DIPARTIMENTO
DI MATEMATICA

SAPIENZA
UNIVERSITÀ DI ROMA

Metodi e modelli matematici per l'ambiente

edizione 2024/2025

Dario Benedetto - `http://brazil.mat.uniroma1.it/dario`

Sapienza Università di Roma
Dipartimento di Matematica
Piazzale Aldo Moro n. 5, 00185 Roma
www.mat.uniroma1.it

MMMA 2024/2025

December 11, 2024

Contents

Introduzione

Uno degli scopi di questo corso è farti conoscere alcuni argomenti della matematica che possono esserti utile come esperto di tematiche ambientali. Una parte del corso sarà dedicata ai metodi della statistica, con esercitazioni con R in laboratorio. In queste note troverai poco di questi argomenti, perché esistono testi soddisfacenti, che elenco nel paragrafo successivo. Qui troverai invece gli appunti sulla parte di modellistica matematica. In estrema sintesi, un modello matematico consiste in

a) identificare variabili e parametri con cui descrivere alcuni aspetti di un fenomeno;

b) inventare/scoprire le relazione matematiche tra queste grandezze;

c) esplorare con strumenti analitici e numerici il comportamento del modello, in particolare

- fare previsioni
- studiare come cambia il modello al cambiare dei parametri
- validare il modello, cioè confrontare gli esiti dell'esplorazione con il comportamento reale del sistema.

Incontrerai, presumibilmente, due tipi di difficoltà. La prima è che potresti far fatica a seguire alcuni passaggi e alcuni ragionamenti, perché hai dimenticato molta della matematica che ti è stata insegnata nella triennale. In questi appunti ci sono esercizi che servono a risvegliare le tue competenze matematiche, ma il corso non tratta di questo, ha ambizioni maggiori che insegnarti di nuovo i logaritmi (ma servirà anche a questo).
La seconda difficoltà è più sottile e ha a che fare con l'interazione tra i due piani descrittivi, quello naturalistico e quello matematico. Questa interazione è indispensabile nei punti **a)** e **b)**, in cui la comprensione degli aspetti biologici e ambientali si deve tradurre in relazioni matematiche (ti aiuterà il fatto che in molti casi queste relazioni sono di pochi tipi differenti). Nel punto **c)**, invece, è solo la matematica che deve parlare, con i suoi metodi deduttivi. La difficoltà consiste nel non confondere la spiegazione matematica con quella naturalistica. L'utilità della matematica è proprio qui: dopo aver stabilito il modello, non serve l'intuito o l'esperienza o qualche conoscenza più profonda: è solo il ragionamento, aiutato da strumenti analitici e numerici, che ci permette di arrivare a una descrizione quantitativa e qualitativa del fenomeno.
Questa relazione tra matematica e natura è chiarissima in fisica (per esempio non possiamo mandare un razzo sulla Luna a "intuito"), mentre solo alcuni aspetti delle scienze naturali sono matematizzati, e questo rende più facile la confusione tra i due piani.
Faccio un esempio in dettaglio: il modello preda-predatore è stato inventato negli anni 20 del 1900, indipendentemente da due scienziati, Lotke (biofisico, chimico, statistico) e Volterra (fisico-matematico di questo ateneo). In particolare Volterra si interessò al fatto

che, durante la prima guerra mondiale, il numero di pesci predatori nel mare Adriatico crebbe in corrispondenza del diminuire della pesca delle specie adatte al consumo umano. Ci occuperemo di questo famoso modello, qui noto solo che, una volte comprese le relazioni tra le grandezze in gioco, si possono matematicamente ottenere varie conclusioni, sulla numerosità dei pesci preda e dei pesci predatori:

i. le due numerosità hanno un andamento periodico nel tempo;

ii. le numerosità medie nel tempo dipendono solo dai parametri del sistema, quindi anche se improvvisamente aggiungessimo prede al sistema, o eliminassimo la metà dei predatori, le medie non cambierebbero;

iii. se cambiamo i parametri del sistema, rendendo più semplice la vita delle prede (simulando in questo modo la riduzione della pesca), allora il numero medio di prede non cambia, mentre il numero medio di predatori aumenta.

Si possono dare delle spiegazioni naturalistiche di queste conclusioni, per esempio

i. se le prende crescono di numero, allora i predatori possono nutrirsi di più, dunque cresceranno di numero facendo decrescere il numero di prede; questo fatto comporterà la diminuzione del numero di predatori che permetterà così l'aumento delle prede, chiudendo il ciclo;

ii. il sistema è in equilibrio ecologico, e non cambia se si cambiano artificialmente le numerosità delle popolazioni (se no non sarebbe un equilibrio);

iii. in una catena trofica, le specie più in alto traggono maggior vantaggio da un aumento delle risorse alla base.

Non c'è nulla di biologicamente errato in questo ragionamenti, ma sono di qualità differente da quelli matematici, infatti sono descrittivi, mentre quelli matematici sono deduttivi. In un ragionamento matematico, le conclusioni sono inevitabili conseguenze delle premesse; per cambiare conclusioni si devono cambiare le ipotesi di partenza del modello, approfondendone la comprensione. Le spiegazioni naturalistiche in questo caso riassumono le conclusioni matematiche, e la loro ragionevolezza ci fa capire che il modello ha una sua solidità dal punto di vista biologico. D'altra parte, la matematica contribuisce ad ampliare le conoscenze non solo con le sue deduzioni, ma anche costringendo lo scienziato a trovarne una sintesi naturalistica (farò esempi a proposito dello shift dei sistemi ecologici).

In questo corso tenterò di insegnarvi a distinguere i ragionamenti matematici da quelli naturalistici, mostrandovi come la matematica permetta a volte di raggiungere conclusioni altrimenti inaccessibili, e come alcuni concetti matematici in realtà siano alla base di descrizioni naturalistiche che oggi ci sembrano ovvie.

Fonti

BDM D. Benedetto, M. Degli Esposti, C. Maffei: Matematica per le scienze della vita, CEA terza edizione 2015 (testo per i corsi di matematica del I anno, include l'introduzione ai test statistici, tranne ANOVA)

WS M.C. Whitlock, D. Schulter: Analisi statistica dei dati biologici, Zanichelli 2010 (testo molto bello, un po' avanzato)

M David S. Moore: Statistica di base, Apogeo Education, seconda edizione 2013 (testo di base di statistica)

R Sheldon M. Ross: Introduzione alla statistica, Apogeo Education, seconda edizione 2014 (testo di base di statistica)

IM Stefano M. Iacus, Guido Masarotto: Laboratorio di statistica con R, McGraw-Hill Education, seconda edizione 2021 (testo di introduzione a R e alla statistica)

Chapter 1

Richiami sulle funzioni elementari

Scopo di questa sezione è rinfrescare qualche nozione di matematica elementare in termini modellistici. Lo farò discutendo alcuni semplici esercizi. Trovate esempi più dettagliati e la teoria su [BDM capp. 5,6,7,8].

1.1 Leggi lineari

Esercizio 1. Pressione

La pressione atmosferica a livello del mare è di (circa) una atmosfera, e cresce di (circa) una atmosfera ogni 10 metri di profondità.

a. Scrivi la legge $P(h)$ che esprime il valore della pressione P in funzione del valore della profondità h.

b. Disegnane il grafico

c. Cosa è proporzionale nella legge che hai scritto?

d. Cosa rappresenta geometricamente il coefficiente di h nel grafico che hai disegnato?

Risposte

a. $P(h) = 1 + h/10$

b. Il grafico è rappresentato da una retta (anzi da una semiretta, perché la legge descritta non ha senso per $h < 0$).

c. Sono proporzionali la **variazione** di pressione ΔP e la variazione di profondità Δh. Più formalmente, dati h_1 e h_2,

$$\Delta h = h_2 - h_1,\ \Delta P = P(h_2) - P(h_1) = \Delta h/10$$

d. Il coefficiente di proporzionalità tra ΔP e Δh è la **velocità media di variazione** di P con h, ed è costante:

$$\frac{\Delta P}{\Delta h} = \frac{1}{10}$$

qualunque siano h_1 e h_2.

Esercizio 2. Crescita di una larva - BDM esempio 5.1.7 e seguenti

La massa m di una larva di insetto pesa alla nascita 10 g, dopo 20 ore pesa 24 g, dopo 30 ore pesa 30 g. Disegna questi dati in un grafico.

a. Quanto pesa la larva dopo 10 ore dalla nascita?

b. E dopo 25?

c. E dopo 40?

(Per risolvere questo esercizio, tra le altre cose hai bisogno di ricordare come si scrive la retta che passa per due punti, vedi [BDM capitolo 5])

Va notato che la velocità media di variazione tra $t = 0$ e $t = 20$ è diversa da quella tra $t = 24$ e $t = 30$. Per dare un valore di $m(10)$ si ricorre all'**interpolazione lineare** tra i dati per $t = 0$ e $t = 20$. Per dare un valore di $m(25)$ si ricorre all'interpolazione tra i dati per $t = 20$ e $t = 30$. Per dare un valore di $m(40)$ si ricorre all'**estrapolazione lineare** usando i dati per $t = 20$ e $t = 30$.

Ricordo che si definisce il concetto fisico e matematico di **velocità istantanea** passando al limite il valore della velocità media mandano a 0 l'incremento. Supponendo di conoscere tutti i valori di $m(t)$, la velocità di variazione istantanea al tempo t è

$$\lim_{\Delta t \to 0} \frac{\Delta m}{\Delta t} = \lim_{\Delta t \to 0} \frac{m(t + \Delta t) - m(t)}{\Delta t}$$

Come è noto, questo valore è per definizione la **derivata** di $m(t)$ (si vedano [BDM capp. 7, 8] per i richiami su limiti e derivate).

La tecnica di interpolazione e estrapolazione e il concetto di derivata sono due dei motivi che spiegano l'abbondanza delle leggi lineari in natura: per piccole variazioni, ogni funzione regolare è ben approssimata da una retta. Vedi [BDM parr. 8.1 e 8.3]

1.2 Leggi esponenziali

Esercizio 3. Duplicazione batterica

Sia $N(t)$ il numero di batteri all'ora t, in una capsula Petri in cui si possono riprodurre senza vincoli. Supponiamo che $N(0) = 10^3$, e che N raddoppi ogni ora.

a. Quanto vale $N(t)$? Quali quantità sono proporzionali?

b. Supponiamo che N raddoppi ogni 3 ore. Quanto vale $N(t)$?

Risposte

a. $N(t) = 10^3 \times 2^t$

b. $N(t) = 10^3 \times 2^{t/3}$

Le due leggi appena scritte sono leggi **esponenziali**. In matematica si usa esprimere le leggi un una base particolare, il numero e, per motivi legati alle proprietà delle derivate. Ricordo che $\ln x$ è la funzione inversa dell'esponenziale e^x. cioè

$$\ln \mathrm{e}^x = x, \quad \mathrm{e}^{\ln y} = y$$

Dunque

$$2^t = \mathrm{e}^{\ln 2^t} = \mathrm{e}^{t \ln 2}$$

dove ho usato la proprietà dei logaritmi

$$\ln a^b = b \ln a$$

In generale, la legge di crescita che abbiamo descritto ha la forma

$$N(t) = N_0 \mathrm{e}^{\alpha t}$$

e non è evidentemente una legge lineare, infatti la velocità di crescita media non è costante. Calcoliamo la velocità istantanea, facendo la derivata:

$$N'(t) = N_0 \alpha \mathrm{e}^{\alpha t} = \alpha N_0 \mathrm{e}^{\alpha t} = \alpha N(t)$$

Dunque in una legge di crescita esponenziale, la velocità istantanea di variazione è proporzionale alla numerosità.
Un altro modo per descrivere questo modello è di ricordare che il **tasso di variazione** è proprio il rapporto tra la velocità di variazione e la quantità che stiamo considerando. Per esempio, tornando all'esempio dei batteri, se la popolazione duplica in un ora, allora il tasso medio di variazione in un'ora è

$$\frac{N(t+1) - N(t)}{N(t)} = 1 = 100\%$$

È da notare che che il tasso istantaneo di variazione è invece $\ln 2 \approx 0.7$. Le leggi esponenziali descrivono fenomeni con tassi di variazione costanti, ma l'intervallo su cui viene misurato il tasso deve essere fisso, al variare dell'ampiezza dell'intervallo cambia anche il tasso.

Siamo passati dalle leggi lineari, in cui la velocità di variazione è costante, a una legge in cui la velocità di variazione non è costante ma proporzionale alla quantità stessa. Torneremo su questo punto.

1.3 Le scale delle grandezze e i logaritmi

Nei paragrafi precedenti ho provato a convincervi che se sto studiando un fenomeno per piccole variazioni delle grandezze in gioco, mi posso aspettare una proporzionalità tra esse, che sarà falsa man mano che le variazioni crescono.
In genere, un fenomeno viene descritto per un intervallo (un "range") di valori delle variabili che fissa la **scala** in cui analizzarlo. Nell'esempio della larva l'intervallo è $[0, 40]$, dunque la scala è quella della decina di gironi. Nell'esempio della pressione non è specificato, ma la legge scritta varrà fino a che l'acqua può essere considerata incomprimibile, anche a 10 000 metri di profondità.

Al cambiare della scala, un particolare aspetto di un fenomeno cambia radicalmente. Per esempio, concentriamoci su cosa c'è intorno a noi alle varie scale di distanze.

1 m una stanza

100 m il quartiere

10 km la città

1000 km la nazione

100 000 km il nostro pianeta e lo spazio fino a circa 1/4 della distanza dalla Luna

La scala è ben descritta dal logaritmo della grandezze. Infatti, in questo esempio, se considero il logaritmo in base 10 ottengo la sequenza 0, 2, 4, 6, che si ottiene con incrementi costanti, ai quali corrispondono grandezze moltiplicate per $100 = 10^2$.
Quando si vogliono rappresentare fenomeni a scale differenti si usano gli **assi logaritmici** (vedi [BDM par. 6.2, in particolare le figure 6.26, 6.27, 6.28])

Un fatto importante è che la legge esponenziale $f(x) = a\mathrm{e}^{\alpha x}$, usando un asse verticale logaritmico è descritta da una retta, infatti

$$\ln f(x) = \ln a + \alpha x$$

Due insegnamenti;
1. nelle leggi esponenziali c'è una relazione di proporzionalità tra la scala del fenomeno e la variabile indipendente;
2. le leggi esponenziali si rappresentano (e si cercano) utilizzando preferibilmente assi verticali logaritmici.

Ci si potrebbe chiedere quali sono i fenomeni naturali in cui bisogna tenere conto della variazione di scala. Faccio due esempi, ma ne faremo altri
1. L'acidità delle acque influisce molto sulla biologia delle specie che le abitano. L'acidità si misura con il pH, che è

$$pH = -\log[H^+]$$

cioè l'opposto del logaritmo in base 10 della concentrazione di ioni idrogeno in moli per litro. Se il pH è 7, la concentrazione è $10^{-7}\,\mathrm{mol}/\,\ell$. Se cambia il pH cambia la scala della concentrazione, e sono questi cambiamenti che hanno realmente effetti biologici (vedi BDM esempi 6.2.11-14).
2. La legge (empirica) di Weber-Fechner asserisce che la variazione della risposta fisiologica a uno stimolo è proporzionale allo stimolo stesso. Per fare un esempio, avvertiamo facilmente la differenza in peso tra 100 grammi e 110 grammi, quella tra 1000 e 1100, ma abbiamo difficoltà a distinguere 1000 grammi da 1010 grammi. In formule, se con S indichiamo lo stimolo, e con p la percezione,

$$\Delta p = k\frac{\Delta S}{S}$$

che possiamo riscrivere come

$$\frac{\Delta S}{\Delta p} = S/k$$

Questa relazione è la stessa che abbiamo provato per le leggi esponenziali. Passando al limite

$$S'(p) = S/k, \quad \text{da cui } S = S_0\mathrm{e}^{(p-p_0)/k}$$

(scritta in questo modo sono sicuro che per $p = p_0$ si ha $S = S_0$). Però sono interessato alla percezione in funzione dello stimolo, cioè alla funzione inversa $p = p(S)$. Passando ai logaritmi si ottiene

$$p(S) = p_0 + k \ln \frac{S}{S_0}$$

In entrambi questi esempi ho descritto dei parametri biologici che dipendono dal logaritmo di quelli fisico/ambientali, cioè dalla loro scala. Per questo il logaritmo dovrebbe essere il migliore amico dello scienziato ambientale.

In questo esempio ho mescolato ragionamenti fisiologici a questione matematiche. Distinguiamole. Questa è una assunzione del modello, che riassume semplificandole, delle osservazioni empiriche di fisiologia:
(P) la variazione di percezione è proporzionale alla variazione specifica dello stimolo.
In simboli matematici

$$\Delta p = k \frac{\Delta S}{S}$$

Con questa equazione abbiamo terminato la formulazione del modello, perché abbiamo trovato la legge che lega le quantità che ci interessano (in questo caso percezione e stimolo). Da questo punto in poi usiamo solo, deduttivamente, la matematica per ottenere informazioni da questo modello. Il primo passaggio che abbiamo fatto è stato di scrivere il rapporto incrementale dello stimolo in funzione della percezione, poi siamo passati al limite, riformulando il modello in termini di velocità istantanea di variazione. Infine abbiamo risolto l'equazione differenziale e abbiamo manipolato la soluzione con le regole degli esponenziali e dei logaritmi:

$$S'(p) = S/k, \quad \textbf{da cui } S(p) = S_0 \mathrm{e}^{(p-p_0)/k}, \quad \textbf{o, equivalentemente } p(S) = p_0 + k \ln \frac{S}{S_0}.$$

Questa espressione ci dice che
(C) la percezione dipende linearmente dal logaritmo dello stimolo.
Si noti che l'affermazione (C), non è quella di partenza: l'ipotesi che usiamo per costruire il modello è l'affermazione (P), l'affermazione (C) si ottiene deduttivamente da (P).

1.4 Leggi a potenza

Le leggi a potenza sono le funzioni del tipo

$$f(x) = ax^\alpha.$$

Al variare di α queste funzioni hanno differenti aspetti. (puoi vedere il loro grafico su BDM cap 5). Hanno una notevole rilevanza in fisiologia e anche in biologia.

Esercizio 4. Formiche giganti
Supponi di ingrandire una formica di un fattore ℓ, cioè di moltiplicare tutte e tre le dimensioni spaziali per ℓ, immaginando di ottenere un organismo con le stesse caratteristiche della formica originaria.

Come varia la massa? Come varia l'area della sezione di una zampa? Come varia la pressione che il peso esercita sulla sezione della zampa?

In questo esercizio si dà per scontato che la formica ingrandita abbia la stessa densità di massa della formica piccola, cioè sia fatta delle stesse sostanze. Il volume scala con ℓ^3, poiché la densità e costante, anche la massa, che è pari alla densità per il volume, scala come ℓ^3. L'area della sezione della zampa, invece, scala come ℓ^2, mentre la pressione, che è pari alla forza (in questo caso la forza peso) diviso l'area della sezione, scala come

$$\ell^3/\ell^2 = \ell$$

Ne segue che portare le dimensioni di una formica da un millimetro a un metro, aumenta la pressione sulle zampe di mille volte.
Questo esempio serve per far notare che la biologica di un organismo ha la sua scala di validità, e per cambiare scala sono necessari adattamenti fisiologici importanti.

Esercizio 5. Superfici e volumi

Nello sviluppo degli organismi, le crescite dei tessuti sono a volte fenomeni di superficie. È dunque utile calcolare come scala la superficie all'aumentare del volume.
Poiché il volume V scala come la lunghezza al cubo, possiamo invertire questa relazione e affermare che la lunghezza scala con $V^{1/3}$; la superficie scala come la lunghezza al quadrato, e dunque come $V^{2/3}$.

L'esponente 2/3 che abbiamo ottenuto nell'esercizio si incontra spesso nelle leggi allometriche, cioè nelle leggi che esprimono delle relazioni quantitative tra le parti degli organismi.
Per esempio, il peso dell'uovo degli uccelli va come il peso dell'uccello elevato alla 2/3, considerando uccelli di specie differenti.

In questo grafico che ho preso da S.J. Gould *Bravo brontosauro* par. 7, "Le uova del kiwi e la campana della libertà", è riportato un tipo grafico il scala logaritmica "dal topo all'elefante" (in questo caso dal colibrì al moa), cioè un grafico in cui sono riportate le dimensioni degli organismi di uno stesso genere ma in scala logaritmica, che viene usata perché tra un specie e l'altra c'è veramente un salto di scala.
Anche l'asse verticale di questo grafico è un asse logaritmico, e questo tipo di grafico viene chiamato log-log. Nell'asse verticale sono rappresenti i pesi tipici delle uova e si nota come la relazione tra le due grandezze sia lineare; in particolare la pendenza è 2/3. Dunque, se

indichiamo con u il peso delle uova e con p il peso dell'uccello, il grafico ci dice che

$$\ln u(p) = c + 2/3 \ln p$$

(la costante c non è determinabile dal grafico, ma non è 0, perché il punto in basso a destra non è l'origine). Determiniamo $u(p)$: passando all'esponenziale

$$\mathrm{e}^{\ln u(p)} = \mathrm{e}^{c+2/3 \ln p} = \mathrm{e}^{c} \times \mathrm{e}^{\ln p^{2/3}} = 2^{c} p^{2/3}$$

Questo esempio serve a ricordare che se il grafico in scala log-log è una retta, allora la relazione tra le grandezze è data da una legge a potenza. Nel caso del grafico in scala log, in cui solo l'asse verticale è logaritmico, si ottiene invece una legge esponenziale.

Concludo per completezza l'esempio del kiwi. Il kiwi ha un uovo di dimensioni spropositate in relazione alla dimensione dell'individuo. Varie spiegazioni "darwiniane' vengono date per questo fatto, Gould e altri suggeriscono che il kiwi sia una versione "nana" di uccelli ora estinti, della dimensione del moa. Per suffragare questa tesi, osservano che l'esponente che lega il peso delle uova e il peso totale tra individui della stessa specie è circa 0.15, e non 2/3 (in generale la potenza delle leggi allometriche intraspecifiche è differente e più piccolo dell'esponente nel caso di leggi interspecifiche). Dal grafico si vede come il punto che rappresenta il kiwi sia, in scala log-log, sulla retta di pendenza 0.15 che passa per il punto che rappresenta il moa.

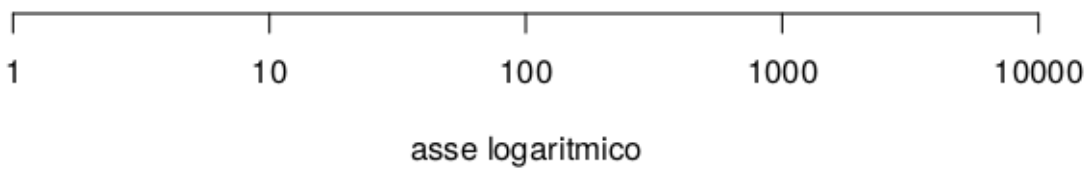
Esercizio 6.

Nell'ultimo esempio ho ipotizzato che l'asse rappresentasse il logaritmo naturale, dunque da $\ln u = c + 2/3 \ln p$ ho ottenuto

$$u = a p^{2/3},$$

con $a = e^{c}$. Come cambia questa legge se suppongo che sugli assi ci fossero i logaritmi in base 10?

Esercizio 7. Assi logaritmici



asse logaritmico

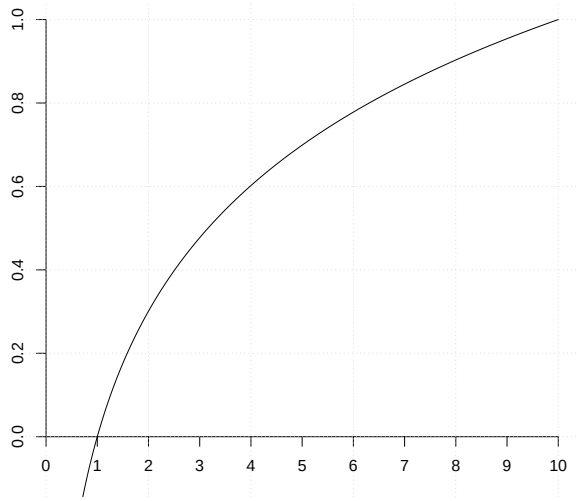
Nel grafico è disegnato una asse orizzontale logaritmico.

Disegna i punti 5, 50, 500.

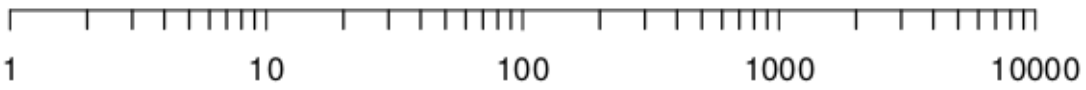
Disegna il punto $\sqrt{10} \approx 3.2$.

Dove andrebbe disegnato lo 0 in questo grafico?

Può essere utile osservare il grafico della funzione log, il logaritmo in base 10.



La soluzione di questo esercizio è parzialmente nel grafico seguente



I segna-numeri tra 1 e 10 corrispondono ai valori 2, 3, ... 9, quelli tra 10 e 100 corrispondono ai valori 20, 30, ... 90, etc.

Esercizio 8. Metabolismo

Negli anni 30 il biologo M. Kleiber studiò la relazione tra il metabolismo basale dei carnivori e il loro peso. In un grafico log-log, in base 10, il valore del metabolismo m espresso in ml di O_2 emessi in un'ora, è legato al peso p espresso in grammi, da una legge lineare di intercetta 0.6 e di coefficiente angolare 0.7.
Scrivere la legge $M(p)$.

Esercizio 9. Cervelli

Supponendo che il peso C del cervello dei primati, espresso in grammi, sia legato con la seguente legge a potenza al peso P dell'animale, espresso in kg:

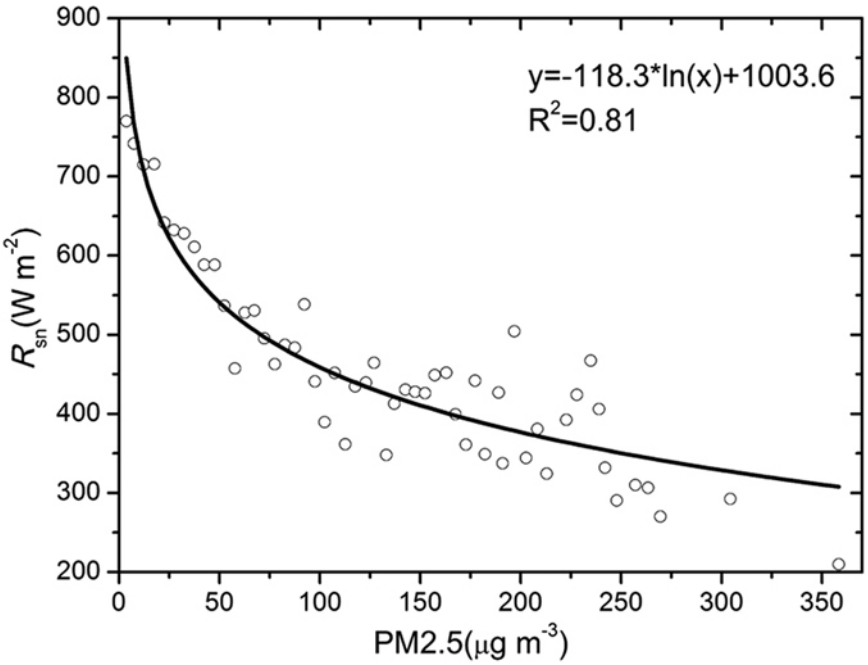
$$C = 60 \times P^{1/4}$$

Calcola i parametri della legge lineare che vedresti in un grafico log-log in base 10.

Come cambia la legge se misuri P in grammi? E come cambia la retta?
Come cambia la legge se misuri C in chilogrammi? E come cambia la retta?
Come cambia la retta se il grafico log-log è in base e?

Esercizio 10. Effetti delle PM2.5

Ho preso la figura che segue da un articolo sugli effetti della presenza di PM2.5 sull'intensità della radiazione solare R_S a banda larga sul cielo di Pechino.



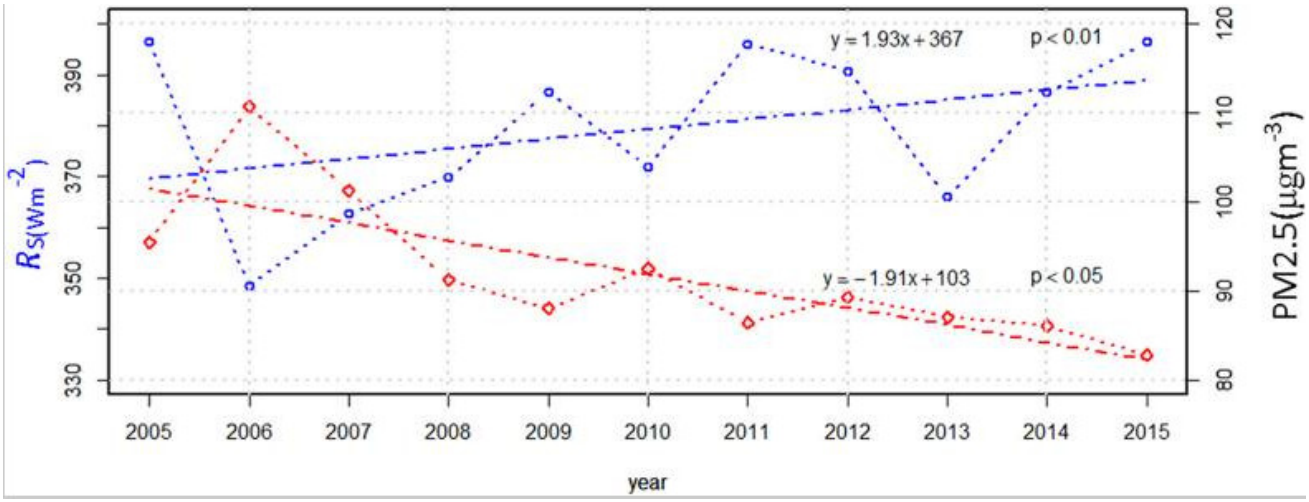
Sono mostrati i dati e la curva che li approssima meglio (vedremo in seguito come si costruisce).

Quale asse dovrebbe essere logaritmico per vedere una retta?

Nella figura seguente sono mostrati i dati medi annui di R_S e $PM2.5$.

Si noti che le due leggi lineari scritte in figura non sembrano coerenti con le scale. Si scrivano le leggi lineari (approssimate) che esprimono R_S in funzione dell'anno A, e $PM2.5$ in funzione dell'anno A. (Suggerimento: si utilizzi l'espressione per la retta tra due punti notando che sulla retta rossa R_S vale circa 370 nel 2005, e circa 350 nel 210; si proceda analogamente per PM2.5).

È coerente questa immagine con i dati dell'immagine precedente? Quale sarebbe l'espressione di R_S in funzione di PM2.5 che si dovrebbe dedurre da questa figura?



1.5 Esercizi di richiamo

Esercizio 11. Leggi lineari

a) Sia $(2-r)/3 = (3s+5)/7$. Esprimi r in funzione di s, e esprimi s in funzione di r.

b) Se $(p+a)/b = (b-2q)/a$

Come si esprime q in funzione di p?

Come si esprime p in funzione di q?

c) Trova p in funzione di q sapendo che se p vale 2, q vale 3, se p vale 3, q vale 2.

d) Trova p in funzione di q sapendo che se p vale $\bar{p}$, q vale $\bar{q}$, se p vale p_0, q vale q_0.

e) Supponi che $p = 2q+3$ e $s = 5q-1$. Esprimi p in funzione di q, ed esprimi q in funzione di p.

f) Supponi che $\Delta p/\Delta q = -2$, e p vale 3 se q vale 4. Scrivi p in funzione di q.

g) Un variazione di un grado Celsius corrisponde a una variazione di 1.8 gradi Fahreneit. Inoltre la temperatura di congelamento dell'acqua, di 0 gradi Celsius, è pari a 32 gradi Fahreneit. Scrivi le formule di trasformazione da gradi Celsisus a Fahreneit e viceversa.

h) Il peso di un neonato aumenta di circa 30 grammi al giorno. Deternina il peso in funzione del tempo, sapendo che nel suo settimo giorno di vita il suo peso è di 3.79 kg. Quanto pesava il terzo giorno? Quanto peserà il decimo? Quando supererà i 4 kg?

Esercizio 12. Proprietà degli esponenziali

Rendi più "semplice" le seguenti espressioni:

a) $10^3 \times 10^5 =$

b) $3^0 =$

c) $6^1 =$

d) $2^{-1} =$

e) $2.5^4 \times 2.5^{-1} =$

f) $4^{2^3} =$

g) $(4^2)^3 =$

h) $5^2 \times 5^{-3}/5^{-1} =$

i) $(ab^2)^5 =$

j) $(a/b^2)^{-1} =$

Esercizio 13. Proprietà dei logaritmi

a) $\log(ab) =$

b) $\log a^b =$

c) $\log_{10} 10 =$

d) $\log_{10} 10^{-3} =$

e) $10^{\log_{10} 14} =$

f) $10^{\log_{10} 14^4} =$

g) $10^{4\times\log_{10} 14} =$

h) $\log 1 =$

i) $\log 10^4 =$

j) $\ln 10^4 =$

k) Esprimi un numero a in funzione di 10^a.

l) Esprimi un numero a in funzione di e^a.

m) Esprimi un numero $a > 0$ in funzione di $\ln a$.

Esercizio 14.

Considera la legge di crescita esponenziale

$$N(t) = 1000 \times 1.1^{t/3}.$$

Riscrivila utilizzando la base e; riscrivila utilizzando la base 10, riscrivila utilizzando la base 2. Determina in quanto tempo N aumenta del 50%, in quanto tempo raddoppia, e in quanto tempo decuplica.
In quanto tempo $N(t)$ diventa 10^6?

Chapter 2

Modelli di evoluzione

In questo capitolo illustrerò i concetti di base della modellistica matematica, attraverso alcuni esempi significativi.
Ritorniamo al semplice modello della duplicazione batterica. Il modello consiste in una legge, espressa in forma matematica, che permette di **determinare il futuro, conoscendo il presente**. Nel caso della duplicazione batterica, l'assunzione del modello è che **il numero di batteri raddoppia ogni ora**. Se conosciamo la numerosità N all'ora t, che indicheremo come $N(t)$ (il "presente") possiamo determinare la numerosità all'ora successiva, cioè $N(t+1)$ (il "futuro"), mediante la formula

$$N(t+1) = 2N(t)$$

Puntualizzo: in questa "legge di aggiornamento" non c'è scritto il valore della numerosità N, c'è solo la regola per determinare il futuro conoscendo il presente. Per conoscere concretamente $N(t)$ è necessario conoscere il **dato iniziale**, per esempio la numerosità al tempo $t = 0$, che chiamo N_0.
Tutte le informazioni sul sistema sono dunque calcolabili a partire da queste due informazioni:

$$\begin{cases} N(t+1) = 2N(t) \\ N(0) = N_0 \end{cases}$$

(in matematica questo sistema prende il nome di "problema ai dati iniziali"). Infatti se vogliamo conoscere $N(4)$ basta raddoppiare per 4 volte il valore di N_0, cioè $N(4) = 2^4 N_0 = 16N_0$. È importante notare che il "dato iniziale" non deve necessariamente essere quello al tempo 0. Per esempio se fissiamo $N(10) = 10\ 000$ siamo comunque in grado di prevedere $N(12) = 40\ 000$, e siamo in grado di **ricostruire il passato**. Per esempio, per calcolare $N(8)$ dovremo dividere due volte per due $N(8) = 10\ 000/4 = 2500$.

Chiameremo **soluzione** la funzione $N(t)$. Dalla legge che governa il modello siamo in grado di trovare una semplice espressione matematica per $N(t)$

$$N(t) = N_0 2^t$$

Se invece fissiamo il dato iniziale N_0 al tempo t_0, la legge diventa

$$N(t) = N_0 2^{t-t_0}$$

Esercizio 15. Modello lineare

Un aspetto quantitativo F di un fenomeno naturale si evolve con velocità di variazione costante a. Ipotizzando che F valga F_0 al tempo t_0, determinare l'espressione di $F(t)$.

Riposta: $F(t) = F_0 + a(t - t_0)$ infatti F al tempo t deve essere uguale a F_0 più la variazione, che è proporzionale al tempo passato $t - t_0$, con coefficiente a.

Modelli alle differenze e modelli differenziali

I due esempi precedenti sono formulati in modo lievemente differente: per il modello di duplicazione ho fornito la regola per calcolare l'avanzamento nel tempo, per il modello lineare ho dato un valore alla velocità di variazione. Il modo più utile di formulare un modello è proprio quest'ultimo.
Torniamo alla duplicazione: invece di scrivere $N(t+1) = 2N(t)$ posso scrivere

$$N(t+1) - N(t) = N(t)$$

a sinistra compare la differenza tra futuro e presente, e a destra c'è solo il presente. Noto anche che $N(t+1) - N(t)$ è la velocità media di variazione in un'ora. Dunque anche il modello di duplicazione si può formulare in termini di velocità di variazione.

Il vantaggio di questo modo di fare è che ci consente di prendere in considerazione anche le velocità istantanee.

Un **modello differenziale** è un modello in cui viene specificata la velocità istantanea di variazione in funzione dello stato presente. Per esempio il modello esponenziale è governato dalla legge

$$N'(t) = \alpha N(t)$$
$$N(t_0) = N_0$$

Abbiamo già trovato la soluzione per un problema di questo tipo: le funzioni che verificano $N' = \alpha N$ sono solo le funzioni $N(t) = c\mathrm{e}^{\alpha t}$, con c costante arbitraria. Imponendo il dato iniziale

$$N_0 = N(t_0) = c\mathrm{e}^{\alpha t_0}$$

e dunque $c = N_0\mathrm{e}^{-\alpha t_0}$, e infine si ottiene

$$N(t) = N_0\mathrm{e}^{\alpha(t-t_0)}.$$

Questo tipo di modello descrive in particolare la crescita malthusiana delle popolazioni: si fissa un intervallo di tempo di riferimento, per esempio un anno, si assume costante il tasso di natalità n =numero di nati/ numerosità della popolazione, e il tasso di mortalità m =numero di morti / numerosità. Dunque

$$N(t+1) - N(t) = (n-m)N(t)$$

Da questa legge si scopre che $N(t) = N(0)(1+n-m)^t = N(0)\mathrm{e}^{t\ln(1-n-m)}$, (questo passaggio l'ho già fatto nel primo paragrafo). Derivando in t si ottiene il modello differenziale con $\alpha = \ln(1+n-m)$ Se $\alpha > 0$ (cioè se il tasso di natalità supera quello di mortalità) la popolazione cresce esponenzialmente. Se accade il contrario, la popolazione decresce esponenzialmente. Se $n = m$, il modello prevede la costanza della numerosità popolazione.

Un altro fenomeno naturale governato da questo modello è il decadimento esponenziale delle sostanze radioattive: in numero di atomi di un isotopo radioattivo che decadono in una unità di tempo è proporzionale al numero complessivo di atomi. In altri termini, il tasso di "mortalità" degli atomi è costante. Più in generale, il modo più semplice e ragionevole di modellizzare quantitativamente fenomeni di mortalità o piú in generale di trasformazione è assumere tassi costanti. Per esempio, nei modelli di crescita tumorale si assume che il tasso di mortalità naturale delle cellule sia costante, nei modelli di epidemia si assume che il tasso di guarigione dei malati sia costante (che vuol dire che il numero di malati che guarisce in un intervallo di tempo fissato è proporzionale al numero di malati).

Faccio un esempio concreto. Il carbonio-14 si dimezza in circa 5730 anni, dunque se $M(t) = M(0)2^{-t/5730}$ (la massa subisce $t/5730$ dimezzamenti in t anni. Riscrivendola come legge esponenziale in base e: $M(t) = M(0)\mathrm{e}^{-at}$, con $a = \ln 5730 = 8.65$, e quindi

$$M'(t) = -aM(t)$$

è la massa al tempo t.

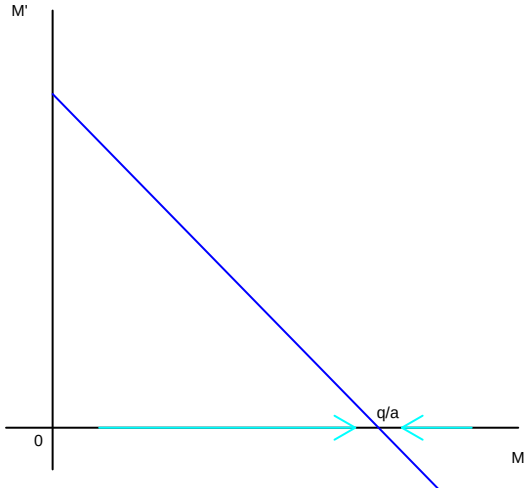
Fin'ora abbiamo analizzato tre leggi di variazione: a velocità costante, a tasso di crescita costante (natalità), a tasso di decrescita costante (mortalità). Ci sono molte situazioni interessanti in cui possono essere coinvolti più fattori (già nel modello di Malthus abbiamo considerato natalità e mortalità).

Per esempio ci possiamo chiedere come varia la quantità di carbonio 14 nell'atmosfera, tenendo presente che si forma per l'interazione tra i raggi cosmici e l'atmosfera (per la precisione con l'azoto-14). Assumiamo l'ipotesi ragionevole che l'azione dei raggi cosmici sia costante, e che la quantità di azoto-14 nell'atmosfera sia costante. Trascuriamo per il momento che il ^{14}C decade: ci sarebbe un aumento della massa a velocità costante dovuto ai raggio cosmici. Indichiamo con q questa velocità costante di accrescimento della massa. Complessivamente, la velocità di variazione di $M(t)$ avrà dunque due contributi: uno di decadimento pari a $-aM$, l'altro di accrescimento a velocità costante, pari a q. Quindi

$$M'(t) = -aM(t) + q$$

Anche di questa **equazione differenziale** si può trovare la soluzione esplicita, ma per ora preferisco provare a studiare l'equazione senza risolverla.

Nel seguente grafico sono riportati in ascissa i possibili valori di M, e in ordinata i corrispondenti valori di M', calcolati con la legge $-aM + q$. Poiché si tratta di una legge lineare, il grafico è una retta, con intercetta q, e con intersezione dell'asse nel punto $\bar{M} = q/a$.



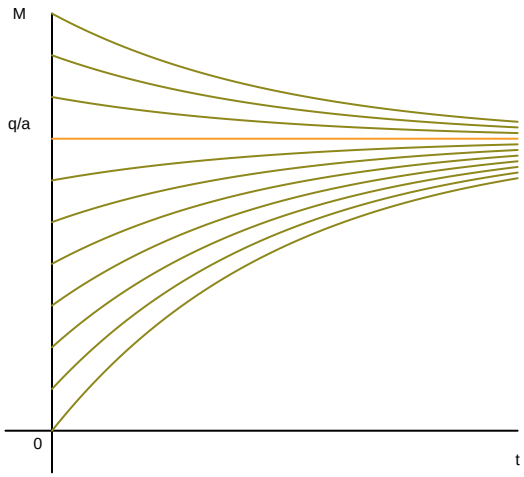
Se M è nell'intervallo $(0, q/a) = (0, \bar{M})$, il valore della funzione è positivo, dunque la velocità di crescita è positiva e quindi M deve aumentare. Al contrario se $M > \bar{M}$, la velocità di crescita è negativa, e dunque M deve diminuire. Se invece $M = \bar{M} = q/a$ la velocità è zero, dunque non ci sarà variazione. Questi punto è un **punti di equilibrio**. Si noti che è anche **attrattivo**, perché se si parte da un valore di M di poco a destra o di poco a sinistra, M si avvicina e tende a $\bar{M}$. Diremo anche che questo equilibrio è **stabile**, proprio perché se si parte a lì vicino, M non si può allontanare.

Cosa vediamo in natura? In genere vediamo gli equilibri stabili (e più in particolare quelli attrattivi). Per esempio, non riuscite a mettere facilmente una penna in verticale sulla punta, perché è un **equilibrio instabile**. In questo esempio del carbonio, vediamo una concentrazione di ^{14}C costante nell'atmosfera. A che serve dunque dunque il modello? Supponiamo, come è accaduto in passato, che ci sia un consistente aumento dell'attività solare per un tempo lungo qualche anno. In tal caso aumentano i raggi cosmici, il coefficiente q aumenta, l'equilibrio $\bar{M}$ cambia, e il modello predice **quantitativamente** come $M(t)$ raggiunge il nuovo equilibrio con il passare del tempo.

Per completare questo esempio, la soluzione esplicita del modello è

$$M(t) = M_0 \mathrm{e}^{-at} + \bar{M}(1 - \mathrm{e}^{-at})$$

Nel grafico seguente sono rappresentate le soluzioni in funzione del tempo, al variare del dato iniziale M. Questa volta sull'asse orizzontale c'è il tempo, su quello verticale il valore di $M(t)$, dunque le curve che vedete disegnate sono le possibili soluzioni in funzione del tempo. Il dato iniziale di ogni curva si legge sull'ordinate dell'asse verticale.



In arancione è disegnata la retta orizzontale che rappresenta la soluzione stazionaria.

Un'ultima importante osservazione. Quando siamo all'equilibrio, matematicamente tutte le funzioni in gioco sono costanti: in questo caso $M' = 0$ e $-aM + q = 0$, cioè $M = q/a$. Dal punto di vista del fenomeno, però l'equilibrio è **dinamico**: il ^{14}C viene creato con velocità q e si autodistrugge con velocità $a\bar{M}$. Il fatto che la velocità di creazione sia uguale a quella di distruzione caratterizza la situazione di equilibrio.

2.1 Modelli a compartimenti

Il modello che abbiamo discusso per il ^{14}C è un esempio semplice di **modello a compartimenti**, che descrive il flusso di una certa sostanza in zone differenti o in ambienti differenti. Per essere più chiaro, nel modello considerato la variabile $M(t)$ è la massa di carbonio ^{14}C nell'atmosfera, che lasciata a se stessa decade con tasso costante, però interagisce con un altro compartimento (non fisico, ma concettuale) che chiamerò genericamente "esterno", in cui invece il ^{14}C viene prodotto.
Consideriamo un altro esempio un po' più complesso, la diffusione del mercurio nei organismi, per esempio pesci. (questo e alcuni degli esempi sono ispirati da J.H. Matis, T.E. Wehrly *Compartmental Models of Ecological and Environmental Systems* in G.P. Patil, C.R. Rao **Environmantal Statitics** handobbok of statistic 12, North-Holland 1994). Si può immaginare di descrivere questo fenomeno con tre compartimenti:

- l'esterno (qui l'acqua), in cui si può pensare ci sia una quantità costante di mercurio
- il compartimento 1, cioè i tessuti dell'organismo che assorbono il mercurio dall'estero (apparato digerente, sangue) e che in parte lo rilasciano all'esterno, e in parte a tessuti più "interni", per esempio le ossa
- il compartimento 2, cioè i tessuti interni, che non interagiscono con l'esterno, ma solo con il compartimento 1

Proviamo a scrivere in astratto il modello differenziale che governa questo fenomeno. Indicherò con $x_1(t)$ la concentrazione di mercurio nel compartimento 1, con $x_2(t)$ la concentrazione di mercurio nel compartimento 2.
La velocità di variazione di x_1 avrà

- un contributo positivo costante dovuto all'esterno, che chiamerò s_1 (s da sorgente);
- un contributo di espulsione del mercurio verso l'esterno che sarà però proporzionale alla quantità di mercurio presente, e dunque sarà della forma $-a_1x_1$
- un contributo di trasferimento al compartimento 2, anche questo proporzionale, del tipo $-a_{12}x_1$
- un contributo di trasferimenti dal compartimento 2, che sarà proporzionale a x_2, e dunque del tipo $a_{21}x_2$

Si noti la scelta della notazione: a_1 è il coefficiente relativo all'interazione con l'esterno, a_{12} quello relativo al contributo del primo compartimento T verso il secondo compartimento, e a_{21} viceversa. Questi numeri non sono necessariamente uguali.
La velocità di variazione di x_2 si determina ragionando nello stesso modo, però manca il contributo di provenienza dall'esterno ($s_2 = 0$), e manca il contributo di trasferimento verso l'esterno ($a_2 = 0$).
In definitiva

$$\begin{cases} x_1' = -(a_1 + a_{12})x_1 + a_{21}x_2 + s_1 \\ x_2' = +a_{12}x_1 - a_{21}x_2 \end{cases}$$

Si noti che poiché il mercurio non viene distrutto nel trasferimento, il termine di crescita del mercurio nel compartimento 1, dovuto al trasferimento dal compartimento 2, deve bilanciare esattamente il contributo di decrescita del mercurio nel compartimento 2, dovuto al trasferimento verso il compartimento 1. Lo stesso vale per il trasferimento dal compartimento 2 al compartimento 1.

Questo sistema è più complesso rispetto all'esempio precedente, perché coinvolge due funzioni. È ancora relativamente semplice perché è un modello con velocità di variazioni lineari (infatti è ancora matematicamente esplicitamente risolubile).

Vediamo se ci sono equilibri, che indicherò con $\bar{x}_1$ e $\bar{x}_2$. Se il sistema è in equilibrio, entrambe le velocità di variazione x_1' e x_2' devono essere zero. Dunque i secondi membri devono essere nulli all'equilibrio, cioè

$$\begin{cases} -(a_1 + a_{12})\bar{x}_1 + a_{21}\bar{x}_2 + s_1 = 0 \\ a_{12}\bar{x}_1 - a_{21}\bar{x}_2 = 0 \end{cases}$$

Guardiamo prima la seconda equazione. Affinché sia verificata, deve accadere

$$a_{12}\bar{x}_1 = a_{21}\bar{x}_2$$

Questa relazione ha un evidente significato: all'equilibrio, la velocità con cui il mercurio passa da 1 a 2, deve essere identica alla velocità con cui il mercurio passa da 2 a 1. Ricaviamo dunque $\bar{x}_2 = \bar{x}_1 a_{12}/a_{21}$. Inserendo questo valore nella prima equazione otteniamo

$$-(a_1 + a_{12})\bar{x}_1 + a_{12}\bar{1}_2 + s_1 = 0$$

cioè

$$-a_1\bar{x}_1 + s_1 = 0, \quad \text{da cui } \bar{x}_1 = s_1/a_1$$

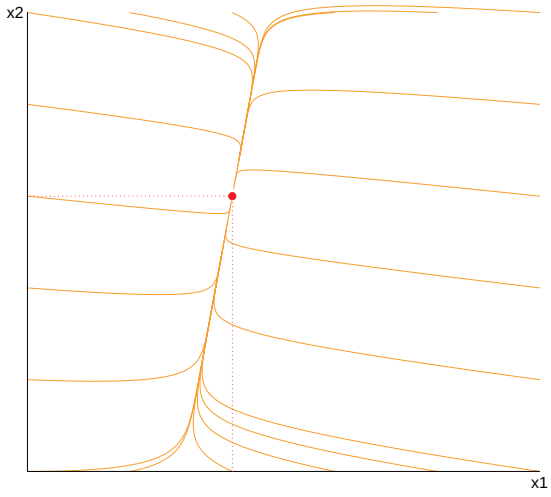
Anche questa uguaglianza ha un chiaro significato: poiché all'equilibrio il flusso netto di mercurio tra i compartimenti 1 e 2 è nullo, il valore di equilibrio della concentrazione di mercurio nel compartimento 1 dipende solo dall'interazione con l'esterno. Infatti è lo stesso equilibrio che si otterrebbe per il solo bilancio di x_1 dato da $x_1' = -a_1x_1 + s_1$.

Riassumendo, abbiamo trovato un solo equilibrio

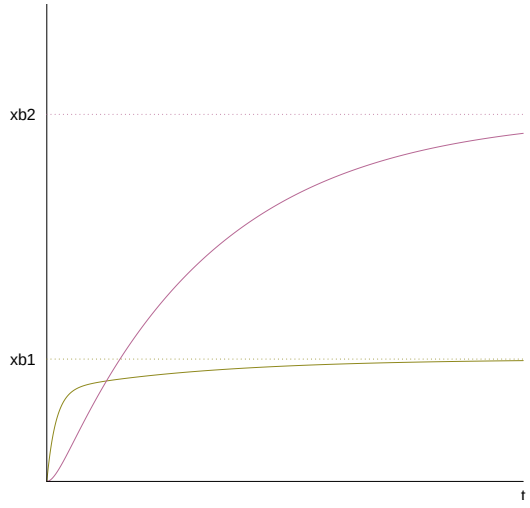
$$\bar{x}_1 = s_1/a_1, \quad \bar{x}_2 = (s_1a_{12})/(a_1a_{21}).$$

Esistono metodi matematici per controllare la stabilità e l'attrattività di questo equilibrio, ma non ne parlerò. Mi limito a far vedere il grafico di diverse soluzioni, che si ottengono cambiando il dato iniziale.

Attenzione: questo grafico è diverso dai due precedenti, perché viene rappresentato il piano delle due variabili x_1 e x_2, Non potendo disegnare il tempo (si potrebbe fare con la terza dimensione, ma non si otterrebbe un grafico più leggibile), per comprendere l'andamento temporale ho disegnato delle frecce.



Questi invece sono i grafici di $x_1(t)$ e $x_2(t)$ con dato iniziale $x_1(0) = 0$, $x_2(0) = 0$. In orizzontale i rispettivi valori di equilibrio.



Esercizio 16. Idrologia delle paludi di Okenfenkee

Lo studio del flusso delle acque in una regione paludosa viene diviso in 4 compartimenti:

- A - l'altopiano, su cui si accumulano le acque piovane, e che può riceve acque di risalita dal sostrato roccioso;
- R - il sottosuolo roccioso dell'altopiano, che riceve acque solo dall'altopiano
- P - la superficie della palude, che riceve acque dall'altopiano, dal sostrato roccioso, e dal sostrato sabbioso sotto la palude
- S - il sostrato sabbioso sotto la palude, che riceve acque dalla palude e dal sostrato roccioso dell'altopiano.

Scrivere un ragionevole sistema differenziale per il flusso delle acque in questo sistema, scegliendo quali coefficienti sono nulli.

2.2 Il modello di Verhulst

In natura non si vedono frequentemente crescite malthusiane di popolazioni. Uno dei casi più evidenti è quello della popolazione umana: a meno di guerre e pestilenze, in epoca storica è in espansione esponenziale. Altri casi si osservano nel caso di colonizzazioni di habitat favorevoli. Per esempio si pensi alla diffusione esponenziale di alcuni virus di altri animali che con mutazioni favorevoli si sono adattai agli ospiti umani. Oppure si pensi all'espansione di popolazioni animali che colonizzano un'isola vulcanica di recente formazione, o ancora alla crescita esponenziale di una "specie aliena" che l'uomo introduce in zone lontane da quelle originarie (una storia estremamente interessante è quella dei conigli in Australia, e delle ulteriori specie aliene introdotte per tentare di controllarne la proliferazione).

Il motivo per cui si osservano raramente crescite esponenziali è che esiste un meccanismo che fa crescere il tasso di mortalità (o decrescere il tasso di natalità) se la popolazione è troppo numerosa e diventa difficile l'accesso alle risorse.

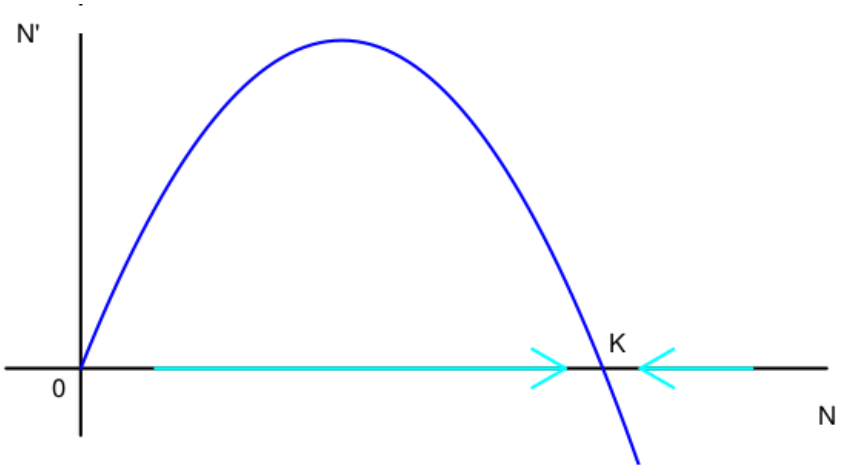
La prima modifica del modello di Malthus che tiene conto di questo effetto è dovuta a Verhulst:

$$\begin{cases} N' = \beta \left(1 - \dfrac{N}{K}\right) N \\ N(0) = N_0 \end{cases}$$

con β e K parametri positivi. Questa volta il tasso di crescita istantaneo N'/N non è costante ma vale

$$N'/N = \beta \left(1 - \frac{N}{K}\right)$$

Vediamo di capire come si comporta questo sistema, aiutandoci con un grafico, in cui in ascissa consideriamo i possibili valori di N, e in ordinata i corrispondenti valori di N', come descritti dalla legge.



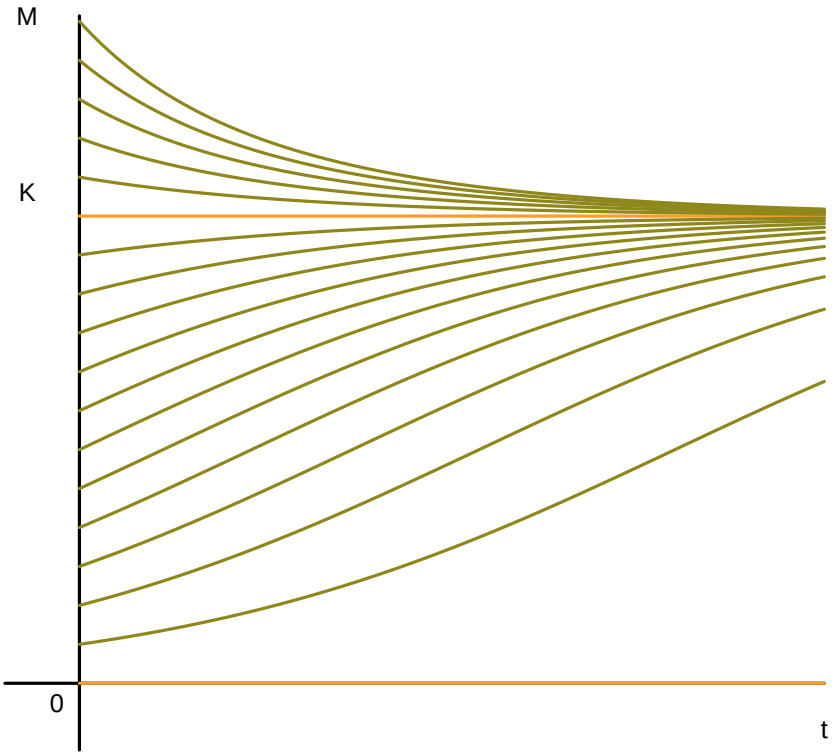
La funzione di N che compare al secondo membro è una parabola, che passa per l'origine, ha la concavità rivolta verso il basso, e si annulla in due punti: $N = 0$ e $N = K$.
Se N è nell'intervallo $(0, K)$, il valore della funzione è positivo, dunque la velocità di crescita è positiva e quindi N deve aumentare. Al contrario se $N > K$, la velocità di crescita è negativa, e dunque N deve diminuire. Se invece $N = 0$ o $N = K$ la velocità è zero, dunque non ci sarà variazione. Questi due punti sono **punti di equilibrio**.

- $N = 0$ è un **equilibrio instabile**, perché appena ci si sposta un po' N comincia a crescere
- $N = K$ è un **equilibrio stabile**, perché sia se ci si sposta di poco a destra, sia se ci si sposta i poco a sinistra, N si riavvicina a K.

Per questo sistema si può scrivere anche l'espressione esplicita della soluzione, che è

$$N(t) = \frac{KN_0}{N_0 + (K - N_0)\mathrm{e}^{-\beta t}}$$

Naturalmente il grafico di questa funzione dipende dal valore di N_0 Lo rappresento nel prossimo grafico, in cui stavolta sull'asse orizzontale c'è il tempo, su quello verticale c'è $N(t)$.



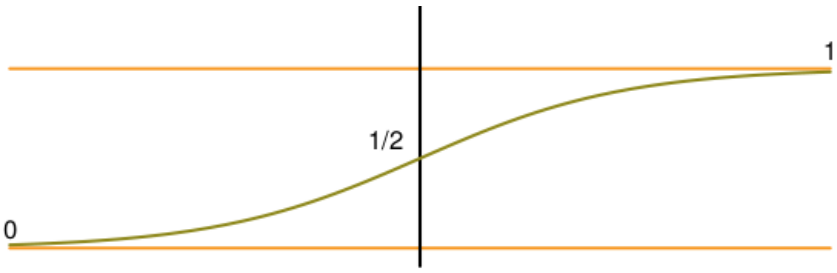
Da questo grafico si possono trarre le stesse conclusioni che abbiamo fatto studiando il grafico di N' in funzione di N. Se $N_0 = 0$ o se $N_0 = K$ il sistema è **in equilibrio**. Se N_0 è tra 0, e K, $N(t)$ cresce (e se è vicino a 0 cresce esponenzialmente), poi la curva si piega e tende a K per $t \to +\infty$. Se invece $N_0 > K$, $N(t)$ decresce e tende a K per $t \to +\infty$.

2.3 La funzione logistica

Studiando per $t < 0$ l'espressione di $N(t)$, si vede che $N(t)$ tende a 0 per $t \to -\infty$. Questa funzione $N(t)$ dunque parte da 0 a $-\infty$ e raggiunge K a $+\infty$, con una forma quasi a S. Fa parte delle cosiddette funzioni logistiche o anche sigmodi, che hanno una grande importanza nella descrizione di vari fenomeni.
L'espressione base di queste funzioni è

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

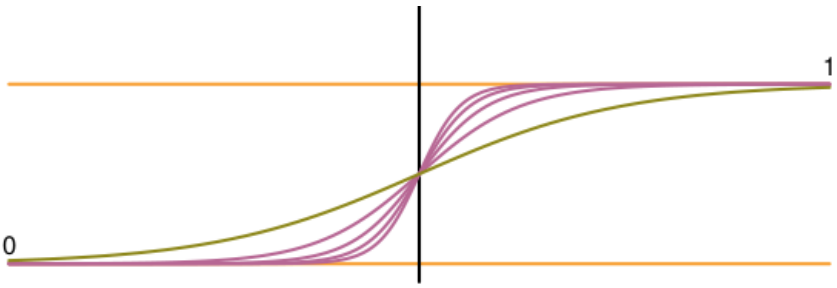
che per unisce 0 (per $x \to -\infty$), a 1 (per $x \to +\infty$), e vale $1/2$ per $x = 0$. Inoltre ha grafico simmetrico rispetto al punto $(0, 1/2)$.



Aggiungiamo ora un parametro, e consideriamo

$$f_\beta(x) = \frac{1}{1 + \mathrm{e}^{-\beta x}}$$

con β numero positivo. I grafici di queste funzioni sono tutti uguali, a parte il **riscalamento** nella variabile x. Per vedere il valore $f(1)$, basta mettere $x = 1$. Ma per vedere il valore $f(1)$ usando la funzione f_β, è necessario porre $\beta x = 1$, cioè $x = 1/\beta$. In pratica dobbiamo rimpicciolire la variabile x di un fattore β.
Nella figura seguente sono riportati i grafici per β da 1 a 5. Come si vede, al crescere di β, il grafico si schiaccia su metà delle rette orizzontali.



Nel limite $\beta \to +\infty$, si ottiene una funzione che in matematica si chiama ϑ, definita in questo modo

$$\vartheta(x) = \begin{cases} 0 \text{ se } x < 0 \\ 1/2 \text{ se } x = 0 \\ 1 \text{ se } x > 0 \end{cases}$$

Questa funzione è il prototipo della modellizzazione dei **fenomeni a soglia**, per esempio l'attivazione di un neurone. Se l'intensità dell'input (cioè del segnale di ingressi) è inferiore al valore di soglia, in questo caso 0, il neurone non emette nulla, appena diventa superiore al valore di soglia, il neurone emette il suo segnale, indipendentemente dall'intensità dell'input.
Si può pensare alla famiglia di funzioni f_β, con β grande, come a una approssimazione *morbida* della funzione di soglia ϑ: il neurone emette sempre un po' di segnale, ma si vede una sensibile differenza solo al passaggio del valore di soglia 0.
Vale la pena fare qualche osservazione sulla simmetria. Il grafico della funzione f_β è evidentemente simmetrica rispetto al punto $(0, 1/2)$. Cerchiamo una prova algebrica di questo fatto valutando la differenza tra la funzione e $1/2$:

$$f_\beta(x) - \frac{1}{2} = \frac{2 - 1 - \mathrm{e}^{-\beta x}}{2(1 + \mathrm{e}^{-\beta x})} = \frac{1}{2}\,\frac{1 - \mathrm{e}^{-\beta x}}{1 + \mathrm{e}^{-\beta x}}$$

Moltiplichiamo numeratore e denominatore dell'ultima frazione per $\mathrm{e}^{\beta x/2}$. Si ottiene

$$f_\beta(x) - \frac{1}{2} = \frac{1}{2} + \frac{1}{2}\,\frac{\mathrm{e}^{\beta x/2} - \mathrm{e}^{-\beta x/2}}{\mathrm{e}^{\beta x/2} + \mathrm{e}^{-\beta x}}. \tag{2.3.1}$$

A questo punto è facile notare che se scambiamo x con $-x$ l'ultima frazione cambia solo di segno, e dunque $f_\beta(x) - \frac{1}{2}$ cambia solo di segno. Questa è appunto la simmetria rispetto al punto $(0, 1/2)$.

È semplice costruire una funzione a soglia che invece di andare da 0 a 1, va da s, valore sinistro, a d, valore destro. Ragioniamo in questo modo: ϑ copre una variazione ampia $1 = 1 - 0$, mentre la funzione che cerchiamo deve coprire una variazione $d - s$. La funzione $(d - s)\vartheta(x)$ fa esattamente questo, ma va da 0 a $d - s$. Per ottenere a sinistra il valore s basta sommarlo. La funzione cercata è

$$s + (d - s)\vartheta(x).$$

Per esercizio, si provi che

$$s + (d - s)f_\beta(x - x_0) = \frac{d + s\mathrm{e}^{-\beta(x-x_0)}}{1 + \mathrm{e}^{-\beta(x-x_0)}}$$

è la versione morbida della funzione di soglia che unisce s a d, con la soglia in x_0. Usando la (2.3.1), si ottiene l'espressione alternativa

$$\frac{s+d}{2} + \frac{d-s}{2}\,\frac{e^{\beta(x-x_0)/2} - e^{-\beta(x-x_0)/2}}{e^{\beta(x-x_0)/2} + e^{-\beta(x-x_0)/2}}.$$

da cui si vede la simmetria rispetto al punto $(x_0, (s + d)/2)$.
Più in generale, una funzione logistica ha l'aspetto

$$g(x) = \frac{d + s\mathrm{e}^{-\beta(x-x_0)}}{a + b\mathrm{e}^{-\beta(x-x_0)}}$$

In questo caso, il valore di soglia è ancora x_0, la funzione va dal valore s/b a $-\infty$ al valore d/a a $+\infty$. Se $a = b$ la funzione è simmetrica rispetto a $(x_0, g(x_0))$ e si riduce all'espressione precedente, dividendo d e s per a; altrimenti la funzione non è simmetrica.

2.4 Cinetica chimica

In questo esempio vedremo per la prima volta come si può modellizzare l'interazione tra due variabili che descrivono un fenomeno. Negli esempi precedenti, o avevamo a che fare con un'unica variabile, oppure, nei modelli a compartimento, le diverse variabili rappresentavano la quantità o la concentrazione di una stessa grandezza, ma all'interno di diversi compartimenti. Supponiamo di sapere che è possibile la reazione chimica che unisce l'elemento X e l'elemento Y per formare un composto XY. Le variabili del sistema sono le concentrazioni dei tre elementi:

$$x = [X],\ y = [Y],\ z = [XY]$$

La variazione della concentrazione di X avrà due termini: uno di accrescimento, dovuto al fatto che XY si decompone nei suoi elementi costitutivi. Come sempre modellizzeremo

questa velocità di decomposizione con un termine proporzionale alla concentrazione $\beta[XY]$. L'altro termine sarà di decrescita, e sarà dovuto al fatto che molecole di X e di Y si incontrano e formano XY. Questo termine deve essere proporzionale a $[X]$, infatti più X c'è, più XY si forma, ma deve anche essere proporzionale a $[Y]$, per lo stesso motivo, dunque

$$[X]' = -\alpha[X][Y] + \beta[XY]$$

A questo punto è semplice scrivere le altre equazioni, infatti anche $[Y]$ varia per gli stessi motivi, e con la stessa legge, infine $[XY]$ si crea a velocità $\alpha[X][Y]$, e si distrugge a velocità $\beta[XY]$. Riassumendo

$$\begin{cases} [X]' = -\alpha[X][Y] + \beta[XY] \\ [Y]' = -\alpha[X][Y] + \beta[XY] \\ [XY]' = \alpha[X][Y] - \beta[XY] \end{cases}$$

Si noti che $[X]'+[XY]' = 0$, così come $[Y]'+[XY]' = 0$. Questo non deve sorprendere, perché $[X] + [XY]$ è la somma della concentrazione di X libera, e di quella di X legata. Il totale, che indicherò con x, non può cambiare nel tempo. Lo stesso accade per $[Y] + [XY] = y$ che è la concentrazione totale di Y. È facile vedere che l'equilibrio si ottiene per

$$\alpha[X][Y] = \beta[XY]$$

Riscrivendo tutto in funzione di $[XY]$:

$$\alpha(x - [XY])(y - [XY]) = \beta[XY]$$

Per esercizio, si risolva questa equazione di secondo grado, mostrando che c'è una sola soluzione positiva minore di x e di y, che è l'unica accettabile. Anche questo equilibrio è **dinamico**: anche se le concentrazioni sono costanti, avvengono continuamente trasformazioni, ma perfettamente bilanciate.
Si osservi infine che usando i valori costanti x e y, si può riscrivere la terza equazione nella sola variabile $[XY]$:

$$[XY]' = \alpha(x - [XY])(y - [XY]) - \beta[XY]$$

Per esercizio si provi che questo sistema ha due equilibri, che quello minore è stabile e attrattivo, che quello maggiore è instabile, e che è fisicamente da scartare perché prevede $[XY]$ superiore al totale di X e di Y.

Ripeto: il motivo di questo esempio è di fare la conoscenza di termini di interazione di tipo prodotto, potete trascurare la descrizione degli equilibri in questo esempio.

2.5 Interazioni di tipo Michaelis-Menten

Un'interessante variazione del modello precedente si ottiene quando si analizzano reazioni in presenza di enzimi. Stavolta le variabili saranno $[S]$, la concentrazione di sostrato, $[E]$, la concentrazione di enzima, il composto SE, che però si trasforma nel prodotto finale P e libera l'enzima E.
Procedendo come nell'esempio precedente, e assumendo che SE in parte decada in $S + E$, e in parte in $P + E$, si ottiene facilmente il sistema

$$\begin{cases} [SE]' = \alpha[S][E] - (\beta + \gamma)[SE] \\ [S]' = -\alpha[S][E] + \beta[SE] \\ [E]' = -\alpha[S][E] + (\beta + \gamma)[SE] \\ [P]' = \gamma[SE] \end{cases}$$

Si noti la differenza tra l'equazione per $[S]$ e quella per $[E]$, dovuta a fatto che il composto $[SE]$ si trasforma sia in S ed E, sia in S e P. Inoltre è utile notare che la quantità $[E]+[SE]$ è pari alla concentrazione totale dell'enzima, che indicherò con e e che non può cambiare nel tempo, come si vede sommando le due corrispondenti equazioni.

Cerchiamo gli equilibri. Guardando l'equazione per $[P]$ si capisce che deve essere $[SE] = 0$, cioè non deve esserci composto. Inserendo questa condizione nelle altre equazioni si ottiene che per l'equilibrio deve valere $[S][E] = 0$, quindi o $[E] = 0$, oppure $[S] = 0$. Si noti che $[SE]+[E] = e$ è concentrazione totale di enzima, perché è la somma di quello libero e di quello legato, mentre $[SE] + [S]$ è la concentrazione totale di sostrato. Possiamo imporre $[E] = 0$ solo impionento che non ci sia enzima, caso che escludiamo in quanto non interessante. Se invece imponiamo $[S] = 0$, stiamo imponendo che non ci sia sostrato. Il valore di $[E] = e$ invece è fissato, e $[P]$ può essere qualunque. Si noti che questi equilibri non sono dinamici: tutti i singoli termini sono nulli. Si noti anche che fuori dall'equilibrio $[P]$ cresce, ma non può crescere all'infinito. Tenderà dunque a una costante, e dunque $[SE]$ tenderà a 0, quindi anche $[S][E]$ deve tendere a 0: in pratica tutto il sostrato viene trasformato in prodotto.
Una notevole semplificazione di questo modello si ottiene se si ipotizza che la reazione avvenga con E in equilibrio dinamico, cioè

$$\alpha[S][E] = (\beta + \gamma)[SE]$$

Poiché $[E] = [SE] - e$, usando queste due equazioni si ottiene $[SE]$ in funzione di $[S]$:

$$[SE] = \frac{\alpha e[S]}{\beta + \gamma + \alpha e[S]}$$

Notando che

$$[S]' = -\alpha[S][E] + \beta[SE] = -\alpha[S][E] + (\beta + \gamma)[SE] - \gamma[SE] = -\gamma[SE]$$

si ottiene infine un sistema in cui non compare più il termine che coinvolge l'enzima ma solo il bilancio tra $[S]$ e $[P]$:

$$\begin{cases} [S]' = -\dfrac{\alpha\gamma e[S]}{\beta + \gamma + \alpha e[S]} \\ [P]' = \dfrac{\alpha\gamma e[S]}{\beta + \gamma + \alpha e[S]} \end{cases}$$

Questo sistema predice che $[S]$ decresce fino a 0, e in corrispondenza $[P]$ cresce fino all'esaurimento del sostrato. Si noti che la decrescita di $[S]$ non è esponenziale: se $[S]$ è grande, la produzione di $[P]$ avviene quasi a velocità costante γ, per poi rallentare e diventare esponenziale quando $[S]$ diventa piccolo. Questo tipo di termine per la velocità di variazione è si chiama proprio Michaelis-Menten, e racchiude in sé la parte di interazione con l'enzima, che non compare più nell'equazione.
Riassumendo: un termini di tipo Michaelis-Menten è un termine di decrescita (o di crescita) per una quantità x del tipo

$$\alpha\frac{x}{a + x}$$

con a parametro positivo. Consideriamo dunque

$$x' = \alpha\frac{x}{a + x}$$

che corrisponde ad assumere che il tasso istantaneo di variazione di x sia $x'/x = \alpha/(a+x)$, dunque decrescente in x. Se x è piccolo, il tasso vale circa α/a, e dunque il modello si riduce al modello esponenziale (decrescente o crescente a seconda del segno di α). Se x è grande, x' è praticamente α cioè c'è una sorgente costante se α è positivo, è un prelievo costante se α è negativo.
Questo tipo di termine può essere generalizzato come segue:

$$\alpha \frac{x^k}{a^k + x^k}$$

con $k \geq 1$. Anche in questo caso per x grande il termine è praticamente α, quindi dà una sorgente o un prelievo costanti a seconda del segno di α. Invece, quando x è piccolo, il termine è praticamente

$$\frac{\alpha}{a^k} x^k.$$

Per capire come si comporta questo termine consideriamo il corrispondente tasso di variazione

$$x'/x = \alpha x^{k-1}/a^k,$$

che, per $k > 1$, tende a 0 se $x \to 0$. Dunque questo termine dà una crescita o decrescita molto lenta se x è piccolo. In particolare, per $\alpha = -1$ con dato iniziale x_0 risulta

$$x(t) = \frac{x_0}{1 + t x_0},$$

che va a zero come $1/t$, molto più lentamente dell'esponenziale negativo.
In sintesi, il termine di Michaelis - Menten generalizzato è un termine con tasso di variazione che tende a 0 per x piccolo, e velocità di variazione costante per x grande.

2.6 Il modello SIR

Il modello SIR è il modello di base per l'evoluzione di un'epidemia "rapida", cioè che si evolva in tempi abbastanza brevi per non dover considerare nascite e morti naturali nella popolazione (al contrario per esempio dell'epidemie di HIV che è in corso da vari decenni). Inoltre si basa sull'ipotesi fondamentale che i guariti non si possano ricontagiare. Infine, l'applicazione di questo modello va limitata ai casi di epidemica che riguarda un solo ospite, e dunque è per esempio inadatto allo studio della diffusione della malaria, che si scambia tra uomo e zanzara.
Le variabili sono: S, il numero di suscettibili, cioè degli individui che non si sono ammalati; I il numero di persone infette e dunque contagiose, R, il numero di guariti ("rimossi"). Alla luce degli esempi precedenti dovrebbe essere chiaro perché il modello ha questa espressione:

$$\begin{cases} S' = -aSI \\ I' = aSI - bI = (aS - b)I \\ R' = bI \end{cases}$$

Si nota subito che ci sono infiniti punti di equilibrio: se $I = 0$, tutte le variabili sono costanti (assenza di epidemia). L'altra osservazione immediata che si può fare è che S è una funzione decrescente, R è una funzione crescente, mentre I decresce se e solo se $aS - b$ è negativo.

Poiché S è decrescente, prima poi $aS-b$ diventerà negativo, e quindi I comincerà a decrescere e l'epidemia si estingue.
In una epidemia all'inizio del suo sviluppo, un ruolo cruciale è giocato dal fattore $aS_0 - b$, dove S_0 è la numerosità della popolazione che si può ammalare. Si noti che $aS_0 - b < 0$ se e solo se

$$\mathcal{R} = \frac{aS_0}{b} < 1$$

Pensando all'epidemia attualmente in corso, il tracciamento dei positivi permette di metterli in isolamento, quindi ai fini del contagio questo equivale a considerarli rimossi. Dunque un buon tracciamento con isolamento domiciliare dai familiari equivale ad aumentare il coefficiente b, e dunque a fare scendere $\mathcal{R}$. Il distanziamento sociale e l'uso delle mascherine si traducono invece in una diminuzione di a, perché rendono improbabile la trasmissione del virus. Infine, la vaccinazione serve a ridurre S_0. L'effetto di tutte queste misure è di ridurre $\mathcal{R}$. Se scende sotto 1, un'epidemia in corso si spegnerà, se non è ancora iniziata non inizierà nemmeno.

Faremo delle simulazioni numeriche su questo modello. Qui riporto solo un conto un po' sofisticato, che spiega la cosiddetta **immunità di gregge**.
Si noti che la prima equazione si può riscrivere come

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln S = -aI$$

e che dalla terza equazione si ottiene $I = R'/b$. Dunque

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\ln S + \frac{a}{b}R\right) = 0$$

Sapere che questa quantità è costante è molto utile. Supponiamo di considerare un'epidemia all'inizio, in cui $R(0) = 0$, e $S(0) = S_0$. Allora

$$\ln S(t) + \frac{a}{b}R(t) = \log S_0$$

cioè

$$\ln \frac{S(t)}{S_0} + \frac{a}{b}R(t) = 0$$

Passando al limite per $t \to +\infty$, S andrà a S_f, cioè al valore finale di quelli che non si sono ammalati, R andrà a R_f, il valore finale degli ammalati. Indico con $r = R_f/S_0$ la frazione complessiva di individui che si ammala. Pensano che il valore iniziale degli infetti sia trascurabile, si può scrivere

$$S_f + R_f = S_0 + I(0) \approx S_0$$

, e dunque

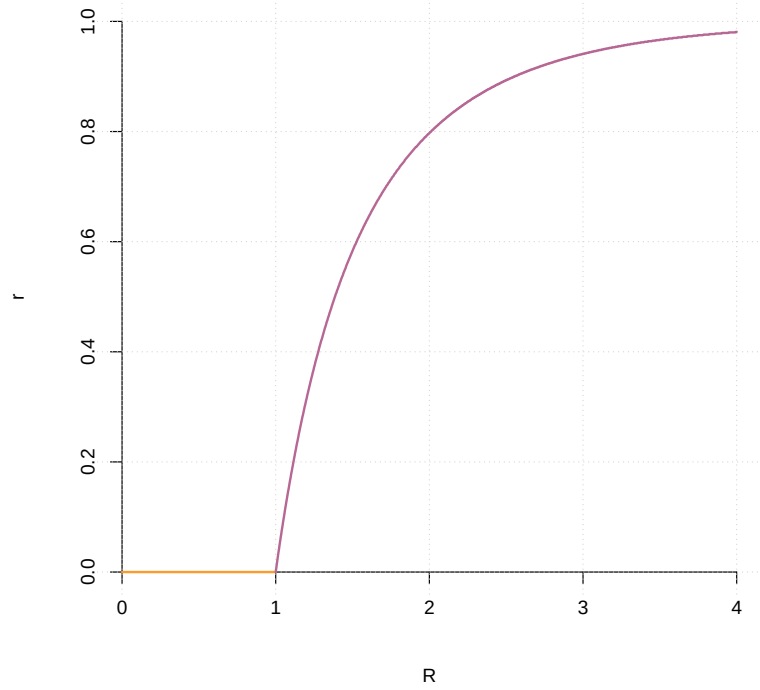
$$S_f/S_0 = 1 - r$$

La relazione scritta sopra diventa dunque un'equazione per r:

$$\ln(1-r) = -\frac{a}{b}R_f = \frac{aS_0}{b}r = -\mathcal{R}r$$

Saper risolvere questa equazione è importante perché ci permette di predire il numero totale di malati in funzione di $\mathcal{R}$. Questa equazione ha la soluzione $r = 0$ per qualunque $\mathcal{R}$ (in

assenza di epidemia non ci sono malati). Però per $\mathcal{R} > 1$ appare un'altra soluzione, che riporto nel grafico.



Come si vede, per $\mathcal{R} < 1$ c'è solo la soluzione nulla e dunque l'epidemia non si puó innescare. Da $\mathcal{R} = 1$ in poi la percentuale finale di contagiati cresce enormemente. Per esempio, se $\mathcal{R} = 1.01$, $r \approx 2\%$, se $\mathcal{R} = 1.1$, $r \approx 18\%$, se $\mathcal{R} = 2$, $r \approx 80\%$. Si capisce dunque la necessità di tenere $\mathcal{R}$ più basso possibile.
Questo grafico spiega l'effetto gregge delle campagne vaccinali: non solo la popolazione vaccinata non si ammala, ma se $\mathcal{R}$ scende sotto 1 non si ammala nemmeno la frazione di popolazione non vaccinata.

2.7 Il modello Lotka-Volterra - orbite periodiche

Il primo modello con interazione che descrivo è il modello preda-predatore, anche detto modello di Lotka-Volterra, dal nome dei due scienziati che lo definirono, indipendentemente, circa un secolo fa.
Le variabili sono $x(t)$, il numero di predatori, e $y(t)$ il numero di prede. Senza prendere in considerazione l'interazione tra queste specie, modellizziamo la velocità di cambiamento di x con tasso costante di decrescita (non avendo accesso alle risorse i predatori si estinguono), e la velocità di cambiamento di y con tasso costante di crescita (in assenza di limitazioni dovute a predatori o scarsezza di risorse, il numero di prede è in crescita malthusiana).
Riflettiamo ora sull'effetto della presenza delle prede nel cambiamento del numero di predatori. Deve trattarsi di un termine di crescita, che è ragionevole supporre proporzionale al numero di prede: se raddoppio le prede, i predatori hanno a disposizione il doppio delle risorse. Inoltre, sarà proporzionale al numero di predatori: se le prede venissero in continuazione rimpiazzate e messe a disposizione dei predatori, anche la numerosità dei predatori dovrebbe una crescita

malthusiana, e questo è fenomeno è equivalente alla proporzionalità della velocità di crescita alla numerosità. Dunque modellizziamo la velocità di variazione del numero di predatori $x(t)$ con l'equazione

$$x' = axy - bx$$

Il secondo termine è quello di decrescita a tasso costante, il primo è quello di crescita a tasso crescente con il numero di prede. Allo stesso modo, modellizziamo la velocità di variazione del numero di prede $y(t)$ con l'equazione

$$x' = -\alpha xy + \beta x$$

In questo caso, il secondo termine è quello di crescita a tasso costante, il primo è un termine di decrescita con tasso che cresce con il numero di predatori.

Osservo che c'è un altro modo per spiegare la presenza dei termini xy: supponendo che le y prede e gli x predatori si muovono casualmente in una stessa area, xy è proporzionale al numero di incontri che possono avvenire nell'unità di tempo, dunque x è proporzionale al numero di predatori che una singola preda può incontrare nell'unità di tempo, e dunque il tasso di estinzione delle prede deve essere proporzionale a x.

Riassumendo

$$\begin{cases} x' = axy - bx = (ay - b)x \\ y' = -\alpha xy + \beta y = (-\alpha x + \beta)y \end{cases}$$

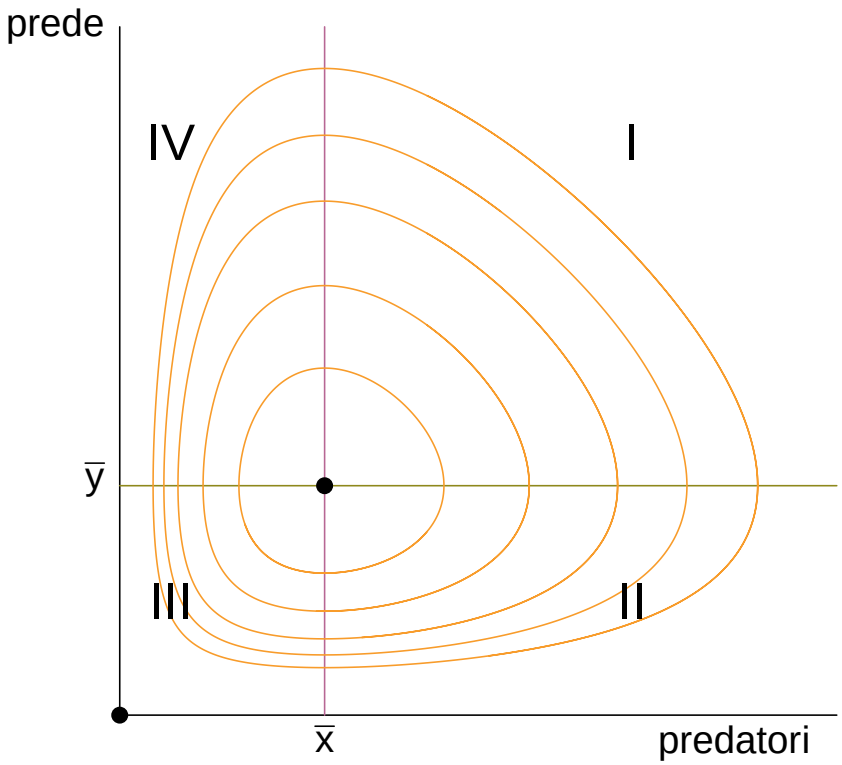
È semplice trovare gli equilibri. Guardando la prima equazione, o $x = 0$, oppure $y = \bar{y} = b/a$. Nel primo caso, inserendo il valore $x = 0$ nella seconda equazione, si ottiene che anche y deve essere 0. Se invece uso $y = \bar{y}$, si ottiene che $x = \bar{x} = \beta/\alpha$.

Come vedremo con le simulazioni, il comportamento del sistema è il seguente:

- $(x, y) = (0, 0)$ è un equilibrio instabile;
- $(x, y) = (\bar{x}, \bar{y})$ è un equilibrio stabile ma non è attrattivo;
- tutte le altre soluzioni sono periodiche

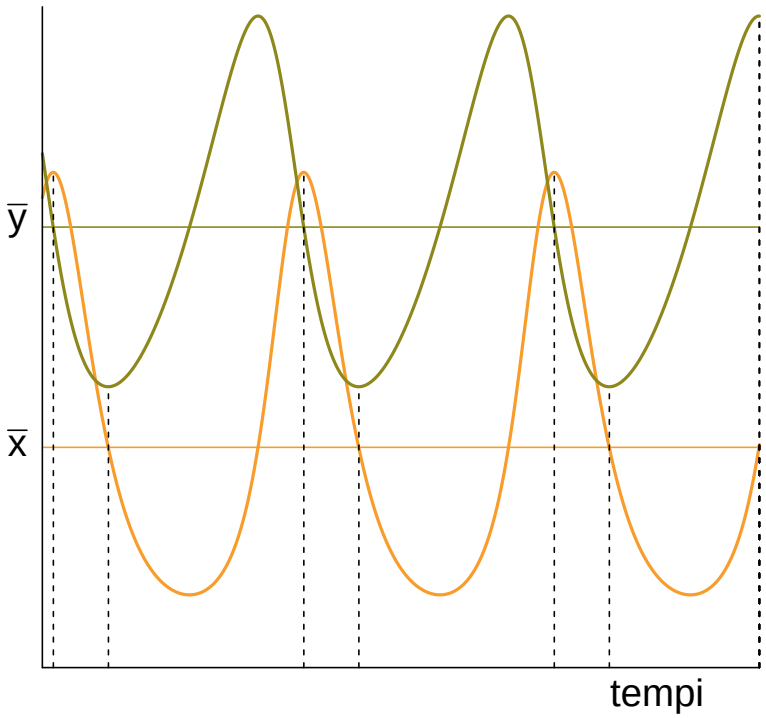
Commentiamo quest'ultima affermazione, riscrivendo il sistema in questo modo

$$\begin{cases} x' = axy - bx = a\left(y - \dfrac{b}{a}\right) = a(y - \bar{y})x \\ y' = -\alpha xy + \beta y = -\alpha\left(x - \dfrac{\beta}{\alpha}\right)y = -\alpha(x - \bar{x})y \end{cases}$$



Osserviamo il grafico in figura, immaginando che il dato iniziale sia il punto (x_0, y_0), che abbiamo scelto con $x_0 > \bar{x}$ e $y_0 > \bar{y}$, cioè nella regione I. Le due equazioni ci dicono che in questa regione x deve crescere e y deve diminuire. Questo andamento continuerà fino a quando y non scende sotto $\bar{y}$, e la soluzione entra nella regione II. Qui y continua a decrescere, ma anche x comincia a decrescere. Quando x passa $\bar{x}$, la soluzione entra nella regione III, in cui y ricomincia a crescere. Infine, quanto y passa $\bar{y}$, la soluzione entra nella regione IV, in cui entrambe le variabili crescono. Quando x sorpassa $\bar{x}$, la soluzione rientra nella regione I, e torna esattamente al punto di partenza.

È istruttivo osservare l'andamento temporale della soluzione su un unico grafico. In ascissa mettiamo il tempo, in ordinata i valori di entrambe le variabili, e disegniamo anche i valori $\bar{x}$ e $\bar{y}$. Come si vede, il massimo e il minimo del numero di predatori (in arancione), si raggiunge nell'istante in cui il numero di prede (in verde), passa il valore di equilibrio. Lo stesso accade al contrario.



Un altro fatto importante che si può dimostrare con un po' più di matematica, è che il valore medio su un periodo della variabile x è sempre esattamente $\bar{x}$, e quello della variabile y è sempre esattamente $\bar{y}$ (si tratta di un fenomeno che riguarda proprio questo modello, e non è per niente generale). Dunque, anche se considero soluzioni che non siano di equilibrio, i valori $(\bar{x}, \bar{y})$ indicano i valori medi delle due variabili.

A questo punto è molto interessante chiedersi come variano $(\bar{x}, \bar{y})$, se cambiano i parametri. Per esempio se aumenta il nutrimento a disposizione delle prede, cresce il parametro β, favorendo in teoria le possibilità delle prede. Ricordando però che

$$\bar{x} = \frac{\beta}{\alpha} \quad \text{e} \quad \bar{y} = \frac{b}{a}$$

si ottiene che il numero medio di prede **non cambia**, mentre aumenta il numero di predatori. Al contrario, una maggior difficoltà di vita per i predatori (che si traduce in un aumento di b), non ne cambia il numero, ma fa aumentare il numero delle prede, che possono prosperare più facilmente. Questo esempio suggerisce che le interazioni ecologiche posso essere complesse, e vanno comprese prima di poter fare valutazioni chiare sul significato ambientale dell'aumento o della diminuzione di una popolazione.

2.8 Il modello di Ross per la malaria

Il modello di Ross per la diffusione della malaria è interessante perché ci mostra altre possibili modellizzazioni dell'interazione tra due specie, e perché suggerisce alcune riflessioni sul controllo biologico.

È un modello per una epidemia che interessa due specie differenti, in cui la guarigione non protegge dalla reinfezione. Ci sono due variabili in questo modello, il numero di zanzare infette Z, il numero di umani infetti U. Ci sono un bel po' di parametri:

- a il numero medio di punture che una zanzara fa nell'unità di tempo;
- p la probabilità di trasmissione del plasmodio a un uomo sano per una puntura di una zanzara infetta;
- q la probabilità di trasmissione del plasmodio a una zanzara sana per una puntura a un uomo infetto;
- b il tasso di guarigione degli uomini;
- β il tasso di guarigione delle zanzare;
- N il numero totale di umani;
- M il numero totale di zanzare

Il numero di contagiati umani cresce con una velocità che ha un contributo negativo bU, e un contributo positivo che si determina in questo modo: le Z zanzare infette pungono aZ volte nell'unità di tempo, quindi il numero medio di punture per singolo umano è aZ/N. Poiché gli umani sani sono $N-U$, il numero $aZ/N(N-U)$ rappresenta il numero di punture subite dagli umani sani nell'unità di tempo. Moltiplicando questo valore per p (la probabilità di trasmissione) si ottiene la prima equazione

$$U' = ap\frac{Z}{N}(N-U) - bU$$

Ragionando nello stesso modo, $M-Z$ zanzare sane pungono $a(M-Z)/N$ volte ogni uomo, dunque

$$Z' = aq\frac{M-Z}{N}U - \beta Z$$

È più utile riscrivere questo sistema per le variabili $z = Z/M$, $u = U/N$, che sono le frazioni si zanzare e uomini infetti rispettivamente. Si ottiene facilmente

$$\begin{cases} u' = apRz(1-u) - bu \\ z' = aq(1-z)u - \beta z \end{cases}$$

dove $R = M/N$ è il rapporto tra numero di zanzare e numero di umani,
Come sempre, si cercano gli equilibri. Aiutati dalla fenomenologia che stiamo descrivendo, siamo portati a supporre che $z = 0$, $u = 0$ sia un punto di equilibrio, che corrisponde all'assenza della malattia. Infatti se si sostituiscono questi valori si ottiene effettivamente

che i membri di destra delle due equazioni sono nulli. Possiamo chiederci se ci sono altri equilibri. Imponendo che siano nulli i membri di destra si ottiene il sistema

$$\begin{cases} apRz(1-u) = bu \\ aq(1-z)u = \beta z \end{cases}$$

Si può risolvere con una piccola fatica che può essere ridotta dividendo tutte e due le equazioni per zu. Si ottiene

$$\begin{cases} apR\left(\frac{1}{u} - 1\right) = b\frac{1}{z} \\ aq\left(\frac{1}{z} - 1\right) = \beta\frac{1}{u} \end{cases}$$

che è facilmente risolubile nelle variabili $1/u$ e $1/z$, essendo un sistema lineare. Si ottiene

$$\begin{cases} u = \dfrac{\frac{a^2pqR}{b\beta} - 1}{\frac{a^2pqR}{b\beta} + \frac{aq}{\beta}} = \dfrac{r-1}{r + \frac{aq}{\beta}} \\ z = \dfrac{\frac{a^2pqR}{b\beta} - 1}{\frac{a^2pqR}{b\beta} + \frac{apR}{b}} = \dfrac{r-1}{r + \frac{apR}{b}} \end{cases}$$

dove $r = \frac{a^2pqR}{b\beta}$. Niente panico: dobbiamo solo capire se questa soluzione esiste e che cosa vuole dire. I numeri che abbiamo ottenuto sono sicuramente minori di 1, però a seconda dei valori dei parametri possono diventare negativi. I due numeratori sono identici, dunque u e z sono positivi se e solo se

$$r > 1$$

Come vedremo con una esplorazione numerica, se questa soluzione esiste, è stabile e attrattiva, mentre l'origine $(u, z) = (0, 0)$ è una soluzione instabile. In questo caso la malattia è **endemica**. Al contrario, se $r \leq 1$, c'è solo l'equilibrio $(u, z) = (0, 0)$ che è stabile e attrattivo, dunque l'epidemia si estingue.
Ross comprese che l'unico parametro su cui si può agire in modo relativamente facilmente è M, il numero di zanzare. Sotto una certa soglia r scende sotto 1, fermando l'epidemia. Naturalmente, è di aiuto anche poter aumentare b (la velocità di guarigione) o usare protezioni che facciano scendere a (il numero di punture), anche in questo caso r si riduce.

Esercizio

Consideriamo un modello per i livelli trofici di un ecosistema, in particolare vogliamo modellizzare l'abbondanza di biomassa, presente all'interno di 5 compartimenti:

- N: disponibile nell'ambiente, sotto forma di componenti di base
- P: nei produttori primari (vegetali);
- H: negli erbivori
- C: nei carnivori
- D: nei decompositori, che ritrasformano biomassa in sostanze di base nell'ambiente

L'esercizio consiste nel provare a capire la relazione tra le variabili, e qual può essere un buon modello che descrive il sistema.

Chapter 3

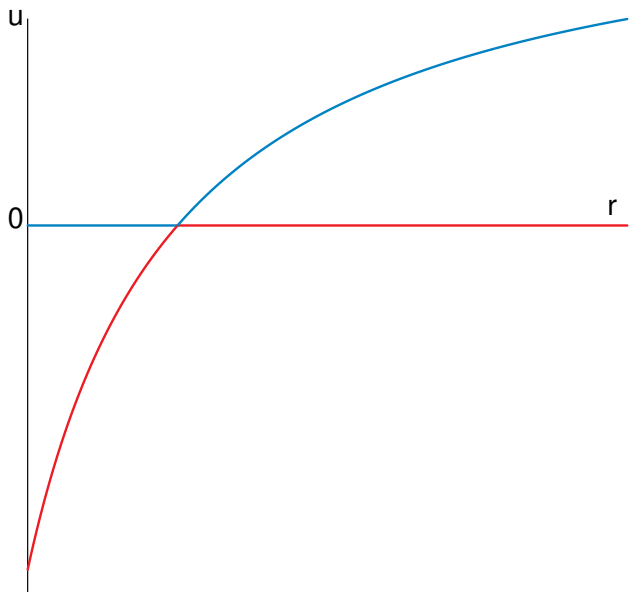
Biforcazioni, catastrofi, caos

Come ho già illustrato, i modelli differenziali permettono di studiare come evolvono nel tempo alcuni fenomeni. In molti degli esempi che abbiamo fatto, al passare del tempo, il sistema raggiunge un equilibrio stabile, anche se ci sono altre possibilità, come nel caso delle orbite periodiche del sistema preda-predatore (nei prossimi paragrafi ne vedremo altre ancora). In questo paragrafo affrontiamo gli effetti sugli equilibri di una modifica dei parametri. Questo argomento è interessante per gli studi ambientali perché in natura spesso alcuni parametri vengono più o meno lentamente modificati, si pensi agli affetti dell'azione umana sul clima e agli effetti dei cambiamenti climatici sugli ecosistemi, In questi esempi, dunque, considereremo dei modelli e ci chiederemo come cambia il loro comportamento al cambiare dei parametri.

Per cominciare, riprendiamo l'analisi delle soluzioni stazionarie del modello di Ross. Indipendentemente dal valore dei parametri, c'è sempre l'equilibrio $(0, 0)$ (assenza di epidemia). Inoltre c'è l'equilibrio

$$\begin{cases} u = \dfrac{\frac{a^2pqR}{b\beta} - 1}{\frac{a^2pqR}{b\beta} + \frac{aq}{\beta}} = \dfrac{r-1}{r + aq/\beta} \\ z = \dfrac{\frac{a^2pqR}{b\beta} - 1}{\frac{a^2pqR}{b\beta} + \frac{apR}{b}} = \dfrac{r-1}{r + apR/\beta} \end{cases}$$

che dipende dal parametro positivo $r = a^2pqR/(b\beta)$. Ricordo che u e z sono, rispettivamente, le frazioni di umani e di zanzare infette, e dunque, anche se questa coppia di valori è sempre una soluzione di equilibrio, essa ha un senso biologico solo se u e z sono compresi nell'intervallo $[0, 1]$, e questo accade se e solo se $r \geq 1$. Ignoriamo questa condizione, e consideriamo tutti i possibili valori di $r \geq 0$. Rappresentiamo in un grafico le soluzioni al variare di r. Considereremo la sola variabile u, ma ricordiamoci che in corrispondenza di u c'è anche a variabile z, cioè che l'equilibrio riguarda la coppia di variabili.



Nell'asse delle ascisse c'è il valore di r, in quello delle ordinate il valore di u. Nel grafico ci sono due curve (che cambiano colore): l'asse delle ascisse, che rappresenta l'equilibrio $u = 0$, e la curva crescente, che rappresenta l'altra soluzione, che cambia con r. Ho colorato in blu la soluzione stabile, e in rosso la soluzione instabile. Come ho già discusso, la soluzione di assenza di epidemia è stabile fino a $r = 1$, dopo diventa instabile, mentre l'altra soluzione, quella endemica, diventa stabile.
Rispetto all'analisi che abbiamo già fatto, aver rappresentato la soluzione endemica anche quando non è naturalisticamente accettabile perché negativa, ci permette di vedere meglio l'aspetto matematico del cambiamento di stabilità. Quello che accade in questo caso è che, attraversandosi, i rami delle due soluzioni si **scambiano** la stabilità.

Mostriamo ora con un esempio che possono accadere fenomeni più complicati, in particolare alcuni modelli esibiscono un comportamenti "catastrofici".

3.1 Un modello per l'eutrofizzazione

Per questo semplice modello mi sono liberamente ispirato all'articolo di Katherine Meyer *Mathematical Review of Resilience in Ecology*, Natural Resource Modeling vol 29, 3 (2016).
L'eutrofizazazione delle acque di un bacino, per esempio un lago, ma anche un mare con poco ricambio, è il fenomeno di accumulo di sostanze nutritive che induce una proliferazione di organismi vegetali, in particolare fitoplancton, che rende inospitale l'ambiente per le specie ittiche. Il modello che mi appresto a descrivere tratta la dinamica delle sostanze nutritive, che si muovono tra tre compartimenti: l'acqua del bacino, le acque reflue (cioè quelle che confluiscono nel bacino), e il fondale del bacino.
Indicherò con $n(t)$ la concentrazione nell'acqua del bacino di una sostanza nutritiva (per esempio l'azoto). Ci sono tre contributi alla variazione di n nel tempo:

- arrivo di nutrimento dalle acque reflue, modelizzato con il termine di flusso costante $+\ell$, con $\ell \geq 0$;
- sedimentazione di n sul fondo, modellizzato con il termine di riduzione a tasso costante $-\alpha n$;
- restituzione di n dal fondo all'acqua, modellizzato con un termine di tipo Michaelis-Menten generalizzato: $+r\frac{n^k}{a^k+n^k}$, con $r, a > 0$. Per semplicità poniamo $k = 2$, $a = 1$.

L'ultimo contributo tiene in conto del fatto che per n piccoli il fenomeno di sedimentazione, con velocità proporzionale a n, deve dominare sul contributo di restituzione, mentre se n è grande, ci si aspetta un velocità di restituzione costante.
Il modello è dunque

$$n' = \ell - \alpha n + r\frac{n^2}{1+n^2} = \ell + \frac{n}{1+n^2}(rn - \alpha(1+n^2))$$

Non è possibile studiare analiticamente questo sistema, se non nel caso $\ell = 0$, cioè in assenza della sorgente di nutrimento. In tal caso, il termine di destra si può scrivere

$$\frac{n}{1+n^2}(-\alpha n^2 + rn - \alpha)$$

il cui numeratore è n per un polinomio di secondo grado. Gli zeri della funzione sono $n = 0$ e

$$n = \frac{1}{2}\left(\frac{r}{\alpha} \pm \sqrt{\left(\frac{r}{\alpha}\right)^2 - 4}\right)$$

che sono numeri reali se $r \geq 2\alpha$. Si noti inoltre che per $n \to +\infty$ la funzione tende a $-\infty$.

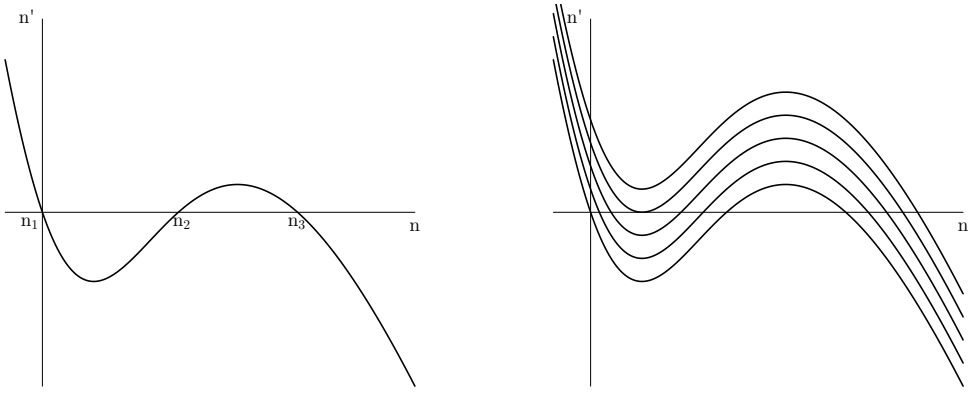


Figure 3.1: Modello per l'eutrofizzazione, con $\ell = 0$ e $\ell > 0$.

Nel primo grafico in figura 3.1 rappresento la velocità di variazione per $\ell = 0$, $r = 4.2$, $\alpha = 2$, considerando anche i valori di n negativi, per chiarire gli aspetti matematici. Questa funzione ha tre zeri, che chiamo, in ordine crescente, n_1, n_2, n_3. Naturalmente $n_1 = 0$. Per questi tre valori, c'è equilibrio tra il flusso con cui n si deposita e quello con cui n torna in circolo dal

fondale. Analizzando il segno di n' in funzione di n notiamo che n_1 e n_3 sono equilibri stabili (e attrattivi), n_2 è un equilibrio instabile. Possiamo immaginare che n_3 corrisponda a una situazione di eutrofizzazione, in cui si è accumulato nel fondale e nell'acqua troppo azoto, mentre n_1 a una situazione normale, in cui il nutrimento è assente (e dunque presumibilmente del tutto assorbito dagli organismi viventi). Nel secondo grafico considero valori di ℓ positivi. Guardando l'espressione di n', si comprende che il grafico si ottiene da quello già disegnato traslandolo verso l'alto esattamente di ℓ.
Si osservi con attenzione cosa accade ai tre equilibri: l'equilibrio n_3 si sposta verso destra, e rimane stabile. Gli equilibrio n_1 e n_2 si spostano l'uno verso l'altro, e per un particolare valore di ℓ arrivano a coincidere. Per valori superiori di ℓ svaniscono entrambi.

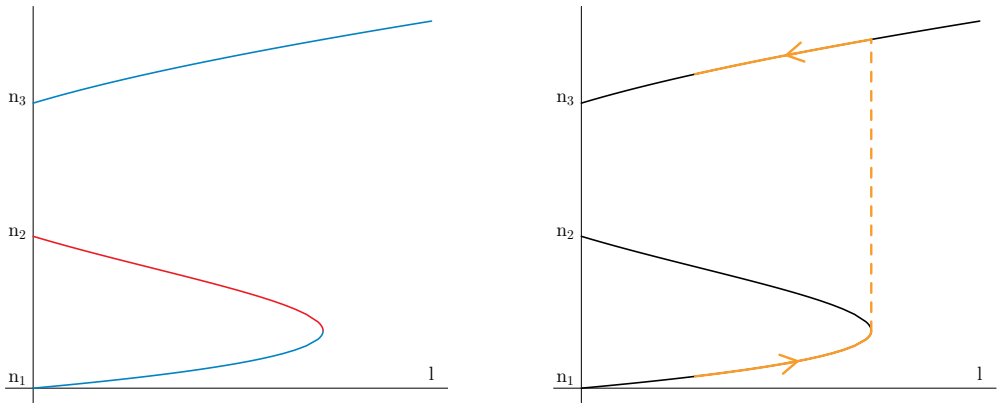


Figure 3.2: Equilibri in funzione di ℓ.

In figura 3.2 rappresento in un grafico come variano gli equilibri al variare di ℓ. In blu ho disegnato gli equilibri stabili n_1 e n_3, in rosso n_2, quello instabile. Come si vede, c'è un valore di soglia per ℓ per il quale n_1 e n_2 si annichilano. Con riferimento al secondo grafico, immaginiamo di considerare una situazione in cui ℓ è piccolo, per esempio $\ell = 0.1$, e il sistema è nell'equilibrio stabile n_1, quindi con poco n disciolto in acqua (assenza di eutrofizzazione). Facciamo crescere ℓ, il sistema resta nell'equilibrio stabile n_1 fino al valore critico di ℓ, passato il quale l'equilibrio non esiste più! L'unica possibilità per il sistema è quella di raggiungere "catastroficamente" l'altro equilibrio stabile, n_3, per cui però il valore di n è grande e il sistema è in una situazione di eutrofizzazione. Nel grafico rappresento in ocra queste modifiche. Al crescere ulteriore di n il valore di n_3 aumenta, ma non ci sono cambiamenti qualitativi. Quello che è accaduto è una "catastrofe": il sistema era in uno stato di equilibrio, ma i parametri sono cambiati e l'equilibrio è scomparso, costringendo il sistema a precipitare in un differente sistema di equilibrio. È da notare che questo cambiamento drammatico non ha avuto segni premonitori! (Da un punto di vista strettamente matematico qualche segno premonitore si potrebbe trovare nella crescita incontrollata della velocità di variazione dell'equilibrio rispetto al parametro, ma non mi aspetto che questa quantità sia misurabile in una osservazione naturalistica).
Naturalisticamente, il crescere del flusso di nutrimento ha ecceduto la capacità del sistema di restare in un equilibrio con piccoli valori di n in acqua, e il sistema ha trovato un altro equilibrio, con n più grande. Potremmo immaginare di voler tornare all'equilibrio n_1 facendo

diminuire ℓ fino al valore di partenza, ma purtroppo non funzionerebbe: se partiamo da un punto di equilibrio n_3, continueremo a rimanere nell'equilibrio n_3, senza riuscire a tornare su n_1. Questa irreversibilità del passaggio da un equilibrio all'altro è cruciale in ecologia: una volta spostato l'equilibrio è molto più difficile tornare indietro, non basta ripristinare i valori originari dei parametri.
Poiché il sistema dipende da più parametri, si può ipotizzare di poter agire, almeno matematicamente, su un altro parametro, in questo caso r, che misura con quanta efficacia il fondale fa rientrare nutrimento nelle acque. Considero dunque $\ell = 0.1$, $\alpha = 2$, e rappresento in figura 3.3 le soluzioni di equilibrio la variare di r tra 3.2 e 4.3. In questo caso, risulta che al decrescere di r gli equilibri n_2 e n_3 si avvicinano e spariscono.
Dunque una strategia per ritornare all'equilibrio n_1 potrebbe essere quella di tornare a $\ell = 0.1$, arrivando però sull'equilibrio n_3. A questo punto si fa decrescere artificialmente r (rimuovendo strati di fondale o altro), fino a che non si produce la catastrofe che fa sparire n_3 e costringe il sistema a tornare sull'equilibrio n_1 (l'unico che c'è). Possiamo poi tornare al valore di r originario.

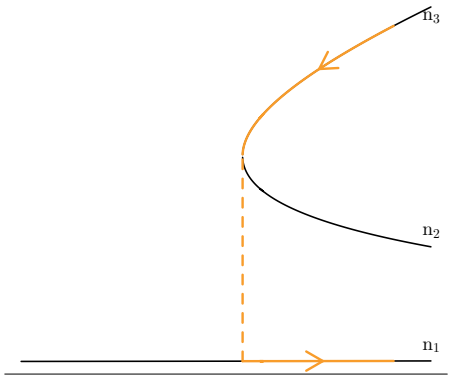


Figure 3.3: Gli equilibri al variare di r.

La situazione descritta nell'esempio può sembrare complessa ma ha aspetto geometrico relativamente semplice e comune a vari fenomeni. Ci sono due parametri che governano gli equilibri, r e ℓ. Per alcuni valori l'equilibrio è unico e stabile, per altri ci sono due equilibri stabili e uno instabile (quello intermedio tra i due).
Nel grafico in figura 3.4 il piano di base è il piano (r, ℓ), l'asse verticale è il valore di n di equilibrio. Come si vede, la superficie disegnata dagli equilibri al variare dei due parametri presenta una "piega": fuori dalla piega c'è un unico equilibrio, stabile. Nel piano (r, ℓ) questa situazione corrisponde alla regione fuori dalle due curve nere. Invece, quanto (r, ℓ) è un punto nella regione tra le due curve nere, ci sono tre posizioni di equilibrio, di cui quella intermedia è instabile.
Le frecce disegnano in questo grafico le situazioni che abbiano descritto precedentemente. Aumentare ℓ fa sparire l'equilibrio inferiore, e fa precipitare il sistema nell'equilibrio superiore, che è uno stato stabile di eutrofizzazione. Tornare indietro con il valore di ℓ non ci riporta allo stato precedente, bisogna prima passare per una catastrofe inversa, in cui sparisce

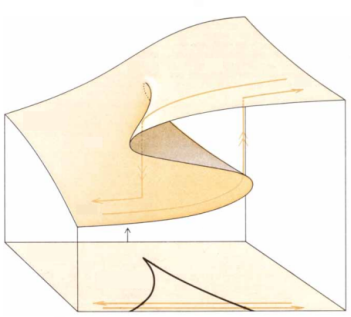


Figure 3.4: La “catastrofe a piega”.

l’equilibrio superiore. Nell’esempio ci riusciamo operando su r. Ci si riesca anche operando sul solo parametro ℓ, ma bisogna considerare valori negativi di ℓ, cioè immaginare di poter filtrare l’acqua per liberarla del nutrimento in eccesso (ma questa situazione non sarebbe modellizzabile con un valore costante, come abbiamo fatto).

3.2 Lo shift di ecosistemi

C’è un altro modo di rappresentare la catastrofe a piega, che viene molto usato anche in studi ecologici. La base di questa descrizione viene dalla fisica, in particolare dalla meccanica. Ricordo in particolare che nei sistemi fisici gioca un ruolo essenziale l’energia potenziale. Supponiamo di considerare un caso unidimensionale e di avere a che fare con una particella che nel punto x ha energia potenziale $V(x)$. La meccanica ci dice che in x la particella sente una forza $f(x) = -V'(x)$. Consideriamo l’energia potenziale in figura. La forza è nulla dove V' è nullo, cioè nei punti di massimo, di minimo, e di flesso a tangente orizzontale di $V(x)$. Per chiarire il moto conviene pensare a una pallina che si muove lungo il grafico di V, soggetta alla gravità. In effetti se mettiamo la pallina ferma in A, B, C, rimane ferma, poiché la forza è 0. D’altra parte se la mettiamo vicino a uno di questi punti si osservano due situazioni differenti: se la mettiamo vicino a A o B, la pallina si allontana; se la mettiamo vicino a C la pallina oscilla intorno a C. Questo vuol dire che A e B sono posizioni di equilibrio instabili, mentre C è una posizione stabile.
Dunque i sistemi meccanici hanno questa proprietà: i punti di minimo relativo dell’energia potenziale sono punti di equilibrio stabile, i punti a tangente orizzontale che non sono di minimo sono punti instabili.

Per analogia, spesso si immagina che in sistema naturale (per esempio un ecosistema) sia in un equilibrio descritto da un punto di minimo di un’opportuna energia potenziale. Al variare dei parametri l’energia potenziale può cambiare e può cambiare la natura degli equilibri.
Considero come esempio l’energia potenziale rappresentata nella figura 3.5, che si modifica al modificarsi di qualche parametro. Partendo dal grafico a destra in alto: inizialmente c’è un solo punto di equilibrio $x = x_1$, che è stabile. Al modificarsi del potenziale l’equilibrio si sposta di poco, in x_2, ma rimane unico e stabile. Nel terzo grafico, l’equilibrio stabile è in x_3 (vicino ai precedenti), ma compare un equilibrio instabile, nel punto di flesso a tangente orizzontale. Nel grafico successivo il punto instabile biforca in un punto instabile (il punto di massimo relativo) e in uno stabile, il punto di minimo relativo a destra. In questo momento dunque il sistema ha due equilibri stabili, ma rimane in quello in cui era, e lo stesso accade in primo grafico in basso a sinistra. Nel grafico successivo la posizione di equilibrio in cui

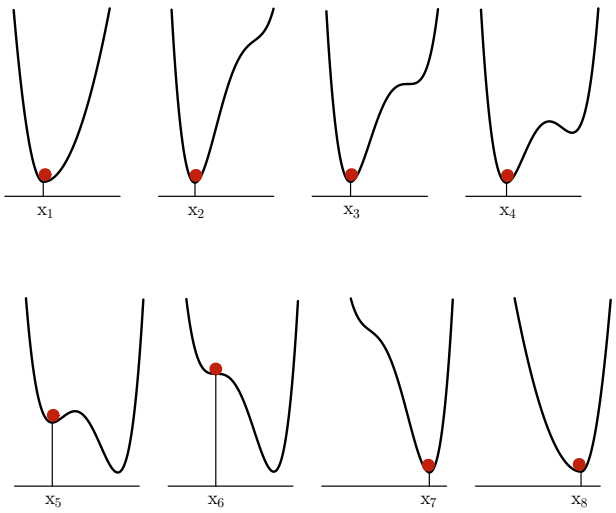


Figure 3.5: Energia potenziale che si modifica provocando uno shift catastrofico dell'equilibrio.

si trova il sistema va a coincidere con l'equilibrio instabile, in un punto di flesso a tangente orizzontale, con ascissa x_6. In questo istante il sistema non è più in equilibrio, e può solo precipitare nell'unico equilibrio stabile a destra, che nei due grafici successivi si sposta in x_7 e x_8. Si noti che mentre i punti da 1 a 6 sono vicini tra loro, e i punti da 7 a 8 sono vicini tra loro, il primo gruppo è distante dal secondo: il sistema di trova ora in un'altra situazione.

Nella figura 3.6 rappresento lo stesso sistema, ma utilizzando il grafico di biforcazione. In ascissa c'è il parametro che viene cambiato, in ordinata le posizioni di equilibrio.
In letteratura trovate facilmente grafici tipo quello in figura 3.5, per esempio nel lavoro di rassegna M. Sheffer, S. Carpenter, J.A. Foley, C. Folke, B. Walker: *Catastrophic shifts in ecosystems* Nature, vol 413, 11 (2001). Un tipico esempio è la transizione di un ecosistema da boscoso a erbaceo, in cui, per il cambiare delle temperature medie e dell'umidità, l'equilibrio dello stato "boscoso" diventa instabile e sparisce, e il sistema raggiunge un altro stato stabile, quello "erbaceo". Un punto di estrema importanza è che anche se i parametri (temperatura e umidità) tornano ai livelli precedenti, il sistema rimane nell'equilibrio "erbaceo" fino a che esiste, o fino a che altri fattori esterni (per esempio un rimboschimento artificiale) non lo cambiano.

Termino questo paragrafo con un esercizio, che serve a far vedere che i sistemi possono cambiare in modi differenti da quelli fin qui discussi.
Mi sono liberamente ispirato al modello studiato nell'articolo "Bautin bifurcations in a forest-grassland ecosystem with human-environment interactions" Scientific Reports (2019) 9:2665

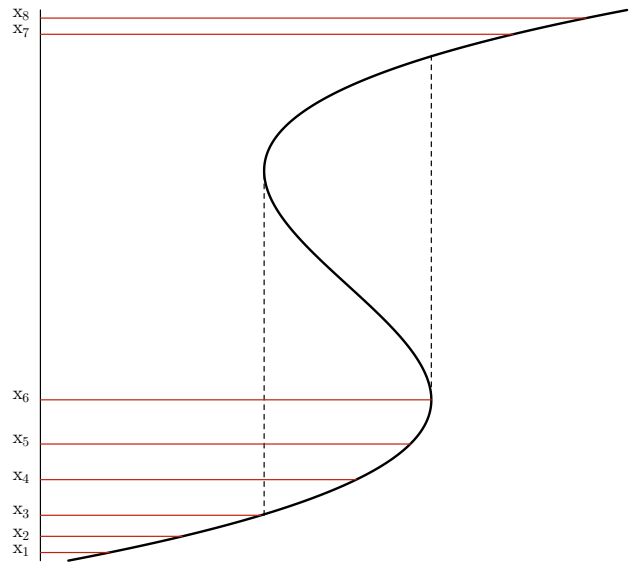


Figure 3.6: Grafico di biforcazione completo per lo shift catastrofico

https://doi.org/10.1038/s41598-019-39296-x semplificandolo.
Si vuole descrivere l'interazione tra una comunità umana e l'estensione di foreste e praterie nella regione di insediamento.
La prima variabile che introduciamo è $f \in [0, 1]$, che è frazione di terreno coperto da foreste. Il valore $1 - f$ è la frazione di terreno coperta da praterie. In assenza di intervento umano, la foresta si riduce, a favore delle praterie, con tasso costante ν. Incendi periodici bruciano le praterie, ma meno gli alberi, sopratutto dove sono più densi e dunque è meno presente il sottobosco. Questo effetto si traduce in un termine di crescita del tipo

$$af(1 - f)w(f)$$

La parte $af(1 - f)$ è un termine di crescita limitata (tipo Verulsth) e non a tasso costante, perché f è una variabile limitata da 1. La funzione $w(f)$ modula il termine di crescita in base alla densità della foresta, e dunque è una funzione crescente. Nell'articolo viene suggerita un'espressione che qui non discuto, per i miei scopi assumo

$$w(f) = f^b$$

L'equazione per f è dunque

$$f' = -\nu f + af(1 - f)f^b$$

L'interazione proposta con la comunità umana è di tipo "dinamica delle opinioni", che riscuote un grande interesse nella modellistica sociale. Indichiamo con x la frazione degli umani "favorevoli" alla foresta. Ci si aspetta che x cresca se f scende sotto una certa soglia, e che x

decresca se f è sopra una certa soglia. Un'espressione semplice con queste caratteristiche è

$$x(1-x)(1-2f)$$

La parte $x(1-x)$ è di nuovo un termine di crescita limitata (x deve essere tra 0 e 1), il termine $1-2f$ è positivo se $f < 1/2$, negativo se $f > 1/2$.
L'effetto di x su f viene modellizzato con il termine

$$hfx$$

che è un contributo di crescita per f con tasso proporzionale alla grandezza di x.

Esercizio 17.

a Scrivi il sistema complessivo

b Poni $\nu = 0.3$, $a = 0.5$, $b = 0.8$, e considera $h = 0$ (cioè non c'è azione umana sulla foresta). Simula il sistema e descrivine il comportamento.

c Ora poni $h = 0.8$. Simula il sistema e descrivine il comportamento.

d Porta il valore di b a 1.2 Simula il sistema e descrivine il comportamento.

e Esplora altri valori dei parametri, e scrivi una relazione conclusiva sul comportamento del modello.

3.3 Modelli differenziali discreti

Vedi [BDM par. 7.4], e in particolare gli esempi 7.4.7, 7.4.8.

3.4 Il modello di May e la transizione al caos

Vedi [BDM par. 7.4] esempio 7.4.12

Chapter 4

Richiami di probabilità e statistica

Pensavo di dare per per scontate le nozioni base, che comunque possono essere trovate su [BDM cap. 10, 11, 12]. Qui faccio una sintesi.

4.1 Mediana, quantili, frequenze cumulate

Vedi [BDM cap 12], o un qualunque testo di statistica elementare, o anche direttamente le esercitazioni con R.

4.2 Proprietà estremali della media

È noto che l'informazione contenuta in una collezione di dati viene spesso sintetizzata con il valore medio, e che la dispersione dei dati viene misurata mediante la deviazione standard. Approfondiamo questi concetti.
Supponiamo di avere una variabile statistica X, che assume i valori dati numerici, $X_1, \dots X_N$. L'idea di fondo è rappresentare la variabile con un solo numero x, con l'espressione

$$X_i = x + \Delta X_i$$

dove $\Delta X_i = X_i - x$ è chiamato, a seconda dei contesti, **scarto** o **errore**. Il numero x che sintetizza i valori della variabile X, deve in un qualche senso essere il più vicino possibile a tutti i dati. Una interessante scelta per la funzione di vicinanza $v(x)$ è data dalla somma dei quadrati delle distanze dei punti da x. Poiché il numero dei dati è fissato, per comodità dividiamo questa somma per N, e in questo modo consideriamo la media delle quadrati delle distanze dei dati da x:

$$v(x) = \frac{1}{N}\sum_{i=1}^{N}(X_i - x)^2$$

Svolgendo i quadrati si ottiene

$$v(x) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - 2\frac{1}{N}\sum_{i=1}^{N} xX_i + \frac{1}{N}\sum_{i=1}^{N} x^2.$$

La seconda somma è pari a $x\overline{X}$, dove $\overline{X}$ è la proprio la media aritmetica, l'ultima somma è pari a x^2. Dunque, cambiando l'ordine della somma,

$$v(x) = x^2 - 2x\overline{X} + \frac{1}{N}\sum X_i^2$$

che è una funzione quadratica, con la concavità rivolta verso l'alto. Il valore di x che minimizza $v(x)$ è dunque l'ascissa del vertice, che risulta essere proprio $\bar{X}$.
Usando la definizione, possiamo notare che $v(\overline{X})$ è proprio la media del quadrato degli scarti da $\overline{X}$, cioè, per definizione è la varianza

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \overline{(X-\overline{X})^2}.$$

Questo valore deve coincidere con l'ordinata del vertice della parabola che è

$$v(\overline{X}) = \overline{X}^2 - 2\overline{X}^2 + \overline{X^2} = \overline{X^2} - \overline{X}^2.$$

Riassumendo: la media aritmetica dei dati è il valore che minimizza la media dei quadrati delle distanze, che in tal caso è pari alla varianza. Inoltre abbiamo mostrato che

$$\sigma_X^2 = \overline{(X-\overline{X})^2} = \overline{X^2} - \overline{X}^2.$$

cioè che la varianza è uguale alla differenza tra la media dei quadrati dei dati e il quadrato della media dei dati.

Ultima osservazione: quando si eleva al quadrato una somma, si ottengono tre termini: i quadrati dei due termini e i doppi prodotti. È una proprietà della media che da

$$X_i = \overline{X} + \Delta X_i$$

quadrando e sommando, i doppi prodotti se ne vanno, cioè

$$\overline{X^2} = \overline{X}^2 + \sigma_X^2.$$

Questa formula è un primo esempio di decomposizione della variabilità (quadratica): la variabilità di X rispetto a 0 si decompone in un termine dovuto alla distanza della media da 0, il termine $\overline{X}^2$, e in un termine di variabilità di X rispetto alla media, il termine σ_X^2.

Un altro modo di vedere questo argomento, forse più sintetico, è il seguente. Torno alla funzione che misura la distanza quadratica di x dai dati:

$$v(x) = \frac{1}{N}\sum_{i=1}^{N}(X_i - x)^2$$

che è lo scarto quadratico medio rispetto a x. Sommo e sottraggo la media dentro, e svolgo i quadrati

$$v(x) = \frac{1}{N}\sum_{i=1}^{N}((X_i - \overline{X}) + (\overline{X} - x))^2 = \sigma_X^2 + 2(\overline{X} - x)\overline{X - \overline{X}} + (\overline{X} - x)^2 = \sigma_X^2 + (\overline{X} - x)^2.$$

Come si può notare, il doppio prodotto scompare. Da questa espressione si capisce che $v(x)$ ha il valore minimo in $\overline{X}$, e in tal caso vale σ_X^2, e che la media ha una proprietà importante: la scarto quadratico medio rispetto a un valore x si decompone in due termini positivi: uno che è la distanza al quadrato tra x e il valore medio, e l'altro che è la varianza, cioè lo scarto quadratico medio rispetto alla media.

4.3 Coppie di variabili statistiche

Per le nozioni introduttive su correlazione d covarianza si veda [BDM cap 12]. Qui presento una versione alternativa della determinazione della retta dei minimi quadrati.
Supponiamo di aver preso N dati per due variabili statistiche, X e Y. Osservando il grafico di dispersione notiamo un buona correlazione, e in effetti il valore del coefficiente di correlazione

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

è in modulo vicino a 1. In tal caso, possiamo pensare che esista una legge lineare $y = ax + b$, che "spiega" y in funzione di x. I dati però non sono perfettamente allineati ($|\rho| < 1$), dunque vale, al variare di i,

$$Y_i = aX_i + b + \varepsilon_i \tag{4.3.1}$$

dove ε_i prende il nome di errore o anche "residuo" (nel senso che il valore di Y "dovrebbe" essere quello teorico, $aX_i + b$, ma c'è una differenza residua rispetto alla "spiegazione").
Il primo problema che ci poniamo è quello di determinare la migliore retta possibile che sia vicina ai dati, cioè dobbiamo trovare a e b. Anche in questo caso è essenziale fissare prima in che senso la retta deve essere migliore. Nell'ottica in cui abbiamo scritto il modello lineare (4.3.1), cerchiamo di minimizzare la somma dei quadrati dei residui: $\frac{1}{N}\sum \varepsilon_i^2$. Come abbiamo già visto nel caso della media, se riusciamo nella nostra operazione di minimizzazione, la media degli errori deve essere 0. Imponendo questo fatto e calcolando la media di entrambi i membri dell'equazione (4.3.1) si ottiene

$$\overline{Y} = a\overline{X} + b$$

Dunque la retta migliore (che è la **retta di regressione** o **retta dei minimi quadrati** passa necessariamente per il baricentro dei dati, cioè per il punto del piano (x, y) che ha come coordinate $(\overline{X}, \overline{Y})$.
Ovviamente questa informazione non basta a trovare la retta: i parametri sono due, e per ora abbiamo ottenuto una sola informazione. Usando questa informazione scriviamo gli scarti di Y:

$$\Delta Y_i = Y_i - \overline{Y} = aX_i + b - (a\overline{X} + b) + \varepsilon_i = a\Delta X_i + \varepsilon_i$$

Dunque

$$\frac{1}{N}\sum \varepsilon_i^2 = \frac{1}{N}\sum (\Delta Y_i - a\Delta X_i)^2$$

che svolgendo il quadrato e sviluppando le somme, è

$$\frac{1}{N}\sum \varepsilon_i^2 = \sigma_Y^2 - 2a\sigma_{XY} + a^2\sigma_X^2$$

Quindi la media dei quadrati dei residui è una funzione quadratica del coefficiente a, che ha come grafico una parabola. Il minimo si ottiene se a è l'ascissa del vertice, cioè

$$a = \frac{\sigma_{XY}}{\sigma_X^2}$$

che si può anche scrivere in termini del coefficiente di correlazione:

$$a = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_X}{\sigma_Y}.$$

Questa uguaglianza ci dice quanto vale a, usando il fatto che la retta passa per il baricentro si determina anche b, risolvendo il problema di partenza.
Usando il valore di a appena trovato, intanto notiamo che

$$a^2\sigma_X^2 = \rho^2\sigma_Y^2.$$

Inoltre possiamo calcolare quanto vale al minino la media del quadrato dei residui, che, poiché i residui hanno media nulla, è la varianza dei residui:

$$\sigma_\varepsilon^2 = \sigma_Y^2 - 2\rho^2\sigma_Y^2 + \rho^2\sigma_Y^2 = \sigma_Y^2 - \rho^2\sigma_Y^2$$

Quindi, anche in questo caso, dalla ottimizzazione del parametro a nel modello

$$\Delta Y_i = a\Delta X_i + \varepsilon_i$$

si ottiene

$$\sigma_Y^2 = a^2\sigma_X^2 + \sigma_\varepsilon^2 = \rho^2\sigma_Y^2 + \sigma_\varepsilon^2$$

Detto in parole: la varianza della variabile Y è la somma di un contributo dovuto alla varianza della variabile X, più un contributo di variabilità dovuto alla varianza dei residui. Dividendo per σ_Y^2 si ottiene

$$1 = \rho^2 + \sigma_\varepsilon^2/\sigma_Y^2$$

che leggiamo in questo modo: ρ^2 è la frazione della variabilità di Y spiegata dal modello lineare, mentre $\sigma_\varepsilon^2/\sigma_Y^2 = 1 - \rho^2$ è la frazione residua. Un modello lineare sarà tanto migliore quanto più ρ^2 si avvicinerà a 1. Il numero ρ^2 è anche detto **coefficiente di determinazione**.

Una ultima osservazione: data la variabile X, la variabile

$$\frac{X - \overline{X}}{\sigma_X}$$

è una variabile adimensionale, a media nulla e con deviazione standard 1. È la **standardizzazione** della variabile X. La retta di regressione ha una forma semplice in termini di variabili standardizzate, infatti

$$\frac{Y - \overline{Y}}{\sigma_Y} = \rho\frac{\sigma_Y}{\sigma_X}\frac{1}{\sigma_Y}(X - \overline{X}) = \rho\frac{X - \overline{X}}{\sigma_X}$$

cioè il coefficiente angolare della retta tra gli scarti standardizzati è proprio il coefficiente di correlazione. Se i dati sono perfettamente allineati con correlazione positiva, gli scarti standardizzati sono perfettamente uguali, cioè $\rho = 1$. In caso di perfetto allineamento con correlazione negativa, gli scarti sono opposti in segno, ma uguali in modulo.

4.4 Probabilità ed eventi

Fenomeni che sono governati da troppe cause, o da cause sconosciute, assumono un aspetto "casuale" (o **aleatorio**) nel loro verificarsi. Esempi semplici e classici sono il lancio di un dado o di una moneta.
La modellizzazione matematica di un evento casuale *discreto* (ovvero di un evento che può verificarsi in un numero finito di varianti), viene fatta in quattro passi:

1) Identificare l'insieme (o spazio) degli **eventi elementari**

$$E = \{e_1,\, e_2,\, \ldots,\, e_n\},$$

cioè l'insieme delle n varianti con cui l'evento può verificarsi. Nel caso della moneta $E = \{T, C\}$, nel caso del dado $E = \{1, 2, 3, 4, 5, 6\}$

2) Assegnare un valore di **probabilità** ad ogni evento elementare. Indicherò con p_i il valore assegnato all'evento e_i. Si assume $0 \leq p_i \leq 1$; il valore 0 è assegnato agli eventi **impossibili**, il valore 1 all'evento **certo**, ovvero che si verifica sicuramente. Inoltre la **somma delle probabilità è 1**:

$$\sum_{i=1}^{n} p_i = 1.$$

Nel modello per la moneta non truccata si considerano equiprobabili gli eventi "T" e "C", dunque si assegna ad essi la probabilità $\frac{1}{2}$. Per una moneta truccata questi numeri saranno diversi: indicando con p_T la probabilità che esca testa, la probabilità che esca croce sarà $p_C = 1 - p_T$. Il caso estremo (una moneta con la "testa" su entrambe le facce) è descritto da $p_T = 1$ e $p_C = 0$.

3) Identificare gli **eventi** che si vogliono descrivere. Chiamerò evento ogni **sottoinsieme dello spazio degli eventi elementari**

$$A \subset E.$$

Il motivo di questa definizione si chiarisce meglio con un esempio: nel lancio di un dado potrebbe interessarci il fatto che esca un numero pari; questo "evento composto" racchiude gli eventi elementari $2, 4, 6$, quindi lo identifichiamo con il sottoinsieme $\{2,\, 4,\, 6\}$ dello spazio degli eventi.

4) Usare la regola di calcolo per le probabilità degli eventi composti: se $A \subset E$, la probabilità che si verifichi A, indicata con $P(A)$, è la somma delle probabilità degli eventi elementari che costituiscono A. Nell'esempio precedente, la probabilità che esca un numero pari è $p_2+p_4+p_6$.

La regola per il calcolo della probabilità di un evento composto è in caso particolare della seguente regola più generale:

A e B sono eventi **incompatibili** se non possono verificarsi contemporaneamente; in tal caso la probabilità dell'evento "A o B" è la somma delle probabilità di A e B

Espressa in termini matematici:

$$\text{se } A \cap B = \emptyset, \quad \text{allora } P(A \cup B) = P(A) + P(B),$$

Infatti gli eventi sono incompatibili se non c'è nessun evento elementare che appartenga ad entrambi (e dunque l'intersezione degli eventi è l'insieme vuoto), mentre l'evento "A o B" è costituito dall'insieme degli eventi elementari per cui si realizza A oppure B, quindi dall'unione insiemistica di A e B.

Esercizio di riepilogo. Un dado tetraedrico irregolare ha le facce numerate da 1 a 4. Si supponga che $p_1 = \frac{1}{3}$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{6}$. Determinare p_4. Descrivere insiemisticamente gli eventi seguenti, e calcolarne la probabilità

- A_1 esce un numero dispari
- A_2 non esce un numero dispari
- A_3 non esce 3
- A_4 esce 2 o un numero dispari

Dire quali sono le coppie di eventi incompatibili.

La "legge dei grandi numeri" dà una motivazione alla costruzione precedente, e lega la probabilità alla statistica.
Supponiamo di avere una moneta non truccata, e facciamo N lanci. Possiamo trattare i dati che abbiamo ottenuto con metodi statistici definendo

$$F_N(T) = \text{frequenza assoluta dell'uscita di T in N lanci.}$$

e

$$f_N(T) = F_N(T)/N = \text{frequenza relativa dell'uscita di T in N lanci.}$$

Se il nostro modello probabilistico è quello giusto per descrivere il fenomeno, ci aspettiamo che al crescere del numero delle prove la frequenza relativa si avvicini al valore assegnato della probabilità.

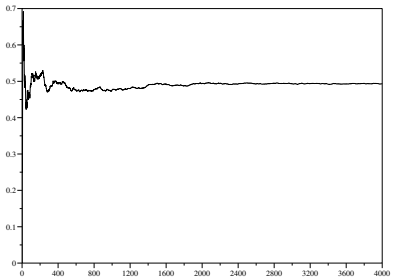


Figure 4.1: Frequenze relative di T al variare del numero di prove

Più in generale, in un modello assegneremo all'evento e_i la probabilità $P(e_i) = p_i$ se ci aspettiamo che p_i sarà il valore asintotico della frequenza relativa dell'evento e_i. In genere la scelta delle p_i viene fatta in base a semplici considerazioni, in cui gioca un ruolo fondamentale la presunzione di **equiprobabilità**: se non abbiamo motivo di pensare che un evento sia favorito rispetto a un altro, gli assegneremo la stessa probabilità. Questo è il caso delle monete, dei dadi, dei giochi di carte. A partire da semplici esempi però si possono costruire modelli molto utili e interessanti.

4.5 Eventi indipendenti

Consideriamo il lancio consecutivo di due monete non truccate. Evidentemente, l'esito del lancio della seconda moneta non dipende dall'esisto del lancio della prima, e viceversa. In particolare l'evento "la prima moneta dà T" è indipendente dall'evento "la seconda moneta dà T". In termini di probabilità, A e B si definiscono **indipendenti** se

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

La ragionevolezza di questa definizione si comprende pensando alle frequenze relative nel caso di molti lanci delle due monete. Supponiamo di fare 1000 lanci. In circa 500 casi la prima moneta dà T; di questi casi, in circa la metà (250) anche la seconda moneta dà T. Dunque la frequenza relativa dell'uscita TT sarà circa un quarto.

Un caso particolare di eventi indipendenti è quello in cui si ripete lo stesso esperimento probabilistico, per esempio il lancio successivo di una moneta, per la quale la probabilità che esca testa è p.
Supponiamo di effettuare 10 lanci. Lo spazio degli eventi è $E^{10} = E \times E \times E \times E \times E \times E \times E \times E \times E \times E$. Per semplicità indicherò gli eventi con sequenze ordinate di T e C. Ad esempio TTCTCCTCCC invece di $(T, T, C, T, C, C, T, C, C, C)$. Resta inteso che il primo simbolo si riferisce al primo lancio, il secondo al secondo, etc. .

Esercizio. Quanti sono gli eventi elementari in E_{10}? [Risposta: $2^{10} = 1024$]

Esercizio. Qual è la probabilità dell'evento TTCTCCTCCC? [Risposta: $p^4(1-p)^6$]
E dell'evento $TTTTCCCCCC$? [Risposta: la stessa]
E dell'evento $CCCCCCTTTT$? [Risposta: la stessa]

Esercizio. Sia N il numero di lanci che viene effettuato; data una stringa (sequenza) qualunque, come determino facilmente la sua probabilità? [Risposta: se k è il numero di T, la probabilità è $p^k(1-p)^{(N-k)}$.

Come si vede dagli esempi precedenti, la probabilità di una sequenza dipende solo dal numero di T che contiene, e non dal loro ordine. La domanda a cui vogliamo rispondere è:
su n lanci, qual è la probabilità che esca k volte T?

Esempio. Con quale probabilità ho due T con due lanci? E una T? E nessuna T?
L'evento "due T" è esattamente l'evento TT, che ha probabilità p^2. L'evento "nessuna T" è esattamente l'evento CC, che ha probabilità $(1-p)^2$. L'evento "una T" è un evento composto: $\{TC, CT\}$. Essendo TC e CT eventi elementari (quindi incompatibili), la probabilità cercata è la somma delle singole probabilità. Esse sono entrambe uguali a $p(1-p)$ (infatti il numero di T è lo stesso). Dunque la probabilità di "una testa" è $2p(1-p)$.

In generale, l'evento "su n lanci esce k volte T" è un evento composto: è l'insieme di tutte le stringhe che contengono esattamente k volte testa; ogni singola stringa con k volte T ha probabilità $p^k(1-p)^{n-k}$.
Dunque per determinarne la probabilità è sufficiente contare il numero di stringhe che hanno esattamente k volte T. La risposta a questa domanda è data dal **coefficiente binomiale**:

$$\binom{N}{k} = \frac{N!}{k!\,(N-k)!},$$

dove con il punto esclamativo si intende il **fattoriale** del numero, cioè il prodotto di tutti gli interi da 1 al numero:

$$a! = 1 \times 2 \times \cdots \times a.$$

In generale, si chiama **distribuzione binomiale** la legge di probabilità che descrive il numero di "successi" su N prove indipendenti, assumendo che la probabilità di successo in una singola prova sia p:

$$P(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

4.6 Probabilità condizionate, formula di Bayes, test diagnostici

Vedi [BDM] cap. 10.

4.7 Variabili aleatorie

Si chiama **variabile aleatoria** una qualunque funzione degli eventi elementari. Questa definizione astratta si concettizza spesso nel caso in cui associamo numeri ad eventi. Per esempio, il numero di successi su N prove indipendenti di un esperimento è una **variabile aleatoria binomiale**.
La scelta di un numero intero a caso tra $1, 2, \ldots N$ è invece una **variabile aleatoria uniforme** in $\{1, \ldots N\}$. Concettualmente non c'è differenza tra trattare eventi che si chiamano A, B, o e, oppure eventi a cui è associato un valore numerico, però il caso di variabili aleatorie numeriche permette di definire delle quantità molto utili, i **valori attesi**.
Li introduco con un esempio. Si supponga di partecipare a un test a risposta multipla. Ogni domanda ha 5 possibili risposte, di cui una sola esatta a cui è assegnato punteggio 1. Alla risposta non data è assegnato il punteggio 0. Per scoraggiare risposte casuali, in genere viene assegnato un punteggio negativo alle risposte errate. Supponiamo che sia -0.25. Chiediamoci cosa accade a uno studente che risponde sempre a caso, su un test fatto di N domande. Il suo punteggio medio per domanda sarà

$$\frac{+1 \times k - 0.25 \times (N-k)}{N}$$

dove k è il numero di risposte esatte che ha dato. Poiché la probabilità di date una risposta esatta scegliendo a caso è $1/5$, invocando la legge dei grandi numeri, ci aspettiamo che il rapporto k/N sia vicino a $1/5$. Analogamente, il rapporto $(N-k)/M$ sarà vicino a $4/5$, che è la probabilità di dare una risposta errata. Dunque il punteggio medio per domanda sarà vicino al valore

$$+1 \times \frac{1}{5} - 0.25 \times \frac{4}{5} = 0$$

"In media", lo studente che risponde a caso riceve 0 punti per esercizio (questo è il motivo della scelta di -0.25 per il punteggio delle risposte errate).
Per predire il punteggio medio su un gran numero di prove, abbiamo sommato i possibili valori della variabile (in questo caso $+1$ e -0.25, pesati con la loro probabilità). Generalizziamo.

Sia X una variabile aleatoria, che può assumere i valori reali $x_1, \dots x_n$, con probabilità $p_1, \dots p_n$. Si chiama **valore atteso** il numero

$$\langle X \rangle = \sum_{i=1}^{n} x_i p_i$$

(nell'esempio precedente, $n = 2$). Supponiamo ora di estrarre N volte questa variabile, ottenendo i valori $X_1, \dots X_N$ (nell'esempio precedente, questi valori sono i punteggi che lo studente ottiene nelle singole domande). Si chiama **media empirica** il valore

$$m_N = \frac{1}{N} \sum_{h=1}^{N} X_h$$

Riorganizziamo i termini della somma, indicando con $F_N(j)$ la frequenza assoluta con cui esce x_j e con $f_N(j)$ la frequenza relativa

$$m_N = \frac{1}{N} \sum_{i=1}^{n} x_i F_i(N) = \sum_{i=1}^{n} x_i f_i(N)$$

Se N è grande, ci si aspetta che $f_i(N) \simeq p_i$, dunque

$$m_N \simeq \langle X \rangle$$

cioè il valore atteso è predittivo del valore della media empirica, per N grande.

Spesso è utile considerare funzioni di variabili aleatorie. In generale, se f è una funzione e X è una variabile aleatoria, $f(X)$ è una variabile aleatoria, e il suo valore atteso è

$$\langle f(X) \rangle = \sum_{i=1}^{n} f(x_i) p_i$$

In particolare, si chiama **varianza** il valore atteso dello scarto quadratico

$$\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle$$

Come vedremo successivamente, questo valore è anche uguale alla differenza tra il valore atteso del quadrato della variabile e il quadrato del valore atteso:

$$\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$$

Come esercizio calcoliamo la varianza del voto dello studente che risponde a caso:

$$\sigma^2 = (1-0)^2 \times \frac{1}{5} + (-0.25 - 0)^2 \times \frac{4}{5} = \frac{1}{5} + \frac{1}{4^2} \times \frac{4}{5} = \frac{1}{4}$$

Calcoliamo anche valore atteso e varianza di una variabile aleatoria X che vale 1 con probabilità p, e 0 con probabilità $1 - p$:

$$\langle X \rangle = 1 \times p + 0 \times (1-p) = p$$

$$\sigma^2 =< (X-p)^2 >= (1-p)^2 \times p + (0-p)^2 \times (1-p) = p(1-p)$$

Sui valori attesi di somma e prodotto valgono i due seguenti importanti fatti:

- il valore atteso della somma di variabili aleatorie è uguale alla somma dei valori attesi delle variabili;
- la varianza della somma di variabili aleatorie **indipendenti** è uguale alla somma delle varianze delle variabili.

La prima asserzione sembra molto naturale, pensando a qualche esempio. Nel caso di risposte casuali a due domande, il valore atteso del voto per la prima domanda è zero, quello per la seconda è zero, il voto atteso complessivo sarà naturalmente $0 + 0 = 0$. In questo esempio, però, le due variabili, voto alla prima domanda, e voto alla seconda domanda, sono indipendenti. Immaginiamo che lo studente scelga a caso quale risposta dare alla prima domanda, e alla seconda risponda esattamente nello stesso modo. In questo caso le due variabili non sono indipendenti, ma il valore atteso è sempre 0. Lo stesso accede se lo studente sceglie a caso la seconda risposta tra quelle diverse dalla risposta che ha dato alla prima domanda. Per esercizio, si calcoli la varianza del voto totale nei tre casi descritti: scelta casuale indipendente, scelta casale identica, scelta casuale differente. Quale sarà la maggiore?
Dimostro che la prima asserzione è vera, anche se le variabili non sono indipendenti. Considero due variabili, X che assume valori x_i, $i = 1, \dots k$, e Y che assume valori y_j, $j = 1 \dots h$. Conoscere la distribuzione di probabilità per X e per Y è una descrizione parziale, perché non prendiamo in considerazione come si accoppiano i valori delle due variabili. Invece è necessario specificare

$$P(X = x_i \text{ e } Y = y_j)$$

che è detta **distribuzione congiunta** delle due variabili. Se le variabili sono indipendenti, allora

$$P(X = x_i \text{ e } Y = y_j) = P(X = x_i)P(Y = y_j)$$

Nel caso generale $P(X = x_i \text{ e } Y = y_j) = p_{ij}$ saranno degli opportuni valori, a somma 1. Sommando su tutti i possibili valori che può assumere la variabile Y, si ottiene la distribuzione della variabile X e viceversa:

$$P(X = x_i) = \sum_{j=1}^{h} P(X = x_i \text{ e } Y = y_j) = \sum_{j=1}^{h} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{k} P(X = x_i \text{ e } Y = y_j) = \sum_{i=1}^{k} p_{ij}$$

A questo punto è facile calcolare

$$\begin{aligned}\langle X + Y \rangle &= \sum_{i=1}^{k}\sum_{j=1}^{h}(x_i + y_j)p_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{h} x_i p_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{h} y_i p_{ij} \\ &= \sum_{i=1}^{k} x_i \sum_{j=1}^{h} p_{ij} + \sum_{j=1}^{h} y_j \sum_{i=1}^{k} p_{ij} \\ &= \sum_{i=1}^{k} x_i P(X = x_i) \sum_{j=1}^{h} y_j P(Y = y_j) = \langle X \rangle + \langle Y \rangle\end{aligned}$$

Vediamo invece perché è vera la seconda asserzione, nel caso in cui

$$P(X = x_i \text{ e } Y = y_j) = P(X = x_i)P(Y = y_j)$$

$$\langle XY\rangle = \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j P(X=x_i)P(Y=y_j)$$
$$= \sum_{i=1}^{k} x_i P(X=x_i) \sum_{j=1}^{h} y_j P(Y=y_j) = \langle X\rangle\langle Y\rangle$$

Come conseguenza, si ha che la varianza della somma di due variabili aleatorie indipendenti è uguale alla somma delle varianze. Infatti lo scarto quadratico della somma è

$$(X+Y-\langle X+Y\rangle)^2 = (X-\langle X\rangle + Y - \langle Y\rangle)^2$$
$$= (X-\langle X\rangle)^2 + 2(X-\langle X\rangle)(Y-\langle Y\rangle) + (Y-\langle Y\rangle)^2$$

Il valore atteso di $X-\langle X\rangle$ e di $Y-\langle Y\rangle$ è zero, dunque, usando l'indipendenza, si ottiene che il valore atteso del termine al centro è 0. I valori attesi degli altri due sono esattamente le varianza di X e Y. In sintesi, per due variabili indipendenti:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

Ora siano in grado di calcolare senza (troppi) sforzi, il valore atteso e la varianza di una variabile aleatoria relativa a N lanci con probabilità di successo p per ogni lancio. Infatti la variabile aleatoria F_N = numero di successi su N esperimenti si ottiene contando il numero di successi, cioè

$$F_N = X_1 + \cdots + X_N$$

dove X_i vale 1 se all'i-esimo esperimento si ottiene il successo, e 0 altrimenti (chiamo F_N questa variabile, perché è la frequenza di successi negli N esperimenti). Le variabili aleatorie X_i sono indipendenti, hanno media p e varianza $p(1-p)$, come abbiamo calcolato negli esercizi precedenti. Usando che il valore atteso della somma è pari alla soma dei valori attesi si ottiene

$$\langle F_N\rangle = Np$$

(come c'era da aspettarsi, pensando alle frequenze). Usando l'indipendenza, anche la varianza è la somma delle varianze, e dunque

$$\sigma^2 = Np(1-p)$$

4.8 Medie empiriche e valori attesi

Torniamo ancora sull'esempio della variabile binomiale F_N, numero di successi in N prove, e chiediamoci che cosa accade al crescere di N. Il suo valore atteso è pN, la sua varianza $Np(1-p)$. Ricordando che la deviazione standard esprime l'ordine di grandezza dello scostamento dal valore atteso, possiamo dire che

$$F_N = pN + \text{ errore, con l'errore dell'ordine di } \sqrt{N}$$

Quindi l'ordine di grandezza dell'errore cresce al crescere di N. Torniamo all'esempio della moneta non truccata. Su 100 lanci mi aspetto circa 50 teste, con un errore dell'ordine di $\sqrt{100/4} \approx 5$. Su 10 000 lanci, l'errore è dell'ordine di 50. Su un milione di lanci l'errore è dell'ordine di 500.

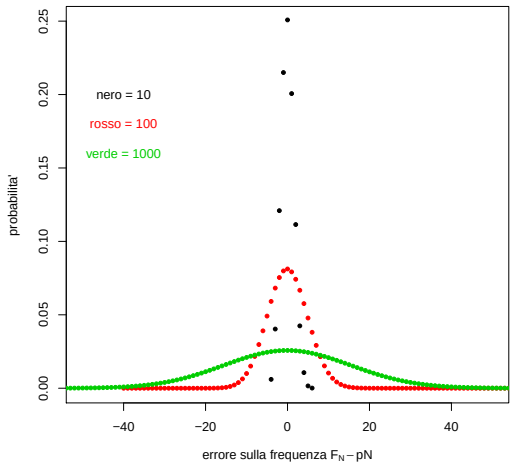


Figure 4.2: Distribuzione di probabilità dell'errore $F_N - pN$

Vediamo graficamente questo fatto, nel caso $p = 0.4$, rappresentando la distribuzione della variabile $F_N - pN$, per alcuni valori di N. Come si nota, al crescere di N la distribuzione si allarga e si abbassa, rendendo probabile che l'errore sia un numero grande.

Chiediamoci invece cosa accade alla media empirica del numero di successi, cioè a $f_N = F_N/N$. Il valore atteso di questa variabile è p, la varianza è $Np(1-p)/N^2 = p(1-p)/N$. Dunque

$$f_N = p + \text{ errore, con l'errore dell'ordine di } \frac{1}{\sqrt{N}}$$

Stavolta, al crescere di N, la taglia dell'errore diminuisce.

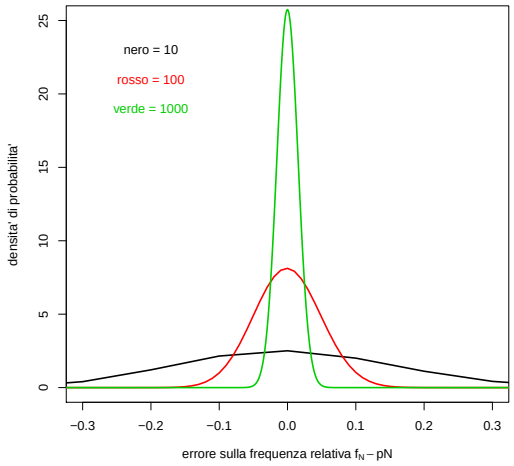


Figure 4.3: Distribuzione di probabilità dell'errore $f_N - p$

In questo secondo grafico non ho rappresentato le probabilità ma le densità, quindi le probabilità sono rappresentate dalle aree.

Stiamo descrivendo con maggior dettaglio la legge dei grandi numeri: per N grande, il valore della frequenza relativa dista dalla probabilità per un errore di taglia $1/\sqrt{N}$, che dunque va a 0 per N che tende a $+\infty$. Graficamente, al crescere di N la densità di probabilità diventa sempre più alta e più stretta, al contrario del caso precedente, in cui diventava più bassa e più larga.

Ricapitolando

$$F_N - pN = \sum_{i=1}^{N}(X_i - p) \text{ è dell'ordine di } \sqrt{N}$$

$$f_N - p = \frac{1}{N}\sum_{i=1}^{N}(X_i - p) \text{ è dell'ordine di } \frac{1}{\sqrt{N}}$$

Osservando queste due relazioni, si comprende che se invece di dividere la somma per per N si divide per $\sqrt{N}$ allora la varianza dell'errore è di ordine 1, non va nè a ∞ nè a 0. Osserviamo graficamente cosa accade alla variabile aleatoria

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(X_i - p)$$

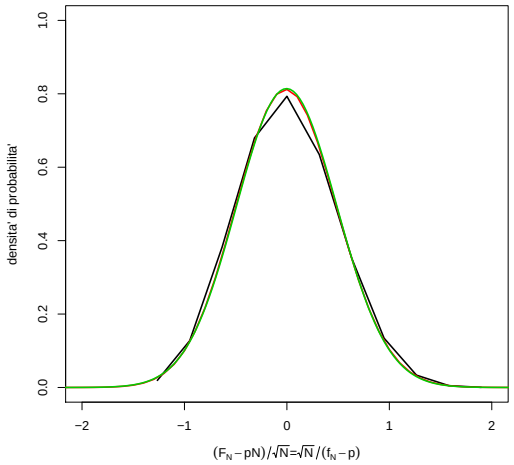


Figure 4.4: Distribuzione di probabilità di $\sum_i (X_i - p)/\sqrt{N}$

La forma della curva si stabilizza rapidamente. Prima di descrivere esattamente cosa accade, richiamo due definizioni.

Una variabile aleatoria **normale standard** è una variabile aleatoria Z che può assumere tutti i valori reali, e ha densità di probabilità

$$\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$$

Il suo valore atteso è 0, la sua deviazione standard è 1 (graficamente, ± 1 sono le ascisse dei punti di flesso).

Consideriamo ora la variabile aleatoria $\sigma Z + \mu$, con $\sigma > 0$ e μ qualunque parametri fissati. Il suo valore atteso è μ, perché Z ha valore atteso nullo. La sua varianza è uguale alla varianza di σZ, perché gli scarti non dipendono da μ, e questa variabile ha varianza σ^2. La variabile $\sigma Z + \mu$ è detta normale o gaussiana, di media μ e deviazione standard σ, e ha densità di probabilità

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

In questo caso, $x = \mu$ è asse di simmetria per il grafico, e μ è l'ascissa del massimo, mentre $\mu \pm \sigma$ sono le ascisse dei punti di flesso. Il grafico si ottiene da quello nella normale standard dilatando di σ le x, dividendo per σ le y, e traslando il grafico a destra di μ. In figura mostro mostro il caso di $\mu = 0$.

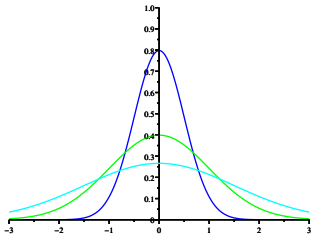


Figure 4.5:

All'aumentare di σ la "campana" si allarga, infatti σ è una misura di dispersione: più è grande, più i dati sono dispersi. In figura $\sigma = 1/2, 1, 3/2$. Se x è una variabile gaussiana di media μ e varianza σ^2,

$$\begin{aligned} P(-\sigma < x - \mu < \sigma) &\simeq 0.6826895 \\ P(-2\sigma < x - \mu < 2\sigma) &\simeq 0.9544997 \\ P(-3\sigma < x - \mu < 3\sigma) &\simeq 0.9973002 \\ P(-1.95996\sigma < x - \mu < 1.95996\sigma) &\simeq 0.95 \end{aligned}$$

Per $N \to +\infty$ la variabile aleatoria

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (X_i - p)$$

tende proprio a una variabile aleatoria gaussiana, di media nulla e di varianza $p(1-p)$. Più in generale vale il **teorema del limite centrale**: siano $X_1, X_2 \ldots$ variabili aleatorie indipendenti e con la stessa distribuzione, di media μ e varianza σ^2. Allora

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i$$

tende a una variabile aleatoria gaussiana di media μ e varianza σ^2.

Il fatto descritto da questo teorema ci permette di precisare meglio il rapporto tra la media empirica in esperimenti ripetuti e il valore atteso: la media empirica

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

è approssimativamente distribuita come una variabile aleatoria gaussiana, di media μ e deviazione standard $\sigma/\sqrt{N}$, e questo fatto è fondamentale nelle applicazioni. La situazione più tipica è quella di voler conoscere una proprietà media di una popolazione statistica. Faccio qualche esempio: in ambito politico sociologico, le intenzioni di voto dei cittadini di uno stato per un referendum, la propensione alla lettura negli adolescenti; nell'ambito delle scienze della vita, in contenuto di sostanze nocive nelle specie dei vari livelli trofici, le dimensioni medie degli adulti di specie animali acquatiche. In tutti questi esempi, non si può ottenere la misura su tutta la popolazione, dunque ci si accontenta di misurare i dati su un **campione**. Non entro nella teoria del campionamento, e mi limito a considerare il caso più semplice, in cui il campione sia ottenuto scegliendo in sequenza un elemento a caso della popolazione. In questo modo siamo esattamente nella condizione descritta dalle ipotesi del teorema del limite centrale, e la media empirica sul campione sarà uno **stimatore** della media vera, con errore che decresce proporzionalmente alla radice quadrata della dimensione del campione.
Nella pratica, sono molto più comuni i campionamenti senza ripetizione, in cui il campione è scelto tutto insieme. Il risultato non cambia, e si tenga presente che se la popolazione è grande rispetto alla numerosità del campione, una strategia di campionamento con ripetizione difficilmente genererà un campione con ripetizioni.

Un'osservazione sulla varianza nei campionamenti. La legge dei grandi numeri e il teorema del limite centrale ci dicono che la media empirica è uno **stimatore** della media di popolazione. In particolare, il suo valore atteso è proprio la media di popolazione.
Se conoscessimo la media di popolazione,

$$\frac{1}{N}\sum_i (X_i - \langle X\rangle)^2$$

sarebbe uno stimatore della varianza. Però $\langle X\rangle$ non è nota, e va sostituita con la media empirica m_N. In questo modo, però, si ottiene uno stimatore non corretto, nel senso che il suo valore atteso non è uguale alla varianza di popolazione.
Si chiama **varianza campionaria** il numero

$$s_N^2 = \frac{1}{N-1}\sum_i (X_i - m_N)^2$$

Si prova che il suo valore atteso, in un campionamento casuale con ripetizioni, è pari alla varianza di popolazione. (Nel caso di campionamento senza ripetizioni, il valore atteso è pari alla varianza campionaria di tutta la popolazione, che, se la popolazione è molto numerosa, differisce di poco da quella di popolazione).

Chapter 5

Indici di diversità

La statistica descrittiva fornisce anche strumenti per la definizione di indici di biodiversità (indici che fanno parte dell'ampia classe degli indici di diversità), che misurano quanto poco omogenea sia una distribuzione.
Per fare un esempio concreto con cui introdurre questi indici, considero una situazione in cui su due territori si misurano le presenze di 3 differenti varietà di una pianta (V1, V2, V3) secondo la seguente tabella di abbondanze.

territorio	V1	V2	V3
A	0.20	0.30	0.50
B	0.30	0.35	0.35
C	0.10	0.20	0.70
D	0.40	0	0.60

Stiamo lavorando sulla tabella di frequenze relative di variabili nominali. Il primo indice utile è la **ricchezza di specie** cioè il **numero di fattori** a frequenza non nulla. Questo indice è 3 per territori A, B, C, ed è 2 per il territorio D, dove dunque c'è meno biodiversità.

Si possono però considerare indici più sofisticati, legati alla descrizione probabilistica dei dati. Supponiamo di essere nel territorio A, e di scegliere casualmente una pianta. La tabella delle frequenze relative ci dà la probabilità con cui osserveremo le tre possibili varietà,

$$p_1 = P(V1) = 0.2,\ p_2 = P(V2) = 0.3,\ p_3 = P(V3) = 0.5$$

Per dare la misura della biodiversità, possiamo calcolare la probabilità che due piante scelte a caso siano uguali. Se questo numero è alto, c'è poca biodiversità, se è basso ce ne è molta. Il calcolo è facile, e si fa usando le regole per il calcolo della probabilità. Sia x_1 la varietà a cui appartiene la prima pianta scelta, e sia x_2 la varietà a cui appartiene la seconda pianta. Vogliamo calcolare

$$P(x_1 = x_2) = P(x_1 = V1, x_2 = V1) + P(x_1 = V3, x_2 = V3) + P(x_1 = V3, x_2 = V3)$$

(si usa la regola della somma perché si tratta di eventi incompatibili). Per calcolare il valore dei singoli addendi possiamo usare la regola del prodotto per il calcolo della probabilità degli eventi indipendenti. Si ottiene

$$P(x_1 = x_2) = p_1^2 + p_2^2 + p_3^2$$

In caso di n varietà,

$$P(x_1 = x_2) = \sum_{i=1}^{n} p_i^2$$

Questo numero si chiama **indice di Simpson**, e misura l'uguaglianza, più che la diversità, nel senso che il valore più alto possibile è 1, e si ottiene se c'è solo una specie, e dunque si ha il minimo possibile di biodiversità e il massimo dell'uguaglianza. Fissato il numero n di specie, il minimo di questo indice si ha quando le abbondanze relative sono uguali, che dunque devono essere pari a $1/n$, e dunque

$$P(x_1 = x_2) = \sum_{i=1}^{n} \frac{1}{n^2} = n\frac{1}{n^2} = \frac{1}{n}$$

Si noti che, nello studio della genetica degli eucarioti, se p_1, p_2, p_3 sono le frequenze dei tre possibili alleli di un gene in una popolazione, allora questo numero è l'**omozigosità**, cioè la frequenza relativa degli omozigoti. L'opposto di questo indice è l'**eterozigosità**, cioè la probabilità di incontrare un eterozigote, che dunque è pari a

$$1 - \sum_{i=1}^{n} p_i^2$$

Nello studio della biodiversità, questo indice è l'**indice di Gini-Simpson**, e vale 0 se non c'è biodivesità, mentre vale $1 - 1/n$, se c'è la massima biodiversità, fissata la ricchezza di specie n.
Per esercizio, si calcoli questo indice per i tre territori della tabella precedente.

Si possono avere indici uguali anche in presenza di situazioni differenti di biodiversità: per esempio avere moltissime specie con bassa frequenza relativa, può essere equivalente ad avere poche specie con frequenze piuttosto differenti.
Per poter distinguere tra queste situazioni si utilizzano delle generalizzazioni di questo indice, che si ottengono a partire da un'interpretazione dell'indice di Simpson. Si può considerare p_i come l'abbondanza relativa a_i della specie i. Dunque

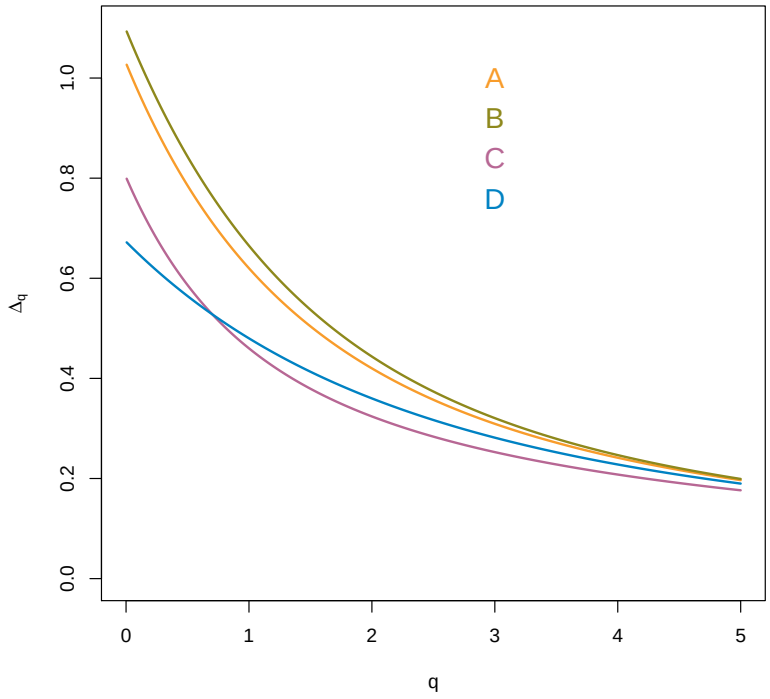
$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{n} p_i p_i = \sum_{i=1}^{n} p_i a_i = \text{ valore atteso dell'abbondanza relativa}$$

Dato $q > 0$, si definisce la famiglia di indice di uguaglianza

$$\lambda_q = \sum_{i=1}^{n} p_i a_i^q = \sum_{i=1}^{n} p_i^{q+1}$$

che è il valore atteso della potenza q dell'abbondanza relativa. A partire da questo indice si può costruire il **profilo di diversità**

$$\Delta_q = \frac{1}{q}(1 - \lambda_q)$$



Per $q = 1$ si ottiene l'indice di Gini-Simpson. Più è grande q, più il contributo delle basse frequenze diminuisce. Per fare un esempio, se ho 10 specie con frequenza $1/100$, nell'indice di Simpson ho un contributo $10/100^2 = 1/1000$, nel caso di λ_2 ho un contributo $10/100^3 = 10^{-5}$ (si veda nel grafico successivo lo scavalcamento delle curve di C e D). Dunque al crescere di q questo indice trascura le abbondanze piccole. Al contrario, al decrescere di q si esalta il contributo alla diversità dovuto alle specie di piccola abbondanza relativa. Vale dunque la pena calcolare il limite per q che tende a 0, ultimo valore possibile per questo indice. Per chi se lo ricorda, questo calcolo si può fare con la regola di de l'Hôpital, osservando che

$$\frac{\mathrm{d}}{\mathrm{d}q} p^{1+q} = \frac{\mathrm{d}}{\mathrm{d}q} \mathrm{e}^{(1+q)\ln p} = p^{1+q} \ln p$$

Dunque

$$\Delta_0 = \lim_{q \to 0} \frac{1}{q}(1 - \lambda_q) = -\sum_{i=1}^{n} p_i \ln p_i$$

Questo indice si chiama **entropia di Shannon** e ha un ruolo cruciale anche in informatica teorica, perché permette di definire il contenuto di informazione di una sequenza di simboli. Più è bassa l'entropia, meno informazione c'è, meno biodiversità c'è.
Si noti che l'entropia di Shannon è il valore atteso di meno il logaritmo della probabilità. Si può pensare che $-\ln p_i$ quantifichi la "sorpresa" di osservare l'evento di frequenza p_i: se $p_i = 1$ la sospresa è 0, se $p_i = 0$ la sorpresa è infinita (si noti la coincidenza formale con la legge di Weber-Fechner: se p_i passa da $1/10$ a $1/100$, la sorpresa raddoppia).
Si faccia attenzione al caso in cui una delle p_i sia nulla: un normale programma di computer non calcola $0 \ln 0$, ma il calcolo del limite ci permette di attribuire il valore 0 a quest'espressione.

A partire da λ_q si può costruire un'ulteriore famiglia di indici. Ricordando che λ_1 è l'indice

di Simpson, che vale 1 se c'è una sola specie, e vale $1/n$ se ce ne sono n, si può definire

$$D_1 = \frac{1}{\lambda_1}$$

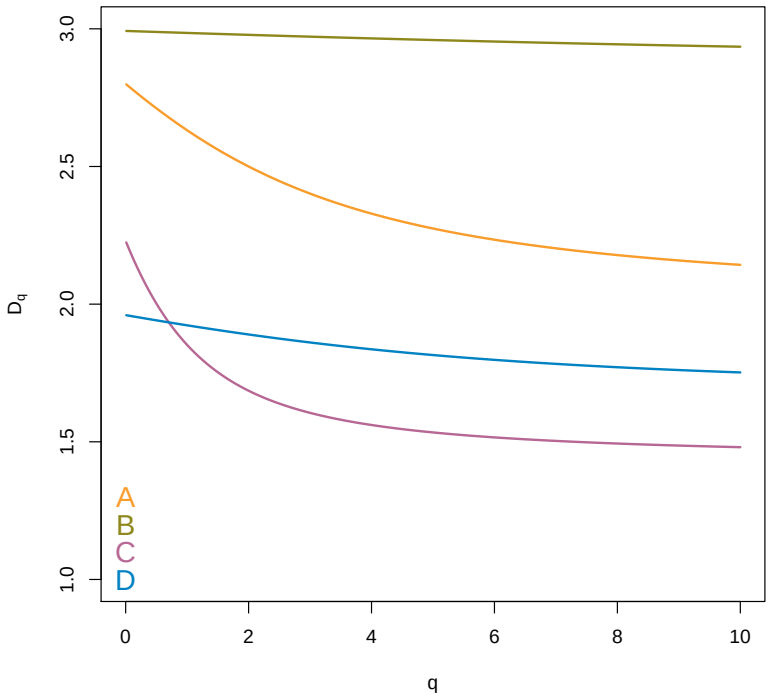
che si può interpretare come il **numero effettivo di specie**, cioè il numero di specie di distribuzione uguale, che darebbe lo stesso valore osservato di λ_1. Si generalizza questo numero considerando

$$D_q = \frac{1}{\lambda_q^{1/q}}$$

che prende il nome di **numero di Hill di ordine** $q+1$, e che si interpreta sempre in termini di numero effettivo di specie. Nel limite $q \to 0$ si ottiene $D_0 = \mathrm{e}^{-\Delta_0}$. Invece nel limite $q \to +\infty$ questo indice tende a

$$\frac{1}{\max p_i}$$

che come si vede dipende solo dalla specie di frequenza massima.
Questi indici sono più leggibili, in quanto rappresentano un "numero di specie", e infatti nell'esempio che stiamo considerando, variano tra 1.5 e 3.



Chapter 6

Introduzione ai test statistici

6.1 Test binomiale esatto

[BDM 12.4] e materiale di laboratorio.

6.2 $z-$test e $t-$test

Introduco questi test con un esempio. Supponiamo che sia noto che la distribuzione delle lunghezze dei pesci di una data specie in laghetto sia ben approssimata da una gaussiana di media 14 cm e deviazione standard 1.2 cm.
In un laghetto vicino viene trovato un pesce simile, ma di lunghezza 14 cm.
Ci chiediamo se il fatto che la lunghezza sia un po' diversa, ci faccia dubitare che il pesce sia della stessa specie di quelli del primo laghetto.
L'ipotesi H_0 di questo test è che il dato sulla lunghezza del pesce si il risultato di una estrazione casuale di una variabile gaussiana Z, di media 14 e deviazione standard 1.2. Il valore p del test è la probabilità delle code:

```
> 2*pnorm(14,mean=12.5,sd=1.2,lower.tail=F)
[1] 0.2112995
```

Il valore è superiore al 20%, dunque non possiamo dubitare dell'ipotesi H_0. Si osservi che è noto che per una gaussiana standard le code oltre 1.96 pesano il 5%. Il valore standardizzato della misura è

$$\frac{14 - 12.5}{1.2} = 1.25 < 1.96$$

e infatti non possiamo rifiutare H_0.

Supponiamo ora di aver trovato sei pesci nel secondo laghetto, e che la media delle loro lunghezze sia 14. In questo caso la variabile aleatoria che rappresenta la misura è la media di 6 misure. È noto dalla teoria che se le misure X_i sono variabili gaussiane, di media m e deviazione standard σ, allora la media empirica

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

è gaussiana di media m e deviazione standard $\sigma/\sqrt{N}$.

Quindi per testare l'ipotesi nulla H_0 che i 6 pesci abbiano lunghezze distribuite come quelle dei pesci del primo laghetto, calcoliamo il valore

$$z = \frac{14 - 12.5}{1.2/\sqrt{6}} = 1.25 \times \sqrt{6} \approx 3$$

valore che supera 2.57 che corrisponde a un valore di soglia per p di 0.01, ma è al di sotto di 3.29 che corrisponde al valore di soglia $p = 0.001$. Infatti

```
> 2*pnorm(14,mean=12.5,sd=1.2/sqrt(6),lower.tail=F)
[1] 0.002199647
```

In questo caso dobbiamo smentire l'ipotesi nulla H_0, e concludere che i pesci del secondo laghetto sono di una specie con una caratteristica differente (la lunghezza).
A questo punto possiamo anche chiederci quant'è la lunghezza media dei pesci del secondo laghetto. Il valore misurato è 14, ma il valore vero sarà presumibilmente un numero differente. Si chiama **intervallo di confidenza** per la media vera l'insieme di tutti i valori m che, assunti come ipotesi nulla, non verrebbero smentiti dal test, al livello di soglia stabilita. In questo caso la soglia è 5%, che viene descritta come **livello di fiducia** del 95%. Quali sono il valori di m, media teorica, che non verrebbero smentiti dal dato osservato? Per le gaussiane questo conto è semplice: deve essere

$$\frac{|m - 14|}{1.2/\sqrt{6}} \leq 1.96$$

cioè

$$|m - 14| \leq 1.96 \times 1.2/\sqrt{6}$$

che corrisponde all'intervallo richiesto

$$m \in (14 - 0.96, 14 + 0.96) = (13.04, 14.96).$$

Il valore della media teorica 12.5 non è in questo intervallo, e infatti l'ipotesi nulla è stata rifiutata alla soglia del 5%. Si noti, infine, che non si può affermare che con probabilità del 5% la media vera è in quell'intervallo, perché la media vera è un numero, non una variabile aleatoria.

Questo facile esempio purtroppo non è realistico, perché in genere non c'è modo di conoscere la deviazione standard di una popolazione. Questa informazione viene in genere ottenuta per campionamento. Ricordo che si chiama varianza campionaria

$$s_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - m_N)^2$$

dove m_N è la media empirica. Si divide per $N - 1$ perché i numeri $X_i - m_N$ non sono tutti indipendenti, infatti la loro somma fa 0. Il numero di quelli indipendenti è $N - 1$, ed è detto **numero dei gradi di libertà**, abbreviato con df ("degrees of freedom").
Quando abbiamo a che fare con dati empirici, dobbiamo tener presente che la varianza non è nota, ma è solo stimata approssimativamente dalla varianza campionaria. Se misuriamo N dati con media empirica (campionaria) m_N e deviazione standard s_N, e assumiamo come H_0 che la media vera (detta anche "teorica") sia il valore assegnato m, allora il numero

$$t = \frac{m_N - m}{s_N/\sqrt{N}}$$

è distribuito non come una variabile aleatoria gaussiana, ma come una variabile "t di Student" a $N-1$ gradi di libertà. Il $t-$test mette alla prova l'ipotesi H_0 che la media sia m, misurando le code della distribuzione di Student rispetto al valore calcolato t. Supponiamo per esempio che le misure delle lunghezze dei sei pesci fossero

$$15.79\ 12.72\ 15.84\ 12.28\ 12.94\ 14.42$$

La media empirica è approssimativamente 14, la deviazione standard campionaria è 1.58. Il **valore della statistica** è

$$t = (14 - 12.5)/1.59 \times \sqrt{6} \approx 2.3225$$

che va valutato rispetto alla distribuzione t di Student a 5 grandi di libertà:

```
2*pt(2.32,df=5,lower.tail=F)
[1] 0.06784195
```

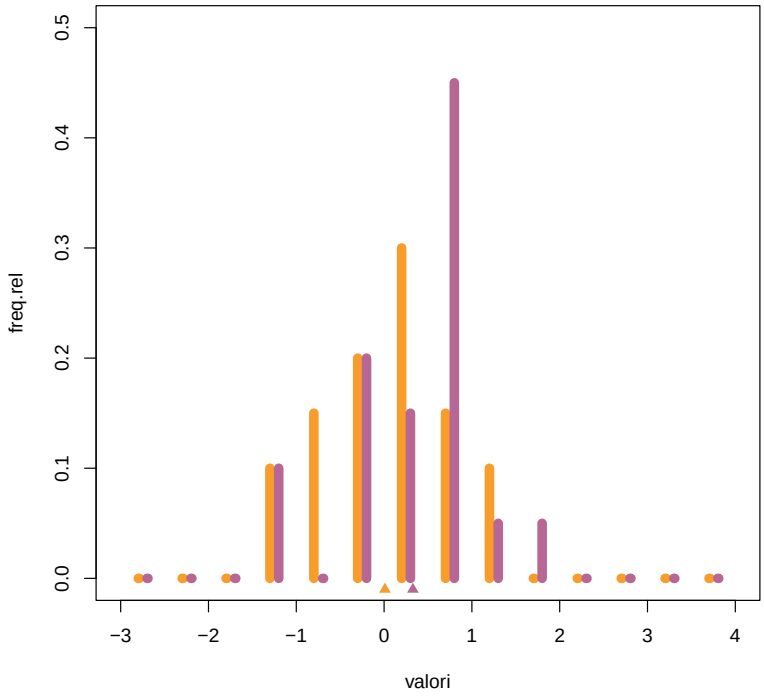
(le istruzioni `pt rt dt qt` sono le analoghe di `pnrom rnorm dnorm qnorm` per la distribuzione t di Student). Anche in questo caso rifiutiamo l'ipotesi nulla. Sia per il $t-$test che per il calcolo degli intervalli di confidenza per la media possiamo ricorrere direttamente a un'istruzione di R:

```
> t.test(y,m=12.5)

        One Sample t-test

data:  y
t = 2.3225, df = 5, p-value = 0.06784
alternative hypothesis: true mean is not equal to 12.5
95 percent confidence interval:
 12.33993 15.65674
sample estimates:
mean of x
 13.99833
```

Condizioni per fare il `t-test`

L'uso del `t-test` prevede che la popolazione da cui è estratto il campione sia gaussiana. La verifica "a occhio" della guassiantià può essere fatta con il comando `qnorm` che confronta i quantili del campione con i quantili di una guassiana standard; se nel grafico si vede, approssimativamente, una retta, allora il campione si può considerare guassiano. Una risposta più affidabile si ottiene con un apposito test `shapiro.test`.

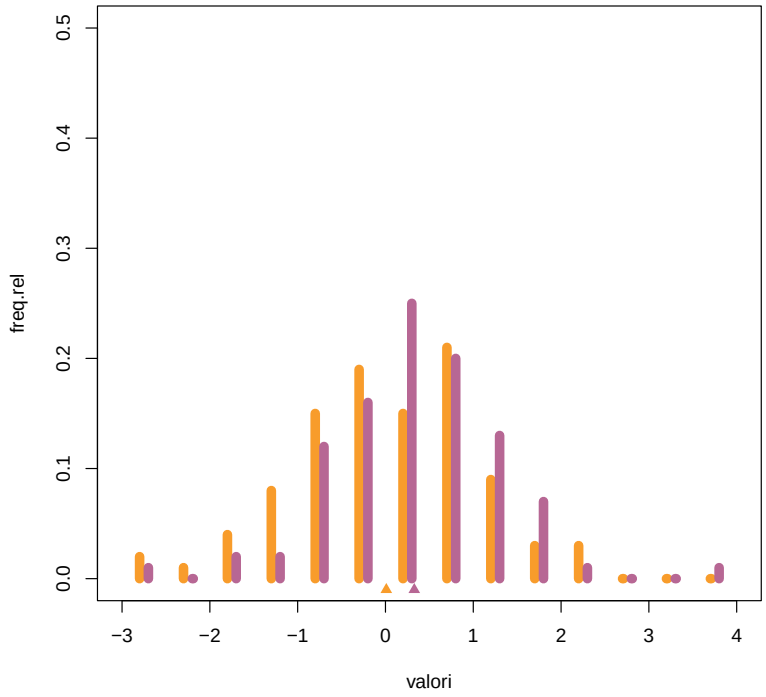
Il $t-$test si usa anche per il confronto tra medie di gruppi. Consideriamo la figura seguente, in cui sono riportati due istogrammi appaiati per 20 valori X (in arancione), e 20 valori Y, in viola. Le due corrispondenti medie sono indicate dai due triangoli vicino a 0, e valgono $m_X \approx 0.01$, e $m_Y \approx 0.33$, e la differenza è $\delta = m_X - m_Y \approx -0.32$.



Ci chiediamo se queste due medie sono differenti.

```
t.test(X,Y)

        Welch Two Sample t-test

data:  x20 and y20
t = -1.4027, df = 38, p-value = 0.1688
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7782064  0.1411719
sample estimates:
  mean of x   mean of y
0.009658318 0.328175561
```

Come si vede la risposta è no, e viene anche calcolato l'intervallo di confidenza per la differenza tra le due medie, che infatti contiene lo 0.
Ripetiamo l'analisi nel caso di 100 dati per X e 100 dati per Y, rappresentati nella seconda figura.



I valori delle medie sono gli stessi di prima: $m_X \approx 0.01$, e $m_Y \approx 0.33$, e la differenza è $\delta = m_X - m_Y \approx -0.32$.

```
t.test(X,Y)

        Welch Two Sample t-test

data:  x100 and y100
t = -2.3209, df = 196.9, p-value = 0.02132
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.58916599 -0.04786849
sample estimates:
  mean of x   mean of y
0.009658318 0.328175561
```

In questo caso, però, il test indica che dobbiamo rifiutare l'ipotesi nulla, nonostante la differenza tra le due medie sia la stessa. Cosa è cambiato?
Il test deve valutare se la differenza delle due medie sia frutto del campionamento (ipotesi H_0) oppure no (ipotesi alternativa). In entrambi i casi la differenza è la stessa, e neanche gli istogrammi sono poi così differenti. Il fatto è che nel secondo caso la differenza è grande rispetto all'errore statistico che si compie su una media di 100 dati.

A rigore, il `t-test` per due campioni richide
- normalità dei campioni
- uguaglianza delle varianze
Ho già segnalato che la normalità si verifica mediante lo `shapiro.test`. Per l'uguaglianza delle varianze il test più consigliato è il test di Levene (che `R` ha su un pacchetto separato). In altri contesti utilizzeremo il test di Barteltt. All'atto pratico non serve la verifica dell'uguaglianza delle varianze, perché esiste una variante del t-test che è in grado di gestire questo caso, ed è il test di Welch. L'istruzione `t.test` in caso di due campioni fa esattamente il test di Welch.
Cosa fare se i dati non sono distribuiti normalmente? Se il numero di dati è grande, il teorema centrale del limite, che vale per i campionamenti casuali, dovrebbe consentire di usare il `t-test`. Come si capisce se il numero di dati è sufficientemente grande? Come indicazioni pratiche, non usate il t-test in caso di distribuzioni fortemente asimmetriche, e in caso di presenza di outlier (*vedi* WS capitolo 13), Altrimenti è il caso di passare a un *test non parametrico*, in cui H_0 non si basa su una distruzione nota (binomiale, gaussiana, etc.). In particolare, l'alternatica non parametrica al `t-test` è il test dei segni dei ranghi di Wilcoxon (oppure la variante Mann-Whitney test). Per questo test H_0 è un'affermazione sulla mediana (o sulla differenza di due mediane), e non usa l'assunzione di normalità.

Un'altro possibile modo di trattare i casi non gaussiani è di "trasformare" i dati. Per esempio i dati di concentrazione non hanno tipicamente distribuzioni normali, ma il loro logaritmo si. In genere si "prova" a trasformare dati con funzioni semplici, come il logartimo o le potenze. Avere una "teoria" su come sono distribuiti i dati può essere di aiuto, come nel caso delle concentrazioni.

Osservo infine che in genere i test non parametrici sono "meno potenti" dei test parametrici. Infatti l'errore di tipo II nei test parametrici, cioè di accettare l'ipotesi nulla anche se è falsa, è in realtà un'affermazione sul parametro che governa una distribuzione (per esempio la media di una variabile gaussiana). Nel caso di un test non parametrico, lo "spazio" delle ipotesi alternative è più largo.

6.3 ANOVA

Tra i test che riguardano variabili statistiche di conteggio ho brevemente illustrato il test delle proporzioni di R, che è un'implementazione del test del χ^2. Consideriamo il seguente esempio. Sono stati sperimentati tre trattamenti riguardo la fertilizzazione su piante di pomodoro: "bat" (batteri), "fert" (chimici), "ctrl" (controllo: nessuna fertilizzante). A fine esperimento ogni pianta è stata classificata con la variabile dicotomica "ad alta vitalità" / "a bassa vitalità". In questa situazione si chiama **variabile di risposta** la variabile con cui viene descritto lo stato vitale della pianta, e **variabile esplicativa** il trattamento. Il motivo di questi nomi è che il trattamento deve "spiegare" come varia nell'esperimento lo stato vitale della pianta, cioè come la pianta "risponde" ai diversi trattamenti.
Questi sono i dati misurati.

	ctrl	bat	fert
alta	25	33	50
bassa	20	30	30

Per verificare l'ipotesi H_0 che i trattamenti non hanno effetto sulla vitalità, si fa un test di

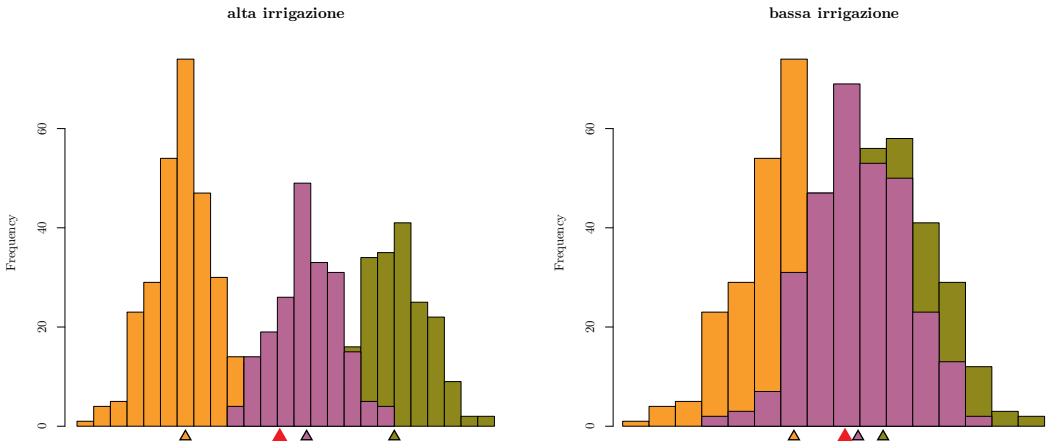
indipendenza del χ^2, o, equivalentemente, un test delle proporzioni:

```
> av <- c(25,33,50)
> bv <- c(20,30,30)
> prop.test(av,av+bv)
     3-sample test for equality of proportions without continuity correction
data:  av out of av + bv
X-squared = 1.5629, df = 2, p-value = 0.4577
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.5555556 0.5238095 0.6250000
```

Fare questo test è concettualmente differente dal confrontare le tre proporzioni a coppie, facendo tre test, uno per ogni coppia. Infatti, fare più test aumenta la probabilità di commettere una errore di tipo I, cioè di rifiutare l'ipotesi nulla se è vera, e dunque si rischia di vedere una differenza dove non c'è. Infatti se in un confronto abbiamo probabilità α di commettere un errore di tipo I, in k confronti la probabilità che almeno un confronto dia un falso positivo è $1 - (1 - \alpha)^k$. Per $\alpha = 0.05$ e $k = 3$ questo valore è 0.14, molto più grande di 0.05. Ci sono dunque tipi strategie per trattare questi casi. Una consiste nel diminuire la soglia di significatività α in funzione del numero k dei livelli della variabile esplicativa. Per la precisione, la correzione di Bonferroni prevede di dividere α per k. Questa correzione è però giudicata troppo "conservativa", cioè sfavorebvole all'individuazione dei veri positivi. La strategia più comune è quella di usare dei test che fanno un confronto complessivo, discriminando tra l'ipotesi H_0 che non ci siano differenze tra i trattamenti, e l'ipotesi alternativa che almento uno dei trattamenti abbia un effetto.

Nel caso di variabili di "risposta" nominali si usa il test del χ^2 che in effetti discrimina l'indipendenza della variabile risposta dalla variabile esplicatica. Nel caso di variabili quantitative, se i livelli della variabile esplicativa fossero due useremmo il `t-test`, per più livelli si ricorre all'ANOVA che ora descrivo.

Consideriamo di nuovo la variabile espicativa data dal trattamento, a 3 livelli, e una variabile di risposta continua, la concentrazione di flavonoidi sulla buccia dei pomodori. Supponiamo di aver fatto due esperimenti in condizioni differenti (per esempio bassa irrigazione e irrigazione nornmale). I dati sono ripostati nei grafici: in arancione i dati di controllo, in viola quelli per fertilizzazione batterica, in verde quelli per fertilizzazzione chimica. I triagoli in basso segnano i valori delle medie nei tre gruppi, il triangolo rosso segna la media complessiva.



Quello che il test deve valutare è se la differenza tra le medie (equivalentemente la differenza con la media complessiva) può essere una fluttazione del caso, e dunque essere dell'ordine della deviazione standard delle medie campionarie, oppure se le medie sono ben separate tra loro. Nelle figure, nel primo caso si osservano medie ben separate, nel secondo un po' meno. In realtà in entrambi i casi H_0 andrà rifiutata, il numero dati della simulazione è elevato, 300 per ogni trattamento, dunque la deviazione standard campionaria è circa un ventesimo di quella delle singole variabili.
Al differenza del t-test che confronta due medie, ora abbiamo a che fare con tre medie. Anche in questo caso non dobbiamo utilizzare tre t-test per confrontare le medie a due a due, ma una procedura complessiva che ci dica se ci sono medie differenti, poi cercheremo di indagare su quali siano queste medie differenti.
Il test il questione prende il nome di ANOVA (Analisys Of VAriance); è un test parametrico che assume la normalità dei dati dei diversi gruppi e anche che la varianza dei gruppi sia la stessa (questa proprietà prende il nome di **omoschedasticità**). Vediamo come funziona.
Vedremo un esempio in laboratorio in cui delle piante di pomodoro vengono sottoposte a tre differenti trattamenti: fertilizzazione mediante fertilizzanti chimici ("fert") o batterici ("bat"), o non fertilizzate (controllo, "cntrl"), e saremo interessati agli effetti di questi trattamenti sulla quantità di flavonoidi sulla buccia, come misura della salute della pianta. In questo caso la variabile esplicativa ha tre possibili valori "fert", "bat", "cntrl". Indicherò con a il numero di livelli della variabile esplicativa, in questo caso $a = 3$. Per ogni trattamento, vengono misurati dei dati. Indico con n_1 il numero di dati relativi al primo trattamento, con n_2 il numero di dati relativi al secondo, etc.. Il numero totale di dati è

$$N = n_1 + \dots n_a.$$

Se tutti gli n_i sono uguali a un valore n, allora l'esperimento si dice **bilanciato**, e in tal caso $N = an$. Indico con X_{ik} il $k-$esimo dato relativo al trattamento $i-$esimo. Indico con

$$X_{i\bullet} = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}$$

la media della variabile per il trattamento $i-$esimo. La media complessiva è

$$X_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{a} \sum_{k=1}^{n_i} X_{ik} = \frac{1}{N} \sum_{i=1}^{a} \left(n_i \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik} \right) = \sum_{i=1}^{a} \frac{n_i}{N} X_{i\bullet}$$

che è anche uguale alla media delle medie relative ai trattamenti, pesate con la dimensione del campione. Come facciamo in generale per una variabile statistica, descriviamo questa variabile come la sua media più uno scarto. Per ogni dato relativo al trattamento i possiamo dunque scrivere

$$X_{ik} = X_{i\bullet} + \varepsilon_{ik}.$$

Si ricordi che a i fissato le variabili ε_{ik} hanno media nulla, Diamo un'espressione dello scarto dalla media complessiva:

$$X_{ik} - X_{\bullet\bullet} = (X_{i\bullet} - X_{\bullet\bullet}) + \varepsilon_{ik}.$$

Dunque lo scarto di un dato dalla media complessiva è pari allo scarto della media del gruppo dalla media complessiva dato da $X_{i\bullet} - X_{i\bullet}$, più lo scarto ε_{ik} del dato rispetto alla media del gruppo. Indico con α_i lo scarto della medie del gruppo dalla media complessiva $\alpha_i = X_{i\bullet} - X_{\bullet\bullet}$

Si chiama **devianza totale** la somma dei quadrati degli scarti tra i dati e la media complessiva

$$SS_{tot} = \sum_{i=1}^{a} \sum_{k=1}^{n_i} (X_{ik} - X_{\bullet\bullet})^2$$

Con "SS" si intende "Sum of Squares", dunque questa è la somma dei quadrati totale. La esplicito rispetto in α_i e ε_{ik}.. Ricordo che ogni volta che se scrivo dei dati come la media più la somma degli scarti, la stessa decomposizione vale per la media del quadrato, che è pari al quadrato della media sommato alla media del quadrato degli scarti. In questo caso, per ogni i, ε_{ik} hanno media nulla in k, e la media di $X_{ik} - X_{\bullet\bullet}$ è α_i. Dunque

$$\frac{1}{n_i} \sum_{k=1}^{n_i} (X_{ik} - X_{\bullet\bullet})^2 = \alpha_i^2 + \frac{1}{n_i} \sum_{k=1}^{n_i} \varepsilon_{ik}^2$$

Moltiplicando questa uguaglianza per n_i e sommando su i si ottiene

$$SS_{tot} = \sum_{i=1}^{a} n_i \alpha_i^2 + \sum_{i=1}^{a} \sum_{k=1}^{n_i} \varepsilon_{ik}^2.$$

L'ultimo termine è la **devianza dentro** i gruppi definiti dai diversi trattamenti:

$$SS_{dentro} = \sum_{i=1}^{a} \sum_{k=1}^{n_i} \varepsilon_{ik}^2.$$

Invece il primo termine è la **devianza tra** i gruppi, perché misura lo scarto delle medie dei gruppi dalla media complessiva:

$$SS_{tra} = \sum_{i=1}^{a} n_i \alpha_i^2 = \sum_{i=1}^{a} n_i (X_{i\bullet} - X_{\bullet\bullet})^2.$$

Se vale l'ipotesi nulla che non c'è differenza tra i gruppi, le differenze tra le medie sono solo il risultato di differenti campionamenti da una stessa distribuzione, dunque la grandezza della variabilità tra i gruppi deve essere simile alla grandezza della variabilità dentro i gruppi. Come esprime questo fatto in termini di SS_{tra} e SS_{dentro}? È necessario calcolare le varianze campionarie, dividendo le devianze per il numero di gradi di libertà. La devianza tra i gruppi

è la somma di a scarti dalla media complessiva, dunque ha $a-1$ gradi di libertà. La devianza nel gruppo i-esimo ha $n_i - 1$ gradi di libertà, dunque la somma ha $N - a$ gradi di libertà. Quindi:

$$s^2_{tot} = \frac{1}{N-1} SS_{tot}$$
$$s^2_{tra} = \frac{1}{a-1} SS_{tra}$$
$$s^2_{dentro} = \frac{1}{N-a} SS_{dentro}$$

La teoria afferma che se vale l'ipotesi nulla, allora s^2_{tra} è uguale a s^2_{dentro} e il rapporto s^2_{tra}/s^2_{dentro} ha una precisa distribuzione, quella di una cosiddetta "variabile F di Fisher a $(a-1, N-a)$ gradi di libertà". Se non vale l'ipotesi nulla, s^2_{tra} sarà maggiore di s^2_{dentro}. Il test è dunque un test monolatero. Il test ANOVA traduce in questo modo il confronto tra più medie, nel confronto tra varianza tra i gruppi e varianza nei gruppi.

ANOVA a due vie

Può accadere che ci sia più di una variabile esplicativa, per esempio potremmo avere delle piante sottoposte a vari trattamenti, in differenti condizioni di irrigazione. In questo caso si parla di ANOVA a due vie, intendendo appunto che ci sono due variabili esplicative. Supponiamo dunque di avere una prima variabile A che assume a distinti valori $A_1, \dots A_a$, e una seconda B che assume b distinti valori $B_1, \dots B_b$. Ci sono dunque ab valori possibili della coppia di variabili (A, B); assumiamo che per ogni coppia abbiano di avere n dati (esperimento bilanciato). Indico con X_{ijk} il valore del $k-$esimo dato per cui $A = A_i$ e $B = B_j$. Il numero totale dei dati sia dunque $N = nab$.
Posso operare esattamente come ho fatto prima, considerando come variabile esplicativa la coppia di variabili (A, B), e dunque scrivo

$$X_{ijk} - X_{\bullet\bullet\bullet} = \delta_{ij} + \varepsilon_{ijk} \tag{6.3.1}$$

dove ε_{ijk} ha media nulla in k,

$$\delta_{ij} = X_{ij\bullet} - X_{\bullet\bullet\bullet}$$

è lo scarto tra le media $X_{ij\bullet}$ della variabile nel gruppo (A_i, B_j) e la media complessiva $X_{\bullet\bullet\bullet}$.
Determiniamo i gradi di libertà delle variabili in gioco. Fissati i e j, ε_{ijk} è una variabile con $n-1$ gradi di libertà, perché la somma in k è 0. Dunque, complessivamente, ε_{ijk} è una variabile a

$$\sum_{i=1}^{a} \sum_{j=1}^{b} (n-1) = abn - ab = N - ab$$

gradi di libertà.
Si può notare che nell'espressione (6.3.1) il membro di sinistra ha $N-1$ gradi di libertà, mentre δ_{ij}, che ha somma nulla in i, j, ha $ab-1$ gradi di libertà. Come verifica di questi conti si nota facilmente che

$$N - 1 = ab - 1 + N - ab.$$

Fin'ora abbiamo trattato le due variabili A e B come un'unica variabile di coppia, mentre vogliamo indagare sugli effetti delle due variabili sulla variabili di risposta. Per fare questo

dobbiamo analizzare il termine di scarto dalle media del gruppo δ_{ij} che riscriviamo nel seguente modo, sommando e sottraendo le medie di X a A fissato e a B fissato:

$$\begin{aligned}\delta_{ij} &= X_{ij\bullet} - X_{\bullet\bullet\bullet} \\ &= X_{i\bullet\bullet} - X_{\bullet\bullet\bullet} + X_{ij\bullet} - X_{i\bullet\bullet} \\ &= (X_{i\bullet\bullet} - X_{\bullet\bullet\bullet}) + (X_{\bullet j\bullet} - X_{\bullet\bullet\bullet}) + (X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})\end{aligned}$$

Chiamo con α_i lo scarto della media del gruppo A_i dalla media complessiva:

$$\alpha_i = X_{i\bullet\bullet} - X_{\bullet\bullet\bullet}$$

con β_j lo lo scarto della media del gruppo B_i dalla media complessiva:

$$X_{\bullet j\bullet} - X_{\bullet\bullet\bullet}$$

Il termine rimanente lo indico con r_{ij}:

$$r_{ij} = X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet}$$

Osservo che α_i ha $a-1$ gradi di libertà, β_j ne ha $b-1$, mentre r_{ij}, che ha media nulla sia in i che in j, ne ha $(a-1)(b-1)$ (come gli scarti di una tipica tabella di contingenza).

Siamo finalmente pronti a mostrare come funziona un test ANOVA a due vie. Abbiamo scritto lo scarto come:

$$X_{ijk} - X_{\bullet\bullet\bullet} = \alpha_i + \beta_j + r_{ij} + \varepsilon_{ijk},$$

cioè come a la somma di 4 contributi: l'effetto α_i del valore della variabile A, l'effetto β_j del valore della variabile B, l'effetto r_{ij} dell'**interazione** tra i valori delle variabili A e B, e una variabilità intrinseca ε_{ijk}. Corrispondentemente, con calcoli analoghi a quelli già fatti per l'ANOVA a una via che non ripeto, decompongo la devianza:

$$SS_{tot} = SS_A + SS_B + SS_{AB} + SS_{dentro}$$

dove SS_A è la devianza tra i gruppi definiti dalla variabile A, dove SS_B è la devianza tra i gruppi definiti dalla variabile B, SS_{AB} è la devianza tra i gruppi definiti dall'interazione delle variabili, SS_{dentro} è la devianza dentro ai gruppi.
Corrispondentemente si determinano le varianze

$$\begin{aligned}s^2_{tot} = \frac{1}{N-1}SS_{tot}, \quad s^2_A = \frac{1}{a-1}SS_A, s^2_B = \frac{1}{b-1}SS_B, \\ s^2_{AB} = \frac{1}{(a-1)(b-1)}SS_{AB}, \quad s^2_{dentro} = \frac{1}{N-ab}SS_{dentro}.\end{aligned}$$

Il test ANOVA a questo punto consiste in tre confronti:

$$s^2_A/s^2_{dentro}, \quad s^2_B/s^2_{dentro}, \quad s^2_{AB}/s^2_{dentro}$$

con lo scopo di scoprire se c'è una variabilità in A, in B, o nell'interazione AB, che eccede la variabilità interna dei dati.

Questo test si chiama ANOVA a due vie con interazione. È possibile rinunciare a osservare l'interazione delle due variabili, considerando la decomposizione

$$X_{ijk} - X_{\bullet\bullet\bullet} = \alpha_i + \beta_j + \bar{\varepsilon}_{ijk},$$

cioè immaginando che lo scarto abbia una componente spiegata da A, una spiegata da B e una parte intrinsecamente variabile. In pratica abbiamo accorpato r_{ij} nella parte variabile:

$$\bar{\varepsilon}_{ijk} = r_{ij} + \varepsilon_{ijk}.$$

Operando in questo modo $\bar{\varepsilon}_{ijk}$ non ha più media nulla in k, ma ha media nulla in k e i, e in k e j. I suoi gradi di libertà sono la somma di quelli di r e di ε, cioè $(a-1)(b-1)+N-ab = N-a-b+1$. La decomposizione della devianza diventa

$$SS_{tot} = SS_A + SS_B + SS_{dentro}$$

Le rispettive varianze sono

$$s^2_{tot} = \frac{1}{N-1}SS_{tot}, \quad s^2_A = \frac{1}{a-1}SS_A, \quad s^2_B = \frac{1}{b-1}SS_B, \quad s^2_{dentro} = \frac{1}{N-a-b+1}SS_{dentro}.$$

Il test ANOVA consiste in due confronti:

$$s^2_A/s^2_{dentro}, \quad s^2_B/s^2_{dentro}.$$

Si noti che rispetto al caso con interazione, cambia s^2_{dentro} dunque valutare se la variabile A ha effetto sulla variabile di risposta può dare risultati differenti, a seconda che consideriamo o meno l'interazione tra le due variabili.

Le ipotesi per poter usare ANOVA sono la normalità dei dati, e l'uguglianza delle varianze. Così come per il t-test, se le distribuzioni sono normali ma le varianze sono differenti, esistono correzioni che permettono di utilizzare ancora ANOVA, in particolare su R utilizzeremo `oneway.test`. Se la normalità è violata, si può utilizzare il test non parametrico di Kruskal-Wallis, che più che un test per le medie è un test che verifica se i dati dei vari gruppi vengono dalla stessa distribuzione. Si osservi che se questo è vero (ipotesi nulla), in particolare le varianze devono essere uguali. Dunque non ha molto senso usare il Kruskal-Wallis in assenza di omoschedasticità, fatto che comunque implica la differenza tra le distribuzioni. Uno dei test post-hoc nel caso non parametrico che può essere usato è il `pairwise.wilcox`.
Si tenga infine presente che ANOVA è considerato un test **robusto**. Cito qui quanto riportato dal testo [WS]. "L'ANOVA è sorprendentemente robusta rispetto alle deviazioni dall'assunzione di normalità, in particolare quando le dimensioni campionarie sono grandi. Questa robustezza deriva dalle proprietà delle medie campionarie descritte dal teorema del limite centrale (...). L'ANOVA è robusta anche rispetto agli scostamenti dall'assunzione di uguale varianza nelle k popolazioni, ma soltanto se i campioni hanno tutti all'incirca la stessa dimensione."

In un contesto reale di ricerca o lavoro, vi consiglio di approfondire le condizioni di utilizzabilità dei test che volete usare (che purtroppo R non riporta nelle sue pagine di manuale).

6.3.1 ANOVA con prove ripetute

Considero i seguenti dati, relativi a 10 soggetti, indicati dal numero progressivo s, di cui vengono misurate le ore di sonno o, dopo $g = 0,\ 30,\ 60$ giorni di un trattamento $t = PT$ oppure FT.

s	g	t	o
1	0	PT	5.6
1	30	PT	3.3
1	60	PT	8.9
2	0	PT	4.8
2	30	PT	4.7
2	60	PT	7.9
3	0	PT	3.9
3	30	PT	5.3
3	60	PT	8.5
4	0	PT	5.8
4	30	PT	4.5
4	60	PT	8.2
5	0	PT	3.5
5	30	PT	3.7
5	60	PT	6.9

s	g	t	o
6	0	FT	3.9
6	30	FT	4.7
6	60	FT	7.0
7	0	FT	4.1
7	30	FT	4.9
7	60	FT	5.6
8	0	FT	5.0
8	30	FT	3.8
8	60	FT	5.3
9	0	FT	4.0
9	30	FT	4.7
9	60	FT	5.3
10	0	FT	4.0
10	30	FT	4.8
10	60	FT	6.3

In quello che segue considero la variabile "giorni" come una variabile nominale. Chiarisco questo punto: se la considerassi come una variabile quantitativa in un modello di regressione mi aspetterei un effetto linerare, dunque passare da 0 a 30 giorni o da 30 a 60 giorni dovrebbe dare la stessa variazione sul numero di ore dormite. Questa ipotesi è fisiologicamente poco consistente: una terapia farmacologica a lunga durata cambia la biochimica, dunque "0", "30", "60" possono essere pensati come tre distinti stati biochimici dei pazienti, senza una relazione quantitativa. Qundi prima di continuare ridefiniamo le variabili `ore` e `soggetto` come variabili nominali.

```
sonno$giorno <- as.factor(sonno$giorno)
sonno$soggetto <- as.factor(sonno$soggetto)
```

Ignoriamo la variabile trattamento e chiediamoci se la variabile "giorno" (cioè da quanto tempo il soggetto è sotto trattamento) influenza il numero di ore dormite.

```
summary(aov(ore~giorno,data=sonno))
            Df Sum Sq Mean Sq F value  Pr(>F)
giorno       2  43.01  21.506   22.77 1.6e-06 ***
Residuals   27  25.50   0.944
```

In questa ANOVA c'è un errore concettuale, perché per ogni soggetto vengono fatte tre misure, dunque parte della devianza residua 36.51 si spiega in termini di devianza dovuta alla variabile "soggetto", e questa devianza va tolta dal confronto per capire se la devianza relativa alla variabile "giorno" sia grande.

Evidenziamo i contributi alla devianza:

```
summary(aov(ore~giorno+soggetto,data=sonno))
            Df Sum Sq Mean Sq F value   Pr(>F)
giorno       2  43.01  21.506  24.133 8.05e-06 ***
soggetto     9   9.46   1.051   1.179    0.364
Residuals   18  16.04   0.891
```

In quest'ANOVA la devianza residua viene decomposta nella parte soggetto, da 9 df, che viene messa da parte, e in una vera parte residua, da 18 df. Il test viene fatto rapportando la varianza tra i giorni alla varianza residua.
In qeusto output c'è un p-value che non ha senso calcolare, quello per la variabile soggetto. Per avere un output più 'piulito' c'è una precisa istruzione di `R`:

```
 summary(aov(ore~giorno+Error(soggetto/giorno),data=sonno))

Error: soggetto
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9  9.456   1.051

Error: soggetto:giorno
          Df Sum Sq Mean Sq F value   Pr(>F)
giorno     2  43.01  21.506   24.13 8.05e-06 ***
Residuals 18  16.04   0.891
```

Osservo che lo stesso output si sarebbe ottenuto con il comando
`summary(aov(ore~giorno+Error(soggetto),data=sonno))`.
La sintassi `soggetto/giorno` mette sono in evidenza che la variabilità di `giorno` è dentro `soggetto`: per ogni soggetto infatti ci sono tutti e tre i livelli di `giorno`. Nel gergo di ANOVA si dice che `giorno` è una variabile *within* la variabile `soggetto`

Trascuriamo ora la variabile giorno, e vediamo se il tipo di trattamento influenza il numero di ore di sonno. Per ogni trattamento ho 15 dati, dunque potrei pensare di fare il seguente test.

```
 summary(aov(ore~trattamento,data=sonno))
            Df Sum Sq Mean Sq F value Pr(>F)
trattamento  1   4.88   4.880   2.148  0.154
Residuals   28  63.63   2.272
```

Anche qui però sto facendo un errore, infatti questa ANOVA interpreta i dati come l'esito dei due trattamenti su due differenti campioni di 15 soggetti, invece ogni trattamento riguarda solo 5 soggetti, con 3 misure per ogni soggetto al variare dei giorni di terapia.

Per provare a capire come procedere, decomponiamo la devianza più che possiamo:

```
summary(aov(ore~trattamento*soggetto*giorno,data=sonno))
                   Df Sum Sq Mean Sq
trattamento         1   4.88   4.880
soggetto            8   4.58   0.572
giorno              2  43.01  21.506
trattamento:giorno  2   7.87   3.936
soggetto:giorno    16   8.17   0.510
```

Si noti che non c'è fariabilità residua, perché una volta specificato il soggetto, il nmero di ore, il trattamento ci siamo ridotto a un solo dato. Inoltre manca la coppia trattamento:soggetto, perché la variabile trattamento è *beetwen* i soggetti, infattei ogni trattamento è stato assegnato a diversi soggetti ma ogni soggetto è stato somministrato un solo trattamento. Ricordo che invece la variabile giorno è *within* i soggetti, cioè "giorno" ha i differenti valori in tutti i soggetti.

Non è facilissimo capire i df della tabella. Sono immediati quelli di trattamento, variabile a due livelli e con 1 df, quelli di giorno, variabile a tre livelli e 2 df, quelli di trattamento:giorno, tabella 3×2 a 2 df. La coppia soggetto:giorno sembrerebbe una tabella 10×3 e dunque dovrebbe avere 18 df. Però i soggetti sono divisi in due gruppi, a seconda del trattamento. Dunque in realtà si tratta di due tabelle 5×3, da 8 df ciascuna, in totale appunto 16. Anche i df di soggetto non sono 9 ma 8, proprio perché si tratta di due gruppi da 5 soggetti.

Se siamo interessati a valutare la devianza dovuta al trattamento, dobbiamo considerare come residua tutta la devianza che non è relativa alla variabile giorno. Rimane dunque solo la variabile soggetto, a 8 df, che in questo senso fa da residuo per il trattamento. Si veda infatti l'output del seguente comando:

```
summary(aov(ore~trattamento+Error(soggetto),data=sonno))

Error: soggetto
            Df Sum Sq Mean Sq F value Pr(>F)
trattamento  1  4.880   4.880   8.532 0.0193 *
Residuals    8  4.576   0.572

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 20  59.05   2.953
```

La devianza residua su cui viene fatto il test per il trattamento è quella degli 8 df di soggetto. La devianza residua trascurata è la somma delle devianze relative alla variabile giorno.

Infine, analizziamo i due fattori trattamento e giorno.

```
summary(aov(ore~trattamento*giorno+Error(soggetto/giorno),data=sonno))

Error: soggetto
            Df Sum Sq Mean Sq F value Pr(>F)
trattamento  1  4.880   4.880   8.532 0.0193 *
Residuals    8  4.576   0.572

Error: soggetto:giorno
                   Df Sum Sq Mean Sq F value   Pr(>F)
giorno              2  43.01  21.506  42.128 4.21e-07 ***
trattamento:giorno  2   7.87   3.936   7.711  0.00452 **
Residuals          16   8.17   0.510
```

Come si vede la prima parte dell'output contiene l'informazione sulla variabile trattamento, che abbiamo già visto con il comando precedente. La seconda parte contiene l'informazione sulla variabile giorno e sulla coppia trattamento giorno. Si noti che il risultato relativo all'effetto della variabile giorno è un po' differente rispetto al risultato del comando
`summary(aov(ore~giorno+Error(soggetto/giorno),data=sonno))`
perché ora parte della devianza è stata spiegata con la variabile trattamento (come accade in genere quanto si passa dall'analisi a una via all'analisi a due vie).

6.3.2 Test per i coefficienti della retta di regressione

Il test dell'ANOVA appartiene ai test per verificare l'esistenza di un **modello lineare** nella relazione tra variabili, dunque ha delle similitudini con la retta di regressione. Infatti, in ANOVA consideriamo

$$X_{ik} = X_{\bullet\bullet} + \alpha_i + \varepsilon_{ik},$$

cioè pensiamo che il dato X_{ik} sia **predetto** dal valore medio complessivo $X_{\bullet\bullet}$, più un effetto α_i dovuto al trattamento, più un errore, ε_{ik} sperabilemnte gaussiano. Il valore di α_i è $X_{i\bullet} - X_{\bullet\bullet}$, e il test ANOVA serve a capire se questa differenza è nulla, in pratica se la variabile trattamento influenza i dati, oppure no.
Nei modelli di regressione lineare scriviamo

$$Y_k = aX_k + b + \varepsilon_k,$$

dove $aX_k + b$ è la **predizione** in base al valore della variabile esplicativa, mentre ε_k è la differenza tra il valore vero e quello predetto, ed è un numero modellizzato con una variabile aleatoria di media nulla, sperabilmente gaussiana. Inoltre poiché la retta di regressione passa per il baricentro, abbiamo scritto anche

$$Y_k = \bar{Y} + a(X_k - \bar{X}) + \varepsilon_k$$

cioè Y_k è predetto dalla media generale $\bar{Y}$ più un effetto dovuto alla variabile esplicativa X, più la variabilità intrinseca ε. Si può testare se c'è l'effetto della variabile esplicativa: l'ipotesi nulla sarà che non c'è effetto, cioè che $a = 0$. Il test è analogo all'ANOVA e si fa decomponendo la devianza.

Ricordo che abbiamo provato che

$$\sigma_Y^2 = a^2\sigma_x^2 + \sigma_\varepsilon^2$$

cioè abbiamo decomposto la varianza della variabile di risposta in una parte spiegata dal modello lineare, e quindi spiegata dalla variabilità della variabile esplicativa, e una variabilità intrinseca, chiamata **residua**. Rileggiamo in termini di devianza questa uguaglianza:

$$\sum_k (\Delta Y_k)^2 = a^2 \sum_k (\Delta X_k)^2 + \sum_k \varepsilon_k^2.$$

La devianza totale, a $N-1$ df, è la somma della devianza "tra i gruppi", che in questo caso è la devianza dovuta alla relazione lineare con la variabile esplicativa, più la varianza residua. Osserviamo che la somma delle ε_k è nulla, dunque sembrerebbe che ci sono $N-1$ gradi di libertà. Però la procedura di ottimizzazione per a, quella che ci permette di trovare la retta di regressione, impone l'indipendenza statistica tra ΔX_k e ε_k, infatti risulta

$$\overline{\varepsilon \Delta X} = \overline{(\Delta Y - a\Delta X)\Delta X} = \sigma_{xy} - a\sigma_X^2$$

che è proprio 0 perché $a = \sigma_{XY}/\sigma_X^2$. Dunque ε_k hanno somma nulla e verificano $\overline{\varepsilon \Delta X} = 0$, quindi è sufficiente conoscerne $N-2$ per determinare gli altri due: (infatti usando una condizione possiamo esprimere uno degli ε_k in funzione degli altri $N-1$, quindi possiamo scrivere la seconda condizione in termini di $N-1$ variabili ε_k e ricavarne una in funzione delle rimanenti $N-2$). Dunque i gradi di libertà sono $N-2$. Ne segue che la devianza dovuta a X ha un solo grado di libertà.
L'istruzione per costruire la retta di regressione della variabile `y` in funzione della variabile `x` è `summary(lm(y~x))`. Per testare l'ipotesi nulla che a sia zero, basta considereare il `summary` del modello. Per esempio, per la variabile spazio di frenata `dist` in funzione della velocità `speed` per il dataset `cars`, il comando `summary(lm(dist~speed,data=cars))` dà

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
```

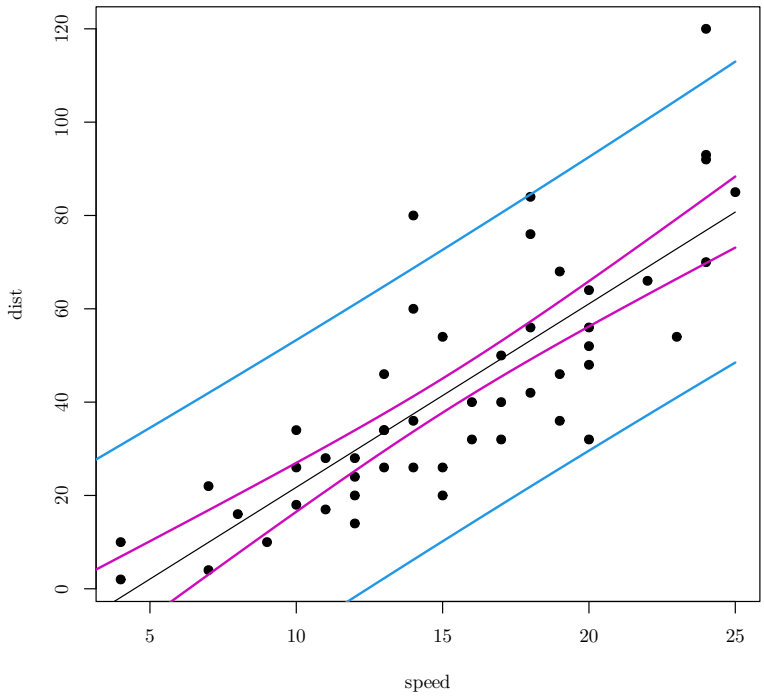
Il secondo p-value è quello relativo alla variabile `speed`.
Si confronti questo output con quello di `summary(aov(dist~speed,data=cars))` che dà

```
            Df Sum Sq Mean Sq F value   Pr(>F)
speed        1  21185   21185   89.57 1.49e-12 ***
Residuals   48  11354     237
```

Come si vede la riga relativa a `dist` dà lo stesso p-value. Si noti che `speed` ha un df, perché è una variabile quantitativa e non qualitativa, infatti il suo effetto è spiegato da un solo numero, io coefficiente a.

Nell'output del test per il modello lineare c'è anche il p-value per l'intrcetta, cioè viene anche testata l'ipotesi nulla che il coefficiente b sia 0. Non entro nel dettaglio di questo test. Osservo solo che, sia per a che per b, attraverso i test si possono ottenere degli intervalli di fiducia. In pratica si possono considerare tutte le rette $aX+b$ che sono compatibili con i dati. In questo modo si possono determinare gli intervalli di fiducia per i valori di $aX+b$; calcolato per ogni possibile valore di X si ottengono in questo modo le **bande di fiducia per la retta di**

regressione. Se invece vogliamo valutare l'intervallo di fiducia della predizione, dobbiamo considerare anche la variabilità dell'errore ε. In questo modo si ottengono le **bande di fiducia per la predizione**.



6.4 Modelli lineari generalizzati e massima verosimiglianza

Rivediamo da un altro punto di vista la costruzione della retta di regressione per due variabili statistiche, X e Y, partendo da un modello, ipotizziamo cioè che

$$Y_i = aX_i + b + \varepsilon_i$$

dove ε_i sono il risultato di un'estrazione di variabili aleatorie gaussiane indipendenti di media nulla e varianza σ^2, non nota.
Vogliamo trovare i "migliori valori" dei coefficienti a e b. Possiamo ragionare in questo modo. Aver visto Y_i equivale ad aver visto l'errore $\varepsilon_i = Y_i - (aX_i + b)$, e, per ipotesi, la densità di probabilità di questo evento è

$$\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(Y_i-(aX_i+b))^2/2\sigma^2}.$$

Poiché abbiamo anche supposto l'indipendenza, la densità di probabilità dell'evento che abbiamo osservato, cioè i valori $Y_1, \dots Y_N$ in corrispondenza di $X_1, \dots X_N$, è il prodotto delle densità:

$$\frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \mathrm{e}^{-(Y_i-(aX_i+b))^2/2\sigma^2}.$$

Un criterio per scegliere i migliori valori di a e b è quello detto di **massima verosimiglianza**, cioè di trovare a e b che rendono massima possibile la probabilità di vedere quello che

effettivamente abbiamo visto (in questo esempio massimizziamo la densità di probabilità). Per poter fare questo conto, notiamo che rendere massima una quantità è la stessa cosa che rendere massimo il suo logaritmo, perché il logaritmo è una funzione crescente. Calcoliamo il logaritmo della densità di probabilità scritta sopra:

$$-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - (aX_i + b))^2.$$

L'unica parte di quest'espressione che dipende da a e b è quella dentro la sommatoria, e compare con un segno meno. Dunque i valori di a e b che realizzano la massima verosimiglianza sono quelli che rendono minima la somma

$$\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - (aX_i + b))^2,$$

e questa condizione è esattamente quella dei minimi quadrati.
In conclusione di questo esempio, la retta dei minimi quadrati è anche la retta che si ottiene con il principio di massima verosimiglianza, nell'ipotesi modellistica che $Y = aX + b + \varepsilon$, dove ε è gaussiana di media nulla e deviazione standard fissata.

Il principio di massima verosimiglianza si usa anche per determinare interessanti dipendenze non lineari. Farò qui l'esempio della regressione "logistica". Prima però rivediamo la relazione

$$Y_i = aX_i + b + \varepsilon_i$$

Il valore $aX_i + b$ predice $< Y_i >$, il valore atteso della variabile Y calcolata nel posto X_i. Dunque ci sono tre elementi:

- un predittore lineare, $aX + b$;
- una relazione tra la media e il predittore lineare, in questo caso l'identità: $< Y_i >= aX_i + b$ (questa funzione si chiama "funzione di link", o funzione di collegamento);
- una assegnata legge di probabilità per la variabile Y_i: in questo caso una gaussiana di deviazione standard fissata, e di media determinata attraverso il predittore lineare.

Consideriamo ora un esempio, tratto da [IM]. Nello studio delle cause della tragica esplosione dello shuttle del 1986 si è compreso che ha giocato un ruolo la rottura degli anelli a "O", e che la probabilità di rottura è una funzione della temperatura al momento del lancio. Ci sei anelli a "O", e questi sono i dati riguardo alla loro rottura (riporto solo quelli relativi ai due lanci avvenuti a temperatura più bassa e quello avvenuto alla massima temperatura; la temperatura è in gradi Fahrenheit):

```
lancio temp  danno
1 53 1
1 53 1
1 53 0
1 53 0
1 53 0
1 53 0
```

```
2 57 1
2 57 0
2 57 0
2 57 0
2 57 0
2 57 0
23 81 0
23 81 0
23 81 0
23 81 0
23 81 0
23 81 0
```

In questo caso, si può ipotizzare che la variabile aleatoria "rottura", che vale 1 se l'anello si rompe, e 0 se non si rompe, sia una bernoulliana b di parametro incognito p (la probabilità di rottura), con p che decresce al crescere della temperatura. Si osservi che $p =< b >$ è esattamente il valore atteso della variabile bernoulliana,
Si può immaginare che p sia 1 per temperature molto basse e sia 0 per temperature molto alte. La relazione tra p e T non può dunque essere lineare. Serve una funzione che passi da un valore all'altro in modo monotono. Ci sono vari candidati per questa funzione di collegamento, e una delle più usate è la funzione **logistica** di cui abbiamo parlato a proposito del modello di Verhulst. Si tratta di funzioni del tipo

$$f(x) = \frac{1}{1 + \mathrm{e}^{-\alpha(x-x_0)}}$$

Si veda [BDM] esempio 7.2.4.
Cerchiamo dunque il miglior predittore lineare tale che

$$p(T) =< b(T) >= \frac{1}{1 + \mathrm{e}^{-(aT+b)}}$$

Questa relazione può essere anche scritta al contrario, con qualche semplice passaggio:

$$\log \frac{p}{1-p} = aT + b$$

La funzione che a p associa $\log \frac{p}{1-p}$ si chiama "logit". In questo caso, dunque, la funzione di link è la funzione logit, e la variabile aleatoria è bernoulliana.
Come si trova il miglior predittore lineare? Si usa la massima verosimiglianza. Se p_i è la probabilità di rottura alla temperatura T_i, e ho visto b_i (cioè 1 se c'è stata rottura, 0 se non c'è stata), la probabilità dell'evento che ho visto è

$$p_i^{b_i}(1-p_i)^{1-b_i},$$

infatti se ho b_i vale 1, $1-b_i = 0$ e dunque l'espressione $p_i^{b_i}(1-p_i)^{1-b_i}$ è uguale a $p_i^1(1-p_i)^0 = p_i$, che è proprio la probabilità di $b_i = 1$. Al contrario, se $b_i = 0$, l'espressione è uguale a $p_i^0(1-p_i)^1 = 1-p_i$, che è la probabilità di $b_i = 0$. Il miglior predittore lineare per $\texttt{logit}(p)$ si ottiene dunque massimizzando

$$\prod_{i=1}^{N} \left(\frac{1}{1 + \mathrm{e}^{-(aT_i+b)}}\right)^{b_i} \left(\frac{1}{1 + \mathrm{e}^{+(aT_i+b)}}\right)^{1-b_i},$$

dove ho usato il fatto che

$$1 - p_i = 1 - \frac{1}{1 + \mathrm{e}^{-(aT_i+b)}} = \frac{1}{1 + \mathrm{e}^{+(aT_i+b)}}.$$

`R` farà il conto per noi, trovando a e b.

In tutti i fenomeni in cui ci si aspetta un valore di soglia si utilizzano logistiche. Sono però possibili altre scelte. Consideriamo ancora una variabile bernoulliana b, ma supponiamo che $p = P(Z > aT + b)$, cioè il parametro p della variabile bernoulliana è governato dal comportamento di una variabile normale standard. In questo caso la funzione che lega p a T è

$$p = \texttt{pnrom}(aT + b)$$

cioè la probabilità cumulata per una gaussiana standard. La funzione inversa che esprime $aT + b$ in termini di p è detta **probit** e ha un andamento analogo alla logistica (ma con transizione più netta, perché `pnorm` ha code meno pesanti della logistica).
C'è un'ulteriore possibile scelta, che calcola il predittore lineare per il **log-log complementare** (abbreviato in `cloglog`) di p, cioè la funzione

$$\ln(-\ln(1 - p)) = aT + b.$$

La funzione inversa è

$$p = 1 - \mathrm{e}^{-\mathrm{e}^{aT+b}},$$

che ha code estremamente più leggere, e dunque descrive transizioni brusche.

Chapter 7

Componenti principali

Per una introduzione all'argomento vedi [BDM 12.5].

7.1 Un esempio

Il metodo delle componenti principali è un metodo che si applica quando si hanno molte variabili statistiche più o meno correlate tra loro, e ci si aspetta che la variabilità dei dati possa però essere descritta con poche variabili.
Supponiamo di considerare K dati, ognuno composto dalla misura di N variabili statistiche differenti. Geometricamente, ogni dato è un punto in uno spazio dimensione N, in cui gli assi perpendicolari rappresentano le N variabili statistiche. Il metodo delle componenti principali permette di trovare dei nuovi assi perpendicolari, intorno al baricentro, in modo che i dati si possono descrivere, con un piccolo errore, usandone solo alcune. In questo modo si riduce la dimensione dello spazio delle variabili, trascurando quelle meno rilevanti.
Invece di dare una spiegazione astratta, passo direttamente all'esempio concreto. La teoria che sto descrivendo è utilissima se N è grande, ma da indicazioni interessanti anche per per piccoli N. In particolare, nell'esempio che considereremo avremo 15 anfore (dunque $K = 15$) e per ogni anfora avremo la misura di 4 differenti caratteristiche geometriche, e dunque $N = 4$:

`h` è l'altezza

`la` è la larghezza dell'apertura

`im` è l'altezza alla quale inizia il manico

`fm` è l'altezza alla quale finisce il manico

Ci si può chiedere quanto queste variabili siano correlate tra loro, e se è sufficiente considerarne meno di 4 per descrivere la geometrica delle anfore.
Per esempio, se ci fosse correlazione massima tra tutte le variabili, vorrebbe dire che esiste un solo modello di anfora, prodotto in varie dimensioni. Se invece di fossero due tipi di anfore, quelle con manico grande e quelle con manico piccolo, indipendentemente dalla dimensione

complessiva, la variabile importante sarebbe `fm-im`, che sarebbe poco correlata con l'altezza e la larghezza. Se esistessero anfore di qualunque altezza e qualunque larghezza, altezza e larghezza dovrebbero essere poco correlate, dunque entrambe le variabili dovrebbero essere separatamente prese in considerazione.
Questi sono i dati, in centimetri

	fm	im	h	la
1	24.4	21.3	30.4	10.7
2	23.6	20.3	32.2	8.8
3	20.9	17.5	29.6	8.9
4	22.7	19.2	28.4	11.0
5	20.7	17.7	27.7	10.1
6	25.6	21.8	33.8	9.4
7	21.4	17.7	30.9	11.6
8	26.9	23.5	33.9	12.6
9	24.5	21.2	32.8	10.1
10	24.1	21.1	33.2	10.0
11	26.6	22.8	33.8	10.4
12	20.9	17.6	27.9	10.4
13	24.6	21.1	34.1	10.4
14	25.0	22.1	35.6	11.6
15	24.5	21.1	34.0	10.5

L'altezza media è 31.89 cm, la larghezza media è 10.43 cm, mentre la media delle variabili im e fm è, rispettivamente, 20.40 e 23.76 centimetri.
Nella tabella seguente sono riportati le differenze dai valori medi.

	fm	im	h	la
1	0.64	0.90	-1.49	0.27
2	-0.16	-0.10	0.31	-1.63
3	-2.86	-2.90	-2.29	-1.53
4	-1.06	-1.20	-3.49	0.57
5	-3.06	-2.70	-4.19	-0.33
6	1.84	1.40	1.91	-1.03
7	-2.36	-2.70	-0.99	1.17
8	3.14	3.10	2.01	2.17
9	0.74	0.80	0.91	-0.33
10	0.34	0.70	1.31	-0.43
11	2.84	2.40	1.91	-0.03
12	-2.86	-2.80	-3.99	-0.03
13	0.84	0.70	2.21	-0.03
14	1.24	1.70	3.71	1.17
15	0.74	0.70	2.11	0.07

Questa è la matrice di covarianza

```
         fm        im         h        la
fm 4.1154286 3.9978571 4.3001429 0.6042857
im 3.9978571 3.9585714 4.2200000 0.6171429
h  4.3001429 4.2200000 6.5126667 0.4890476
la 0.6042857 0.6171429 0.4890476 1.0223810
```

da cui si nota la piccola variabilità della larghezza rispetto alle altre variabili, mentre questa è la matrice di correlazione

```
          fm        im         h        la
fm 1.0000000 0.9904909 0.8306076 0.2945971
im 0.9904909 1.0000000 0.8311201 0.3067680
h  0.8306076 0.8311201 1.0000000 0.1895245
la 0.2945971 0.3067680 0.1895245 1.0000000
```

Come si vede, c'è una grande correlazione tra inizio e fine dell'altezza del manico, e queste due variabili sono anche abbastanza correlate all'altezza, mentre la larghezza è poco correlata con tutte le variabili.

L'istruzione di `R` che trova le componenti principali è `prcomp`. Vediamo l'output del comando nel caso delle anfore.

```
Rotation:
           PC1         PC2         PC3          PC4
fm -0.53442345 -0.3883972  0.2791895  0.696844594
im -0.52435231 -0.3883532  0.2454524 -0.716930628
h  -0.65829082  0.6819521 -0.3187255  0.002937617
la -0.07809888 -0.4829811 -0.8719062  0.020235025
```

La "rotazione" permette di passare dalle vecchie coordinate `fm, im, h, la`, alle nuove `PC1,...PC4` (ricordo che il baricentro è posto in 0). In particolare, se un dato aveva coordinate fm, im, h, la (rispetto al baricentro) la nuova prima coordinata del dato si ottiene moltiplicando questi valori per la prima colonna e poi sommando

$$-0.574fm - 0.576im - 0.531h - 0.237la$$

Analogamente per le altre.

Questa è la matrice di covarianza nelle nuove coordinate

```
             PC1           PC2           PC3           PC4
PC1  1.342307e+01  1.636774e-15 -9.102342e-16 -3.010668e-15
PC2  1.636774e-15  1.314039e+00  5.290519e-17 -1.687205e-17
PC3 -9.102342e-16  5.290519e-17  8.339232e-01 -2.647550e-17
PC4 -3.010668e-15 -1.687205e-17 -2.647550e-17  3.801110e-02
```

Come si vede le varianza sono in ordine decrescente, e le covarianze sono praticamente nulle: le componenti principali sono variabili statisticamente indipendenti.

```
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     3.6638 1.14632 0.91319 0.19496
Proportion of Variance 0.8599 0.08418 0.05343 0.00244
Cumulative Proportion  0.8599 0.94414 0.99756 1.00000
```

Si legge qui quanto la varianza di ogni nuova variabile contribuisce al totale, cioè quanta parte della varianza complessiva è "spiegata" dalle varie componenti. Poiché sono in ordine decrescente, è utile guardare anche la proporzione cumulata, che ci dice che le prime due componenti spiegano circa il 95% della variabilità dei dati, mentre la quarta componente è piuttosto inutile.
Guardando la matrice di rotazione, si nota che la prima componente principale è fatta in ugual misura dalle prime tre altezze, e in misura minore dalla larghezza (tutte con lo stesso segno). Quindi è una misura della dimensione dell'oggetto. Poiché tutti i segni sono negativi, più è alto PC1, più l'oggetto è piccolo. La dimensione delle anfore spiega circa l'85% della variabilità dei dati.
La seconda componente è composta sostanzialmente dalla larghezza e, in misura minore, ma con segno opposto, dalle altre tre variabili. Dunque è una misura della larghezza, ma anche dello schiacciamento dell'oggetto: più PC2 è grande, più l'anfora è tozza e i manici sono bassi.
La componente PC3 ha fm e im uguali ma opposti in segno all'altezza e larghezza dunque rappresenta la posizione relativa del manico rispetto alle dimensioni principali
Infine, PC4 misura la larghezza del manico rispetto alla media, infatti è praticamente $0.7(im - fm)$.