) infinite soluzioni. Il software cerca quella che rende minima la somma dei residui $\sigma^2_{\varepsilon_1} + \sigma^2_{\varepsilon_2}$.

Se però erano tre output ed un solo fattore, cioè il modello

$$\begin{aligned} Y_1 &= a_1 X + b_1 + \varepsilon_1 \\ Y_2 &= a_2 X + b_2 + \varepsilon_2 \\ Y_3 &= a_3 X + b_3 + \varepsilon_3 \end{aligned}$$

avevamo

$$\begin{aligned} Cov\,(Y_1, Y_2) &= a_1 a_2 \\ Cov\,(Y_1, Y_3) &= a_1 a_3 \\ Cov\,(Y_2, Y_3) &= a_2 a_3 \\ \sigma^2_{Y_1} &= a_1^2 + \sigma^2_{\varepsilon_1} \\ \sigma^2_{Y_2} &= a_2^2 + \sigma^2_{\varepsilon_2} \\ \sigma^2_{Y_3} &= a_3^2 + \sigma^2_{\varepsilon_3}. \end{aligned}$$

Sono 6 equazioni nelle 6 incognite $(a_1, a_2, a_3, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3})$ per cui in linea di principio c'è una sola soluzione. Con 4 output certamente è sovradeterminato; in questi casi, di non risolubilità, il criterio è costruire con i parametri $(a_1, a_2, a_3, \sigma_{\varepsilon_1}, \sigma_{\varepsilon_2}, \sigma_{\varepsilon_3})$ una matrice di covarianza più vicina possibile (in una certa metrica) alla matrice Q_Y.

Vediamo anche il caso di due fattori e tre output:

$$\begin{aligned} Y_1 &= a_{11} X_1 + a_{12} X_2 + b_1 + \varepsilon_1 \\ Y_2 &= a_{21} X_1 + a_{22} X_2 + b_2 + \varepsilon_2 \\ Y_3 &= a_{31} X_1 + a_{32} X_2 + b_3 + \varepsilon_3 \end{aligned}$$

Qui vale, sempre prendendo i fattori standardizzati, e supponendoli indipendenti tra loro e dagli errori,

$$\begin{aligned} Cov\,(Y_1, Y_2) &= a_{11} a_{21} + a_{12} a_{22} \\ Cov\,(Y_1, Y_3) &= a_{11} a_{31} + a_{12} a_{32} \\ &ecc. \end{aligned}$$

cioè 6 equazioni in 9 incognite. Il principio è sempre lo stesso.

6.3.6 Forma matriciale del problema

Si può sintetizzare tutto con le matrici. Immaginiamo le variabili Y_i raccolte nel vettore aleatorio $Y = (Y_1, ..., Y_n)$, le X_i nel vettore $X = (X_1, ..., X_d)$, gli errori ε_i nel vettore $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$, quindi

$$Y = AX + b + \varepsilon.$$

Con calcoli simili a quelli precedenti o a quelli del capitolo 1 delle note in inglese, nelle ipotesi di indipendenza e standardizzazione di X dette sopra, si ottiene la relazione

$$Q_Y = AA^T + Q_\varepsilon$$

che ricorda la ben nota relazione $Q_Y = AQ_X A^T$. Qui Q_X è l'identità in quanto X ha componenti indipedenti e standard (è come un vettore gaussiano standard). Invece Q_ε è la covarianza del rumore, matrice diagonale (a causa dell'indipendenza dei rumori)

$$Q_\varepsilon = \begin{pmatrix} \sigma^2_{\varepsilon_1} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sigma^2_{\varepsilon_n} \end{pmatrix}.$$

Si noti che, nella bae di partenza, la matrice Q_Y non è diagonale, mentre Q_ε sì. Se cambiassimo base diagonalizzando Q_Y, perderemmo la diagonalità di Q_ε, quindi questo non è un passaggio utile.

Il problema visto sopra allora si può riassumere così: data Q_Y e date le dimensioni $d < n$, trovare una matrice A con n righe e d colonne ed una matrice diagonale Q_ε tali che $Q_Y = AA^T + Q_\varepsilon$. Se accettassimo $d = n$, basterebbe prendere $A = \sqrt{Q_Y}$, $Q_\varepsilon = 0$. Ma questo non è possibile: $d < n$.

A seconda dei valori di d ed n, il problema è risolubile univocamente, oppure per infinite matrici A (ed in tal caso si minimizza la varianza globale dell'errore), oppure non è risolubile esattamente, nel qual caso si cercano A e Q_ε tali che

$$d\left(Q_Y, AA^T + Q_\varepsilon\right)$$

sia minima, avendo indicato con $d(.,.)$ un'opportuna distanza tra matrici.

6.3.7 Loadings, rotazioni, interpretazioni

La matrice A è detta matrice dei *loadings*, esattamente come per PCA.

Supponiamo di aver trovato una soluzione (A, Q_ε) del problema $Q_Y = AA^T + Q_\varepsilon$. Sia U una matrice ortogonale, un cambio di base, una *rotazione*, tale che UU^T è l'identità. Allora (AU, Q_ε) è un'altra soluzione:

$$(AU)(AU)^T + Q_\varepsilon = AUU^TA^T + Q_\varepsilon = AA^T + Q_\varepsilon = Q_Y.$$

In termini di modello $Y = AX + b + \varepsilon$ si tratta di averlo scritto nella forma

$$\begin{aligned} Y &= (AU)X' + b + \varepsilon \\ X' &= U^TX. \end{aligned}$$

In altre parole, "ruotando" i fattori e modificando A, si risolve il problema nello stesso modo. Che vantaggio può avere una soluzione rispetto ad un'altra, che differiscano per una rotazione? Se si riesce a trovare una rotazione in cui A sia particolarmente ricca di zeri (o valori molto piccoli), questo può venire a vantaggio di una buona *interpretazione* dei fattori, dell'attribuire un significato ai fattori. Ragioniamo su un esempio.

Si pensi all'esempio della lezione 22. Suggestionati dall'analisi svolta con PCA, che suggerisce la presenza di due fattori, immaginiamo ci siano appunto due fattori X_1, X_2 che influenzano le quattro variabili TD, RD, PE, HC, che spiegano la particolare struttura di variabilità di queste grandezze tra le nazioni europee:

$$\begin{aligned} TD &= a_{11}X_1 + a_{12}X_2 + b_1 + \varepsilon_1 \\ RD &= a_{21}X_1 + a_{22}X_2 + b_2 + \varepsilon_2 \\ &\textit{ecc.} \end{aligned}$$

Immaginiamo di eseguire una FA e di trovare una soluzione (A, Q_ε). Il software, nel calcolare (A, Q_ε), ovviamente ignora ogni possibile interpretazione applicativa (per il SW, che si parli di TD o di risultati calcistici è la stessa cosa). Quindi, a priori, il SW non aiuta lo studioso

a dare un'interpretazione dei risultati. Ma supponiamo che tramite una rotazione si ottenga una matrice A con numeri nettamente distinti, in numeri grandi e numeri piccoli. Un loading piccolo significa che c'è poca relazione tra il fattore e la variabile che esso lega. Ad esempio, se venisse che a_{11} è piccolo, vorrebbe dire che il fattore X_1 non è legato a TD, ma serve per spiegare le altre variabili. Questo minore o maggiore grado di associazione di un fattore a certe variabili può contribuire a dare un nome, un significato, a quel fattore.

6.3.8 FA e PCA

Di fatto, PCA effettua operazioni vagamente simili a quelle di FA, e quindi risulta un buon strumento per l'identificazione di fattori comuni. Il metodo PCA si può vedere come la diagonalizzazione della matrice di covarianza Q_Y:

$$Q_Y = UDU^T$$

dove

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}.$$

Ora, delle n dimensioni, prendiamo le prime d, prime secondo l'ordine degli autovalori λ_i, cioè le prime d componenti principali (es. $d = 2$). Indichiamo con $V = (V_1, ..., V_d)$ le coordinate rispetto alla base $e_1, ..., e_d$ dei primi d autovettori di Q_Y, e con $\varepsilon = (\varepsilon_{d+1}, ..., \varepsilon_n)$ le rimanenti coordinate. Vale (la prima uguaglianza è il semplice cambio di base)

$$Y = U \begin{pmatrix} X \\ \varepsilon \end{pmatrix} = AX + B\varepsilon$$

dove A è la matrice fatta dalle prime d colonne di U, B dalle ultime $n - d$. In altre parole, Anche PCA può essere visto come un modello lineare, della forma

$$\begin{aligned} Y_1 &= a_{11}V_1 + ... + a_{1d}V_d + r_1 \\ &... \\ Y_n &= a_{n1}V_1 + ... + a_{nd}V_d + r_n \end{aligned}$$

dove il vettore $r = (r_1, ..., r_n)$ dei residui è dato da

$$r = B\varepsilon.$$

Le prime componenti principali $V = (V_1, ..., V_d)$ giocano lo stesso ruolo dei fattori della FA. Il resto $r = B\varepsilon$ è, in molti esempi, piccolo, in quanto le ultime componenti principali sono quelle con minore varianza.

Ricordiamo che PCA si applica a dati standardizzati. Per questo non compare b.

L'unica differenza tra il modello $Y = AV + B\varepsilon$ offerto da PCA ed il modello $Y = AX + \varepsilon$ della FA (omettiamo b anche in FA per fare il confronto), sta nel fatto che nella FA si richiede che gli errori siano indipendenti. Invece gli errori che saltano fuori da PCA sono piccoli ma non necessariamente indipendenti. Per questo, i risultati dei due metodi non coincidono. Però è difficile incontrare problemi in cui differiscano in modo nettissimo.

6.3.9 I comandi di R. Linguaggio

Si cerchino i comandi `factanal` e `varimax`.

Communality of variable Y_1: $a_{11}^2 + ... + a_{1d}^2$. Uniqueness of variable Y_1: $\sigma_{\varepsilon_1}^2$.

Quindi la varianza di Y_1 è la somma dei due. In altre parole, il rapporto tra la communality e la varianza di Y_1 è la varianza spiegata dal modello, relativamente alla variable Y_1.

Lo stesso vale per tutte le variabili in output.

6.4 Metodi di classificazione e clustering

6.4.1 Regressione logistica

Definizione 57 *Un modello di regressione logistica tra p fattori $X_1, ..., X_p$ ed un output Y è una relazione del tipo*

$$Y \sim B(1, p)$$
$$g(p) = a_1 X_1 + ... + a_p X_p + b.$$

Abbiamo sintetizzato il modello in una definizione concisa perché questo concetto risulta in genere particolarmente oscuro e impreciso. Ora però cerchiamo di capirlo in modo più progressivo.

Come in tutti i modelli regressivi, anche nella regressione logistica ci sono dei fattori $X_1, ..., X_p$ misurabili, ed un output Y anch'esso misurabile, tutti relativamente ad un insieme di unità sperimentali. Tuttavia, nella regressione logistica l'output Y è dicotomico: 0 o 1, mentre i predittori assumono valori reali generici, come nella regressione lineare multipla tradizionale.

Si pensa che, dati i valori dei predittori, l'output $Y \in \{0, 1\}$ sia casuale ma con legge univocamente determinata dai predittori. Y è una v.a. di Bernoulli, quindi la sua legge è identificata dal parametro $p = P(Y = 1)$. Questo numero è univocamente determinato dai predittori, è funzione deterministica dei valori assunti dai predittori.

Inoltre, il modo di dipendere dai predittori, nel modello di regressione logistica, non è qualsiasi ma avviene solo attraverso una loro combinazione affine, detta *predittore lineare*

$$\eta = a_1 X_1 + ... + a_p X_p + b.$$

Non stiamo affermando che $p = \eta$, ma che p dipende da $X_1, ..., X_p$ solo attraverso una combinazione affine η di questo tipo, e non tramite espressioni magari quadratiche o altro.

Mettiamo a confronto regressione logistica (RLog) e regressione lineare multipla (RLM) tradizionale, per spiegare meglio il modello RLog. Nella RLM, dati i valori $x_1, ..., x_p$ dei predittori, noti i coefficienti $a_1, ..., a_p, b$, l'output è una v.a. gaussiana Y di media $\mu = \eta$ (media uguale al predittore lineare) e varianza σ^2, quindi rappresentabile nella forma

$$Y = a_1 x_1 + ... + a_p x_p + b + \varepsilon$$

con $\varepsilon \sim N(0, \sigma^2)$. Invece, nella RLog, dati i valori $x_1, ..., x_p$ dei predittori, noti i coefficienti $a_1, ..., a_p, b$, l'output è una v.a. di Bernoulli Y, di parametro p che dipende da η attraverso una certa funzione.

Pensiamo ad un esempio: supponiamo che gli individui siano le nazioni europee e che, per una certa nazione, sia $Y = 1$ se la nazione migliora la propria condizione economica ($Y = 0$ altrimenti) durante l'anno 2011. I predittori potrebbero essere gli investimenti in ricerca, e così via del 2010. Noti i valori dei predittori, la casualità non è certo esaurita, quindi Y resta aleatorio, ma la sua legge (cioè p) è ora determinata, nota. Nel modello RLog si suppone che la probabilità p di miglioramento sia nota quando sono noti i predittori. Inoltre si suppone che p dipenda dai predittori solo attraverso la loro combinazione affine η.

Un altro esempio: gli individui sono esemplari di complessi sistemi meccanici o elettronici, $Y = 1$ se il sistema funziona per un anno, i predittori possono essere valori misurati di caratteristiche meccaniche ecc. di sottoparti, del materiale ecc.

Essendo p una probabilità, non possiamo pensare che la relazione tra p ed η sia del tipo $p = \eta$, cioè

$$p = a_1 x_1 + ... + a_p x_p + b$$

altrimenti otterremmo per p valori anche esterni a $[0, 1]$. Si deve adottare un modello del tipo

$$g(p) = a_1 x_1 + ... + a_p x_p + b$$

dove g è una funzione definita in $[0, 1]$ a valori reali, invertibile. In modo che sia

$$p = g^{-1}(a_1 x_1 + ... + a_p x_p + b).$$

Una scelta molto comune è la funzione detta *logit*

$$g(p) = \log\left(\frac{p}{1-p}\right).$$

Per $p \to 0$ essa tende a $-\infty$, mentre per $p \to 1$ tende a $+\infty$; ed è strettamente crescente, oltre che regolare. La sua funzione inversa è

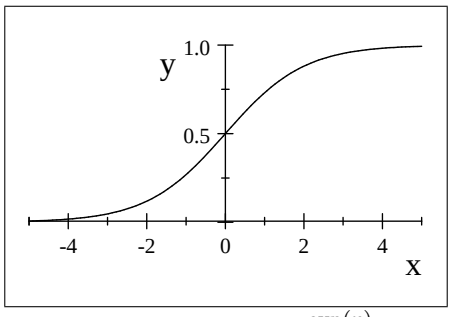
$$g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

[Infatti $\log\left(\frac{p}{1-p}\right) = \eta$, $\frac{p}{1-p} = \exp(\eta)$, $p = (1-p)\exp(\eta)$, $p(1 + \exp(\eta)) = \exp(\eta)$, $p = \frac{\exp(\eta)}{1+\exp(\eta)}$.] In definitiva, il modello è

$$Y \sim B(1, p) \text{ con } p = \frac{\exp(\eta)}{1 + \exp(\eta)} \text{ dove } \eta = a_1 x_1 + ... + a_p x_p + b.$$

Quando i coefficienti $a_1, ..., a_p, b$ sono divenuti noti, preso un nuovo individuo, calcolati i valori dei suoi predittori $x_1, ..., x_p$, si calcola la probabilità p relativa a quell'individuo (probabilità di questo o quell'accadimento, dipende dal problema). Se p è molto elevata, siamo abbastanza sicuri che per quell'individuo sarà $Y = 1$, mentre se è molto bassa, conteremo che sia $Y = 0$; nel mezzo ovviamente c'è molta indecisione sul valore di Y di quell'individuo, pur valendo comunque che se $p > 1/2$ è più probabile $Y = 1$ e viceversa.

Nella teoria generale dei modelli lineari generalizzati, il numero $\eta = a_1 x_1 + ... + a_p x_p + b$ viene detto *predittore lineare*, la funzione g^{-1} viene detta *link function* e la funzione g viene detta *mean function*. Nella regressione logistica, la link function è la funzione logistica, rappresentata in figura.



Funzione logistica $\frac{\exp(\eta)}{1+\exp(\eta)}$.

Resta il problema di trovare i coefficienti. Si devono avere n individui di cui si conoscano i valori dei predittori X_i e di Y. Si usa il metodo della *massima verosimiglianza*. Noti i valori $x_1, ..., x_p$ dei predittori di un individuo, abbiamo detto che Y è $B(1,p)$, con $p = g^{-1}(\eta)$, $\eta = a_1x_1 + ... + a_px_p + b$. Quindi $P(Y=1) = p$, $P(Y=0) = 1-p$. Se indichiamo uno dei due numeri 0 o 1 con y, si può scrivere in una sola formula

$$P(Y=y) = p^y (1-p)^{1-y}.$$

Supponiamo come abbiamo detto che, per un individuo noto, sia noto anche il valore di Y, che chiamiamo con y. Il numero $p^y (1-p)^{1-y}$ è la verosimiglianza relativa a quell'individuo. In astratto, la verosimiglianza è funzione di molte grandezze: $x_1, ..., x_p$, y, $a_1, ..., a_p, b$. Trattandosi di un individuo con $x_1, ..., x_p$, y noti, ben precisi, la verosimiglianza è funzione di $a_1, ..., a_p, b$. Se poi consideriamo gli n individui indipendenti, ed indichiamo con $x_1^{(i)}, ..., x_p^{(i)}$, $y^{(i)}$ i loro valori noti, vale

$$P\left(Y^{(1)} = y^{(1)}, ..., Y^{(n)} = y^{(n)}\right) = \prod_{i=1}^{n} \left(p^{(i)}\right)^{y^{(i)}} \left(1-p^{(i)}\right)^{1-y^{(i)}}$$

dove

$$p^{(i)} = g^{-1}\left(\eta^{(i)}\right), \qquad \eta^{(i)} = a_1x_1^{(i)} + ... + a_px_p^{(i)} + b.$$

Questa è la verosimiglianza del campione sperimentale, funzione di $a_1, ..., a_p, b$. Il metodo di massima verosimiglianza consiste nel cercare i valori di $a_1, ..., a_p, b$ che rendono massima la verosimiglianza, cioè $\prod_{i=1}^{n} \left(p^{(i)}\right)^{y^{(i)}} \left(1-p^{(i)}\right)^{1-y^{(i)}}$. Tecnicamente, conviene massimizzare il logaritmo della verosimiglianza (è equivalente), cioè

$$\sum_{i=1}^{n} \left(y^{(i)} \log p^{(i)} + \left(1-y^{(i)}\right) \log\left(1-p^{(i)}\right)\right).$$

Il software esegue la massimizzazione con un procedimento iterativo.

Classificazione tramite regressione logistica

Il metodo della regressione logistica serve ad esempio per effettuare una classificazione *non perentoria*. L'output Y può assumere due valori, che per comodità espositiva chiamiamo A e B. Assume A con probabilità p.

A partire da un set di dati, cioè di individui di cui si conoscano sia i predittori sia la classe, A o B, a cui appartengono, si calcolano i coefficienti del modello. Poi, esaminando nuovi individui che vogliamo classificare sulla base della sola conoscenza dei predittori, calcoliamo il numero p di un individuo, ed assegnamolo alla classe che ha probabilità maggiore (A se $p > 1/2$). Eseguita così è una classificazione perentoria, ma è corredata dal numero p stesso, che fornisce un'indicazione del grado di sicurezza che abbiamo, nella classificazione appena eseguita.

E' la stessa logica della predizione tramite modello regressivo, Paragrafo 6.3.4. Invece che desiderare una predizione numerica di una grandezza Y associata a certe unità sperimentali, desideriamo sapere se quelle unità appartengono ad una categoria o ad un'altra.

Modelli lineari generalizzati

Sono una generalizzazione del modello base, gaussiano,

$$Y = a_1 X_1 + ... + a_p X_p + b + \varepsilon$$

e di quello bernoulliano (detto modello binomiale, in questo contensto) appena visto

$$Y \sim B(1, p) \text{ con } p = \frac{\exp(\eta)}{1 + \exp(\eta)} \text{ dove } \eta = a_1 x_1 + ... + a_p x_p + b.$$

In generale, si ipotizza che l'output Y abbia distribuzione di una certa classe, ad esempio appunto gaussiana, Bernoulli, Poisson, ecc., e si ipotizza che un suo parametro fondamentale θ, di solito la media (μ per la gaussiana, p per la Bernoulli, λ per la Poisson) sia legato ai fattori attraverso una formula del tipo

$$\theta = g^{-1}(\eta)$$

dove

$$\eta = a_1 x_1 + ... + a_p x_p + b$$

è chiamiato *predittore lineare*. Chiamiamo *link function* la funzione g. Nella regressione logistica si prende come g la funzione logit. Nella regressione tradizionale, g è l'identità.

Il comando di R che esegue la regressione per i modelli lineari generalizzati è `glm`.

6.4.2 Formulazione probabilistica del problema decisionale e regola di Bayes

Per capire il prossimo metodo di classificazione, è utile qualche premessa di teoria delle decisioni.

L'idea base della teoria delle decisioni si può descrivere tramite le nozioni forndamentali del calcolo delle probabilità: l'universo degli eventi, le partizioni, la formula di fattorizzazione e quella di Bayes.

Supponiamo di avere un universo Ω, una partizione (C_k) (ad esempio la suddivisione di Ω in un insieme C_1 ed il suo complementare $C_2 = C_1^c$), e dobbiamo prendere una decisione: quale degli eventi C_k si è verificato (o si verificherà)? Abbiamo usato la lettera C come "classe", immaginando di voler effettuare una classificazione.

Supponiamo di conoscere le (cosidette) *probabilità a priori* dei C_k, i numeri $P(C_k)$. A volte sono note da statistiche precedenti (come nell'esempio 1 che vedremo tra poco), altre volte, più che conoscerle, le si ipotizza. Ad esempio, a volte si suppongono tutte uguali (C_k equiprobabili a priori) per sottolineare il nostro grado di ignoranza iniziale circa quale dei C_k si quello giusto.

Ipotizziamo che gli eventi C_k influiscano su (o comunque siano collegati ad) un evento A che possiamo osservare e che vediamo che si è verificato. Supponiamo di conoscere le probabilità condizionali

$$P(A|C_k)$$

per tutti i k. Tramite il teorema di Bayes, allora, possiamo calcolare le *probabilità a posteriori* dei B_k, i numeri

$$P(C_i|A) = \frac{P(A|C_i)\,P(C_i)}{\sum_k P(A|C_k)\,P(C_k)}.$$

Queste sono le probabilità dei C_k nel momento in cui sappiamo che l'evento A si è verificato.

La *regola decisionale di Bayes* è: scegliere tra i C_k quello con la *massima probabilità a posteriori*. In simboli: $C_i^{opt} := \arg\max_{C_i} P(C_i|A)$, ovvero

$$C_i^{opt} := \arg\max_{C_i} P(A|C_i)\,P(C_i)$$

in quanto il denominatore è uguale per tutti i $P(C_i|A)$. Va notato che, se pur in casi plausibilmente rari, potrebbero esistere due diversi C_i che massimizzano questa espressione. In questo caso il metodo non è in grado di prendere una decisione e si può ad esempio dire (anche se in un senso lievemente improprio) che il metodo ha commesso un errore, per cui includeremo questa eventualità negli eventi di errore studiati sotto.

Esempio 1. Si sa a priori che lo 0.2% della popolazione soffre di una certa malattia dopo i 50 anni. Quella malattia non è ovvia da diagnosticare. Se la malattia è presente, una certa analisi la evidenzia nel 90% dei casi. Se non è presente, l'analisi produce un falso positivo nel 15% dei casi. Un medico esegue l'analisi a un paziente, che risulta positivo. Il medico che decisione prende? (intendiamo: è più propenso a credere che il paziente abbia o non abbia la malattia?). Soluzione: indichiamo con C_1 l'evento: ha la malattia, con A l'evento: risulta positivo all'analisi; conosciamo: $P(C_1) = 0.002$, $P(C_2) = 0.998$, $P(A|C_1) = 0.9$, $P(A|C_2) = 0.15$, quindi calcoliamo

$$P(A|C_1)\,P(C_1) = 0.9 \cdot 0.002 = 0.001\,8$$
$$P(A|C_2)\,P(C_2) = 0.15 \cdot 0.998 = 0.149\,7.$$

la conclusione è che il medico è ancora più propenso a credere che il paziente sia sano. Quell'analisi è poco discriminante. Non si deve però pensare che l'analisi non sia servita a niente. Ora, per la prossima analisi, si parte da una probabilità a priori diversa: il paziente cade in una categoria di persone che ha probabilità $\frac{0.001\,8}{0.001\,8+0.149\,7} = 0.01$ di essere ammalata, $\frac{0.149\,7}{0.001\,8+0.149\,7} = 0.99$ di essere sana (proporzioni ben diverse da quelle iniziali).

Osservazione: nel caso equiprobabile, essendo $P(C_i)$ uguale per tutti, il criterio diventa semplicemente

$$C_i^{opt} := \arg\max_{C_i} P(A|C_i)\,.$$

Esempio 2. Una rete di trasmissione invia messaggi codificati con 0 e 1. Sulla rete c'è un disturbo, che con probabilità 0.1 modifica 1 in 0 e con probabilità 0.1 modifica 0 in 1. Se riceviamo un 1, cosa decidiamo che sia stato spedito? Soluzione. Per ignoranza, supponiamo che siano equiprobabili l'invio di 0 o di 1. Indichiamo con C_1 l'evento: è stato inviato 1, con A l'evento: abbiamo ricevuto 1; conosciamo: $P(C_1) = P(C_2) = 0.5$, $P(A|C_1) = 0.9$, $P(A|C_2) = 0.1$. Siccome le alternative C_1 e C_2 sono equiprobabili, basta confrontare $P(A|C_1)$ con $P(A|C_2)$ e scegliere il più grande. Quindi ovviamente decidiamo che è stato spedito 1. Questo esempio, così formulato, appare ovvio e poco istruttivo; interessante sarebbe proseguirne l'analisi in un'altra direzione: la probabilità di errore, data da

$$P_{err} = P(A|C_2)\,P(C_2) + P(A^c|C_1)\,P(C_1)$$

è piuttosto alta (vale $P_{err} = 0.1$) e renderebbe troppo incerta la trasmissione di messaggi, quindi bisogna inventare procedimenti per limitare la possibilità di sbagliare. Da qui nascono i codici di correzione d'errore.

6.4.3 Classificazione: idee generali

Abbiamo un insieme di osservazioni possibili, che per fissare le idee supponiamo sia $\mathbb{R}^p$. L'interpretazione è che di ciascun individuo (nazione, provincia ecc.) osserviamo il valore di p fattori $X_1, ..., X_p$, quindi un individuo è rappresentato da un punto $x = (x_1, ..., x_p) \in \mathbb{R}^p$.

Abbiamo due *classi*, C_1 e C_2. Ogni individuo appartiene all'una o all'altra.

Gli individui si dividono in due *gruppi*, il gruppo detto di training (*training set*) ed il gruppo detto di test (*test set*). Di tutti gli individui conosciamo la stringa $(x_1, ..., x_p)$, ma solo degli individui del training set conosciamo la classe. Vorremmo inventare un procedimento per classificare gli individui del test set. Naturalmente vorremmo poter fare questa classificazione sulla base dei valori $(x_1, ..., x_p)$ di questi individui, e sulla base dell'analogia con la classificazione, nota, degli altri individui (quelli training).

Si tratta allora di suddividere $\mathbb{R}^p$ in due regioni, che chiamiamo A_1 e A_2, un po analogamente alle classi corrispondenti C_1 e C_2. Tutti gli individui test la cui stringa $(x_1, ..., x_p)$ cade in A_1, vengono classificati di classe C_1, gli altri di classe C_2:

$$(x_1, ..., x_p) \in A_1 \longrightarrow \text{ classe } C_1$$
$$(x_1, ..., x_p) \in A_2 \longrightarrow \text{ classe } C_2.$$

Come effettuare la suddivisione di $\mathbb{R}^p$ in due regioni? Dobbiamo basarci sui dati noti, cioè sul training set. Immaginiamo: abbiamo in $\mathbb{R}^p$ due insieme di punti, tutti relativi a individui del training set: i punti $P_1, ..., P_k$ degli individui di classe A_1 e quelli $P_{k+1}, ..., P_n$ di quelli di classe A_2. Abbiamo indicato con n il numero di individui del training set. L'ideale sarebbe, o potrebbe sembrare che sia, dividere $\mathbb{R}^p$ in due regioni A_1 e A_2 tali che A_1 contenga tutti i punti $P_1, ..., P_k$ ed A_2 tutti i punti $P_{k+1}, ..., P_n$. Questa strategia ha vari difetti:

- non è univoca (infinite regioni hanno questa proprietà ed è proprio del tutto arbitrario sceglierne una);

- non tiene conto del fatto che le sole variabili $X_1, ..., X_p$ non dovrebbero permettere una classificazione sicura (salvo problemi molto particolari e privi di aleatorietà), quindi deve essere possibile che un individuo di classe A_1 stia nella regione A_2 e viceversa;

- è facile immaginare disposizioni dei punti P_i tali che, per dividerli come detto sopra, siamo costretti a immaginare regioni A molto contorte; se immaginiamo che dietro il nostro tentativo di classificazione ci sia una realtà "fisica", una *struttura*, un legame reale tra le variabili $X_1, ..., X_p$ e la classe (a meno di errore ed altre variabili non identificate o considerate), è molto strano che questo legame passi attreverso complicate formule matematiche (quelle necessarie a descrivere una regione molto contorta); di solito i legami fisici tra grandezze hanno natura polinomiale o comunque abbastanza semplice.

Quindi si rinuncia al requisito che A_1 contenga tutti i punti $P_1, ..., P_k$ e A_2 tutti gli altri. Si vuole che ciò avvenga per la maggior parte dei punti, salvaguardando contemporaneamente qualche criterio di *struttura* e *semplicità geometrica* delle due regioni. Una scelta molto comune, che vedremo, è che le due regioni siano dei semispazi, cioè la divisione in due di $\mathbb{R}^p$ sia realizzata da un iperpiano.

Il discorso astratto si estende al caso di più classi, senza modifiche particolarmente rilevanti, se non notazionali. Le suddivisioni però saranno più complicate.

6.4.4 Classificazione bayesiana

Quello descritto fino ad ora è il problema e lo schema di classificazione in generale. Discutiamo ora l'approccio bayesiano, ispirato alla teoria bayesiana delle decisioni. Supponiamo che sia nota la distribuzione di probabilità congiunta del vettore aleatorio $X = (X_1, ..., X_p)$, condizionata all'essere di classe C_1 o C_2: indichiamo con $f_X(x|C_i)$ la densità congiunta di X, $x = (x_1, ..., x_p) \in \mathbb{R}^p$, quando l'individuo in esame è di classe C_i. Queste densità devono essere note, eventualmente ricavate dal training set nel seguente modo: si prendono i punti $P_1, ..., P_k$ e si cerca di estrapolare da essi una densità di probabilità, $f_X(x|C_1)$; e lo stesso di fa per i punti $P_{k+1}, ..., P_n$, trovando una $f_X(x|C_2)$.

Per Bayes (immaginiamo di usare un analogo della formula di Bayes nel caso di densità)

$$P(C_1|x) = \frac{f_X(x|C_1)\, P(C_1)}{f_X(x|C_1)\, P(C_1) + f_X(x|C_2)\, P(C_2)}$$

ed analogamente per $P(C_2|x)$. Bisogna conoscere o aver fissato a priori le due probabilità $P(C_i)$.

Quindi, il metodo di classificazione bayesiano funziona così: se di un nuovo individuo misuriamo $x = (x_1, ..., x_p)$, gli associamo la classe C_i che massimizza $f_X(x|C_1)\, P(C_1)$. Nel caso di più di due classi, il discorso è identico.

Date le classi C_i e le densità $f_X(x|C_i)$, per ogni i resta definita la regione A_i data da tutti i punti $x = (x_1, ..., x_p) \in \mathbb{R}^p$ che portano alla classificazione C_i. Nel caso di due classi,

$$A_1 = \{x \in \mathbb{R}^p : f_X(x|C_1)\, P(C_1) > f_X(x|C_2)\, P(C_2)\}$$
$$A_2 = \{x \in \mathbb{R}^p : f_X(x|C_2)\, P(C_2) > f_X(x|C_1)\, P(C_1)\}.$$

6.4.5 Il caso gaussiano e la Linear Discriminant Analysis

Se ad esempio si usano le gaussiane come modello per le densità congiunte $f_X(x|C_i)$, cioè

$$f_X(x|C_i) = \frac{1}{\sqrt{(2\pi)^n \det Q_i}} \exp\left(-\frac{1}{2}(x-\mu_i)^T Q_i^{-1}(x-\mu_i)\right)$$

allora la disuguaglianza $f_X(x|C_1)P(C_1) > f_X(x|C_2)P(C_2)$ diventa, passando ai logaritmi,

$$(x-\mu_2)^T Q_2^{-1}(x-\mu_2) - (x-\mu_1)^T Q_1^{-1}(x-\mu_1) > \log(\det Q_1) - \log(\det Q_2).$$

Si trova una condizione della forma

$$x^T\left(Q_2^{-1} - Q_1^{-1}\right)x + ...$$

con termini lineari e costanti, quindi, a seconda delle proprietà della matrice $Q_2^{-1} - Q_1^{-1}$, si trovano regioni curvilinee di vario tipo. Questa è la *Quadratic Discriminant Analysis*, su cui non entriamo in dettaglio.

Se però si ipotizza che le due gaussiane abbiano la stessa covarianza Q (il legame tra i predittori sia lo stesso per le due classi), e che differiscano solo i valori medi dei predittori, i termini quadratici si semplificano e troviamo

$$2x^T Q^{-1}(\mu_1 - \mu_2) > \mu_1^T Q^{-1}\mu_1 - \mu_2^T Q^{-1}\mu_2.$$

Si tratta di un semispazio. In altre parole, $\mathbb{R}^p$ viene suddiviso dall'iperpiano

$$x \cdot v = \alpha$$
$$v = Q^{-1}(\mu_1 - \mu_2), \qquad \alpha = \frac{1}{2}\left(\mu_1^T Q^{-1}\mu_1 - \mu_2^T Q^{-1}\mu_2\right).$$

La matrice Q ed i vettori μ_i si stimano dai dati:

1. prima si usano i punti $P_1, ..., P_k$ per stimare μ_1 (si prende come stimatore il punto medio di $P_1, ..., P_k$) e si usano i punti $P_{k+1}, ..., P_n$ per stimare μ_2;
2. poi si centrano i punti, cioè si calcolano i punti $P_i' = P_i - \mu_1$ per $i = 1, ..., k$, $P_i' = P_i - \mu_2$ per $i = k+1, ..., p$;
3. infine si calcola la matrice di covarianza empirica usando tutti i punti P_i'.

L'immagine dell'iperpiano serve solo idealmente per capire il risultato. Infatti, per eseguire la classificazione di un nuovo individuo, rappresentato da una nuova stringa $x = (x_1, ..., x_p)$, basta:

1. calcolare $v = Q^{-1}(\mu_1 - \mu_2)$ e $\alpha = \frac{1}{2}\left(\mu_1^T Q^{-1}\mu_1 - \mu_2^T Q^{-1}\mu_2\right)$ (usando gli oggetti stimati)
2. assegnare il nuovo individuo alla classe C_1 se $x \cdot v > \alpha$, mentre alla classe C_2 se $x \cdot v < \alpha$.

Naturalmente la prima delle due operazioni si può svolgere una volta per tutte, essendo la stessa per tutti i nuovi individui. Osserviamo che la situazione di mancata classificazione, cioè il caso $x \cdot v = \alpha$, in pratica non può mai avvenire.

Quella appena descritta è la *Linear Discriminant Analysis*. Per utilizzarla col software `R` bisogna prima caricare il package MASS, col comando `require(MASS)`, poi usare il comando `lda` (si consiglia di leggerlo con `?lda`).

6.4.6 Clustering

Le tecniche di classificazione appena descritte partono dall'esistenza di classi prestabilite e si pongono il problema di assegnare nuovi individui alle classi (*classificare* nuovi individui). Essi però inglobano già una sorta di clustering, nella fase di creazione delle classi. Ad esempio, nella regressione logistica, gli individui di cui è noto tutto (valore delle variabili che fungono da predittori, e della variabile di classe, cioè 0 o 1) vengono usati per determinare il modello (i coefficienti della parte lineare regressiva), che poi verrà usato per classificare nuovi individui di cui siano noti solo i valori dei predittori. Ma la creazione del modello in pratica è la creazione di due classi che separano il meglio possibile gli individui noti, quindi è un'operazione di clustering. C'è però una differenza concettuale ripetto al clustering che stiamo per descrivere: nel creare un modello di regressione logistica, quindi nel creare due classi, si usano individui di cui è noto il valore della classe (0 o 1). Invece, nei metodi che descriveremo ora, a priori nulla distingue gli individui in classi. Si immagina però che essi possano essere membri di classi differenti; allora il metodo dovrà identificare le classi e attribuire ad esse gli individui; infine, il metodo dovrebbe fornire un giudizio sull'appartenenza di un individuo ad una classe, cioè dovrebbe dare una dichiarazione di quanto è sicura la sua classificazione, oppure è vaga.

Si pensi ad un insieme W di punti Q del piano ($Q \in W$), sparpagliati, ciascuno rappresentante un individuo (descritto quindi da due variabili, due predittori). Ci saranno casi in cui i punti sono un po' separati in due gruppi, o più di due gruppi, pur essendo vaga la separazione. Si pensi alle case di due città limitrofe in zone molto abitate: si va da una città all'altra quasi senza soluzione di continuità, però il grado di addensamento è diverso nelle due zone proprie delle città rispetto alla parte intermedia, dove c'è ancora un po' di campagna qua e là. Abbiamo quindi questo insieme di punti. Ipotizziamo che esso sia suddividibile in due classi (il caso con tre o più classi è simile, ma torneremo su questo punto). Vediamo alcune idde generali per trovare una buona suddivisione.

Alcune idee dei paragrafi precedenti sarebbero perfettamente adatte: cercare una retta, o una parabola (linear o quadratic discriminant analysis) che separa bene l'insieme dei punti. Sviluppiamo altre idee.

Immaginiamo che le due classi siano come due nuvole un po' ellittiche, pur con vaghezza (magari senza una vera soluzione di continuità tra le nuvole). Iniziamo col cercare i *centri* delle nuvole. Avendo deciso che sono due, si cercano due centri, M_1 e M_2 (qui entra in gioco il numero di classi deciso a priori: se avessimo deciso di dividere in tre classi, avremmo cercato tre centri). Si inizi mettendo a caso due punti M_1 e M_2 nel piano, in assenza di suggerimenti migliori (se invece c'è un'idea migliore la si usi). Poi si trovino gli *insiemi di Voronoi* di questi due punti, che chiamiamo V_1 e V_2: V_i è l'insieme dei punti del piano che distano da M_i meno che dall'altro centro. Sono due semipiani. Se partivamo da tre centri M_1, M_2, M_3 trovavamo una divisione in tre "angoli", e così via. Poi, chiamiamo W_1 e W_2 gli insiemi dei punti originari che cadono in V_1 e V_2 rispettivamente: W_i è l'insieme dei punti $Q \in W$ che appartengono a V_i, quindi che distano da M_i meno che dall'altro centro. Questa è già una suddivisione possibile, però relativa ad una scelta iniziale dei centri, fatta a caso o comunque non ancora ottimizzata in alcun modo.

Diamo un punteggio alla suddivisione trovata: calcoliamo la somma delle distanze al

quadrato di tutti i punti di W_1 da M_1

$$d_1^2 = \sum_{Q \in W_1} d^2 (Q, M_1)$$

ed analogamente per W_2: $d_2^2 = \sum_{Q \in W_2} d^2 (Q, M_2)$. Questa suddivisione è caratterizzata dal numero $d_1^2 + d_2^2$; se tale numero è alto, la suddivisione viene considerata poco buona (i punti di ciascun gruppo distano troppo dal loro centro). In generale, per k gruppi, il numero da calcolare è

$$\sum_{i=1}^{k} \sum_{Q \in W_i} d^2 (Q, M_i) .$$

Si vorrebbero trovare i punti M_i che rendono minima questa espressione. Si possono inventare vari algoritmi che cercano di trovare dei buoni centri M_i. L'algoritmo k-means lavora su centri M_i che vengono presi, ad ogni passo dell'algoritmo iterativo, pari alla media aritmetica dei punti di W_i (poi vengono ricalcolati i W_i, poi i loro punti medi M_i e così via). L'algoritmo k-medoids utilizza invece come centri alcuni dei punti di W stesso, aggiornando iterativamente i medoidi (alla ricerca dei migliori) attravero scambi causali tra i medoidi e gli altri punti di W. Gli algoritmi differiscono poi, tra altre cose, per la distanza $d (Q, M_i)$ che viene utilizzata (rimandiamo alla letteratura specializzata per questi ed altri dettagli).

Questi algoritmi hanno un difetto: raggruppano secondo la minima distanza dai centri, quindi tendono a costruire dei raggruppamenti equilibrati, della stessa grandezza. Questa simmetria può essere poco adatta a certe applicazioni, in cui si capisce ad occhio che i punti $Q \in W$ sono divisi in gruppi di ampiezza differente, per esempio una grossa nuvola con una piccola nuvola satellite. Gli algoritmi descritti fino ad ora forzerebbero la suddivisione ad essere abbastanza simmetrica, attribuendo una parte di punti della grossa nuvola alla parte W_i relativa al piccolo satellite. C'è allora una variante, detta algoritmo EM (Expectation-Maximization) basata sulle misture di gaussiane e la massima verosimiglianza, che permette di trovare partizioni diseguali, più aderenti a certe dituazioni pratiche.

In genere il software, come input di un particolare metodo di clustering (k-means ecc.), chiede i punti $Q \in W$ (una tabella di dati come quella di PCA) ed il numero di classi k in cui vogliamo suddividerli. Come output fornisce le classi trovate, in genere elencando gli elementi delle classi, e fornendo una raffigurazione grafica dei punti separati in gruppi, raffigurazione spesso legata a PCA. Infatti, se i punti $Q \in W$ stanno in uno spazio a dimensione maggiore di 2, il modo più naturale è innanzi tutto mostrare questi punti attraverso una visione che li distingua il più possibile (e questo è svolto da PCA), sovrapponendo poi ad essa la suddivisione in gruppi. Esistono anche visualizzazioni tridimensionali a colori.

Oltre a questo, il software fornisce in output dei parametri numerici che servono a giudicare la suddivisione ottenuta, il più comune dei quali è la *silhoutte*. Tramite questi numeri abbiamo una quantificazione della bontà o vaghezza dei cluster ottenuti che, oltre ad essere un metro di giudizio di tipo assoluto, può essere utilizzato in modo comparativo per decidere il numero k. Esso era stato scelto a priori, ma con quale criterio? Ci saranno casi in cui, o per ragioni di evidenza grafica o per motivi applicativi, sapremo come decidere k a priori; altri in cui si va per tentativi e si sceglie k a posteriori: quello che massimizza la silhouette.

Descriviamo la silhouette secondo una delle sue possibili definizioni. La silhouette di un singolo individuo $Q \in W$, relativa alla partizione $W_1, ..., W_k$ trovata con un qualsiasi metodo tipo k-means ecc., è data dall'espressione

$$s(Q) = \frac{b(Q) - a(Q)}{\max(a(Q), b(Q))}.$$

Indicando con $W(Q)$ il cluster, tra i vari $W_1, ..., W_k$, che contiene il punto Q, il numero $a(Q)$ è la distanza media quadratica di Q dagli altri punti del proprio cluster $W(Q)$:

$$a(Q) = \sum_{Q' \in W(Q)} d(Q, Q')^2.$$

Il numero $b(Q)$ invece è la distanza media quadratica di Q dai punti del cluster "successivo", così definito: si calcolano i numeri

$$\sum_{Q' \in W_i} d(Q, Q')^2$$

per ogni $W_i \neq W(Q)$ e si prende il minimo; questo è $b(Q)$. Si verifica che il numero $s(Q)$ soddisfa

$$-1 \leq s(Q) \leq 1.$$

Più $s(Q)$ è vicino a 1, più si ritiene che la clusterizzazione di Q sia buona. Infatti, supponiamo che $s(Q)$ sia vicino a 1. Innanzi tutto questo implica che $b(Q) - a(Q)$ è positivo, quindi $\max(a(Q), b(Q)) = b(Q)$ e vale

$$s(Q) = \frac{b(Q) - a(Q)}{b(Q)} = 1 - \frac{a(Q)}{b(Q)}.$$

Ora, se questo rapporto vale quasi 1, significa che $a(Q)$ è molto piccolo rispetto a $b(Q)$, cioè che la distanza media di Q dai suoi compagni di gruppo è decisamente minore di quella dai membri del gruppo "successivo". Questo è sintomo di buona clusterizzazione di Q.

La silhouette di un singolo individuo Q serve a giudicare quali individui sono stati raggruppati bene e quali no. Poi, mediando sugli individui di un gruppo W_i si ottiene la silhouette media di W_i, che descrive quanto preciso o vago sia il gruppo W_i. Infine, mediando sui gruppi si ottiene una silhouette media complessiva della clusterizzazione $W_1, ..., W_k$, che può essere utilizzata per confrontare vari k tra loro (oltre che vari metodi anche di natura diversa).

Si suggerisce, col software R, l'uso del comando `pam`, che svolge la cluster analysis con metodo dei medoidi.

6.5 Esercizi

6.5.1 Esercizio n. 1

- *Problema: cosa incide sul tasso di disoccupazione (TD)?* Vorremmo creare una tabella con alcune colonne $X_1, ..., X_n$ (fattori che forse influiscono sul TD) e la colonna Y=TD, e come righe (unità sperimentali) le diverse nazioni europee. Dalla sua analisi speriamo di comprendere le cause di una maggiore o minore disoccupazione. Bisogna allora prendere il TD ad un certo tempo, es. anno 2009.

- *Percorso*: Eurostat, Statistics Database, Statistics A - Z, Unemployment, Database, LFS series - Detailed annual survey results, Total unemployment - LFS series, Unemployment rates by sex, age groups and nationality, select data, age: 25-64, citizen: total, geo: all; sex: total, time 2009; poi update ecc. come negli esercizi sulle serie storiche, scaricare e salvare su file Excel.

  Cosa può influire sulla disoccupazione? La spesa in research and developement?

- *Percorso*: Eurostat, Statistics Database, Statistics A - Z, Research and development, Database, Research and development, Statistics on research and development, R&D expenditure at national and regional level, Total intramural R&D expenditure (GERD) by sectors of performance, time: geo: all, sector: Higher education sector , 2005, unit: Percentage of GDP. Update ecc, scaricare e salvare su file Excel.

  Che altro?

- *Percorso*: Eurostat, High-tech industry and knowledge-intensive services (stessa pagina di Research and development), High-tech industries and knowledge-intensive services: economic statistics at national level, Venture capital investments, Economic statistics on high-tech industries and Knowledge Intensive Services at the national level (htec_eco_sbs) , geo: all, indicator: Number of enterprises, nace: High-technology sectors, time: 2005. Update ecc, scaricare e salvare su file Excel.

- Creare su Excel una tabella con le nazioni di cui si hanno tutti i dati, eliminando i riassunti europei, con i nomi delle nazioni nella prima colonna ed i nomi delle tre variabili (abbreviate in TD, RD, I) nella prima riga. Accorgimenti: mettere un nome fittizio anche in cima alle nazioni; usare nomi brevi per le nazioni e soprattutto senza separatore.

- Creare un file *EsercizioMultiv1.RData* (seguire in generale lo standard degli esercizi sulle serie storiche), caricare la tabella col comando (per indicazioni vedere l'appendice agli esercizi sulle serie storiche; in breve, conviene salvare la tabella su file di testo con l'opzione limiti di tabulazione, salvare tale file nella cartella dell'esercizio, da R cambiare directory portandosi in tale cartella, poi eseguire il seguente comando):

```
U<-read.table(clipboard,dec=',',header=T,row.names=1)
```

Digitando `U` e invio, si vede la tabella su R. Riportiamo qui la tabella scomposta in due

parti, per motivi di spazio:

	TD	RD	I
Belg	6.6	0.41	16943
Bulg	6.0	0.05	5274
Czec	5.9	0.23	33179
Denm	5.1	0.60	10202
Germ	7.3	0.41	81825
Esto	12.3	0.39	1396
Gree	8.4	0.28	11330
Spai	16.0	0.33	44985
Fran	7.5	0.40	77990
Ital	6.5	0.33	136767
Latv	15.3	0.23	1736
Lith	12.2	0.41	2184

	TD	RD	I
Luxe	4.1	0.02	1231
Hung	8.8	0.24	34104
Neth	2.8	0.54	26300
Aust	4.0	0.61	15700
Pola	6.8	0.18	47776
Portu	9.0	0.29	17288
Roma	5.7	0.06	16214
Sloven	5.2	0.24	4232
Slovak	10.5	0.10	2071
Finl	6.5	0.66	6823
Swed	6.0	0.79	37834
UK	5.6	0.45	132887
Norw	2.2	0.47	12116

Calcoliamo la matrice di correlazione tra le tre variabili:

```
cor(U)
```

	TD	RD	I
TD	1	-0.197	-0.109
RD	-0.197	1	0.148
I	-0.109	0.148	1

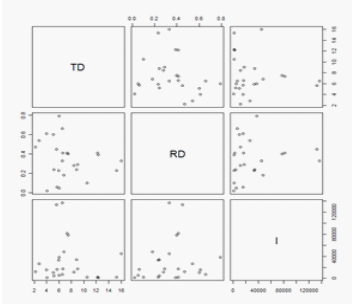
Risultato pessimo: sostanziale scorrelazione tra le tre variabili!

Ma c'è un errore banale, che potrebbe essere alla base di questo fallimento: la variabile I non è una percentuale, ma un totale. E' vagamente proporzionale alla dimensione della nazione. Va percentualizzato. Bisogna trovare la popolazione e dividere per essa (moltiplicando per un fattore che riporti a numeri circa unitari).

Errore a parte, con

```
plot(U)
```

si ottiene il seguente disegno. I disegni relativi ad I sono ovviamente insensati. Il disegno tra TD e RD invece è corretto (corrisponde a dati utilizzati nel modo giusto) ma è assai deludente. Pensavamo che una maggiore spesa in R&D provocasse un minore TD. Un po' è vero, il coefficiente di correlazione vale circa -2 e dal grafico si vede una lieve struttura a retta decrescente. La è molto vaga, il risultato non è netto; giocano molti altri fattori, c'è un'enorme variabilità attorno alla retta di regressione, variabilità inspiegata.



Esercizio per il futuro: quali altri fattori concorrono?

Insoddisfatti per l'insuccesso, cerchiamo di vedere I dati più da vicino. Un modo (un po' esotico) è il seguente, che anticipa un metodo che vedremo estensivamente.

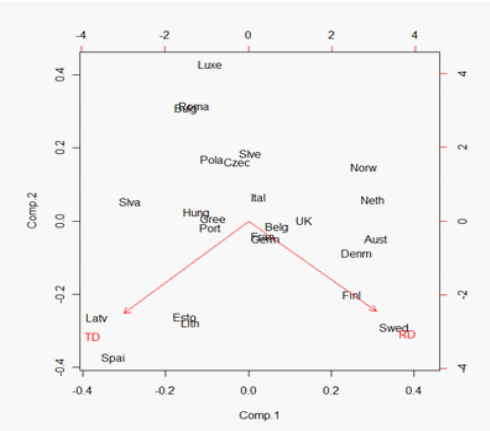
Riprendiamo il foglio Excel, standardizziamo i dati, solo relativi a TD e RD. La standardizzazione su Excel è facile e sicura: basta inserire due righe al fondo della tabella, nella prima delle quali mettere le medie delle colonne, nella seconda le deviazioni standard, poi si costruisce una seconda tabella delle dimensioni della prima, in cui si mettono i valori standardizzati, cioè depurati della media e divisi per la deviazione standard (conviene fare la verifica mettendo anche qui le due righe in fondo con media e deviazione standard, che ora devono essere 0 ed 1, numericamente parlando, per tutte le colonne).

Carichiamo la tabella come

```
US<-read.table(clipboard,dec=',',header=T,row.names=1)
```

Poi eseguiamo:

```
PCA<-princomp(US); biplot(PCA)
```



Il SW mette in orizzontale la linea principale lungo cui si sviluppano I dati. Fuori da essa emergono gli "outliers". Sono paesi particolari, a cui forse vanno applicati ragionamenti a parte.

Esercizio: eliminare un po' di outliers (a mano su Excel e ricaricare) e calcolare la matrice di correlazione.

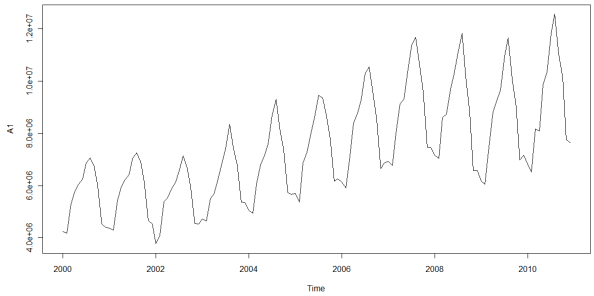
6.5.2 Esercizio n. 2

In attesa di idee sui fattori della disoccupazione, approfondiamo a livello tecnico serie temporali e regressione insieme. Usiamo la regressione multipla per creare modelli di serie temporali.

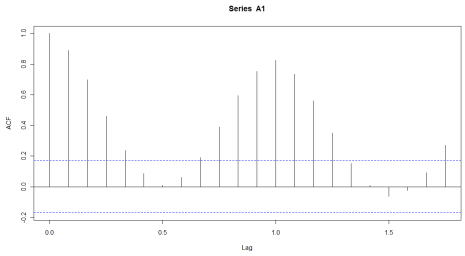
- *Percorso*: Eurostat, Statistics Database, Statistics A – Z, Transport, Database, Air Transport, Air transport measurement – passengers, Overview of the air passenger transport by country and airports, Air passenger transport by reporting country, select data: geo: Italy, schedule:total, time: all, ecc. Total; update, ecc., scaricare e salvare su file Excel.

- Sul file Excel, depurare di valori annuali o quadrimestrali. Tenere solo i dati dal 2000 al 2010 inclusi. Riempire l'anno mancante 2001 con i valori 2000 proporzionati tramite il valore annuale 2001, che è disponibile (valore 2000 per totale 2001 diviso totale 2000).

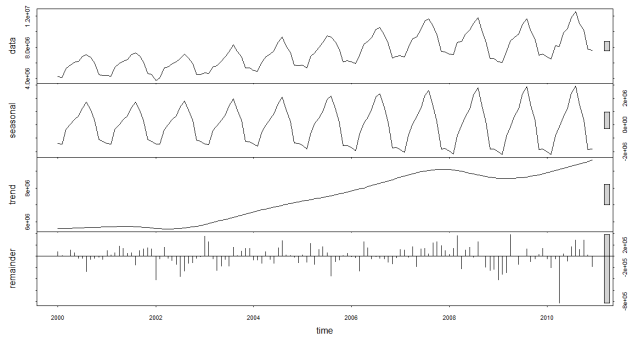
```
A<-scan(clipboard,dec=',')
```
Salvare il file *EsercizioMultiv2.RData*.
```
A1 <- ts(A, frequency=12,start=c(2000,1)); ts.plot(A1)
```



Eseguire prime analisi:
```
acf(A1)
```




```
plot(stl(A1,6))
```



che confermano l'elevata periodicità, del resto evidente, nonché una flessione intorno al 2008. I residui sono piccolissimi (vedi barra).

Possiamo tranquillamente applicare HW, AR ottimizzato ecc. Applichiamo comunque a priori l'ipotesi che il modello sia del tipo

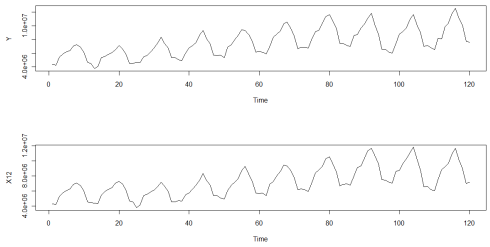
$$X_n = aX_{n-1} + bX_{n-12} + c$$

ed utilizziamo la regressione lineare, per far questo. Creiamo tre variabili X1, X12, Y estraendo le finestre giuste:

```
L<-length(A1); Y<-A1[13:L]; X1<-A1[12:(L-1)]; X12<-A1[1:(L-12)]
```

Per curiosità e rassicurazione, si vedano

```
par(mfrow=c(2,1)); ts.plot(Y); ts.plot(X12)
```



Eseguiamo ora la regressione lineare multipla:

```
mod1 <- lm(Y ~X1+X12)
```

Vediamo i coefficienti a, b, c:

```
summary(mod1)
```

Call:

lm(formula = Y ~X1 + X12)

Residuals:

Min 1Q Median 3Q Max

-1508798 -293779 32857 348593 957401

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.413e+05 1.820e+05 0.776 0.439

X1 2.417e-01 4.536e-02 5.328 4.88e-07 ***

X12 7.777e-01 4.755e-02 16.356 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 498100 on 117 degrees of freedom

Multiple R-squared: 0.9393, Adjusted R-squared: 0.9383

F-statistic: 905.3 on 2 and 117 DF, p-value: < 2.2e-16

Entrambi a e b sono molto significativi. La sua varianza spiegata R^2 è elevatissima, quindi il modello è molto buono.

Il suo utilizzo predittivo richiede un po' di fatica. Indichiamo con P il vettore delle previsioni dei due anni successivi, in realtà (per motivi che si capiranno nel ciclo di for) arricchito dei dati storici:

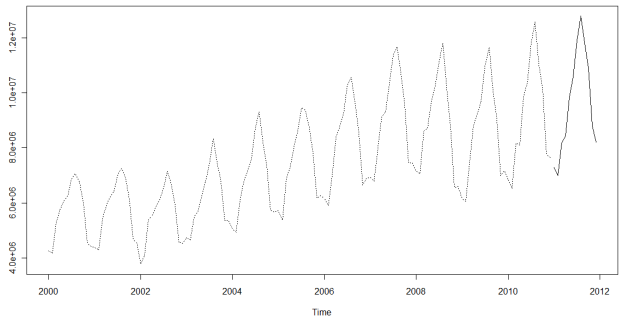
```
P<-1:(L+12); P[1:L]<-A1
```

Estraiamo i coefficienti:

```
a<-mod1$coefficient[2]; b<-mod1$coefficient[3]; c<-mod1$coefficient[1]
```

(verificare che sono quelli giusti). Calcoliamo iterativamente le previsioni:

```
for (n in (L+1):(L+12)) { P[n] <- a*P[n-1] +b*P[n-12] + c }
P1 <- ts(P, frequency=12,start=c(2000,1))
par(mfrow=c(1,1)); ts.plot(A1, window(P1, start=c(2011,1)) , lty=c(3,1))
```



Esercizio: con lo stesso metodo, posizionarsi a dicembre 2007 (come fosse il presente) ed eseguire la previsione dell'anno successivo, confrontandola graficamente con i valori reali.

Esercizio: applicare HW ed AR ottimizzato (ed eventualmente stl) ed osservare i risultati, traendo qualche conclusione.

Domanda: cosa può far preferire il risultato ottenuto con il modello dell'esercitazione odierna?

6.5.3 Esercizio n. 3

Dal sito ISTAT, precisamente da http://sitis.istat.it/sitis/html/ (SITIS è il Sistema Indicatori Territoriali), preleviamo i dati regionali relativi ai seguenti indicatori (da Consulta i dati, sotto le voci Sanità assistenza e previdenza, Condizioni economiche delle famiglie, Mercato del lavoro:

PLIC = numero di posti letto in istituti di cura

SC = spese generali

SA.SC = spese per alimenti rispetto alle spese generali

TD = tasso di disoccupazione

TMI = tasso di mortalità infantile.

Tutti i dati sono già percentualizzati rispetto al numero di abitanti. Li abbiamo inoltre standardizzati, direttamente su Excel.

Esercizio. I dati riportati nel seguito sono stati prelevati negli anni scorsi. Rintracciare i dati nuovi, più recenti, e ripetere con essi le analisi.

	PLIC	SC	SA.SC	TD	TMI
Piem	0.088	0.471	-0.707	-0.607	-0.395
Vaos	-1.545	0.348	-0.642	-0.813	1.578
Lomb	0.202	1.397	-0.836	-0.790	-0.538
TrAA	0.677	0.435	-1.269	-0.966	-0.075
Vene	0.088	1.334	-1.210	-0.848	-0.497
FrVG	0.639	-0.005	-1.028	-0.804	-1.301
Ligu	1.190	-0.247	0.470	-0.429	-0.354
EmRo	0.658	1.177	-1.315	-0.863	-0.347
Tosc	0.126	1.092	-0.795	-0.644	-1.355
Umbr	-1.431	0.675	-0.140	-0.524	-1.287
Marc	0.278	1.090	-0.265	-0.702	-0.0006
Lazi	2.329	0.546	-0.080	-0.113	-0.014
Abru	0.335	-0.373	0.402	-0.456	0.040
Moli	0.658	-1.289	0.065	0.451	-1.151
Camp	-1.811	-1.314	2.031	1.664	0.414
Pugl	-0.766	-0.926	1.038	0.648	1.109
Basi	-0.747	-1.154	0.661	0.844	2.001
Cala	-0.500	-1.727	1.571	2.153	0.632
Sici	-0.918	-1.130	1.332	1.517	1.783
Sard	0.449	-0.403	0.717	1.285	-0.238

Questa tabella è stata prima costruita su Excel copiando i dati da Istat (abbreviando i nomi per comodità grafiche successive, evitando gli spazi tra più parole di un nome), poi è stata standardizzata direttamente su Excel, come spiegato nell'esercizio 1.

Domanda: evidenziare pregi e difetti della tabella standardizzata rispetto a quella originaria.

Carichiamo i dati in R con la procedura riassunta nell'esercizio 1: si mette il file di testo *indicatori_ benessere.txt* (salvato da Excel) nella cartella del presente esercizio, si esegue "cambia cartella" dal menu "file", posizionandosi nella cartella giusta, poi si usa

```
IB <- read.table(file=indicatori_benessere.txt)
```

Oppure si provi il comando

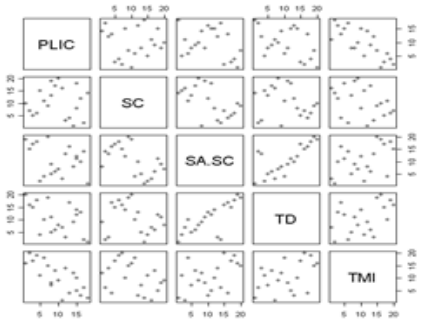
```
IB<-read.table(clipboard,dec=',',header=T,row.names=1)
```

Osservare la tabella su R (con IB invio). Fare

```
cor(IB)
```

	PLIC	SC	SA.SC	TD	TMI
PLIC	1	0.32	-0.41	-0.36	-0.44
SC	0.32	1	-0.84	-0.85	-0.48
SA.SC	-0.41	-0.84	1	0.90	0.51
TD	-0.36	-0.85	0.90	1	0.48
TMI	-0.44	-0.48	0.51	0.48	1

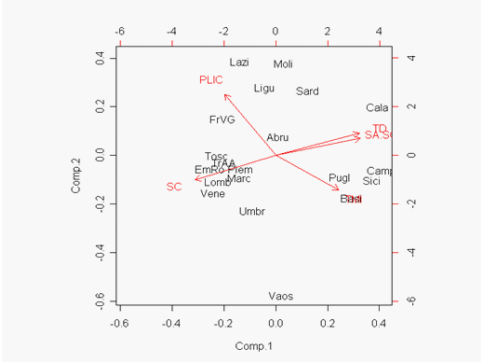
```
plot(IB)
```



Le correlazioni sono abbastanza buone, se confrontate coi valori descritti nella prima sezione di questo capitolo.

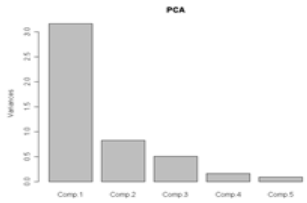
Ha senso fare una regressione tipo $TD = a * SA.SC + b$? E viceversa? Ragionare sul significato della regressione. Eseguiamo

`PCA <- princomp(IB), poi biplot(PCA)`



Questo disegno è estremamente istruttuivo: si rivedano le numerose considerazioni descritte nella sezione teorica su PCA.

`plot(PCA)`



`summary(PCA)`

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.778	0.910	0.711	0.401	0.297
Proportion of Variance	0.666	0.174	0.106	0.033	0.018
Cumulative Proportion	0.666	0.840	0.947	0.981	1

6.5.4 Esercizio n. 4

Questo esercizio ed il seguente sono dovuti agli studenti del corso 2010-11, come ricordato anche nella prefazione. In questi esercizi si suggerisce o di reperire autonomamente dati simili facendo opportune ricerche, oppure semplicemente di copiare queste tabelle su Excel.

Esaminiamo i seguenti dati relativi alle variabili TD = tasso di disoccupazione, RD = Research & Developement,PE = Spesa nella pubblica istruzione (rispetto al PIL), HC = Health Care.

	TD	RD	PE	HC
Belg	6,6	0,41	5,93	9,99
Bulg	6,0	0,05	4,51	7,09
Czec	5,9	0,23	4,26	6,96
Denm	5,1	0,6	8,30	9,60
Germ	7,3	0,41	4,53	10,54
Esto	12,3	0,39	4,88	5,08
Spai	16,0	0,33	4,23	8,36
Fran	7,5	0,4	5,65	11,03
Latv	15,3	0,23	5,06	6,76
Lith	12,2	0,41	4,90	6,23

	TD	RD	PE	HC
Luxe	4,1	0,02	3,78	7,62
Hung	8,8	0,24	5,47	8,09
Neth	2,8	0,54	5,48	9,72
Aust	4,0	0,61	5,48	10,26
Pola	6,8	0,18	5,47	6,20
Portu	9,0	0,29	5,39	9,63
Roma	5,7	0,06	3,48	5,10
Solven	5,2	0,24	5,67	8,23
Slovak	10,5	0,1	3,85	7,34
Finl	6,5	0,66	6,31	8,39
Swed	6,0	0,79	6,97	8,93
Norw	2,2	0,47	7,02	8,65

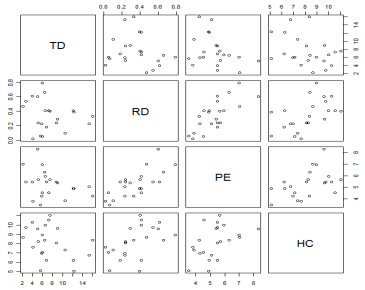
Dopo aver salvato questa tabella su un file di testo (anche con le virgole), la si copia, si scrive

```
M<-read.table(clipboard,dec=',',header=T,row.names=1)
```

su R e si da invio. Controllare chiedendo `M`.

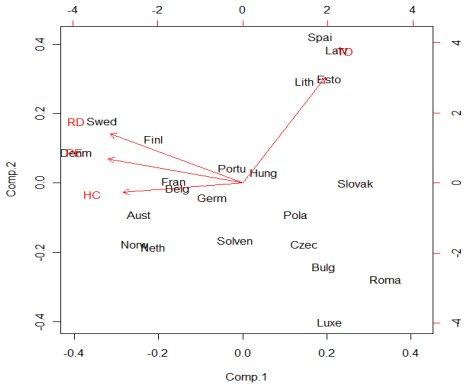
```
cor(M); plot(M)
```

	TD	RD	PE	HC
TD	1	-0.186	-0.336	-0.343
RD	-0.186	1	0.741	0.531
PE	-0.336	0.741	1	0.466
HC	-0.343	0.531	0.466	1



Abbiamo sempre standardizzato i dati su Excel ma questo si può fare anche su R:

```
Ms<-M; for (i in 1:4) { Ms[,i]<-( M[,i]-mean(M[,i]))/sd(M[,i]) }
PCA <- princomp(Ms); biplot(PCA)
```



Il risultato è molto interessante. Un utile esercizio potrebbe essere quello di mettere insieme questi indicatori con altri ancora trovati in altri momenti. Con

```
summary(PCA)
```

si legge:

Cumulative Proportion	0.587	0.803	0.942	1

Eseguire anche

```
plot(PCA)
```

Se interpretiamo l'asse orizzontale come una variabile del tipo: "spese per il bene pubblico", possiamo chiederci la classifica delle nazioni. Chiediamo:

```
PCA$loadings
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
TD	0.345	0.884	0.247	-0.194
RD	-0.556	0.413	-0.147	0.706
PE	-0.565	0.203	-0.454	-0.658
HC	-0.502		0.844	-0.174

Queste sono le componenti di una base rispetto all'altra. Proviamo:

```
0.345^2+0.884^2+0.247^2+0.194^2
```

[1] 0.999126

```
0.345^2+0.556^2+0.565^2+0.502^2
```

[1] 0.99939

E così via. L'ampiezza (e segno) dei numeri danno indicazione del legame. Es: TD è rappresentato più che altro da Comp.2; Comp.1 cattura soprattutto RD, PE, HC (in egual misura), un po' meno TD; e così via. Tutte cose del resto chiare dal disegno. Non visibili nel disegno sono i ruoli di Comp.3 e Comp. 4.

La classifica è data dalle proiezioni delle nazioni su Comp.1, cambiate di segno perché, come si vede dai loadings e dal disegno, Comp.1 è orientata nel verso opposto al significato intuitivo di "ampia spesa". La proiezione di un vettore su uno normalizzato è il prodotto scalare dei due. Il vettore Comp.1 ha componenti (nella base canonica) 0.345, -0.556 ecc. mentre le componenti delle nazioni (nella base canonica) sono i numeri della tabella (standardizzata). Pertanto, ad esempio, i punteggi di Belg e Bulg sono:

```
-sum(Ms[1,]* PCA$loadings[,1])
```

[1] 1.093552

```
-sum(Ms[2,]* PCA$loadings[,1])
```

[1] -1.355473

Possiamo introdurre un vettore punteggio, P,

```
P <- 1:nrow(Ms)
```

e riempirlo con

```
for (i in 1:nrow(Ms)) { P[i] <- -sum(Ms[i,]* PCA$loadings[,1]) }
```

Digitando poi `P` si vede che così si perdono i nomi delle nazioni. Un trucco per leggere i punteggi a fianco delle nazioni è

```
P<-Ms; for (i in 1:nrow(Ms)) { P[i,1] <- -sum(Ms[i,]* PCA$loadings[,1]) };
P
```

ed ora nella prima colonna ci sono i punteggi, che qui ricopiamo cambiando a mano il nome alla colonna:

	punteggio
Belg	1.09
Bulg	-1.35
Czec	-1.03
Denm	2.78
Germ	0.51
Esto	-1.45
Spai	-1.30
Fran	1.15
Latv	-1.57
Lith	-1.04

	punteggio
Luxe	-1.46
Hung	-0.35
Neth	1.49
Aust	1.73
Pola	-0.88
Portu	0.18
Roma	-2.39
Solven	0.13
Slovak	-1.89
Finl	1.48
Swed	2.36
Norw	1.80

Con pazienza si possono ordinare in una classifica.

6.5.5 Esercizio n. 5

Esaminiamo i seguenti dati. Premessa: per capire le leggi che possono regolare il mercato del lavoro, conviene mettersi in un periodo neutro, 2002-07. Idea originale: esaminare gli incrementi nel tempo piuttosto che i valori assoluti (anch'essi sono stati esaminati). I dati che seguono sono gli incrementi 2002-07 di 6 variabili, standardizzati, depurati di due degli

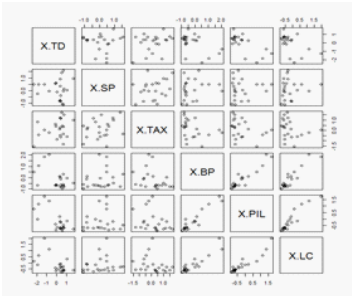
outliers (Slovacchia e Romania).

	X.TD	X.SP	X.TAX	X.BP	X.PIL	X.LC
Belg	0,70	-0,03	0,34	-0,71	-0,68	-0,62
Bulg	-2,32	0,47	-1,54	0,52	0,77	0,12
Czec	0,30	-0,83	0,12	0,13	0,19	-0,12
Denm	0,38	-0,80	0,28	-0,79	-0,68	-0,55
Germ	0,56	-1,07	0,45	-0,93	-0,86	-0,76
Esto	-0,75	-0,03	-0,22	2,03	1,38	1,13
Gree	0,35	0,90	1,22	0,03	-0,20	-0,34
Spai	-0,14	0,54	0,61	-0,68	-0,40	-0,43
Fran	0,38	0,34	-0,49	-0,78	-0,75	-0,55
Ital	-0,08	0,60	0,39	-0,63	-0,84	-0,66
Latv	-1,24	0,47	0,06	1,84	1,73	1,98
Lith	-1,83	0,47	-0,44	1,05	1,15	0,61

	X.TD	X.SP	X.TAX	X.BP	X.PIL	X.LC
Ire	0,70	1,57	-0,33	0,31	-0,45	-0,32
Hung	1,08	0,03	-0,60	-0,04	-0,16	0,31
Neth	0,76	0,13	1,50	-0,79	-0,68	-0,58
Aust	0,41	-0,40	1,17	-0,64	-0,69	-0,70
Pola	-1,91	-0,27	0,83	-0,68	-0,01	-0,17
Portu	1,69	0,94	1,00	-0,49	-0,76	-0,61
Sloven	0,41	-0,87	-0,66	-0,12	-0,25	-0,07
Finl	0,06	-0,13	-0,88	-0,66	-0,63	-0,54
Swed	0,70	-1,10	-1,32	-0,60	-0,64	-0,55
UK	0,53	1,40	1,72	-0,79	-0,78	-0,35

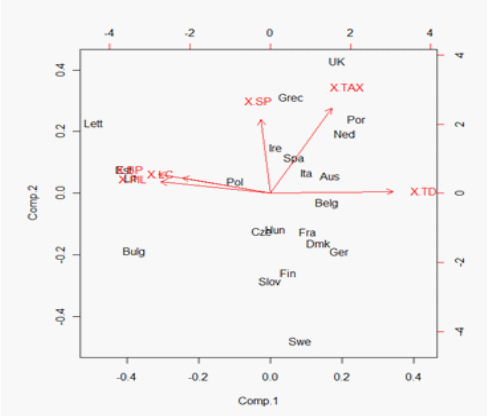
dove:
X.TD Tasso di disoccupazione 25-64 anni
X.SP Rapporto spesa pubblica/pil
X.TAX Percentuale di tasse sul lavoro sui bassi salari
X.BP Busta paga netta media annuale per un lavoratore single senza figli
X.PIL Pil pro-capite a parità di potere di acquisto
X.LC Aumento proporzionale del costo del lavoro

```
plot(D); cor(D)
```

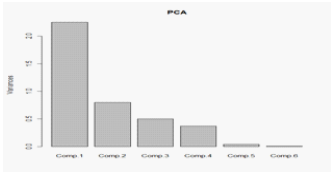


	X.TD	X.SP	X.TAX	X.BP	X.PIL	X.LC
X.TD	1	-0.040	0.295	-0.515	-0.719	-0.545
X.SP	-0.040	1	0.276	0.190	0.079	0.124
X.TAX	0.295	0.276	1	-0.328	-0.3249	-0.264
X.BP	-0.515	0.190	-0.328	1	0.938	0.917
X.PIL	-0.719	0.079	-0.3249	0.938	1	0.936
X.LC	-0.545	0.124	-0.264	0.917	0.936	1

```
PCA <- princomp(D); biplot(PCA)
```



(le tre frecce a sinistra sono BP, PIL, LC). plot(PCA):




```
PCA$loadings
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
X.TD	0.544		0.806	0.143		-0.183
X.SP		0.636	0.130	-0.752		
X.TAX	0.272	0.735	-0.284	0.547		
X.BP	-0.495	0.166	0.423	0.141	-0.596	0.417
X.PIL	-0.485			0.208		-0.837
X.LC	-0.386	0.135	0.253	0.226	0.796	0.291

Molti loadings sono pari a zero. Non è uno zero in senso stretto: recita l'help:

```
?loadings
```

Small loadings are conventionally not printed (replaced by spaces), to draw the eye to the pattern of the larger loadings.

In questo modo l'interpretazione delle componenti può risultare più facile.

6.5.6 Esercizio n. 6

Pare che certe agenzie di rating usino (o abbiano utilizzato in passato) i seguenti indicatori, tra altri, per assegnare le nazioni in questa o quella categoria:

1) PIL (Prodotto Interno Lordo – valore complessivo dei beni e servizi prodotti)

2) debito pubblico (debito dello Stato nei confronti di chi ha sottoscritto obbligazioni – quali, in Italia, BOT e CCT – destinate a coprire il fabbisogno finanziario) sullo stesso

3) deficit del bilancio pubblico.

Più di recente, pare vengano utilizzati anche parametri come la differenza tra attività e passività finanziarie e l'entità dei debiti delle famiglie e delle imprese in relazione al PIL.

Si trovino i dati di alcuni di questi indicatori in rete, relativamente al periodo che si ritiene più opportuno, o usando dati incrementali, o medie o altro secondo la propria intuizione.

Si costruisca una tabella con le nazioni scelte come righe, gli indicatori scelti come prime p colonne, e come colonna (p+1)-esima una colonna di 0 e 1 così pensata.

Si possono eseguire vari esercizi. Si può prendere come ultima colonna una classificazione binaria proposta, relativamente ad un certo anno, da una agenzia di rating. Oppure si può scegliere la classificazione in nazioni che hanno già subito bancarotta rispetto a quelle che non l'hanno subita. Oppure una classificazione ideale in cui ad esempio Grecia e Irlanda hanno 1, in quanto sono le nazioni verso cui l'Europa sta già effettuando operazioni massicce di aiuto economico.

Si utilizzino poi i comandi della regressione logistica per assegnare una probabilità di "fallimento" alle varie nazioni esaminate.

Esempio artificiale di prova. Partiamo dalla solita tabella degli indicatori di benessere (per lo scopo che abbiamo, la loro standardizzazione non era necessaria). Assegnamo punteggio 1

alle nazioni del Nord Italia: Piem, Vaos, Lomb, TrAA, Vene, FrVG, Ligu. EmRo.

	PLIC	SC	SA.SC	TD	TMI	Geo
Piem	0.088	0.471	-0.707	-0.607	-0.395	1
Vaos	-1.545	0.348	-0.642	-0.813	1.578	1
Lomb	0.202	1.397	-0.836	-0.790	-0.538	1
TrAA	0.677	0.435	-1.269	-0.966	-0.075	1
Vene	0.088	1.334	-1.210	-0.848	-0.497	1
FrVG	0.639	-0.005	-1.028	-0.804	-1.301	1
Ligu	1.190	-0.247	0.470	-0.429	-0.354	1
EmRo	0.658	1.177	-1.315	-0.863	-0.347	1
Tosc	0.126	1.092	-0.795	-0.644	-1.355	0
Umbr	-1.431	0.675	-0.140	-0.524	-1.287	0
Marc	0.278	1.090	-0.265	-0.702	-0.0006	0
Lazi	2.329	0.546	-0.080	-0.113	-0.014	0
Abru	0.335	-0.373	0.402	-0.456	0.040	0
Moli	0.658	-1.289	0.065	0.451	-1.151	0
Camp	-1.811	-1.314	2.031	1.664	0.414	0
Pugl	-0.766	-0.926	1.038	0.648	1.109	0
Basi	-0.747	-1.154	0.661	0.844	2.001	0
Cala	-0.500	-1.727	1.571	2.153	0.632	0
Sici	-0.918	-1.130	1.332	1.517	1.783	0
Sard	0.449	-0.403	0.717	1.285	-0.238	0

Copiare la tabella su un file txt, salvarlo col nome IBplus nella cartella della lezione, da R cambiare directory e caricarlo col comando

```
IB <- read.table(file=IBplus.txt,header=T)
```

Su R, scrivere

```
IB
```

per verificare.

Costruiamo i vettori con le singole variabili:

```
PLIC<-IB[,1]; SC<-IB[,2]; SA.SC<-IB[,3]; TD<-IB[,4]; TMI<-IB[,5]; Nord<-IB[,6]
```

Provare, a titolo di esempio, ad eseguire la regressione:

```
reg<- lm(TD ~PLIC+SC+SA.SC+TMI)
```

e poi chiedere informazioni con `summary(reg)`. In sisntesi, l'esito è:

Estimate Pr(>|t|)

(Intercept) 1.065e- 1.00000

PLIC 6.308e-04 0.99576

SC -3.006e- 0.13320

SA.SC 6.481e-01 0.00496 **

TMI 8.899e-03 0.94400

Multiple R-squared: 0.8465, Adjusted R-squared: 0.8055

p-value: 5.793e-06

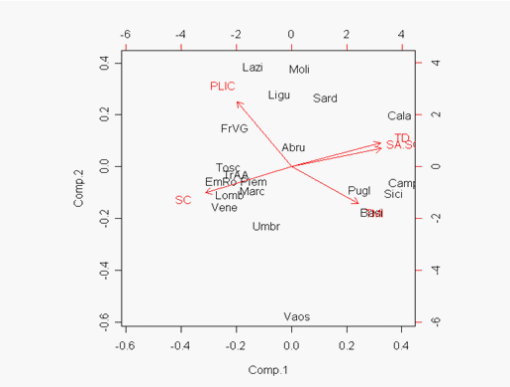
Si vede che, pur non essendoci un chiaro significato causa/effetto, come metodo previsivo può funzionare (salvo il fatto che I dati erano riferiti allo stesso anno; andrebbe rifatto usando

anni diversi). Veniamo alla regressione logistica: si usano i Generalized Linear Models con distribuzione in uscita binomiale

```
Nordism<-glm(Nord ~SC+SA.SC+TD,family=binomial)
predict(Nordism,type = response)
```

	Nordism	Geo
Piem	0.88	1
Vaos	0.99	1
Lomb	0.67	1
TrAA	0.99	1
Vene	0.97	1
FrVG	0.99	1
Ligu	0.24	1
EmRo	0.99	1
Tosc	0.27	0
Umbr	0.040	0
Marc	0.24	0
Lazi	2.7 e-06	0
Abru	0.65	0
Moli	3.3 e-07	0
Camp	2.2 e-16	0
Pugl	1.0 e-11	0
Basi	1.1 e-12	0
Cala	2.2 e-16	0
Sici	2.2 e-16	0
Sard	2.2 e-16	0

Si tenga presente che non sono punteggi, non seguono una scala lineare, ma rappresentano la probabilità di essere classificati Regione del Nord Italia. Come verifica osserviamo PCA svolta a suo tempo:



Esercizio: eseguire la regressione `Nord~SC` + `SA.SC` + `TD` e fare predict.